Notes on CSPs and Polymorphisms

Contents

0.1   General Outline

These notes were born from a multi-year learning seminar at MIT. Each section corresponds roughly to a one-hour talk from the seminar, with details filled in. The subsections that occur after some of the sections consist of optional extra material that wasn't covered in the learning seminar due to time constraints. The appendices consist of longer portions of optional material - summaries of famous universal algebraic theories that are useful to know about in order to navigate the literature.

In the next section we give a teaser for the sorts of results we'll try to prove here, mainly to convince the reader that there are highly nontrivial results in this area, and that it is not just abstract nonsense. The text proper begins in Chapter 1, which consists of the foundational

abstractions we need later, together with several fundamental examples illustrating three different behaviors of CSPs that need to be understood in order to understand the general case.

In Chapter 2, we go over the breakthrough theory of algebras few subpowers, which lead to the first truly nontrivial algorithmic result in this area. In Chapter 3, we go over the more technically challenging theory of absorbing subalgebras and its application to CSPs of "bounded width" - although the algorithms used to solve CSPs in this chapter are much simpler than the ones from the previous chapter, the algebraic machinery necessary to prove that these algorithms always succeed is much more difficult (but more broadly applicable). Chapters 2 and 3 do not necessarily need to be read in order, as the algebraic approaches used are quite different. In Chapter 4 we move to trying to understand the general case of finite Taylor algebras, starting with the simpler case of conservative Taylor algebras to introduce a few of the new ideas that will be necessary to handle the general case.

Currently these notes are in an unfinished state - maybe half way through the material needed for the CSP dichotomy for finite structures, with much more planned if that is ever finished.

0.2 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 1.1.

The story starts with a result of Schaefer [130] on a problem he called "Generalized Satisfiability".

Definition 0.2.1. If Γ is a set of relations on $\{0, 1\}$, then GenSAT(Γ) 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 Γ 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 satifies all of the given constraints.

Theorem 0.2.2 (Schaefer [130]). *If* GenSAT(Γ) *is not NP-complete, then Γ 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 $\mathbb{F}_2$.*

3

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

The next result of this form is due to Hell and Nešetril [70], on a generalization of n-coloring which they call "H-coloring".

Definition 0.2.3. 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 0.2.4 (Hell, Nešetril [70]). *H-coloring is in P if H is bipartite, and it is NP-complete otherwise.*

These two results led Feder and Vardi [61] to ask whether there is a general dichotomy between P and NP. However, any such dichotomy has to avoid Ladner's [99] anti-dichotomy result.

Theorem 0.2.5 (Ladner [99]). *If $P \neq NP$, then there are problems in NP which are neither in P nor NP-complete.*

In order to avoid Ladner's result, Feder and Vardi focused on a special type of problem: "constraint satisfaction problems" (abbreviated as CSPs) with a fixed "template".

Definition 0.2.6. A CSP-*template* T consists of a finite set D together with a finite collection $\Gamma = (R_1, ..., R_n)$ of relations on D - equivalently, we can think of T as a relational structure $(D, R_1, ..., R_n)$. 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.

Example 0.2.1. The problem k-COLORING (given a graph, determine if it can be colored with k colors) is equivalent to CSP($\{1, ..., k\}, \neq$) = CSP(K_k), where K_k is the complete graph of k vertices (considered as a relational structure).

Example 0.2.2. 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 0.2.3. The problem 3SAT can be thought of as CSP($\{0, 1\}, R_{(0,0,0)}, ..., R_{(1,1,1)}$), 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 0.2.4. The problem NAE-SAT is CSP($\{0, 1\}, NAE$), where $NAE = \{0, 1\}^3 \setminus \{(0, 0, 0), (1, 1, 1)\}$ 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 0.2.5. 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 0.2.6. The problem HORN-SAT is CSP($\{0, 1\}, \{0\}, \{1\}, \{0, 1\}^3 \setminus \{(1, 1, 0)\}$) (the third constraint is $(x \wedge y) \implies z$). This problem is known to be P-complete, and it can be solved in linear time.

Example 0.2.7. 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 [138] (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 0.2.8. 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 [61] 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 0.2.7. 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 [39], 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 0.2.8. 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 [61] 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 [61] introduced the concept of the *ability to count*. Their definition of this concept is quite technical, and it was later realized that it's enough to focus on a simpler case: the affine CSP over an abelian group.

Definition 0.2.9. For every abelian group A, we define the associated *affine CSP* to be the CSP with domain A, with the ternary relation $\{(x, y, z) \mid x + y + z = a\}$ and the unary singleton relation $\{a\}$ for each element $a \in A$.

In case the reader wants to see the general definition of the ability to count, we have reproduced it below.

Definition 0.2.10. 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 any instance of the CSP which satisfies the following properties has no solution:

- the instance only uses the constraints C, Z,

- 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

- A contains exactly one more copy of the constraint C than B does.

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

Theorem 0.2.11 (Feder, Vardi [61])**.** *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 0.2.1. Any CSP which can't "simulate" a CSP which has the ability to count *does* have bounded width.

Conjecture 0.2.2. 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*! In fact, the conjecture about the ability to count holds even if we only require that our CSP can't simulate any affine CSP.

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

Problem 0.2.1. What is the largest possible generalization of the Gaussian elimination algorithm?

Feder and Vardi [61] 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 [60] 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 [47] studied the "expressive rate" of a constraint language Γ, 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 Γ. 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 Γ is "polynomially learnable". The same conjecture occurs in chapter 10 of Víctor Dalmau's thesis [50], in an algebraic form.

Conjecture 0.2.3. For any constraint language Γ, the logarithm of the number of distinct n-variable relations which can be defined by primitive positive formulas over Γ 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 [77], who reinterpreted an instance of a CSP as a homomorphism problem between relational structures.

Definition 0.2.12. An instance of the *general combinatorial problem*, or GCP, is a pair of relational structures $\langle \mathbf{A}, \mathbf{B} \rangle$ 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 from $\mathbf{A}$ to $\mathbf{B}$.

Example 0.2.9. Suppose that $\mathbf{T}$ 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 $\mathbf{X} = (V, C_1, ..., C_n)$, 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 $\mathbf{X} \rightarrow \mathbf{T}$ 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 $\langle \mathbf{X}, \mathbf{T} \rangle$ is equivalent to the instance of CSP($\mathbf{T}$) corresponding to $\mathbf{X}$.

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

Example 0.2.10. 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 0.2.11. 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 [67]. We won't discuss this research direction much.

After demonstrating the generality of the framework, Jeavons [77] restricts to studying homomorphism problems with a fixed target structure $\mathbf{T}$. He calls this GCP(Γ), where Γ is the list of relations of $\mathbf{T}$, but we will call it CSP($\mathbf{T}$) 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.

Definition 0.2.13. Two relational structures $\mathbf{A}, \mathbf{B}$ with the same signature are *homomorphically equivalent* if there exist homomorphisms $\mathbf{A} \rightarrow \mathbf{B}, \mathbf{B} \rightarrow \mathbf{A}$.

The homomorphism point of view now makes it obvious that if $\mathbf{A}$ and $\mathbf{B}$ are homomorphically equivalent, then CSP($\mathbf{A}$) and CSP($\mathbf{B}$) 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 [77] 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 Γ 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 CSP(Γ), 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 Γ (and equality), and conjunctions $\wedge$, but which does not involve negation, disjunction, implication, or universal quantification (such a formula is called a *conjunctive query* in database theory).

Example 0.2.12. The template we gave for HORN-SAT did not contain all possible Horn clauses - it stopped at the 3-ary Horn clause $x \wedge y \implies z$. The 4-ary Horn clause $x \wedge y \wedge z \implies w$ can be represented by the following primitive positive formula:

$$\exists u \ (x \wedge y \implies u) \wedge (u \wedge z \implies w).$$

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

Definition 0.2.14. 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, 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 0.2.15. *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 Γ.*

In order to understand this result we must define clones, polymorphisms, and the concept of a function preserving a relation.

Definition 0.2.16. 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 $\pi_i^k(x_1, ..., x_k) = x_i$ (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 $g_1, ..., g_k$ to the function

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

The reader should play with the above definition in order to convince themself 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 π_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 0.2.17. A k-ary function f is said to *preserve* an m-ary relation R, written $f \rhd R$, 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

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

A function f is a *polymorphism* of a relational structure (D, Γ) or of a relational clone Γ 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 $f \rhd R$ 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 kth power of the relational structure (D, R) (the kth 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 ith coordinates is in R for each $i \leq k$). From the algebraic point of view, we have $f \rhd R$ iff the subset $R \subseteq D^m$ is a subalgebra of the algebraic structure $(D, f)^m$, where $(D, f)^m$ is the categorical mth 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 ones 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 0.2.13. 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 $\pmod{p}$.

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 [77] was able to give a new and relatively simple proof of Schaefer's dichotomy theorem [130]. The algebraic structures corresponding to the basic polynomial time solvable problems are as follows.

Example 0.2.14. 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 $\min : \{0, 1\}^2 \to \{0, 1\}$. This operation is an example of a *semilattice* operation.

Example 0.2.15. If $\Gamma = (\leq, \neq)$ is the template corresponding to 2SAT, then the clone of polymorphisms is generated by the *majority* (or *median*) function $\mathrm{maj} : \{0, 1\}^3 \to \{0, 1\}$.

Example 0.2.16. 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 Γ are unary, then $\mathrm{CSP}(\Gamma)$ 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 [54] 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 [61], 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 [36] 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 Γ 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 [37] 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 $\mathrm{CSP}(\Gamma)$ 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 0.2.18 (Few Subpowers [25], [76])**.** *The following are equivalent for an algebraic structure $\mathbb{A}$ on a finite domain:*

- *the number of subalgebras of $\mathbb{A}^n$ grows like $2^{O(n^k)}$ for some fixed k,*

- *every subalgebra of $\mathbb{A}^n$ has a (nice) generating set of size $O(n^k)$ for some fixed k, called a compact representation,*

- $\mathbb{A}$ *has a k-edge term for some k, that is, a term satisfying the identity*

$$
f \left(\begin{bmatrix} y & y & x & x & \cdots & x \\ y & x & y & x & \cdots & x \\ x & x & x & y & \cdots & x \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x & x & x & x & \cdots & y \end{bmatrix} \right) = \begin{bmatrix} x \\ x \\ x \\ \vdots \\ x \end{bmatrix},
$$

  where all but the first column has exactly one y and $k-1$ xs, and

- *a Gaussian-elimination-like algorithm solves $\mathrm{CSP}(\mathbb{A})$ 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 (up to permuting the inputs). 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 0.2.2 (Subpower Membership Problem)**.** Given a finite subset $S \subseteq \mathbb{A}^n$ and an element $x \in \mathbb{A}^n$, determine if x is in the subalgebra of $\mathbb{A}^n$ generated by S.

A result of Kozik [93] shows that for general algebraic structures, this problem can be EXPTIME-complete. A recent result of Shriner [131] 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 0.2.4. If $\mathbb{A}$ has few subpowers, then the subpower membership problem can be solved in polynomial time.

Peter Mayr [108] 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 [38] has proved the conjecture in the special case that the algebra $\mathbb{A}$ 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 0.2.19 (Bounded Width Algebras [40], [12], [18], [100], [19], [96], [94])**.** *For an idempotent algebra on a finite domain, the following are equivalent:*

- CSP($\mathbb{A}$) *has bounded width,*

- CSP($\mathbb{A}$) *can't simulate any CSP which has the ability to count,*

- CSP($\mathbb{A}$) *has relational width* $(2,3)$,

- CSP($\mathbb{A}$) *can be solved by a "cycle consistency" algorithm, which checks arc-consistency and checks that every "cycle" of constraints has a valid solution extending each possible value of every variable in the cycle,*

- $\mathbb{A}$ *has terms* f, g *of arity* 3 *satisfying the identities*

$$g(x,x,y) = g(x,y,x) = g(y,x,x) = f(x,x,y) = f(x,y,x) = f(x,y,y),$$

- CSP($\mathbb{A}$) *can be "robustly" solved by the basic semidefinite programming relaxation (i.e., if an ϵ-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 0.2.20 (Cyclic Terms for Taylor Algebras [17]). *For an idempotent algebraic structure $\mathbb{A}$ on a finite domain, the following are equivalent:*

- *there is a finite set of identities satisfied by the operations of $\mathbb{A}$ which can't be satisfied by essentially unary functions,*

- *for every prime $p > |\mathbb{A}|$, $\mathbb{A}$ has a "cyclic term" f of arity p, that is, a term which satisfies the identity*
$$f(x_1, ..., x_{p-1}, x_p) = f(x_2, ..., x_p, x_1),$$

- *$\mathbb{A}$ 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($\mathbb{A}$) is in P iff $\mathbb{A}$ has a cyclic term."

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

Definition 0.2.21. 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.

13

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

$$\tfrac{1}{2}c(A) + \tfrac{1}{2}c(B) \geq \tfrac{1}{2}c(A \vee B) + \tfrac{1}{2}c(A \wedge B).$$

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

Theorem 0.2.22 (VCSP Dichotomy [97], [91]). *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 0.2.23 (CSP Dichotomy [44], [142]). *A finite algebra $\mathbb{A}$ has a cyclic term iff* CSP($\mathbb{A}$) *is in P.*

A major open problem is whether one can test if CSP(Γ) is in P in time polynomial in the size of the description of the constraints of Γ (given that Γ 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 0.2.24. 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 [31] 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, ω-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 [31] are the following:

- the automorphism group of $(\mathbb{Q}, <)$, 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 $(L, |)$, where $|$ is a 3-ary relation where $ab|c$ 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 [32]. 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 0.2.5 ([24]). Let $\mathbf{A}$ be the core of a reduct of a finitely bounded homogeneous structure. Then CSP($\mathbf{A}$) is in P iff $\mathbf{A}$ has a 6-ary polymorphism s and unary polymorphisms α, β 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($\mathbf{A}$) 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 0.2.25. If $\mathbf{A}, \mathbf{B}$ are relational structures such that a homomorphism $\mathbf{A} \to \mathbf{B}$ exists, then PCSP($\mathbf{A}, \mathbf{B}$) is the following problem. The input is a relational structure $\mathbf{C}$ s.t. there exists a homomorphism $\mathbf{C} \to \mathbf{A}$ (the promise), although this map is not revealed to us. The desired output is a homomorphism from $\mathbf{C}$ to $\mathbf{B}$.

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

Example 0.2.17. 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** → **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 $\mathbf{X} = (\mathbb{Z}, x + y + z = 1)$, note that the inclusion map **A** → **X** is a homomorphism, and that the map sgn : $\mathbb{Z} \to \{0, 1\}$ given by

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

defines a homomorphism **X** → **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 [9] for a proof).

The relevant algebraic object in this context is Pol(**A**, **B**), the set of homomorphisms $\mathbf{A}^k \to \mathbf{B}$. At first this structure doesn't seem algebraic at all, since there is no way to compose elements of Pol(**A**, **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 0.2.6. For any $k \geq l \geq 3$, the promise problem $\mathrm{PCSP}(K_l, K_k)$ 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 [6], and this result was slightly generalized and put into the PCSP framework in [33].

Recent results in the study of PCSPs include a result of Barto, Bulin, Opršal, and Krokhin [9] in which they used minion techniques to show that $\mathrm{PCSP}(K_d, K_{2d-1})$ is NP-hard for every $d \geq 3$, reducing from the hypergraph promise problem $\mathrm{PCSP}(\mathrm{NAE}_2, \mathrm{NAE}_k)$. The hypergraph coloring problem $\mathrm{PCSP}(\mathrm{NAE}_2, \mathrm{NAE}_k)$ was itself shown to be hard via a reduction from a variant of the PCP theorem [57]. The big result in PCSPs is the following result which connects computational complexity to height 1 identities satisfied by the minion of polymorphisms.

Theorem 0.2.26 (Barto, Bulin, Opršal, Krokhin [9])**.** *If there is a "minion homomorphism" from* $\mathrm{Pol}(\mathbf{A}_1, \mathbf{B}_1)$ *to* $\mathrm{Pol}(\mathbf{A}_2, \mathbf{B}_2)$*, then* $\mathrm{PCSP}(\mathbf{A}_2, \mathbf{B}_2)$ *has a logspace reduction to* $\mathrm{PCSP}(\mathbf{A}_1, \mathbf{B}_1)$*.*

Remark 0.2.1. 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 $I \mapsto \mathrm{Hom}(A^I, B)$.

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 $\mathrm{Pol}(\mathbf{A}, \mathbf{B})$ has an empty limit ("no" instances) (this is the "promise satisfaction of a minor condition" problem of [9]).

0.3 Incomplete list of Notation and Definitions

Most of the notation is either standard, or will be defined as it is introduced. In this section we record some of the definitions so that the reader can refer back to it if necessary.

Definition 0.3.1. A (first order) *structure* $\mathbf{A}$ is a tuple $(A, \{f_i\}, \{R_j\})$ such that A is a set (called the *underlying set* of $\mathbf{A}$), each f_i is a function of some arity n_i on A, i.e. $f_i : A^{n_i} \to A$, and each R_j is a relation of some arity m_j on A, i.e. $R_j \subseteq A^{m_j}$. The *signature* of a first order structure is the assignment of each function symbol f_i to an arity n_i together with an assignment of each relation symbol R_j to some arity m_j. If two structures $\mathbf{A}, \mathbf{B}$ share the same signature, then we sometimes write $f_i^{\mathbf{A}}$, $R_j^{\mathbf{A}}$ to refer to the interpretations of the function and relation symbols in $\mathbf{A}$, and $f_i^{\mathbf{B}}, R_j^{\mathbf{B}}$ to refer to the interpretations of the same function and relation symbols in $\mathbf{B}$.

If $\mathbf{A}, \mathbf{B}$ are structures with the same signature, then a *homomorphism* $\varphi : \mathbf{A} \to \mathbf{B}$ is a map $\varphi : A \to B$ of the underlying sets, such that for each function symbol f_i we have

$$\varphi(f_i^{\mathbf{A}}(a_1, ..., a_{n_i})) = f_i^{\mathbf{B}}(\varphi(a_1), ..., \varphi(a_{n_i}))$$

for all $a_1, ..., a_{n_i} \in A$, and such that for each relation symbol R_j we have

$$(a_1, ..., a_{m_j}) \in R_j^{\mathbf{A}} \quad \implies \quad (\varphi(a_1), ..., \varphi(a_{m_j})) \in R_j^{\mathbf{B}}$$

for all $a_1, ..., a_{m_j} \in A$. A homomorphism $\varphi : \mathbf{A} \to \mathbf{B}$ is called an *isomorphism* if there is a homomorphism $\psi : \mathbf{B} \to \mathbf{A}$ such that $\psi \circ \varphi = \mathrm{id}_{\mathbf{A}}$ and $\varphi \circ \psi = \mathrm{id}_{\mathbf{B}}$.

A structure is called a *relational structure* if it has no functions in its signature, and a structure is called an *algebraic structure* or an *algebra* if it is nonempty and has no relations in its signature. We usually write a relational structure with the bold font, i.e. $\mathbf{A}$, while we usually write an algebraic structure with the blackboard bold font, i.e. $\mathbb{A}$.

We define the *total size* of a relational structure $\mathbf{A} = (A, \{R_j\})$, written $\|\mathbf{A}\|$, to be

$$\|\mathbf{A}\| := \sum_j |R_j|.$$

Note that the number of bits needed to describe $\mathbf{A}$ is larger than $\|\mathbf{A}\|$ by a factor of about $\log_2 |A|$, ignoring the overhead needed to describe the signature of $\mathbf{A}$.

When convenient, we often abuse notation to treat a structure $\mathbf{A}$ like its underlying set A: we write $a \in \mathbf{A}$ to mean that $a \in A$, we write $|\mathbf{A}|$ for $|A|$, etc.

Definition 0.3.2. A *subalgebra* of an algebra $\mathbb{A}$ is an algebraic structure $\mathbb{B}$ with the same signature such that the underlying set B of $\mathbb{B}$ is a subset of the underlying set A of $\mathbb{A}$, and such that the inclusion map $\iota : B \hookrightarrow A$ defines a homomorphism $\iota : \mathbb{B} \to \mathbb{A}$. In this case we write $\mathbb{B} \leq \mathbb{A}$. If $S \subseteq A$, then we define $\mathrm{Sg}_{\mathbb{A}}(S)$, the *subalgebra generated by* S, to be the smallest subalgebra of $\mathbb{A}$ whose underlying set contains S.

If $\mathbb{A}, \mathbb{B}$ are structures with the same signature, then we define their *product* $\mathbb{A} \times \mathbb{B}$ to be the structure with underlying set $A \times B$ where A, B are the underlying sets of $\mathbb{A}, \mathbb{B}$, with each function symbol f_i interpreted by

$$f_i^{\mathbb{A} \times \mathbb{B}} \left(\begin{bmatrix} a_1 \\ b_1 \end{bmatrix}, ..., \begin{bmatrix} a_{n_i} \\ b_{n_i} \end{bmatrix} \right) = \begin{bmatrix} f_i^{\mathbb{A}}(a_1, ..., a_{n_i}) \\ f_i^{\mathbb{B}}(b_1, ..., b_{n_i}) \end{bmatrix},$$

and with each relation symbol R_j interpreted by

$$\left(\begin{bmatrix} a_1 \\ b_1 \end{bmatrix}, ..., \begin{bmatrix} a_{m_j} \\ b_{m_j} \end{bmatrix} \right) \in R_j^{\mathbb{A} \times \mathbb{B}} \iff (a_1, ..., a_{m_j}) \in R_j^{\mathbb{A}} \wedge (b_1, ..., b_{m_j}) \in R_j^{\mathbb{B}}.$$

Arbitrarily large products $\prod_{i \in I} \mathbb{A}_i$ of structures $\mathbb{A}_i$ all having the same signature are defined similarly.

A *homomorphic image* of $\mathbb{A}$ is defined to be any $\mathbb{B}$ of the same signature such that there is a surjective homomorphism $\varphi : \mathbb{A} \twoheadrightarrow \mathbb{B}$. An algebraic structure $\mathbb{A}$ is called *simple* if every surjective homomorphism $\mathbb{A} \twoheadrightarrow \mathbb{B}$ is either an isomorphism, or has $|\mathbb{B}| = 1$.

Definition 0.3.3. A *congruence* on an algebraic structure $\mathbb{A}$ is an equivalence relation θ on $\mathbb{A}$ which is also a subalgebra: $\theta \leq \mathbb{A}^2$. The set of all congruences of $\mathbb{A}$ is written as $\mathrm{Con}(\mathbb{A})$. If S is some collection of ordered pairs $(a, b) \in \mathbb{A}^2$, then the *congruence generated by* S, written $\mathrm{Cg}_{\mathbb{A}}(S)$, is the least congruence of $\mathbb{A}$ which contains S. We generally use greek letters for congruences.

If $\theta \in \mathrm{Con}(\mathbb{A})$ and $a \in \mathbb{A}$, then we write a/θ for the *congruence class* of a with respect to θ, that is, $a/\theta = \{b \mid (a, b) \in \theta\}$. We say that a, b are *congruent* with respect to θ if $(a, b) \in \theta$, and we may write this in symbols in several different ways:

$$(a, b) \in \theta \iff b \in a/\theta \iff a \equiv b \pmod{\theta} \iff a \equiv_\theta b \iff a\ \theta\ b.$$

If $\theta \in \mathrm{Con}(\mathbb{A})$, then we write $\mathbb{A}/\theta$ for the set of equivalence classes of θ considered as an algebraic structure, with

$$f_i^{\mathbb{A}/\theta}(a_1/\theta, ..., a_{n_i}/\theta) = f_i^{\mathbb{A}}(a_1, ..., a_{n_i})/\theta.$$

The algebra $\mathbb{A}/\theta$ is called a *quotient* of $\mathbb{A}$, and there is a canonical *quotient map* $\mathbb{A} \twoheadrightarrow \mathbb{A}/\theta$ which takes each element of $\mathbb{A}$ to its equivalence class in θ.

If $\varphi : \mathbb{A} \to \mathbb{B}$ is a homomorphism, then we define the *kernel* of φ, written $\ker \varphi$, as the equivalence relation on $\mathbb{A}$ given by $(x, y) \in \ker \varphi$ iff $\varphi(x) = \varphi(y)$. The kernel of a homomorphism is always a congruence, and if φ is surjective then $\mathbb{A}/\ker \varphi$ is isomorphic to $\mathbb{B}$. In particular, every homomorphic image of $\mathbb{A}$ is isomorphic to a quotient of $\mathbb{A}$.

For $\alpha, \beta \in \mathrm{Con}(\mathbb{A})$, we define their *meet* $\alpha \wedge \beta$ to be their intersection, and we define their *join* $\alpha \vee \beta$ to be the congruence generated by $\alpha \cup \beta$. We define the least congruence $0_{\mathbb{A}}$ to be the equivalence relation on $\mathbb{A}$ where each equivalence class is a singleton, and we define the greatest congruence $1_{\mathbb{A}}$ to be the equivalence relation on $\mathbb{A}$ consisting of just one equivalence class.

Definition 0.3.4. If $R \subseteq A_1 \times \cdots \times A_n$ is a (possibly multisorted) relation, then for any $I \subseteq \{1, ..., n\}$ we define the projection map $\pi_I : R \to \prod_{i \in I} A_i$ in the obvious way. A relation $R \subseteq A_1 \times \cdots \times A_n$ is *subdirect* if $\pi_i(R) = A_i$ for all i.

If $\mathbb{A}_1, ..., \mathbb{A}_n$ are algebraic structures with the same signature, then any subalgebra $\mathbb{R} \leq \mathbb{A}_1 \times \cdots \times \mathbb{A}_n$ is called a (multisorted) *relation* on the $\mathbb{A}_i$s. If $\pi_i(\mathbb{R}) = \mathbb{A}_i$ for all i then we say that $\mathbb{R}$ is *subdirect*, and we write this in symbols as $\mathbb{R} \leq_{sd} \mathbb{A}_1 \times \cdots \times \mathbb{A}_n$.

Definition 0.3.5. If $R \subseteq A \times B$ is a binary relation, then we define its *reverse* $R^- \subseteq B \times A$ by

$$R^- = \{(b, a) \mid (a, b) \in R\}.$$

If $R \subseteq A \times B$ and $S \subseteq B \times C$, then we define their *relational composition* $R \circ S \subseteq A \times C$ by

$$R \circ S = \{(a, c) \mid \exists b \in B \text{ s.t. } (a, b) \in R \wedge (b, c) \in S\}.$$

If $R \subseteq A \times B$ and $U \subseteq A$, then we define the *sum* $U + R \subseteq B$ by

$$U + R = \{b \mid \exists a \in U \text{ s.t. } (a, b) \in R\},$$

and for $V \subseteq B$ we define the difference $V - R \subseteq A$ by $V - R = V + R^-$. Note that we have $(U + R) + S = U + (R \circ S)$ for $U \subseteq A$, $R \subseteq A \times B$, $S \subseteq B \times C$.

Definition 0.3.6. If $\{R_j\}$ is some collection of relations, then a *primitive positive formula* over the R_js is defined to be a formula of the form

$$\exists y_1 \cdots \exists y_n \text{ s.t. } \bigwedge_{i \in I} \varphi_i(x_1, ..., x_m, y_1, ..., y_n),$$

where I is a finite set, and each φ_i is either some relation R_j applied to some of the variables $x_1, ..., x_m, y_1, ..., y_n$, or is the equality relation applied to some pair of variables.

A relation R is *primitively positively definable* over the R_js if there is a primitive positive formula φ such that the elements of R are exactly the tuples of values $(x_1, ..., x_m)$ that satisfy φ. For instance, the relational composition $R \circ S$ is primitively positively definable over R and S.

Chapter 1

Initial Intuition

1.1   The Inv-Pol **Galois connection**

We begin by recalling some definitions from the introduction.

Definition 1.1.1. 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 1.1.2. 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 $\pi_i^k(x_1, ..., x_k) = x_i$ (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 $g_1, ..., g_k$ to the function

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

Definition 1.1.3. A k-ary function f is said to *preserve* an m-ary relation R, written $f \rhd R$, 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

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

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

We can write the condition for $f \rhd R$ more compactly as $M \in R^k \implies f(M) \in R$, where $M \in R^k$ 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 1.1.4. If Γ is any set of relations on a domain D, then we define $\langle \Gamma \rangle$ to be the relational clone generated by Γ (that is, $\langle \Gamma \rangle$ is the smallest relational clone which contains Γ). 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 $\mathbb{A} = (D, \mathcal{O})$ is an algebraic structure, we let $\mathrm{Clo}(\mathbb{A})$ be the clone generated by the basic operations of $\mathbb{A}$, so $\mathrm{Clo}(\mathbb{A}) = \langle \mathcal{O} \rangle$.

Definition 1.1.5. If Γ is any set of relations on a domain D, then we define $\mathrm{Pol}(\Gamma)$ to be the the set of operations on D that preserve every relation of Γ. If $\mathcal{O}$ is any set of operations on D, we define $\mathrm{Inv}(\mathcal{O})$ 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

$$\mathrm{Pol}_k(\Gamma) = \{f : D^k \to D \mid \forall R \in \Gamma, \ f \triangleright R\},$$
$$\mathrm{Inv}_m(\mathcal{O}) = \{R \subseteq D^m \mid \forall f \in \mathcal{O}, \ f \triangleright R\}.$$

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

Example 1.1.1. If $\mathbb{A} = (D, \mathcal{O})$ is an algebraic structure, then $\mathrm{Inv}_2(\mathcal{O})$ determines (among other things)

- the lattice of subalgebras of $\mathbb{A}$,

- $\mathrm{Aut}(\mathbb{A})$, the automorphism group of $\mathbb{A}$,

- $\mathrm{End}(\mathbb{A})$, the semigroup of endomorphisms of $\mathbb{A}$,

- $\mathrm{Con}(\mathbb{A})$, the lattice of congruences on $\mathbb{A}$,

- the set of partial orders on D which are compatible with the operations of $\mathbb{A}$, and

- $\mathrm{Inv}_2(\mathbb{B})$ for any subalgebra $\mathbb{B} \subset \mathbb{A}$ or quotient $\mathbb{B} = \mathbb{A}/\sim$.

It is easy to see that for all Γ, $\mathrm{Pol}(\Gamma)$ will be a clone, and that for all $\mathcal{O}$, $\mathrm{Inv}(\mathcal{O})$ will be a relational clone. As a consequence, we have $\langle \Gamma \rangle \subseteq \mathrm{Inv}(\mathrm{Pol}(\Gamma))$ and $\langle \mathcal{O} \rangle \subseteq \mathrm{Pol}(\mathrm{Inv}(\mathcal{O}))$. 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

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix} \in \leq \quad \Longrightarrow \quad \begin{bmatrix} f(0,0,1) \\ f(0,1,1) \end{bmatrix} \in \leq,$$

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

$$\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \in \leq^3 \quad \Longrightarrow \quad f\left(\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}\right) \in \leq.$$

Theorem 1.1.6. *If Γ is a set of relations on a finite domain D, then $\mathrm{Inv}(\mathrm{Pol}(\Gamma)) = \langle \Gamma \rangle$. In fact, if a relation $S \subseteq D^m$ is preserved by $\mathrm{Pol}(\Gamma)$ and can be generated by k elements of D^m (using operations of $\mathrm{Pol}(\Gamma)$), then S can be defined by a primitive positive formula over Γ which involves at most $|D|^k$ auxiliary variables.*

Proof. Suppose that S is generated by elements $x_1, ..., x_k \in D^m$, and let X be the matrix having the x_is as columns. Then $S = \{f(X) \mid f \in \operatorname{Pol}_k(\Gamma)\}$, so as a starting point we will construct a primitive positive formula Φ that describes $\operatorname{Pol}_k(\Gamma)$.

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 $(a_1, ..., a_k)$-coordinate of f is $f(a_1, ..., a_k)$. We can now give a positive primitive formula for $\operatorname{Pol}_k(\Gamma) \subseteq D^{D^k}$:

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

If Γ 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 Φ' of the inner conjunctions will define the same subset of D^{D^k}.

Finally, to define S we use the primitive positive formula

$$S(a) := \exists f \in D^{D^k} \ \Phi'(f) \wedge (f(X) = a). \qquad \square$$

Theorem 1.1.7. *If $\mathcal{O}$ is a set of operations on a finite domain D, then $\operatorname{Pol}(\operatorname{Inv}(\mathcal{O})) = \langle \mathcal{O} \rangle$.*

Proof. Suppose that $f \in \operatorname{Pol}(\operatorname{Inv}(\mathcal{O}))$ 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 $\pi_i : D^k \to D$, $\pi_i(x_1, ..., x_k) = x_i$. 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 $f \in \operatorname{Pol}(\operatorname{Inv}(\mathcal{O}))$ and $\mathcal{F}(k) \in \operatorname{Inv}(\mathcal{O})$, we must have $f \triangleright \mathcal{F}(k)$, so in particular we must have $f(\pi_1, ..., \pi_k) \in \mathcal{F}(k)$. But $f(\pi_1, ..., \pi_k)$ is exactly f thought of as an element of D^{D^k}, so this means that $f \in \langle \mathcal{O} \rangle$. $\qquad \square$

Corollary 1.1.8. *There is an order reversing bijection between clones and relational clones, given by the operations* Inv *and* Pol.

Remark 1.1.1. The map $\{1, ..., k\} \to D^{D^k}$ given by $i \mapsto \pi_i$, where $\pi_i : D^k \to D$ is given by $\pi_i(x_1, ..., x_k) = x_i$, shows up in the theory of approximation algorithms as the *long code*, which is the longest way of encoding $\{1, ..., k\}$ over the alphabet D that doesn't have any redundant coordinates.

Example 1.1.2. In the next section we will prove the following three correspondences between clones and relational clones on the domain $\{0, 1\}$:

- $\langle 2\text{SAT} \rangle = \langle \leq, \neq \rangle$ corresponds to $\langle \text{maj} \rangle$ (the majority function on three inputs),

- $\langle \text{HORN-SAT} \rangle = \langle \{0\}, \{1\}, x \wedge y \implies z \rangle$ corresponds to $\langle \min \rangle$ (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 1.1.9. If $\mathbb{A}, \mathbb{A}'$ are two algebraic structures on the same domain such that every basic operation of $\mathbb{A}'$ is in $\operatorname{Clo}(\mathbb{A})$, then we say that $\mathbb{A}'$ is a *reduct* of $\mathbb{A}$ and that $\mathbb{A}$ is an *expansion* of $\mathbb{A}'$. If $\operatorname{Clo}(\mathbb{A}) = \operatorname{Clo}(\mathbb{A}')$, then $\mathbb{A}$ and $\mathbb{A}'$ 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 [122] (see also chapter II.3 of [101]). It is known that on a domain of size ≥ 3, there are uncountably many clones [140], [141] (see also chapter II.8 of [101]). In particular, we see that most clones and relational clones can't be generated by finitely many functions or relations.

Definition 1.1.10. A clone $\mathcal{O}$ is said to be *finitely generated* if there is a finite set S of operations such that $\mathcal{O} = \langle S \rangle$. It is said to be *finitely related* if there is a finite set of relations Γ such that $\mathcal{O} = \mathrm{Pol}(\Gamma)$.

Example 1.1.3. The clone on $\{0,1\}$ generated by the binary implication function $\rightarrow$, given by $\rightarrow(x,y) = \neg x \vee y$, 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 $\langle \rightarrow \rangle$, but every way of identifying two coordinates of t_2^n gives a function which is in $\langle \rightarrow \rangle$ (exercise: why does this prove that $\langle \rightarrow \rangle$ can not be finitely related?). $\mathrm{Inv}(\rightarrow)$ is generated by the infinite sequence of relations $R_1, R_2, ...$ given by $R_n = \{0,1\}^n \setminus \{(0, ..., 0)\}$, and $\langle \rightarrow \rangle$ consists of all functions of the form $f(x_1, ..., x_n) \vee x_i$.

Matthew Moore [112] has shown that determining whether a given finitely generated clone is finitely related is a Turing-complete problem, and therefore undecidable in general. It is conjectured that determining whether a given finitely related clone is finitely generated is also undecidable in general.

Remark 1.1.2. The Galois connection between relational clones and clones on a finite set was originally discovered by Geiger in 1968 [64], and Reinhard Pöschel investigated the general case (where the domain may be infinite) in [121] - 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 [77].)

Remark 1.1.3. The Galois connection presented here, between operations and relations, can be viewed as being induced by the two-sorted preservation relation $\rhd$. 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

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

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

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

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 $\rhd$, 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 $G \circ F = \mathrm{Pol} \circ \mathrm{Inv}$ and $F \circ G = \mathrm{Inv} \circ \mathrm{Pol}$ 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).

1.2 Three basic examples

We start with the correspondence between 2SAT and majority.

Theorem 1.2.1. *Suppose that a relation $R \subseteq \{0, 1\}^m$ is preserved by the majority function $\mathrm{maj} : \{0, 1\}^3 \to \{0, 1\}$. 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 ith. We will show that R is equivalent to

$$\Phi(x_1, ..., x_m) := R_1(x_2, x_3, ..., x_m) \wedge R_2(x_1, x_3, ..., x_m) \wedge R_3(x_1, x_2, ..., x_m),$$

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

Definition 1.2.2. 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 1.2.3. *Suppose that a function* $f : \{0,1\}^k \to \{0,1\}$ *is monotone and self-dual. Then* $f \in \langle \mathrm{maj} \rangle$.

Proof. We prove this by induction on k. It's easy to check that there are no monotone self-dual functions of arity ≤ 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 $\langle \mathrm{maj} \rangle$. We claim that we have

$$f(x, y, z, ...) = \mathrm{maj}(f(x, y, y, ...), f(z, y, z, ...), f(x, x, z, ...)),$$

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

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

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

Examining the proof, we see that every n-ary monotone self-dual function f can be written in terms of maj as a term of depth at most $n - 2$, such that every subterm is obtained by identifying some of the variables of f.

Corollary 1.2.4. *For any odd n, the n-ary function m_n given by*

$$m_n(x_1, ..., x_n) := \begin{cases} 1 & \sum_i x_i > \frac{n}{2}, \\ 0 & \sum_i x_i < \frac{n}{2} \end{cases}$$

is in the clone generated by maj. *In fact, we may write m_n as a term of depth at most $n - 2$, such that every subterm is also a linear threshhold function, where for $a \in \mathbb{N}^n$ with $\sum_i a_i = n$ we define the n-ary linear threshhold function t_a by*

$$t_a(x_1, ..., x_n) := \begin{cases} 1 & \sum_i a_i x_i > \frac{n}{2}, \\ 0 & \sum_i a_i x_i < \frac{n}{2}. \end{cases}$$

For the majority function m_n, we can actually find a substantially smaller term using a probabilistic construction. (A deterministic construction, based on sorting networks, can be found in [92].)

Proposition 1.2.5 (Valiant [137])**.** *For any odd n, the majority function m_n can be represented by a term of depth $O(\log(n))$ and size $O(n^{4.3})$.*

Proof. We'll follow Goldreich's exposition [65]. Consider the completely generic formula $f_\ell(y_1, ..., y_{3^\ell})$ of depth ℓ, defined recursively by $f_0 = \pi_1$, $f_1 = \text{maj}$, and

$$f_{\ell+1}(y) := \text{maj}(f_\ell(y_1, ..., y_{3^\ell}), f_\ell(y_{3^\ell+1}, ..., y_{2\cdot 3^\ell}), f_\ell(y_{2\cdot 3^\ell+1}, ..., y_{3^{\ell+1}})).$$

Then define a random function $g_\ell(x_1, ..., x_n)$ by replacing each y_i in f_ℓ with a random choice of x_{j_i}, where the j_i are independently and uniformly randomly chosen from the set $\{1, ..., n\}$. For any particular input $x \in \{0, 1\}^n$, if p_i is the probability that $g_i(x) = m_n(x)$, then we have

$$p_0 \geq \frac{1}{2} + \frac{1}{2n}$$

and

$$\begin{aligned}
p_{i+1} &= 3(1-p_i)p_i^2 + p_i^3 \\
&= 0.5 + (1.5 - 2(p_i - 0.5)^2)(p_i - 0.5) \\
&= 1 - (3 - 2(1 - p_i))(1 - p_i)^2.
\end{aligned}$$

A little computation then shows that for $\ell \approx (1 + 1/\log_2(1.5))\log_2(n) \approx 2.71\log_2(n)$ we have $1 - p_\ell < 2^{-n}$, so a union bound shows that for this choice of ℓ at least one assignment to the y_is has $g_\ell(x) = m_n(x)$ for all $x \in \{0, 1\}^n$. $\qquad\square$

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 1.2.6. 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 $A \cap B \neq 0$.

Proposition 1.2.7. *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 $\{1, ..., n\}$ 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 1.2.8. For every n, we define an undirected graph $\mathcal{M}_n$ whose vertices are the maximal intersecting families of subsets of $\{1, ..., n\}$, and whose edges are the pairs of families $\mathcal{F}, \mathcal{G}$ such that $|\mathcal{F} \setminus \mathcal{G}| = 1$.

The graph $\mathcal{M}_4$ is depicted below, with vertices labeled by the minimal elements of the corresponding intersecting families as well as the corresponding monotone self-dual functions (written in terms of the majority function, which we abbreviate as m).

26

$\{2\}$
y

$\{1,2\},\{1,3\},\{2,3\}$
$m(x,y,z)$

$\{1,2\},\{2,3\},\{2,4\},\{1,3,4\}$
$m(x,y,m(y,z,w))$

$\{1,3\},\{2,3\},\{3,4\},\{1,2,4\}$
$m(z,w,m(x,y,z))$

$\{2,3\},\{2,4\},\{3,4\}$
$m(y,z,w)$

$\{3\}$
z

$\{4\}$
w

$\{1,3\},\{1,4\},\{3,4\}$
$m(x,z,w)$

$\{1,4\},\{2,4\},\{3,4\},\{1,2,3\}$
$m(z,w,m(x,y,w))$

$\{1,2\},\{1,3\},\{1,4\},\{2,3,4\}$
$m(x,y,m(x,z,w))$

$\{1,4\},\{2,4\},\{3,4\}$
$m(x,y,w)$

$\{1\}$
x

The graph $\mathcal{M}_n$ is always connected: given two maximal intersecting families $\mathcal{F},\mathcal{G}$, there will always be some minimal element of $\mathcal{F}$ which is not contained in $\mathcal{G}$, and switching this set with its complement gives us a maximal intersecting family $\mathcal{F}'$ which is adjacent to $\mathcal{F}$ and has one more element in common with $\mathcal{G}$ than $\mathcal{F}$ does. For more about maximal intersecting families of sets, see [111].

Next we move to the correspondence between HORN-SAT and the minimum operation.

Theorem 1.2.9. *Suppose that a relation $R \subseteq \{0,1\}^m$ is preserved by the minimum function* $\min : \{0,1\}^2 \to \{0,1\}$. *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 $C_i = x_1 \vee \cdots \vee x_p \vee \bar{x}_{p+1} \vee \cdots \vee \bar{x}_{p+q}$, $p \geq 2$. 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 1.2.10. *Suppose that a function $f : \{0,1\}^k \to \{0,1\}$ preserves the relations $\{0\}, \{1\}$, and $x \wedge y \implies z$. Then there is a nonempty subset $I \subseteq \{1,...,k\}$ such that $f(x_1,...,x_k) = \min_{i \in I} x_i$.*

Proof. Since f preserves $\{0\}$ and $\{1\}$, we have $f(0,...,0) = 0$ and $f(1,...,1) = 1$. Since $\leq$ is in the relational clone generated by $x \wedge y \implies z$, f must be monotone.

For each subset $I \subseteq \{1,...,k\}$, let χ_I be the indicator vector of I. Suppose that I, J have $f(\chi_I) = f(\chi_J) = 1$, then from $\chi_I \wedge \chi_J \implies \chi_{I \cap J}$ (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 $f(\chi_J) = 1 \iff J \supseteq I^*$. $\square$

27

Remark 1.2.1. The fact that min-closed relations on the domain $\{0, 1\}$ can always be written as intersections of Horn clauses has the following useful consequence in logic.

Suppose that $P_1, ..., P_m$ is a list of logical statements about some type of structure M in some collection of structures $\mathcal{M}$. Suppose that for every pair of structures $M_1, M_2 \in \mathcal{M}$ there is a structure $M' \in \mathcal{M}$ such that for each i, P_i holds in M' iff P_i holds in both M_1 and M_2. Then there is a collection of Horn clauses $\phi_1, ..., \phi_n$ in the propositions $P_1, ..., P_m$ such that an assignment of true/false values to the P_is can be realized by some $M \in \mathcal{M}$ iff the assignment satisfies the collection of Horn-clauses $\phi_1, ..., \phi_n$.

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

Theorem 1.2.11. *Suppose that a relation $R \subseteq (\mathbb{Z}/p)^m$ is preserved by the ternary operation $x - y + z$ (mod 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 1.2.12. *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 $\langle x - y + z \pmod{p} \rangle = \langle \frac{x+y}{2} \pmod{p} \rangle$.*

1.3 Varieties, Birkhoff's HSP theorem, and the hardness proof

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

Remark 1.3.1. Suppose $\mathbb{A}, \mathbb{B}$ are two algebraic structures with associated relational structures $\mathbf{A}, \mathbf{B}$. It is tempting to think that a homomorphism $\mathbb{A} \to \mathbb{B}$ will correspond to a homomorphism $\mathbf{A} \to \mathbf{B}$, or vice versa. Unfortunately, this is total nonsense - if the (functional) signatures of $\mathbb{A}$ and $\mathbb{B}$ match, the (relational) signatures of $\mathbf{A}$ and $\mathbf{B}$ will likely have nothing to do with each other!

In a similar vein, the automorphism groups $\text{Aut}(\mathbb{A})$ and $\text{Aut}(\mathbf{A})$ have almost nothing to do with eachother. A trivial but illuminating example is the case where $\mathbb{A}$ has no functions at all, so that $\text{Aut}(\mathbb{A})$ is the full symmetric group - in this case, $\mathbf{A}$ 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 $\mathbf{A}$ is *rigid*, with $\text{Aut}(\mathbf{A}) = \{1\}$.

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 1.3.1. *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 $\mathbb{B}$ is a subalgebra of $\mathbb{A}$, we can convert any instance of $\text{CSP}(\mathbb{B})$ into an instance of $\text{CSP}(\mathbb{A})$ by simply adding an extra unary constraint for each variable corresponding to the relation $\mathbb{B} \subseteq \mathbb{A}^1$.

If $\mathbb{B} = \mathbb{A}^n$ for some n, then we can convert an instance of $\text{CSP}(\mathbb{B})$ to an instance of $\text{CSP}(\mathbb{A})$ by replacing each variable with an n-tuple of variables, and using the fact that every subalgebra of $(\mathbb{A}^n)^m$ is a subalgebra of $\mathbb{A}^{mn}$.

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

More generally, if we have several algebras $\mathbb{A}_1, \mathbb{A}_2, ...$ in the same (functional) signature, we can define $\text{CSP}(\{\mathbb{A}_1, \mathbb{A}_2, ...\})$ to be the problem where each variable comes with a *sort* - that is, a specific algebra $\mathbb{A}_i$ 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 $\mathbb{A}_i$s. This sort of multisorted relation was considered by Bulatov and Jeavons [37]. In this framework, there is a logspace equivalence between $\text{CSP}(\mathbb{A}_1 \times \mathbb{A}_2)$ and $\text{CSP}(\{\mathbb{A}_1, \mathbb{A}_2\})$.

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 1.3.2. 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 $\mathbb{A}$ *satisfies* an identity $s \approx t$, written $\mathbb{A} \models s \approx t$, if

$$\forall x_1, ..., x_n \in \mathbb{A} \; s(x_1, ..., x_n) = t(x_1, ..., x_n)$$

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

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 1.3.3. The *variety* $\mathcal{V}(\mathcal{T})$ determined by a set of identities $\mathcal{T}$ is the set of algebras that satisfy all of the identities in $\mathcal{T}$. If $\mathbb{A}_1, \mathbb{A}_2, ...$ is a collection of algebras, then $\mathcal{V}(\mathbb{A}_1, \mathbb{A}_2, ...)$ is the variety associated to the set of all identities that hold simultaneously in all of the algebras $\mathbb{A}_i$.

Birkhoff [26] introduced a convenient notation for manipulating sets (strictly speaking these are classes, not sets) of algebras: if $\mathcal{S}$ is a set of algebras, then $P\mathcal{S}$ is the set of products of algebras from $\mathcal{S}$ (possibly infinite - P_{fin} is the notation if one restricts to finite products), $S\mathcal{S}$ is the set of subalgebras of algebras from $\mathcal{S}$, and $H\mathcal{S}$ is the set of quotients (homomorphic images) of algebras from $\mathcal{S}$.

Theorem 1.3.4 (Birkhoff's HSP Theorem [26]). *For any collection $\mathcal{S}$ of algebras, we have $\mathcal{V}(\mathcal{S}) = HSP(\mathcal{S})$, that is, an algebra $\mathbb{A}$ satisfies every identity which is satisfied in every element of $\mathcal{S}$ if and only if it is the homomorphic image of a subalgebra of a product of elements of $\mathcal{S}$. Furthermore, if $\mathcal{S}$ is a finite collection of finite algebras, then the set of finite algebras in $\mathcal{V}(\mathcal{S})$ is equal to $HSP_{fin}(\mathcal{S})$.*

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

For the other containment, suppose that $\mathbb{A} \in \mathcal{V}(\mathcal{S})$, and suppose that $\mathbb{A}$ is generated by a subset $I \subseteq \mathbb{A}$. We let $\mathbb{P}$ be the product of all the algebras of $\mathcal{S}$, and define the "free algebra" $\mathcal{F}(I)$ be the subalgebra of $\mathbb{P}^{\mathbb{P}^I}$ which is generated by the projection functions π_i for $i \in I$, given by $\pi_i(x) = x_i$. We claim that there is a surjective homomorphism $h : \mathcal{F}(I) \twoheadrightarrow \mathbb{A}$ 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 $\mathbb{P}$, but $s(i_1, ..., i_n) \neq t(i_1, ..., i_n)$ in $\mathbb{A}$. But then $s \approx t$ is satisfied by $\mathbb{P}$, and hence by every algebra in $\mathcal{S}$, and is not satisfied in $\mathbb{A}$, contradicting our assumption that $\mathbb{A} \in \mathcal{V}(\mathcal{S})$.

For the last claim, note that if $\mathbb{A}, \mathcal{S}$, and every element of $\mathcal{S}$ are finite, then so are $I, \mathbb{P}, \mathbb{P}^I$, and $\mathbb{P}^{\mathbb{P}^I}$. $\square$

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 1.3.5. *If $\mathcal{T}$ is a family of identities, then the set of identities which hold in $\mathcal{V}(\mathcal{T})$ is equal to the closure of $\mathcal{T} \cup \{x \approx x\}$ 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 $\mathcal{F}_{\mathcal{T}}(x_1, ...)$ 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 $\mathcal{T}$. The result will be an algebra satisfying all of the identities of $\mathcal{T}$, 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. $\square$

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

Theorem 1.3.6. *If $CSP(\mathbb{A})$ is not NP-complete, then there is a finite set of identities $s_i \approx t_i$ which are satisfied by $\mathbb{A}$, which can't be satisfied by assigning each function symbol to a projection of the same arity.*

Proof. If $\mathcal{V}(\mathbb{A})$ contains an algebra $\mathbb{B}$ of size at least 2 where each function symbol acts as a projection, then $CSP(\mathbb{B})$ is NP-complete and has a logspace reduction to $CSP(\mathbb{A})$. Such an algebra $\mathbb{B}$ will exist if there is a way to assign the function symbols to projections that satisfies *every* identity satisfied by $\mathbb{A}$. 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 Kőnig's Lemma). $\square$

Example 1.3.1. Consider the algebra $\mathbb{A} = (\{0, 1\}, \min)$, and use the binary function symbol s to abbreviate min. Then $\mathcal{V}(\mathbb{A}) = SP(\mathbb{A}) = \mathcal{V}(\mathcal{T}_{semi})$, where $\mathcal{T}_{semi}$ 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 π_1, π_2.

An algebra in $\mathcal{V}(\mathcal{T}_{semi})$ 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 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 1.3.2. Consider the algebra $\mathbb{A} = (\{0, 1\}, \mathrm{maj})$, and use the ternary function symbol m to abbreviate maj. Then $\mathcal{V}(\mathbb{A}) = SP(\mathbb{A}) = \mathcal{V}(\mathcal{T}_{med})$, where $\mathcal{T}_{med}$ 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 π_1, π_2, π_3.

An algebra in $\mathcal{V}(\mathcal{T}_{med})$ 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 $\mathcal{M}_n$ of maximal intersecting families from the last section. For more about the theory of median algebras, see [127].

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:

$$(x \wedge y) \vee (y \wedge z) \vee (z \wedge x) \approx (x \vee y) \wedge (y \vee z) \wedge (z \vee x),$$

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 $\mathcal{T}_{med}$. 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 $x \wedge y = m(0, x, y)$ and $x \vee y = m(x, y, 1)$, and the median operation m can be recovered from $\wedge, \vee$ via the above formula [29].

Example 1.3.3. The operation $m(x, y, z) = x - y + z \pmod{p}$ 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 π_1, π_2, π_3.

Similarly, if p is odd, the operation $m(x, y) = \frac{x+y}{2} \pmod{p}$ 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 1.3.7. A variety $\mathcal{V}$ is *finitely generated* if there is a finite list of finite algebras $\mathbb{A}_1, ..., \mathbb{A}_n$ such that $\mathcal{V} = \mathcal{V}(\mathbb{A}_1, ..., \mathbb{A}_n)$. A variety $\mathcal{V}$ is *locally finite* if the free algebra on n generators $\mathcal{F}_{\mathcal{V}}(x_1, ..., x_n)$ is finite for every n. A variety $\mathcal{V}$ is *finitely based* if there is a finite set of equations $\mathcal{T}$ such that $\mathcal{V} = \mathcal{V}(\mathcal{T})$.

A variety $\mathcal{V}$ is locally finite iff for all $\mathbb{A} \in \mathcal{V}$ and for all finite subsets $\{a_1, ..., a_n\} \subseteq \mathbb{A}$, the subalgebra of $\mathbb{A}$ generated by $a_1, ..., a_n$ 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 $x^n \approx e$ is locally finite.

Remark 1.3.2. 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 [59] show that a pseudovariety is determined by an infinite sequence of identities, such that a finite algebra is contained in the pseudovariety iff it satisfies *all but finitely many* of the identities in the sequence. The trick is to sort the isomorphism classes of finite algebras by their sizes, and for each size k write down a finite set of identities in k variables which characterizes the free algebra on k generators in the subvariety generated by the set of algebras of size at most k.

Reiterman [126] 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^{∞} 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 [126] 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.

1.4 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 1.4.1. Two relational structures $\mathbf{A}, \mathbf{B}$ with the same signature are *homomorphically equivalent* if there exist homomorphisms $\mathbf{A} \to \mathbf{B}, \mathbf{B} \to \mathbf{A}$.

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 in the book by Hobby and McKenzie [74].

Example 1.4.1. Consider the relational structure $\mathbf{A}$ corresponding to the binary implication algebra $\mathbb{A} = (\{0,1\}, \to)$. This relational structure has as basic relations $R_n = \{0,1\}^n \setminus \{(0,...,0)\} = x_1 \vee \cdots \vee x_n$. The unary algebraic operation $\to(x,x)$ of $\mathbb{A}$ takes every element to 1, and defines an endomorphism of relational structures $\mathbf{A} \to \mathbf{A}$ whose image is $\{1\}$. Together with the inclusion relation, we get a homomorphic equivalence between $\mathbf{A}$ and the one-element relational structure with domain $\{1\}$ and relations $R_n \mid_{\{1\}^n} = \{1\}^n$, whose CSP is clearly trivial.

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

Definition 1.4.2. A relational structure on a finite domain $\mathbf{A}$ is called a *core* if every endomorphism of $\mathbf{A}$ is also an automorphism of $\mathbf{A}$. If $\mathbf{A}$ is not a core, then $\mathbf{B}$ is called a *core of* $\mathbf{A}$ if $\mathbf{B}$ is a core and $\mathbf{B}$ is homomorphically equivalent to $\mathbf{A}$.

Remark 1.4.1. 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 $(\mathbb{Q}, <)$. See section 3.6 of [31] for more information about cores of infinite structures.

Proposition 1.4.3. *Every relational structure on a finite domain $\mathbf{A}$ has a core. Any two cores of $\mathbf{A}$ are isomorphic.*

Proof. The first statement follows directly from induction on the size of $\mathbf{A}$: if $\mathbf{A}$ 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 $\mathbf{B}, \mathbf{B}'$ are two cores of $\mathbf{A}$ then they are homomorphically equivalent, and composing the maps $\mathbf{B} \to \mathbf{B}'$, $\mathbf{B}' \to \mathbf{B}$ gives us endomorphisms of $\mathbf{B}, \mathbf{B}'$ which must both be invertible by the definition of a core. $\square$

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

Problem 1.4.1. Given a finite relational structure $\mathbf{A}$ as input, determine whether or not $\mathbf{A}$ 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 1.4.4 (Hell, Nešetřil [71])**.** *Determining whether a given undirected graph is a core is 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 1.4.5. 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 1.4.6. *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: $\{a\} \in \langle \mathbf{A} \rangle$ iff $\{a\}$ is closed under $\mathrm{Pol}(\mathbf{A})$, and since $\{a\}$ is generated by a single element, we only need to check that it is closed under $\mathrm{Pol}_1(\mathbf{A})$, 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 $(a_1, ..., a_m) \in R$, we impose the constraint $R(f_{a_1}, ..., f_{a_m})$ 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 $\mathrm{CSP}(\{1, ..., k\}, \neq)$ (the k-coloring problem) is logspace equivalent to $\mathrm{CSP}(\{1, ..., k\}, \neq, \{0\}, ..., \{k\})$ - 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 1.4.7. *Suppose that $\mathbf{A} = (D, \Gamma)$ is a core on a finite domain D, and let $\mathbf{A}^{rig}$ be the rigid core obtained by adjoining all singleton unary relations to **A**. Then $\mathrm{CSP}(\mathbf{A})$ is equivalent to $\mathrm{CSP}(\mathbf{A}^{rig})$ under logspace reductions.*

Proof. We need to find a way to convert an instance of $\mathrm{CSP}(\mathbf{A}^{rig})$ to an instance of $\mathrm{CSP}(\mathbf{A})$ 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 Φ by

$$\Phi(f) \coloneqq \bigwedge_{R \in \Gamma} \bigwedge_{(a_1, ..., a_m) \in R} R(f_{a_1}, ..., f_{a_m}).$$

34

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

$$\Psi(x) = \exists x_{n+1}, ..., x_{n+m} \ \Psi_0(x) \wedge \bigwedge_{(i,a)\in E} x_i \in \{a\},$$

where Ψ_0 is a primitive positive formula using the relations of Γ, and E is a set describing the additional unary singleton constraints of Ψ. Let Ψ' be the following formula:

$$\Psi'(x) := \exists f \ \exists x_{n+1}, ..., x_{n+m} \ \Phi(f) \wedge \Psi_0(x) \wedge \bigwedge_{(i,a)\in E} x_i = f_a.$$

We claim that the instance Ψ' of $\mathrm{CSP}(\mathbf{A})$ has a solution iff the instance Ψ of $\mathrm{CSP}(\mathbf{A}^{rig})$ has a solution. Suppose that f, x solves Ψ', then by the construction of $\Phi(f)$ f describes an endomorphism $f : \mathbf{A} \to \mathbf{A}$, and since $\mathbf{A}$ is a core this endomorphism must have an inverse f^{-1}. Then $f^{-1}(x)$ satisfies Ψ_0 (since f^{-1} is an endomorphism of $\mathbf{A}$), 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 $\Psi(x)$. $\qquad\square$

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

Definition 1.4.8. A function $f : D^k \to D$ is *idempotent* if it satisfies the identity $f(x, x, ..., x) \approx x$. An algebraic structure $\mathbb{A} = (D, \mathcal{O})$ is *idempotent* if every $f \in \mathcal{O}$ is idempotent. Equivalently, $\mathbb{A}$ is idempotent if every singleton subset of D is a subalgebra of $\mathbb{A}$.

Definition 1.4.9. If $\mathbb{A} = (D, \mathcal{O})$ is an algebraic structure, then the *idempotent reduct* $\mathbb{A}^{id}$ of $\mathbb{A}$ has the same domain, and has as its operations the set of all idempotent functions $f \in \langle \mathcal{O} \rangle$ (or, alternatively, some smaller generating set of idempotent functions).

Example 1.4.2. If $\mathbb{A} = (\mathbb{Z}/p, +, 0, 1)$, then $\mathbb{A}^{id}$ 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).

Proposition 1.4.10. *If $\mathbf{A}$ is a core corresponding to the algebraic structure $\mathbb{A}$, then the rigid core $\mathbf{A}^{rig}$ corresponds to the idempotent reduct $\mathbb{A}^{id}$. In particular, every CSP is equivalent up to logspace reductions to $\mathrm{CSP}(\mathbb{A})$ for some idempotent algebra $\mathbb{A}$.*

The reader might be worried that there is no obvious way to generate the collection of all idempotent operations contained in a given clone. For core structures this is not an issue: the polymorphisms of a core structure always decompose neatly into idempotent parts and invertible unary parts.

Proposition 1.4.11. *Suppose that $\mathcal{O}$ is a clone such that all of the unary operations in $\mathcal{O}$ are invertible. Then for every k-ary function $f \in \mathcal{O}$, if we define the unary function f_{un} by*

$$f_{un}(x) := f(x, ..., x)$$

and the k-ary function f_{id} by

$$f_{id}(x_1, ..., x_k) := f_{un}^{-1}(f(x_1, ..., x_k)),$$

then f_{id} is idempotent and

$$f = f_{un} \circ f_{id}.$$

In particular, if G is the group of unary operations in $\mathcal{O}$, then for every k there are precisely $|G|$ times as many k-ary operations in $\mathcal{O}$ as there are k-ary idempotent operations in $\mathcal{O}$.

If $\mathcal{O}$ is generated by the functions $f_1, ..., f_m$ of arities $k_1, ..., k_m$, then the set of idempotent operations of $\mathcal{O}$ is generated by the functions

$$(f_i \circ (g_1, ..., g_{k_i}))_{id},$$

over all choices of i and all choices of $g_1, ..., g_{k_i} \in G$. In particular, the set of idempotent operations of $\mathcal{O}$ is finitely generated if and only if the full clone $\mathcal{O}$ is finitely generated.

Example 1.4.3. There is an example of a core structure $\mathbf{A}$ which has polymorphisms satisfying a nontrivial system of identities, but such that its rigidification $\mathbf{A}^{rig}$ has no such polymorphisms and is therefore NP-complete. This example is due to Ross Willard and can be found in [24].

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

$$R(a_i, b_j) := (i = j) \wedge (a \neq b),$$
$$S(a_i, b_j) := i \neq j.$$

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

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

$$s(a_i, b_j, c_k) = \begin{cases} c_k & i = j, \\ a_i & i \neq j. \end{cases}$$

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 $\mathbf{A}^{rig}$, we see that polymorphisms of $\mathbf{A}^{rig}$ 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

$$f(1, 1, ..., 1) = 1, f(2, 1, ..., 1) = 2.$$

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

$$f(3, 2, ..., 2) = f(3, 3, ..., 3) = 3.$$

These imply that

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

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

$$f(3, z_2, ..., z_n) = 3$$

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

Alternatively, we could have shown that $\text{Pol}(K_3^{rig})$ 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 $(x = a) \wedge (y = b) \implies (z = c)$, it's easy to construct the rest).

1.4.1 Reflections and Height 1 Identities

Barto, Opršal, and Pinsker [23] 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 $\text{CSP}(\mathbf{A})$ and $\text{CSP}(\mathbf{B})$ introduced so far can be described by combining just two basic cases:

- if $\mathbf{B}$ is a "pp-power" (defined below) of $\mathbf{A}$, then $\text{CSP}(\mathbf{B})$ has a logspace reduction to $\text{CSP}(\mathbf{A})$, and

- if $\mathbf{B}$ is homomorphically equivalent to $\mathbf{A}$ then $\text{CSP}(\mathbf{B}) = \text{CSP}(\mathbf{A})$.

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

Definition 1.4.12. A *pp-power* of a relational structure $\mathbf{A}$ is a relational structure $\mathbf{B}$ with domain $\mathbf{A}^n$ for some n, such that every relation of $\mathbf{B}$ can be defined by a primitive positive formula using the relations of $\mathbf{A}$ (note that the signatures of $\mathbf{A}$ and $\mathbf{B}$ will generally be different).

Proposition 1.4.13. *If $\mathbf{B}$ is homomorphically equivalent to a pp-power of $\mathbf{A}$, then there is a reduction from $\text{CSP}(\mathbf{B})$ to $\text{CSP}(\mathbf{A})$ which can be computed in linear time and logarithmic space.*

Definition 1.4.14. We say that $\mathbf{A}$ *pp-constructs* $\mathbf{B}$ if $\mathbf{B}$ is homomorphically equivalent to some pp-power of $\mathbf{A}$.

For instance, here is how we can go about adding a singleton unary relation $\{a\}$ to a core $\mathbf{A}$ in the pp-constructability framework. Let $\mathbf{B}$ be the relational structure which has the new unary relation $\{a\}$ (along with all of the original relations which $\mathbf{A}$ had). We will define a relational structure $\mathbf{C}$ which will be a pp-power of $\mathbf{A}$ having domain $\mathbf{A}^2$, and show that $\mathbf{C}$ is homomorphically equivalent to $\mathbf{B}$.

Let O be the orbit of a under $\text{Aut}(\mathbf{A})$ - note that O is in the relational clone defined by $\mathbf{A}$ - and for every m-ary relation R of $\mathbf{A}^2$, make a corresponding relation $\tilde{R}$ of $\mathbf{C}$ by

$$((x_1, y_1), ..., (x_m, y_m)) \in \tilde{R} \iff (x_1, ..., x_m) \in R \wedge y_1 = \cdots = y_m \in O.$$

For the relation $\{a\}$ of $\mathbf{B}$, we make a corresponding relation S of $\mathbf{C}$ given by

$$(x, y) \in S \iff x = y \in O.$$

To show that $\mathbf{B}$ and $\mathbf{C}$ are homomorphically equivalent, we just need to exhibit a pair of homomorphisms between them. The homomorphism $\mathbf{B} \to \mathbf{C}$ is given by $x \mapsto (x, a)$. To define the

homomorphism from $\mathbf{C}$ to $\mathbf{B}$, we need to choose an automorphism g_y of $\mathbf{A}$ with $g_y(y) = a$ for every $y \in O$. Then the homomorphism $\mathbf{C} \to \mathbf{B}$ is given by $(x, y) \mapsto g_y(x)$ if $y \in O$ (and (x, y) maps to an arbitrary element if $y \notin O$).

Barto, Opršal, and Pinsker [23] also characterize what happens on the algebraic side of the picture when one relates two relational structures by a pp-power or a homomorphic equivalence. The new thing here is really the homomorphic equivalence: if $g : \mathbf{A} \to \mathbf{B}$ and $h : \mathbf{B} \to \mathbf{A}$, then there is a relationship between $\mathrm{Pol}(\mathbf{A})$ and $\mathrm{Pol}(\mathbf{B})$ which they call a *reflection*, which takes a function $f \in \mathrm{Pol}_k(\mathbf{A})$ to the operation

$$\xi(f) : (x_1, ..., x_k) \mapsto g(f(h(x_1), h(x_2), ..., h(x_k)))$$

in $\mathrm{Pol}_k(\mathbf{B})$. Note that ξ does not respect composition: $\xi(f_0 \circ (f_1, ..., f_k))$ is not in general equal to $\xi(f_0) \circ (\xi(f_1), ..., \xi(f_k))$. However, ξ *does* preserve *height 1 identities*.

Definition 1.4.15. 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_is and y_js are (not necessarily distinct) variables. A map $\mathrm{Pol}(\mathbf{A}) \to \mathrm{Pol}(\mathbf{B})$ (taking functions to functions) which respects height 1 identities is called a *height 1 clone homomorphism* or a *minion homomorphism*.

Definition 1.4.16. 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 $g \circ f \circ (h, ..., h)$ on B corresponding to the operation $f \in \mathcal{O}$.

Proposition 1.4.17. $\mathbf{B}$ *is homomorphically equivalent to a pp-power of* $\mathbf{A}$ *iff* $\mathrm{Pol}(\mathbf{B})$ *contains a reflection of* $\mathrm{Pol}(\mathbf{A})^n$ *for some* n *(by* $\mathrm{Pol}(\mathbf{A})^n$ *we mean the clone of operations of* $\mathrm{Pol}(\mathbf{A})$ *acting on a power of the* domain*).*

Proof. We prove the non-obvious direction. Let A, B be the underlying sets of $\mathbf{A}, \mathbf{B}$, and suppose that $g : A^n \to B$ and $h : B \to A^n$ induce a reflection $\xi : \mathrm{Pol}(\mathbf{A})^n \to \mathrm{Pol}(\mathbf{B})$. We will construct a pp-power $\mathbf{C}$ of $\mathbf{A}$ with underlying set A^n which is homomorphically equivalent to $\mathbf{B}$. For every relation R of $\mathbf{B}$, let $\tilde{R}$ be the relation

$$\tilde{R} := \{f(h(r_1), ..., h(r_k)) \mid f \in \mathrm{Pol}_k(\mathbf{A}), r_1, ..., r_k \in R\}.$$

By definition, $\tilde{R}$ is the closure of $h(R)$ under $\mathrm{Pol}(\mathbf{A})$, so $\tilde{R}$ is defined by a primitive positive formula over $\mathbf{A}$. We use $\tilde{R}$ as the relation corresponding to R in $\mathbf{C}$. Finally, we just need to check that $g : \mathbf{C} \to \mathbf{B}$ and $h : \mathbf{B} \to \mathbf{C}$ are homomorphisms. That h is a homomorphism follows from $h(R) \subseteq \tilde{R}$. To check that g is a homomorphism, note that if $x = f(h(r_1), ..., h(r_k)) \in \tilde{R}$ with $r_1, ..., r_k \in R$, then $g(x) = \xi(f)(r_1, ..., r_k)$ is an element of R since $\xi(f) \in \mathrm{Pol}(\mathbf{B})$ by assumption. $\square$

Theorem 1.4.18 (ERP Theorem [23]). $\mathrm{Pol}(\mathbf{B})$ *contains a reflection of* $\mathrm{Pol}(\mathbf{A})^n$ *for some* n *iff there is a height 1 clone homomorphism* $\mathrm{Pol}(\mathbf{A}) \to \mathrm{Pol}(\mathbf{B})$.

Proof. We prove the non-obvious direction. Let $\mathbb{A} = (A, \mathrm{Pol}(\mathbf{A}))$ be the algebraic structure corresponding to $\mathbf{A}$, and suppose $\xi : \mathrm{Pol}(\mathbf{A}) \to \mathrm{Pol}(\mathbf{B})$ is a height 1 clone homomorphism. Let $\mathcal{F}$ be the subalgebra of $\mathbb{A}^{\mathbb{A}^B}$ generated by the functions $\pi_b : \mathbb{A}^B \to \mathbb{A}$ given by $\pi_b : f \mapsto f(b)$. Note that $\mathcal{F}$ is secretly the free algebra over $\mathbb{A}$ on $|B|$ generators.

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

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

Remark 1.4.2. A height 1 clone homomorphism $\mathrm{Pol}(\mathbf{A}) \to \mathrm{Pol}(\mathbf{B})$ is completely determined by its restriction to polymorphisms of $\mathbf{A}$ of arity at most $|\mathbf{B}|$. So there are only finitely many candidates for maps that show $\mathbf{A}$ pp-constructs $\mathbf{B}$: if the underlying sets are A, B, then there are at most $\left| B^{\mathrm{Pol}_{|B|}(\mathbf{A})} \right| \le \left| B^{A^{A^B}} \right|$ candidates. If $\mathbf{B}$ has only finitely many basic relations, then we can test each candidate pp-construction in finite time, so whether or not $\mathbf{A}$ pp-constructs $\mathbf{B}$ is decidable.

1.5 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 1.5.1. An algebra $\mathbb{A}$ is a *Taylor algebra* if it has an idempotent term t that satisfies a system of identities of the form

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

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

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

Theorem 1.5.2 (Taylor [135]). *If an idempotent algebra $\mathbb{A}$ 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 1.5.1. Let f be an idempotent ternary term satisfying the identity

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

39

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

Definition 1.5.3. If $f : D^k \to D$ and $g : D^l \to D$, we define the *star composition* $f * g : D^{kl} \to D$ to be $f \circ (g, g, ..., g)$.

Proposition 1.5.4. *If f, g are idempotent, then $f, g \in \langle f * g \rangle$.*

Proof. $f(x_1, ..., x_k) \approx f(g(x_1, ..., x_1), ..., g(x_k, ..., x_k))$ and $g(x_1, ..., x_l) \approx f(g(x_1, ..., x_l), ..., g(x_1, ..., x_l))$.
□

Definition 1.5.5. 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_is and y_js are (not necessarily distinct) variables.

Proposition 1.5.6. *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 $s = f_1 * \cdots * f_n$. 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 occuring in $\mathcal{T}'$, and let $t = s * \cdots * s$, 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
$$x^k_{(i_1, ..., i_m)} = y_{i_k}$$
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 $t = \pi_{(i, ..., i)}$.
□

Corollary 1.5.7. *If $\mathbb{A}$ is an idempotent algebra and $\mathrm{CSP}(\mathbb{A})$ is not NP-complete, then $\mathbb{A}$ has a Taylor term.*

When $\mathbb{A}$ is not Taylor, the above result lets us conclude that there is some two element algebra $\mathbb{B} \in HSP(\mathbb{A})$ with no nontrivial operations, but it doesn't give us a good bound on how large a power of $\mathbb{A}$ we will need to take to find $\mathbb{B}$. It turns out that, in fact, if such a $\mathbb{B}$ exists then it already can be found inside $HS(\mathbb{A})$. We will prove a slight generalization of this fact, which applies to *strictly simple* algebras.

Definition 1.5.8. An algebra $\mathbb{A}$ is called *strictly simple* if $\mathbb{A}$ is simple and every subalgebra of $\mathbb{A}$ either has size 1 or is equal to $\mathbb{A}$.

Lemma 1.5.9. *If $\mathbb{A}$ is an idempotent algebra and $\mathbb{B} \in HSP(\mathbb{A})$ is strictly simple, then $\mathbb{B} \in HS(\mathbb{A})$.*

Proof. (Following Zhuk [143]) Pick n minimal such that there is some $\mathbb{S} \leq \mathbb{A}^n$ and some $\sigma \in \mathrm{Con}(\mathbb{S})$ with $\mathbb{S}/\sigma \cong \mathbb{B}$. If there is any pair $r, s \in \mathbb{S}$ such that $\pi_1(r) = \pi_1(s)$ but $r/\sigma \neq s/\sigma$, then if we set $a = \pi_1(r)$ and
$$\mathbb{S}' = \pi_{\{2,\dots,n\}}(\mathbb{S} \cap (\{a\} \times \mathbb{A}^{n-1})) \leq \mathbb{A}^{n-1},$$
and define $\sigma' \in \mathrm{Con}(\mathbb{S}')$ by restricting σ in the obvious way, then $r' = \pi_{\{2,\dots,n\}}(r)$ and $s' = \pi_{\{2,\dots,n\}}(s)$ have $r'/\sigma' \neq s'/\sigma'$. Thus $\mathbb{S}'/\sigma'$ is isomorphic to a subalgebra of $\mathbb{B}$ of size at least 2, so $\mathbb{S}'/\sigma' \cong \mathbb{B}$, contradicting the minimality of n.

Otherwise, if there is no such pair r, s, then there is a congruence $\sigma_1 \in \mathrm{Con}(\pi_1(\mathbb{S}))$ such that for all $r \in \mathbb{S}$, the congruence class r/σ is completely determined by $\pi_1(r)/\sigma_1$. But then we have
$$\mathbb{B} \cong \mathbb{S}/\sigma \cong \pi_1(\mathbb{S})/\sigma_1 \in HS(\mathbb{A}). \qquad \square$$

Corollary 1.5.10. *If $\mathbb{A}$ is idempotent, then either $\mathbb{A}$ has a Taylor term, or there is some two element algebra $\mathbb{B} \in HS(\mathbb{A})$ with no nontrivial operations.*

Corollary 1.5.11. *If $\mathbb{A}$ is idempotent and has no Taylor term, then there are nonempty subalgebras $\mathbb{B}, \mathbb{C} \leq \mathbb{A}$ such that $\mathbb{B} \cap \mathbb{C} = \emptyset$ and $(\mathbb{B} \cup \mathbb{C})^3 \setminus (\mathbb{B}^3 \cup \mathbb{C}^3) \leq \mathbb{A}^3$. In particular, $\mathrm{CSP}(\mathbb{A})$ can simulate NAE-SAT in a trivial way.*

Remark 1.5.1. A recent result of Olšák simplifies the identities we need to consider even further. Olšák [113] proves that in any Taylor algebra, whether finite or infinite, there is always a 6-ary *weak 3-cube* term t, that is, an idempotent term satisfying the identity
$$t(x, y, y, y, x, x) \approx t(y, x, y, x, y, x) \approx t(y, y, x, x, x, y).$$

The weak 3-cube term may be understood as saying that the ternary relation on the free algebra $\mathcal{F}_\mathcal{V}(x, y)$ 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 1.5.2. 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 $\mathcal{T}$ on ternary functions $f_1, \dots, f_n$, and that you want to determine whether these identities rule out projections.

Define a set of binary functions $f_i^j(x, y)$, $j \leq 3$, by $f_i^1(x, y) = f_i(x, y, y), f_i^2 = f_i(y, x, y), f_i^3(x, y) = f_i(y, y, x)$, and identify any pair of f_i^js which are indentified by $\mathcal{T}$. Make a drawing of a hypergraph

with a vertex for every equivalence class of f_i^js, with an edge connecting any pair of vertices g, h with $g(x, y) \approx h(y, x)$ under $\mathcal{T}$, and with a hyperedge for each f_i connecting it to f_i^1, f_i^2, f_i^3. An assignment of projections π_j to the functions f_i is the same as a choice j of 1-IN-3 of the vertices on the hyperedge f_i to be granted the value π_1, while every edge of the hypergraph corresponds to a $\neq$ constraint. (Olšák's paper [113] has one such picture, and I've found the technique enormously helpful for visualizing large systems of identities on ternary functions.)

As a concrete example, take following collection of four ternary terms p, q, r, s defined in terms of Olšák's weak 3-cube term:

$$p(x, y, z) = t(y, z, z, x, x, x),$$
$$q(x, y, z) = t(x, z, y, z, y, z),$$
$$r(x, y, z) = t(y, x, x, z, y, y),$$
$$s(x, y, z) = t(x, x, y, z, y, x).$$

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

$$p(x, y, x) \approx q(y, x, x),$$
$$p(x, x, y) \approx r(y, x, x),$$
$$q(x, y, x) \approx s(x, y, x),$$
$$r(x, x, y) \approx s(x, x, y),$$
$$s(x, y, y) \approx q(x, x, y) \approx r(x, y, x).$$

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 occurences of the 1-IN-3 SAT constraint (for the four ternary terms p, q, r, s), and one occurence of the $\neq$ constraint (coming from the fact that the last identity above relates $s(x, y, y)$ to $q(x, x, y)$):
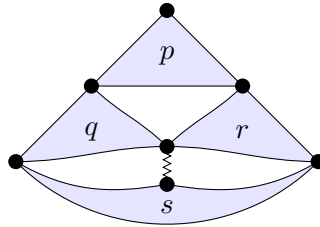
It is now easy (well, as easy as solving a small instance of 1-IN-3 SAT) to verify that the associated 1-IN-3 SAT instance has no solution, so this system of identities can't be satisfied by projections.

Remark 1.5.3. Taylor's original reason for studying Taylor algebras was to try to deeply understand the reason that π_1 of a topological group is always abelian. Taylor [135] 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 the π_1s of topological $\mathcal{V}$-objects will share a nontrivial property iff $\mathcal{V}$ has a Taylor term, and that this occurs iff π_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 [58] 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 $0 \in \mathbb{A}$ with $m(0, x) = m(x, 0) = x$ for all x, then m is commutative and associative.

Note that our assumption on m implies that $m * m$ satisfies the identities

$$m(x, y) \approx m * m(x, y, 0, 0) \approx m * m(x, 0, y, 0) \approx \cdots \approx m * m(0, 0, x, y),$$

where in each $m * m$ we always have the x occuring 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 es, xs, and ys 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 xs in the middle two rows and the ys in the top or bottom, and to move a y up a column whenever that column is free of xs. To temporarily move xs out of the way, we apply the Taylor identities for t to swap them with 0s, possibly shifting the xs 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:

$$m\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = (m * m) * t\left(\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ x & 0 & 0 & 0 & x & x \\ 0 & x & x & x & 0 & 0 \\ y & y & y & y & y & y \end{bmatrix}\right) = (m * m) * t\left(\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & x & x & x & 0 \\ 0 & x & x & x & 0 & 0 \\ y & y & y & y & y & y \end{bmatrix}\right)$$

$$= (m * m) * t\left(\begin{bmatrix} y & 0 & 0 & 0 & 0 & y \\ 0 & 0 & x & x & x & 0 \\ 0 & x & x & x & 0 & 0 \\ 0 & y & y & y & y & 0 \end{bmatrix}\right) = (m * m) * t\left(\begin{bmatrix} y & 0 & 0 & 0 & 0 & y \\ x & 0 & 0 & 0 & x & x \\ 0 & x & x & x & 0 & 0 \\ 0 & y & y & y & y & 0 \end{bmatrix}\right)$$

$$= \cdots = (m * m) * t\left(\begin{bmatrix} y & y & y & y & y & y \\ 0 & 0 & x & x & x & 0 \\ x & x & 0 & 0 & 0 & x \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}\right) = m\left(\begin{bmatrix} y \\ x \end{bmatrix}\right),$$

where we have used the Taylor identity $t(x, 0, 0, 0, x, x) \approx t(0, 0, x, x, x, 0)$ satisfied by a weak 3-cube term to temporarily move the first and last x out of the way.

1.6 Two simple algorithms (width 1 and bounded strict width)

Definition 1.6.1. A CSP template $\mathbf{A} = (D, \Gamma)$ 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 $S_{v_i} \leftarrow \pi_i(X \cap (S_{v_1} \times \cdots \times S_{v_k}))$ for each $i \le k$.
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. (After this section, I'll usually refer to generalized arc-consistency as just "arc-consistency" to save space.)

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 $\pi_j(R \cap (S_{v_1} \times \cdots \times S_{v_m})) \ne S_{v_j}$ **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 1.6.2 (Feder, Vardi [61]). *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** iff **B** has a map to **A** (alternatively, if Γ 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 1.6.3. A connected relational structure is a *tree* if every collection of occurences of relations with arities $r_1, ..., r_k$ 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** → **A** iff every tree which maps to **B** has a map to **A**.

Proposition 1.6.4. **A** *has width 1 iff it has tree duality.*

Proof. If generalized arc-consistency shows that there is no homomorphism **B** → **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**. □

Remark 1.6.1. 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 1.6.2. Dalmau has shown that any CSP with relational width $(2, 2)$ is also solved by generalized arc-consistency [52]. 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 1.6.5. A subalgebra $\mathbb{R} \leq \mathbb{A}_1 \times \cdots \times \mathbb{A}_n$ is called a *subdirect product*, written $\mathbb{R} \leq_{sd} \mathbb{A}_1 \times \cdots \times \mathbb{A}_n$, if $\pi_i(\mathbb{R}) = \mathbb{A}_i$ 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 1.6.6. For $\mathbf{A} = (D, \Gamma)$ a relational structure, define $\mathcal{P}_\emptyset(\mathbf{A})$ to be the structure with ground set $\mathcal{P}(D) \setminus \{\emptyset\}$, and for every m-ary relation $R \in \Gamma$ let the corresponding relation $\mathcal{P}_\emptyset(R)$ be the set of all m-tuples $S_1, ..., S_m \in \mathcal{P}(D) \setminus \{\emptyset\}$ such that there is some nonempty $X \subseteq R$ with $\pi_i(X) = S_i$ for each i.

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

Definition 1.6.7. A homomorphism $\mathcal{P}_\emptyset(\mathbf{A}) \to \mathbf{A}$ is called a *set polymorphism* of **A**.

Definition 1.6.8. A function $f : D^k \to D$ is called *totally symmetric* if the value of $f(a_1, ..., a_k)$ only depends on $\{a_1, ..., a_k\}$. Note that this is stronger than being symmetric, since the multiplicity of the a_is is also ignored.

Theorem 1.6.9. *The following are equivalent:*

- **A** *has width 1,*

- **A** *has a set polymorphism, and*

- **A** *has totally symmetric polymorphisms of every arity.*

Proof. If $\mathbf{A}$ has width 1, then generalized arc-consistency applied to $\mathcal{P}_\emptyset(\mathbf{A})$ shows that there is a homomorphism $f : \mathcal{P}_\emptyset(\mathbf{A}) \to \mathbf{A}$, 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 $\mathcal{P}_\emptyset(R)$). So suppose that f is a set polymorphism, and for every $k \geq 1$, let f_k be the totally symmetric function

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

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

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

Corollary 1.6.10. *A relational structure $\mathbf{A}$ has width 1 iff it is homomorphically equivalent to a pp-power of HORN-SAT.*

Proof. Let f be a set polymorphism of $\mathbf{A}$, and let f_k be the associated totally symmetric polymorphism of arity k. We define a height 1 clone homomorphism from $\langle \min \rangle \to \operatorname{Pol}(\mathbf{A})$ by sending $\min(x_1, ..., x_k)$ to $f_k(x_1, ..., x_k)$. Now apply the ERP Theorem 1.4.18 and Proposition 1.4.17 from the subsection on reflections. $\square$

Example 1.6.1. Suppose that $\mathbf{A}$ 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 $s_n(x_1, ..., x_n) = s(s_{n-1}(x_1, ..., x_{n-1}), x_n)$, and s_n will be totally symmetric for every n. Thus, every relational structure with a semilattice polymorphism has width 1.

Example 1.6.2. 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

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

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 $\operatorname{Inv}(f_3)$ is generated by the ternary relations R_{ab}, R_{bc}, R_{ca}, where R_{ab} is defined by

$$R_{ab}(x, y, z) \coloneqq (x \in \{a, b\}) \wedge (x = a \implies y = z),$$

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

Example 1.6.3. 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

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

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

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

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

$$R_k(x_0, ..., x_k) \wedge R_k^-(x_0, ..., x_k)$$

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

The relational clone $\mathrm{Inv}(\mathcal{O})$ corresponding to this example is generated by the unary relation $\{+\}$, the binary relations $x = -y$ and $x \leq y$, and the ternary relation $(x \geq 0) \wedge (x = 0 \implies y = z)$. The clone $\mathcal{O}$ 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 $\mathbf{A}$ is to apply the arc-consistency algorithm. If $\mathbf{A}$ is a rigid core, we can use the self-reducibility of $\mathrm{CSP}(\mathbf{A})$ to find a solution to every solvable instance $\mathbf{B}$ of $\mathrm{CSP}(\mathbf{A})$ in polynomial time. By applying this to $\mathcal{P}_\emptyset(\mathbf{A})$, we can then *find* a set polymorphism of $\mathbf{A}$ - in time polynomial in the size of $\mathcal{P}_\emptyset(\mathbf{A})$, which is sadly exponential in the size of $\mathbf{A}$. The following problem is currently open.

Problem 1.6.1. Given a rigid core $\mathbf{A}$, can we determine whether it has width 1 in time polynomial in the size of the description of $\mathbf{A}$?

Now we move to the case of bounded strict width. This has a connection to an intriguing paper of Dechter [55] which predates the algebraic approach to the CSP. The next definition follows Dechter [55].

Definition 1.6.11. 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 1.6.12 (Dechter [55]). *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_is 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. $\qquad\square$

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 1.6.13. A CSP template $\mathbf{A} = (D, \Gamma)$ has *strict width* l if every strong $(l+1)$-consistent instance of $\mathrm{CSP}(D, \langle \Gamma \rangle)$ 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 Γ, a hint that it is properly viewed as an algebraic condition. Algebraically, the relevant result is the Baker-Pixley theorem [7].

Theorem 1.6.14 (Baker, Pixley [7]). *The following are equivalent for an algebraic structure $\mathbb{A}$:*

- *every subalgebra of $\mathbb{A}^n$ is equal to the intersection of its projections onto sets of at most l coordinates, and*

- *$\mathbb{A}$ 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 $\mathcal{F}_{\mathbb{A}}(l+1) \subseteq \mathbb{A}^{\mathbb{A}^{l+1}}$ which is generated by the projections $\pi_i : \mathbb{A}^{l+1} \to \mathbb{A}$. Let A_{nu}^{l+1} 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 $X \subseteq \mathbb{A}^{A_{nu}^{l+1}}$ 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 $x_1, ..., x_l \in A_{nu}^{l+1}$, 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 t is an $(l+1)$-ary near-unanimity term, and suppose that $\mathbb{B} \subseteq \mathbb{A}^n$. Let $b \in \mathbb{A}^n$ 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 $\pi_J(b) \in \pi_J(\mathbb{B})$ 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 $b_J = t(b_{J_1}, ..., b_{J_{l+1}}) \in \mathbb{B}$ and has $\pi_J(b_J) = \pi_J(b)$ 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 $b \in \bigcap_{|I|=l} \pi_I(\mathbb{B})$, 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 $\mathbb{B}' = \bigcap_{|I|=l} \pi_I(\mathbb{B}')$, while the assumption on subalgebras of $\mathbb{A}^{l+1}$ gives $\pi_I(b) \in \pi_I(\mathbb{B}')$ for every I with $|I| = l$. $\qquad \square$

Theorem 1.6.15. *A relational structure* **A** *has strict width l iff 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 $\mathbb{B} = \bigcap_{|I| \leq l} \pi_I(\mathbb{B})$, 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 $\mathrm{CSP}(\mathbb{A})$, 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_{k+1}^i 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 $(a_1, ..., a_i', ..., a_{k+1}^i)$ 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. $\square$

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 1.6.4. 2SAT has the ternary polymorphism maj, which is a near-unanimity operation. Therefore 2SAT has strict width 2, a fact which also follows from Dechter's result above [55].

Example 1.6.5. Generalizing 2SAT, let D be any domain, and let $d : D^3 \to D$ 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 $D \sqcup D$, 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.
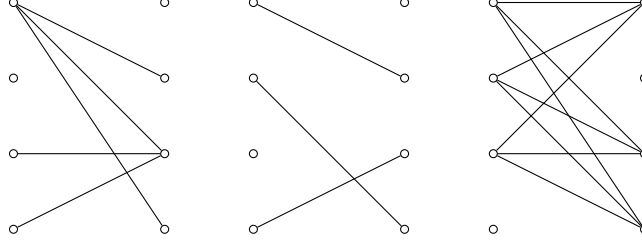
For any a in D, a generating set of binary relations for $\mathrm{Inv}(d)$ is given by the graphs of a pair of bijections which generate the symmetric group on $|D|$ elements, the unary relation $D \setminus \{a\}$, and the binary relation $x = a \vee y = a$.

Example 1.6.6. For every n, the relational structure $(\{0,1\}, \{0\}, \leq, \{0,1\}^n \setminus \{(0,...,0)\})$ 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 $\mathrm{Pol}(\mathbf{A})$ is equivalent to the solvability of the CSP instance Φ (of $\mathbf{A}$ together with singleton unary relations) with variables indexed by elements of $\mathbf{A}^{l+1}$ described by the primitive positive formula

$$\Phi(t) := \bigwedge_{R \in \Gamma} \bigwedge_{M \in R^{l+1}} t(M) \in R \wedge \bigwedge_{a,b \in \mathbf{A}} t(b,a,...,a) = t(a,b,...,a) = \cdots = t(a,a,...,b) \in \{a\}.$$

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 $\mathbf{A}$ has bounded strict width, allowing l to be arbitrarily large?

Problem 1.6.2. Given a relational structure $\mathbf{A}$, 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* [105], [11], [144]. The bad news is that the minimal arity of a near-unanimity term may be very large.

Theorem 1.6.16 (Zhuk, Barto, Draganov [14], [144])**.** *Every relational structure $\mathbf{A}$ with $|\mathbf{A}| = n$ and maximal arity of any relation m which has bounded strict width, has strict width at most $\frac{1}{2}(2m-2)^{3^n}$. 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 $(m-1)^{2^{n-2}}$, and for $m = 2, n \geq 3$ there is an example with no near-unanimity polymorphism of arity at most $2^{2^{n-3}}$.*

Luckily, it is possible to determine whether a relational structure has bounded strict width without actually exhibiting a near-unanimity polymorphism. For instance, in [13] 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.

1.6.1 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

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

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 $R \subseteq D_{v_1} \times \cdots \times D_{v_m}$ 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 1.6.17 (Kun, O'Donnell, Tamaki, Yoshida, Zhou [98])**.** *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 $\mathrm{CSP}(\mathbf{A})$, and consider the (by now standard) instance Φ that describes the existance of a symmetric polymorphism of arity n:

$$\Phi(s) := \bigwedge_{R \in \Gamma} \bigwedge_{M \in R^n} s(M) \in R \wedge \bigwedge_{a_1,...,a_n \in \mathbb{A}} \bigwedge_{\sigma \in S_n} s(a_1,...,a_n) = s(a_{\sigma(1)},...,a_{\sigma(n)}).$$

By the assumption that the basic LP decides $\mathrm{CSP}(\mathbf{A})$, 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 $p_{v,a}$ 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 $\mathbf{A}$ 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 $\mathbf{A}$ to elements of $\mathbf{A}$.

Applying s to each $p_{v,\cdot}$ gives an element $a_v \in \mathbf{A}$, and applying it to each probability distribution $p_{C,\cdot}$ gives a tuple r_C in the associated relation R (since s is a polymorphism). Furthermore, the compatibility equations between the distributions $p_{v_i,\cdot}$ and $p_{C,\cdot}$ that we get when v_i is the ith coordinate of the constraint C, together with the symmetry of s, imply that $a_{v_i} = (r_C)_i$ for each i, so $(a_{v_1},...,a_{v_m}) = r_C \in R$. Thus the a_vs form a valid solution to the CSP instance.

Finally, assume that $\mathbf{A}$ 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 $\theta \leq \frac{1}{|\mathbf{A}|}$, and for each variable v we assign the value

$$a_v = f(\{a \in \mathbf{A} \mid p_{v,a} \geq \theta\}).$$

Note that the restriction $\theta \leq \frac{1}{|\mathbf{A}|}$ ensures that the sets on the right hand side are nonempty. We will show that if θ 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 θ works at least as well. For this we need the following claim.

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

$$\theta \notin (p_{v_i,a}/(2|R|), p_{v_i,a}]$$

for any pair $i \leq m, a \in \mathbf{A}$, then $(a_{v_1},...,a_{v_m})$ 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 $(a_{v_1},...,a_{v_m})$ satisfies R, we just need to check that this colection 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

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

52

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 θ 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 θ uniformly at random from the set $\{\frac{1}{|\mathbf{A}|}, \frac{1}{|\mathbf{A}|T}, \cdots \frac{1}{|\mathbf{A}|T^b}\}$, 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 θ where the claim doesn't apply), and asymptotically that looks like $1 - O(1/\log(1/\epsilon))$. $\qquad\square$

Remark 1.6.3. The dependence of the error in $1 - O(1/\log(1/\epsilon))$ on ϵ in the previous theorem is best possible in the case of HORN-SAT: Guruswami and Zhou [69] show that there are integrality gap instances even for the SDP relaxation, and by a fundamental result of Raghavendra [124] 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 1.6.4. In [98], it is also claimed that the basic LP solves every instance of $\mathrm{CSP}(\mathbf{A})$ if and *only if* $\mathbf{A}$ has width 1. The proof has a subtle error, however. The following counterexample, due to Kun, can be found in [53].

Example 1.6.7. Let $\mathbf{A} = (\{-1, 0, 1\}, R_+, R_-)$, 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

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

is a symmetric polymorphism of $\mathbf{A}$. Thus $\mathrm{CSP}(\mathbf{A})$ is solved by the basic LP relaxation. However, $\mathbf{A}$ 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 $\mathbf{A}$ does not have width 1.

Example 1.6.8. 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 $s_n \in \mathrm{Clo}(s_2)$ for all n (hint: start by defining $t_n(x_1, ..., x_n) = s_2(x_1, s_{n-1}(x_2, ..., x_n))$), so $\mathrm{Inv}(s_2)$ also defines a CSP template which is solved by the basic LP relaxation.

s_2	$-$	0	$+$
$-$	$-$	$-$	0
0	$-$	0	$+$
$+$	0	$+$	$+$

$\mathrm{Inv}(s_2)$ 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

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

(I found this set of generating relations by a technique I learned from Zhuk [145], 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 $\mathrm{Inv}(s_2)$ 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

$$s_n'(x_1, ..., x_n) = \begin{cases} s_{0,n}(x_1, ..., x_n) & \text{if some } x_i = 0, \\ s_{1,n}(x_1, ..., x_n) & \text{if all } x_i \in \{-1, 1\}. \end{cases}$$

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 $s_n'(x, x, y_3, ..., y_n)$ we get by identifying two of its inputs *is* in $\langle s_2 \rangle$ (exercise for the reader), so it preserves every relation in $\mathrm{Inv}(s_2)$ which contains strictly less than n tuples.

The clone $\langle s_2 \rangle$ is strictly contained in the width 1 clone from Example 1.6.3, 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 1.6.3. Given a finite relational structure $\mathbf{A}$, determine if it has symmetric polymorphisms of every arity.

An interesting result in this direction is proved in [46]: an algebraic structure $\mathbb{A}$ has symmetric polymorphisms of all arities iff there is no $\mathbb{B} \in HSP(\mathbb{A})$ which has a pair of automorphisms in $\mathrm{Aut}(\mathbb{B})$ having no common fixed point (in fact, if $\mathbb{A}$ has no symmetric polymorphism of arity n, we can take $\mathbb{B}$ to be the free algebra on n variables in the variety generated by $\mathbb{A}$). If $HSP(\mathbb{A})$ could be replaced by $HS(\mathbb{A})$ in their result, then this would imply that it is enough to check for the existence of symmetric polymorphisms of arities up to $|\mathbf{A}|$.

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 [46], an example is given of a relational structure which has cyclic polymorphisms of every arity, but which has no symmetric polymorphism of arity 5.

1.7 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 1.7.1. *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 $HH^{-1}H = hU(hU)^{-1}hU = hUUh^{-1}hU = hUUU = hU = H$. $\qquad\square$

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 α, β on G, then we can view this equality as the statement that $\alpha \circ \beta = \beta \circ \alpha = \alpha \vee \beta$, where composition of binary relations is defined as follows.

Definition 1.7.2. 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

$$R \circ S(a, c) := \exists b \in B \ R(a, b) \wedge S(b, c).$$

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 1.7.3. *If α, β are congruences on an algebraic structure $\mathbb{A}$, 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 α, β 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 1.7.4. We say that a variety $\mathcal{V}$ is *congruence permutable* if for all $\mathbb{A} \in \mathcal{V}$ and all $\alpha, \beta \in \mathrm{Con}(\mathbb{A})$ we have $\alpha \circ \beta = \beta \circ \alpha$.

Theorem 1.7.5. *A variety $\mathcal{V}$ is congruence permutable iff $\mathcal{V}$ has a ternary term p which satisfies the identity*

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

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

Then $(x, z) \in \alpha \circ \beta$, so if $\mathcal{V}$ has commuting congruences, then there must be some $p(x, y, z) \in \mathcal{F}$ 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 $\mathbb{A} \in \mathcal{V}$ and $\alpha, \beta \in \mathrm{Con}(\mathbb{A})$. 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 $(a, c) \in \beta \circ \alpha$. $\qquad\square$

Definition 1.7.6. 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 1.7.7. A lattice $\mathcal{L}$ is *modular* if for all $\alpha, \beta, \gamma \in \mathcal{L}$, we have

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

Equivalently, a lattice $\mathcal{L}$ is modular if it has no five element sublattice isomorphic to the lattice $\mathcal{N}_5$ whose Hasse diagram is a pentagon: consider the sublattice generated by $\alpha' = \alpha \vee (\gamma \wedge \beta), \beta' = (\alpha \vee \gamma) \wedge \beta$, and γ, with top element $\alpha \vee \gamma = \alpha' \vee \gamma$ and bottom element $\gamma \wedge \beta = \gamma \wedge \beta'$, and note that we always have $\alpha' \leq \beta'$.
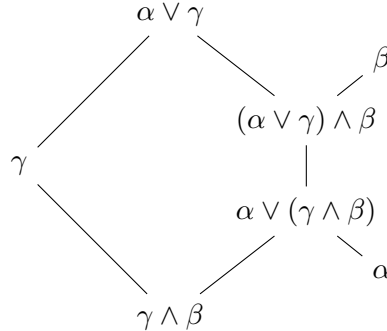
Proposition 1.7.8. *If* $\mathbb{A}$ *has permuting congruences, then* $\mathrm{Con}(\mathbb{A})$ *is a modular lattice.*

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

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

Definition 1.7.9. 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 $\cdot, /, \backslash$, 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 $\cdot$ is a quasigroup operation, then the operations $/, \backslash$ can be defined in terms of $\cdot$ 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 1.7.10. *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

$$p(x, y, y) = (x/y) \cdot ((y/y)\backslash y) \approx (x/y) \cdot ((y/y)\backslash((y/y) \cdot y)) \approx (x/y) \cdot y \approx x. \qquad \square$$

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 1.7.11. 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 1.7.12. *A finite algebraic structure $\mathbb{A}$ has a Mal'cev term p iff every relation $\mathbb{R} \in \mathrm{Inv}(\mathbb{A})$ has the parallelogram property.*

Proof. Suppose first that $\mathbb{A}$ has a Mal'cev term p, let $\mathbb{B}, \mathbb{C} \in \mathcal{V}(\mathbb{A})$ and let $\mathbb{R} \leq \mathbb{B} \times \mathbb{C}$ be a subalgebra of their product. Suppose that $(a, b), (c, b), (c, d) \in \mathbb{R}$. Then

$$\begin{bmatrix} a \\ d \end{bmatrix} = p\left(\begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} c \\ b \end{bmatrix}, \begin{bmatrix} c \\ d \end{bmatrix} \right) \in \mathbb{R},$$

so $\mathbb{R}$ has the parallelogram property.

Conversely, suppose that every relation in $\mathrm{Inv}(\mathbb{A})$ has the parallelogram property. Let $\pi_1, \pi_2 \in \mathbb{A}^{\mathbb{A}^2}$ be the elements corresponding to the functions $\pi_i : (a_1, a_2) \mapsto a_i$. Let $\mathbb{R} \leq (\mathbb{A}^{\mathbb{A}^2})^2$ be the subalgebra generated by the three pairs $(\pi_1, \pi_1), (\pi_2, \pi_1), (\pi_2, \pi_2)$. Then since $\mathbb{R}$ has the parallelogram property, we must have $(\pi_1, \pi_2) \in \mathbb{R}$, so there must be a ternary term p such that

$$\begin{bmatrix} \pi_1 \\ \pi_2 \end{bmatrix} = p\left(\begin{bmatrix} \pi_1 \\ \pi_1 \end{bmatrix}, \begin{bmatrix} \pi_2 \\ \pi_1 \end{bmatrix}, \begin{bmatrix} \pi_2 \\ \pi_2 \end{bmatrix} \right).$$

But then this p is a Mal'cev term for $\mathbb{A}$. $\qquad \square$

If we want to test whether an algebra has a Mal'cev term, then the above result would make it seem like we need to test whether relations of arbitrarily large arity have the parallelogram property. As it turns out, for idempotent algebras we only need to test whether all binary relations have the parallelogram property.

Theorem 1.7.13 (Zhuk [143]). *A finite idempotent algebra $\mathbb{A}$ has a Mal'cev term if and only if every binary relation $\mathbb{R} \in \mathrm{Inv}_2(\mathbb{A})$ has the parallelogram property. More explicitly, this occurs if and only if we have*

$$\begin{bmatrix} a \\ d \end{bmatrix} \in \mathrm{Sg}_{\mathbb{A}^2} \left\{ \begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} c \\ b \end{bmatrix}, \begin{bmatrix} c \\ d \end{bmatrix} \right\}$$

for all $a, b, c, d \in \mathbb{A}$.

Proof. Suppose that $\mathbb{A}$ is not Mal'cev, and consider a relation $\mathbb{R} \in \mathrm{Inv}(\mathbb{A})$ of minimal arity n among those which do not have the parallelogram property. If $n = 2$, we are done. Otherwise, we will try to use $\mathbb{R}$ to define a relation of lower arity which also fails to have the parallelogram property.

Suppose that $\mathbb{R}$ fails to have the parallelogram property when considered as a binary relation on $\mathbb{A}^k \times \mathbb{A}^{n-k}$, and assume without loss of generality that $k \geq 2$. Since $\mathbb{R}$ fails to have the parallelogram property, there are tuples $a, c \in \mathbb{A}^k$ and tuples $b, d \in \mathbb{A}^{n-k}$ such that $(a, b), (c, b), (c, d) \in \mathbb{R}$ but $(a, d) \notin \mathbb{R}$. Write $a_1 = \pi_1(a)$ and $a' = \pi_{2,\dots,k}(a)$, and define c_1, c' similarly. Define $\mathbb{R}' \leq \mathbb{A}^{k-1} \times \mathbb{A}^{n-k}$ by

$$(x', y) \in \mathbb{R}' \quad \Longleftrightarrow \quad \exists x_1 \in \mathbb{A} \; ((x_1, x'), y) \in \mathbb{R} \land ((x_1, x'), b) \in \mathbb{R}.$$

Then we have $(a', b), (c', b), (c', d) \in \mathbb{R}'$, so if $\mathbb{R}'$ has the parallelogram property then we must have $(a', d) \in \mathbb{R}'$. Thus there is some $e_1 \in \mathbb{A}$ such that $((e_1, a'), b), ((e_1, a'), d) \in \mathbb{R}$. Now define $\mathbb{R}'' \leq \mathbb{A} \times \mathbb{A}^{n-k}$ by

$$(x_1, y) \in \mathbb{R}'' \quad \Longleftrightarrow \quad ((x_1, a'), y) \in \mathbb{R}.$$

Then we have $(a_1, b), (e_1, b), (e_1, d) \in \mathbb{R}''$, so if $\mathbb{R}''$ has the parallelogram property then we must have $(a_1, d) \in \mathbb{R}''$, which means that $(a, d) \in \mathbb{R}$, a contradiction. $\square$

Definition 1.7.14. If $\mathbb{R} \leq_{sd} \mathbb{A} \times \mathbb{B}$ is a subdirect binary relation, then the *linking congruence* of $\mathbb{R}$ can refer to any of the following three congruences: the congruence $\ker \pi_1 \vee \ker \pi_2$ on $\mathbb{R}$, the congruence α on $\mathbb{A}$ generated by pairs $a, a' \in \mathbb{A}$ such that there exists a $b \in \mathbb{B}$ with $(a, b), (a', b) \in \mathbb{R}$, or the similar congruence β defined on $\mathbb{B}$. The relation $\mathbb{R}$ is called *linked* if these congruences are full.

Note that in the above definition, we have $\alpha = \pi_1(\ker \pi_1 \vee \ker \pi_2), \beta = \pi_2(\ker \pi_1 \vee \ker \pi_2)$, and

$$\mathbb{A}/\alpha \cong \mathbb{R}/(\ker \pi_1 \vee \ker \pi_2) \cong \mathbb{B}/\beta.$$

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

Proposition 1.7.15 (Goursat's Lemma). *A subdirect binary relation $\mathbb{R} \leq_{sd} \mathbb{A} \times \mathbb{B}$ has the parallelogram property iff there are congruences α, β on $\mathbb{A}, \mathbb{B}$ respectively and an isomorphism $f : \mathbb{A}/\alpha \to \mathbb{B}/\beta$ such that, writing π_α, π_β for the quotient maps, we have $\mathbb{R} = \pi_\alpha^{-1} \circ f^{-1} \circ \pi_\beta$ (treating π_α, π_β, f as binary relations with inputs on the right and outputs on the left).*

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

$$\begin{bmatrix} a_1 \\ b_i \end{bmatrix}, \begin{bmatrix} a_{i+1} \\ b_i \end{bmatrix}, \begin{bmatrix} a_{i+1} \\ b_{i+1} \end{bmatrix} \in \mathbb{R} \implies \begin{bmatrix} a_1 \\ b_{i+1} \end{bmatrix} \in \mathbb{R}. \qquad \square$$

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

Example 1.7.1. Consider the affine algebra $\mathbb{A} = (\mathbb{Z}/p, x - y + z)$, and let $\mathbb{R} \leq_{sd} \mathbb{A}^n$ be the relation $x_1 + \cdots + x_n \equiv 0 \pmod{p}$. Then if we think of $\mathbb{R}$ as a (subdirect) binary relation on $\mathbb{A} \times \mathbb{A}^{n-1}$, it is the graph of the homomorphism $\mathbb{A}^{n-1} \to \mathbb{A}$ given by $(x_2, ..., x_n) \mapsto -x_2 - \cdots - x_n \pmod{p}$.

More generally, for any i, if we think of $\mathbb{R}$ as a subdirect binary relation on $\mathbb{A}^i \times \mathbb{A}^{n-i}$, then the linking congruence gives homomorphisms $\mathbb{A}^i \to \mathbb{A} \leftarrow \mathbb{A}^{n-i}$: $(x_1, ..., x_i) \mapsto x_1 + \cdots + x_i \pmod{p}$ and $(x_{n-i+1}, ..., x_n) \mapsto -x_{n-i+1} - \cdots - x_n \pmod{p}$.

Ternary relations on simple Mal'cev algebras have a particularly interesting structure.

Proposition 1.7.16. *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. $\square$

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 1.7.2. Let $\mathbb{A}_n = (\{0, ..., n-1\}, p)$, where p is the ternary Mal'cev operation defined by

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

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 $\mathbb{A}_n^2$). $\mathrm{Inv}(\mathbb{A}_n)$ 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

$$(x, y, z \in \{0, 1\}) \wedge (x + y + z \equiv 0 \pmod 2).$$

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

Example 1.7.3. 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 θ 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 $\mathbb{A}$, there is also the ternary relation

$$(x = y = z = *) \vee (x, y, z \in \{0, 1\} \wedge x + y + z \equiv 0 \pmod 2)),$$

whose elements correspond to the columns of the matrix

$$\begin{bmatrix} * & 0 & 0 & 1 & 1 \\ * & 0 & 1 & 0 & 1 \\ * & 0 & 1 & 1 & 0 \end{bmatrix}.$$

That this relation forms a subalgebra of $\mathbb{A}^3$ is related to the fact that θ can be considered to be an "abelian congruence" (in a sense we will define later).

1.8 Mal'cev algorithm and compact representations

The algorithm for solving CSPs invariant under a Mal'cev operation, due to Bulatov and Dalmau [36], 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 $\mathbb{R}$ which has the same projection to each factor and contains representatives of all of the "forks" of $\mathbb{R}$ actually generates $\mathbb{R}$.

Definition 1.8.1. If $R \subseteq \mathbb{A}_1 \times \cdots \times \mathbb{A}_n$, then we define the *signature* of R, written $\text{Sig}(R)$, to be the set of triples (i, a, b) with $i \in \{1, ..., n\}, a, b, \in \mathbb{A}_i$ such that there are some $t_a, t_b \in R$ with $\pi_{1,...,i-1}(t_a) = \pi_{1,...,i-1}(t_b)$ 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 1.8.2. *Suppose that a relation* $\mathbb{R} \leq \mathbb{A}_1 \times \cdots \times \mathbb{A}_n$ *is preserved by a Mal'cev term* p*, and that* $S \subseteq \mathbb{R}$ *is a subset with* $\text{Sig}(S) = \text{Sig}(\mathbb{R})$*. Then* $\mathbb{R}$ *is generated by* S *(using only* p*).*

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

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

$$t'' = p(t', t_a, t_b).$$

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

$$\pi_{1,...,i-1}(t'') = \pi_{1,...,i-1}(p(t', t_a, t_a)) = \pi_{1,...,i-1}(t') = \pi_{1,...,i-1}(t).$$

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 $\pi_{1,...,i}(t'') = \pi_{1,...,i}(t)$. $\qquad\square$

Definition 1.8.3. A subset $S \subseteq \mathbb{R}$ is called a *compact representation* of a Mal'cev relation $\mathbb{R}$ if $\mathrm{Sig}(S) = \mathrm{Sig}(\mathbb{R})$ and $|S| \leq 2|\mathrm{Sig}(\mathbb{R})|$.

Proposition 1.8.4. *Every Mal'cev relation $\mathbb{R} \leq \mathbb{A}_1 \times \cdots \times \mathbb{A}_n$ 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 `Nonempty`: it takes as input a compact representation R of a relation $\mathbb{R} \leq \mathbb{A}_1 \times \cdots \times \mathbb{A}_n$ and any description of a relation $\mathbb{S} \leq \mathbb{A}_{i_1} \times \cdots \times \mathbb{A}_{i_k}$ on a small subset $\{i_1, ..., i_k\}$ of the indices, and it tells us whether $\mathbb{R} \cap \mathbb{S} \neq \emptyset$. In the case $\mathbb{R} \cap \mathbb{S} \neq \emptyset$, `Nonempty` returns an element of the intersection.

Algorithm 4 $\texttt{Nonempty}(R, i_1, ..., i_k, \mathbb{S})$, p a Mal'cev term, R a compact representation of $\mathbb{R} \leq \mathbb{A}_1 \times \cdots \times \mathbb{A}_n$, $\mathbb{S} \leq \mathbb{A}_{i_1} \times \cdots \times \mathbb{A}_{i_k}$.

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

Proposition 1.8.5. *`Nonempty` correctly determines whether $\mathbb{R} \cap \mathbb{S} \neq \emptyset$ in time polynomial in n, $|R|$, and $|\pi_{i_1,...,i_k}(\mathbb{R})| \leq \prod_{j \leq k} |\mathbb{A}_{i_j}|$.*

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

The next subroutine for manipulating compact representations is `Fix-values`. `Fix-values` converts a compact representation R of $\mathbb{R} \leq \mathbb{A}_1 \times \cdots \times \mathbb{A}_n$ to a compact representation of

$$\mathbb{R} \wedge (x_1 = a_1) \wedge \cdots \wedge (x_m = a_m),$$

for any choice of $m \leq n$ and $a_i \in \mathbb{A}_i$ for all i. `Fix-values` is really the core of the algorithm, the other steps are mostly formal (in fact, `Nonempty` and `Fix-values` are the only two subroutines which use the Mal'cev term p).

Proposition 1.8.6. *`Fix-values` correctly returns a compact representation of $\mathbb{R}_m = \mathbb{R} \wedge (x_1 = a_1) \wedge \cdots \wedge (x_m = a_m)$ in polynomial time.*

Algorithm 5 Fix-values$(R, a_1, ..., a_m)$, p a Mal'cev 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:      **if** $(j, a_j, a_j) \notin \mathrm{Sig}(R_{j-1})$ **then**
4:          **return** $\emptyset$.
5:      **else**
6:          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:      **for all** $(i, a, b) \in \mathrm{Sig}(R_{j-1})$ with $i > j$ **do**
8:          Let $t_a, t_b \in R_{j-1}$ witness the triple (i, a, b).
9:          Let $t \leftarrow \mathtt{Nonempty}(R_{j-1}, j, i, \{(a_j, a)\})$.
10:         **if** $t \neq \emptyset$ **then**
11:            Set $R_j \leftarrow R_j \cup \{t, p(t, t_a, t_b)\}$.
12: **return** R_m.

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 $(i, a, b) \in \mathrm{Sig}(R_j)$, if $a \neq b$ then we must have $i > j$. For $i \leq j$, we have $(i, a_i, a_i) \in \mathrm{Sig}(R_j)$ iff $\mathbb{R}_j \neq \emptyset$ by how we initialize R_j.

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

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

so from $i > j$ we have $p(t, t_a, t_b) \in \mathbb{R}_j$. 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 Fix-values runs in polynomial time, note that every call to $\mathtt{Nonempty}$ involves a constraint on two variables. $\qquad\square$

Corollary 1.8.7. *Given a compact representation R of a relation $\mathbb{R} \leq \mathbb{A}_1 \times \cdots \times \mathbb{A}_n$ which is preserved by a given Mal'cev operation p, and given a tuple $t \in \mathbb{A}_1 \times \cdots \times \mathbb{A}_n$, we can determine whether $t \in \mathbb{R}$ in polynomial time.*

The next subroutine will give a compact representation for the intersection of a relation $\mathbb{R}$ given by a compact representation R and a relation $\mathbb{S}$ of small arity. In [36] this subroutine was called Next-beta, so we will copy that notation here.

Proposition 1.8.8. *Next-beta correctly finds a compact representation of $\mathbb{R} \cap \mathbb{S}$ in time polynomial in n, $|R|$, and $|\pi_{i_1,\dots,i_k}(\mathbb{R})| \cdot \max_i |\mathbb{A}_i| \leq \prod_{j \leq k} |\mathbb{A}_{i_j}| \cdot \max_i |\mathbb{A}_i|$.*

Bulatov and Dalmau [36] then go on to define a subroutine Next which calls Next-beta on larger and larger projections of $\mathbb{S}$, ensuring that $|\pi_{i_1,\dots,i_k}(\mathbb{R})| \leq |\mathbb{S}| \cdot \max_i |\mathbb{A}_i|$ every time that Next-beta is called. A better approach, leading to a more powerful algorithm, was found by Maróti [106]. The subroutine 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.

Algorithm 6 Next-beta$(R, i_1, ..., i_k, \mathbb{S})$, R a compact representation of $\mathbb{R} \leq \mathbb{A}_1 \times \cdots \times \mathbb{A}_n$, $\mathbb{S} \leq \mathbb{A}_{i_1} \times \cdots \times \mathbb{A}_{i_k}$.

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

Algorithm 7 Intersect$(R, i_1, ..., i_k, S)$, R a compact representation of $\mathbb{R} \leq \mathbb{A}_1 \times \cdots \times \mathbb{A}_n$, S a compact representation of $\mathbb{S} \leq \mathbb{A}_{i_1} \times \cdots \times \mathbb{A}_{i_k}$.

1: Let $t_R \in R$ and $t_S \in S$ be any tuples.
2: Set $R' \leftarrow (R \times \{t_S\}) \cup (\{t_R\} \times S) \subseteq \mathbb{A}_1 \times \cdots \mathbb{A}_n \times \mathbb{A}_{i_1} \times \cdots \times \mathbb{A}_{i_k}$.
3: **for** $j \leq k$ **do**
4:      Set $R' \leftarrow$ Next-beta$(R', i_j, n+j, =_{\mathbb{A}_{i_j}})$.
5: **return** a minimal subset of $\pi_{1,...,n}(R')$ which witnesses every triple $(i, a, b) \in \mathrm{Sig}(\pi_{1,...,n}(R'))$.

Theorem 1.8.9. *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.*

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

Corollary 1.8.10. *For any primitive positive formula φ 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 φ.*

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 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. $\qquad\square$

     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 $\mathrm{Sg}_{\mathbb{A}^n}(S)$ for any small set $S \subseteq \mathbb{A}^n$, and a little thought shows that this can be reduced to the following problem.

Problem 1.8.1. 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 $t \in \mathrm{Sg}_{\mathbb{A}^n}(S)$ in time polynomial in $|S|$ and n?

This is a special case of the Subpower Membership Problem 2.4.1. 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 [63] for a straightforward exposition).

Remark 1.8.1. The proof of correctness of the subroutine `Nonempty` and the algorithm for `Fix-values` are both directly connected to the proof of Theorem 1.8.2. 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 1.8.2.

1.8.1 Near-subgroups

In this subsection, we will describe the maximal polynomial-time solvable extension $\mathbf{G}^*$ of the relational clone $\mathbf{G}$ of cosets of subgroups of $\mathbb{G}^m$, where $\mathbb{G}$ is a finite group. The relational clone $\mathbf{G}^*$ 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 $\mathbf{G}$.

First, consider the simple case where $\mathbb{G} = \mathbb{Z}/n$ 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

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

Using an inductive argument with this as the base case, Feder and Vardi [61] 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 1.8.11 (Feder, Vardi [61])**.** *If we adjoin any unary relation K to the relational structure $\mathbf{G} = (\mathbb{Z}/n, \{1\}, x + y = z)$, 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 $\mathbf{G}$. 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 $(K - 1) \cap \langle i - 1 \rangle$ are subgroups, so $2i, 2i - 1 \in K$ and we 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 $K = \mathbb{Z}/n$). $\square$

Next, consider the case where $\mathbb{G}$ 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

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

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 $\mathbb{G}$ is any finite abelian group, and $K \subseteq \mathbb{G}$ 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 $\mathbb{N} \leq \mathbb{M} \leq \mathbb{G}$ with $\mathbb{M}/\mathbb{N} \cong (\mathbb{Z}/2)^2$, then if K meets any three elements of $\mathbb{M}/\mathbb{N}$ it must also meet the fourth.

Proposition 1.8.12. *Suppose that $\mathbb{G}$ is an abelian group and that $K \subseteq \mathbb{G}$ has $0 \in K$, has the property that if $a, a+b \in K$ then $a + \langle b \rangle \subseteq K$, and the property that for any subgroups $\mathbb{N} \leq \mathbb{M} \leq \mathbb{G}$ with $\mathbb{M}/\mathbb{N} \cong (\mathbb{Z}/2)^2$, we have $|(K \cap \mathbb{M})/\mathbb{N}| \neq 3$. Then K must be a subgroup of $\mathbb{G}$.*

Proof. From the first assumption, for any $a, b \in K$ we must 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 $\mathbb{M} = \langle a, b \rangle$ and $\mathbb{N} = \langle 2a, 2b \rangle$, we see that either $|\mathbb{M}/\mathbb{N}| < 4$ in which case $a+b \in \langle a, b \rangle \subseteq K$, or $\mathbb{M}/\mathbb{N} \cong (\mathbb{Z}/2)^2$ and $|(K \cap \mathbb{M})/\mathbb{N}| \geq 3$. 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. $\square$

Corollary 1.8.13. *If $\mathbb{G}$ is a finite abelian group and $\mathbf{G}$ the associated relational structure, then for any m-ary relation K which is not a coset of a subgroup of $\mathbb{G}^m$, the CSP we get by adding K to $\mathbf{G}$ is NP-complete.*

Proof. Apply the previous proposition to the abelian group $\mathbb{G}^m$. $\square$

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 $b \in \mathbb{G}$ the relation $y = bx$ is a left coset of the diagonal subgroup of $\mathbb{G}^2$. So by the abelian case, the only possibilities for new relations are those described by the following definition.

Definition 1.8.14. A subset $K \subseteq \mathbb{G}$ is a *near subgroup* of $\mathbb{G}$ if it contains 1, and for any $b \in K^{-1}$, any $\mathbb{M} \leq \mathbb{G}$ and any $\mathbb{N} \lhd \mathbb{M}$ with $\mathbb{M}/\mathbb{N}$ abelian, the quotient set $(bK \cap \mathbb{M})/\mathbb{N}$ is a subgroup of $\mathbb{M}/\mathbb{N}$.

Proposition 1.8.15. *If $\mathbb{H} \leq \mathbb{G}$ is a subgroup and $K \subseteq \mathbb{H}$ is a near subgroup of $\mathbb{H}$, then K is a near subgroup of $\mathbb{G}$. Similarly, if $\varphi : \mathbb{G} \twoheadrightarrow \mathbb{H}$ is a surjective group homomorphism and $K \subseteq \mathbb{H}$ is a near subgroup of $\mathbb{H}$, then $\varphi^{-1}(K)$ is a near subgroup of $\mathbb{H}$.*

Theorem 1.8.16 (Aschbacher [4]). *The intersection of two near subgroups of a finite group is a near subgroup.*

Corollary 1.8.17 (Feder [60]). *Let $\mathbb{G}$ be a finite group, and let $\mathbf{G}^*$ be the relational structure on the underlying set of $\mathbb{G}$ having as relations all cosets of all near subgroups of $\mathbb{G}^n$. Then $\mathbf{G}^*$ has a Mal'cev polymorphism.*

Proof. Consider the "free near subgroup generated by two elements", that is, the smallest near subgroup K of $\mathbb{G}^{\mathbb{G}^2}$ which contains π_1, π_2 (a smallest such near subgroup exists since the intersection of all of them is guaranteed to be a near subgroup as well). Let $\mathbb{N}$ be the commutator subgroup of the group generated by π_1, π_2. Since $\langle \pi_1, \pi_2 \rangle / \mathbb{N}$ is abelian, there must be some $c \in \mathbb{N}$ with $\pi_1 \pi_2 c \in K$ by the definition of a near subgroup.

We define a binary polymorphism g by $g = \pi_1 \pi_2 c$, that is, $g(x, y) = xyc(x, y)$, where $c \in \mathbb{G}^{\mathbb{G}^2}$ is interpreted as a function $c : \mathbb{G}^2 \to \mathbb{G}$. 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

$$p(x, y, z) = yg(y^{-1}x, y^{-1}z) = xy^{-1}zc(y^{-1}x, y^{-1}z).$$

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 $\mathbf{G}^*$, let X be any coset of any near subgroup of $\mathbb{G}^n$, and let $x, y, z \in X$. Then $y^{-1}X$ is a near subgroup of $\mathbb{G}^n$. Since $g = \pi_1 \pi_2 c \in K$, g preserves every near subgroup of $\mathbb{G}^n$ (since for any $a, b \in \mathbb{G}^n$, K is contained in the near subgroup of $\mathbb{G}^{\mathbb{G}^2}$ obtained by taking the preimage of the map φ from the subgroup generated by π_1, π_2 to $\mathbb{G}^n$ which sends $\pi_1 \mapsto a, \pi_2 \mapsto b$). 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. $\qquad\square$

In order to prove Aschbacher's Theorem 1.8.16, we first need a more convenient characterization of near-subgroups.

Definition 1.8.18. A subset K of a finite group $\mathbb{G}$ is a *twisted subgroup* if $1 \in K$ and $x, y \in K \implies xyx \in K$.

Proposition 1.8.19. *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 $x^k \cdot 1 \cdot x^k, x^k \cdot x \cdot x^k \in K$ 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 $b^{-1}(x \cdot y \cdot x) = (b^{-1}x) \cdot (b \cdot b^{-1}y \cdot b) \cdot (b^{-1}x) \in K$, so $xyx \in bK$. $\qquad\square$

Proposition 1.8.20. *A subset $K \subseteq \mathbb{G}$ is a near-subgroup iff it is a twisted subgroup such that for any $b \in K^{-1}$, any $\mathbb{M} \leq \mathbb{G}$ and any $\mathbb{N} \lhd \mathbb{M}$ with $\mathbb{M}/\mathbb{N}$ isomorphic to the Klein four-group, $|(bK \cap \mathbb{M})/\mathbb{N}| \neq 3$.*

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 $yx \in yK$ and $y^{-1} \in \langle y \rangle \subseteq K$, so $(yx)^2 \in \langle yx \rangle \subseteq yK$, which is equivalent to $xyx = y^{-1}(yx)^2 \in K$. $\qquad\square$

Example 1.8.1. An explicit example of a near subgroup which is not a subgroup is given in [61]: Let $\mathbb{G}$ be the group of order p^3 (p odd) with elements indexed by triples $(i, j, k) \in (\mathbb{Z}/p)^3$ and multiplication given by

$$(i, j, k)(i', j', k') := (i + i' + jk', j + j', k + k').$$

Let $K = \{(\frac{1}{2}jk, j, k) \mid j, k \in \mathbb{Z}/p\}$. Since $\mathbb{G}$ 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

$$(\tfrac{1}{2}jk, j, k)(\tfrac{1}{2}j'k', j', k')(\tfrac{1}{2}jk, j, k) = (\tfrac{1}{2}(2jk + j'k') + jk' + jk + j'k, 2j + j', 2k + k') \in K.$$

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

That we needed to take p odd in the above example is no coincidence, as the next proposition shows.

Proposition 1.8.21. *If $\mathbb{G}$ is a 2-group, then any near-subgroup of $\mathbb{G}$ is a subgroup of $\mathbb{G}$.*

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

Let $z \in Z(\mathbb{G})$ be a nontrivial involution in the center of $\mathbb{G}$, 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 $K/\langle z \rangle = \mathbb{G}/\langle z \rangle$.

If $z \in K$ then for any $g \in \mathbb{G} \setminus K$, 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 $\mathbb{G} = K$.

Thus we may assume that $z \notin K$, and in fact that $Z(\mathbb{G}) \cap K = 1$. Let $\mathbb{M}$ be a maximal subgroup of $\mathbb{G}$ containing $\langle z \rangle$, then by induction we have $\mathbb{M} \cap K$ a subgroup of $\mathbb{M}$. Since $(\mathbb{M} \cap K)/\langle z \rangle = \mathbb{M}/\langle z \rangle$, we have $\mathbb{M} \cong (\mathbb{M} \cap K) \times \langle z \rangle$.

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

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 1.8.22. If K is a twisted subgroup, we define the *K-radical* Ξ_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 1.8.23. *If K is a twisted subgroup and Ξ_K is the K-radical, then Ξ_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 Ξ_K is normal in $\langle K \rangle$, just note that for any $b \in K$ we have

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

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

$$x(k_1 \cdots k_n) = (k_n \cdots k_1)x(k_1 \cdots k_n) = k_n(\cdots (k_1 x k_1) \cdots)k_n \in K. \qquad \square$$

Proposition 1.8.24. *If K is a twisted subgroup with $\Xi_K = 1$, and if τ satisfies $\tau^2 = 1$, $\tau k \tau = k^{-1}$ for $k \in K$, then τK is preserved under conjugation by elements of $\langle K, \tau \rangle$.*

Proof. If $x, y \in K$, then $x^{-1}\tau yx = \tau xyx \in \tau K$, and $\tau^{-1}\tau y\tau = \tau y^{-1} \in \tau K$. $\qquad\square$

Proposition 1.8.25. *A twisted subgroup $K \subseteq \mathbb{G}$ is a near-subgroup of $\mathbb{G}$ iff the intersection $bK \cap \mathbb{S}$ is a subgroup of $\mathbb{S}$ for every 2-Sylow subgroup $\mathbb{S}$ of $\mathbb{G}$ and every $b \in K^{-1}$.*

Proof. Suppose for contradiction that $\mathbb{M} \leq \mathbb{G}$, $\mathbb{N} \triangleleft \mathbb{M}$ with $\mathbb{M}/\mathbb{N}$ isomorphic to the Klein four-group, and $b \in K^{-1}$ with $|(bK \cap \mathbb{M})/\mathbb{N}| = 3$. We may assume without loss of generality that $\mathbb{M} = \langle K \rangle \cap \mathbb{M}$ and $\mathbb{N} = \langle K \rangle \cap \mathbb{N}$, that $\mathbb{G} = \langle K \rangle$, that $b = 1$, and that $\Xi_K = 1$. From $\Xi_K = 1$, we see that there is an order 2 automorphism τ of $\mathbb{G} = \langle K \rangle$ with $k^\tau = k^{-1}$ for all $k \in K$, so we work in the semidirect product of $\mathbb{G}$ and $\langle \tau \rangle$, with $\tau^2 = 1$ and $\tau g\tau = g^\tau$ for $g \in \mathbb{G}$.

Let $x, y \in K$ be representatives of the nontrivial elements of $(K \cap \mathbb{M})/\mathbb{N}$. 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 $\mathbb{S}_x, \mathbb{S}_y$ be 2-Sylow subgroups of $\mathbb{M}\langle \tau \rangle$ containing $\langle x, \tau \rangle, \langle y, \tau \rangle$, respectively, then by the Sylow theorems there is some $g \in \mathbb{M}\langle \tau \rangle$ with $g^{-1}\mathbb{S}_y g = \mathbb{S}_x$. Then $x, \tau, g^{-1}yg, g^{-1}\tau g \in \mathbb{S}_x$, and our strategy is to show that $x, g^{-1}yg \in K \cap \mathbb{S}_x$.

We have $g^{-1}\tau yg \in \tau K$ by the previous proposition, so $\tau g^{-1}\tau yg \in K \cap \mathbb{S}_x$, and similarly $\tau g^{-1}\tau g \in K \cap \mathbb{S}_x$. Since $K \cap \mathbb{S}_x$ is assumed to be a subgroup, we have

$$xg^{-1}yg = x(\tau g^{-1}\tau g)^{-1}(\tau g^{-1}\tau yg) \in K \cap \mathbb{S}_x.$$

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

Proof of Theorem 1.8.16. If K, K' are near-subgroups of $\mathbb{G}$, then they are both twisted subgroups and so their intersection $K \cap K'$ is also a twisted subgroup. Now let $\mathbb{S}$ be any 2-group contained in $\mathbb{G}$, then for any $b \in K \cap K'$ we have $bK \cap bK' \cap \mathbb{S} = (bK \cap \mathbb{S}) \cap (bK' \cap \mathbb{S})$ is an intersection of subgroups of $\mathbb{S}$, so it is a subgroup of $\mathbb{S}$, and the previous proposition shows that this implies that $K \cap K'$ is a near-subgroup of $\mathbb{G}$. $\qquad\square$

1.9 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 1.9.1. An algebra $\mathbb{A}$ is called *quasi-affine* if there is an abelian group $\mathbb{G} = (G, 0, +, -)$ with underlying set G containing the underlying set of $\mathbb{A}$, such that the restriction of the 4-ary relation $x + y = z + w$ to $\mathbb{A}$ is preserved by all the operations of $\mathbb{A}$.

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

Definition 1.9.2. If $\mathbb{R}$ is a ring and $\mathbb{M}$ is a module over $\mathbb{R}$ with underlying group $(M, 0, +, -)$, then we consider $\mathbb{M}$ to be a universal algebraic object $(M, 0, +, -, \{\phi_r\}_{r \in \mathbb{R}})$, where for each $r \in \mathbb{R}$ the unary operation $\phi_r : \mathbb{M} \to \mathbb{M}$ 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 1.9.3. If $\mathcal{O}$ is any set of operations, then the *polynomial clone* generated by $\mathcal{O}$ is the clone generated by $\mathcal{O}$ 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 1.9.4. *An algebra $\mathbb{A}$ is quasi-affine iff it is a subalgebra of a reduct of the polynomial clone of a module.*

Proof. Let $\mathbb{A}$ be a quasi-affine algebra, and let $\mathbb{G} = (G, 0, +, -)$ be the corresponding group. We may assume without loss of generality that $0 \in \mathbb{A}$, and that $\mathbb{G}$ is the abelian group with the following presentation: the generators are the elements of $\mathbb{A} \setminus \{0\}$, and the relations are given by $x + y - z - w = 0$ for every quadruple of elements $x, y, z, w \in \mathbb{A}$ such that $x + y = z + w$ in $\mathbb{G}$.

Suppose that f is any n-ary operation of $\mathbb{A}$, and for each $i \leq n$ let $\phi_i : \mathbb{A} \to G$ be the unary operation given by
$$\phi_i(x) = f(0, ..., 0, x, 0, ..., 0) - f(0, ..., 0),$$
with the x in the ith position. Since f preserves the relation $x + y = z + w$ on $\mathbb{A}$, we have
$$\phi_i(x) + \phi_i(y) = \phi_i(z) + \phi_i(w)$$
for any $x, y, z, w \in \mathbb{A}$ such that $x + y = z + w$ in $\mathbb{G}$. Thus ϕ_i is compatible with the defining relations of $\mathbb{G}$, so the map $\phi_i : \mathbb{A} \to G$ extends to a unique homomorphism $\phi_i : \mathbb{G} \to \mathbb{G}$.

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 $x_1, ..., x_n \in \mathbb{A}$, 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 ϕ_i. For the inductive step, assume without loss of generality that the nonzero values of the x_is 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 ϕ_{k+1} we have
$$f(x_1, ..., x_{k+1}, 0, ..., 0) = \phi_1(x_1) + \cdots + \phi_k(x_k) + f(0, ..., 0) + \phi_{k+1}(x_{k+1}). \qquad \square$$

Definition 1.9.5. An algebra $\mathbb{A}$ is called *affine* if it is polynomially equivalent to a module.

Proposition 1.9.6. *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 $\mathbb{A}$ has a Mal'cev term. Since $\mathbb{A}$ is polynomially equivalent to a module, there must be some $n + 3$-ary term t and some constants $a_1, ..., a_n \in \mathbb{A}$ 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 ϕ_i and some constant c. Define $p(x, y, z)$ by

$$p(x, y, z) = t(x, t(y, y, y, x, ..., x), z, x, ..., x).$$

Then p is a term of $\mathbb{A}$, and we have

$$p(x, y, z) = x - \Big(y - y + y + \sum_i \phi_i(x) + c\Big) + z + \sum_i \phi_i(x) + c = x - y + z,$$

so p is Mal'cev.

For the converse, if $\mathbb{A}$ 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 $\mathbb{A}$, and therefore for each term f of $\mathbb{A}$ the unary function $\phi(x) = f(x, 0, ..., 0) - f(0, ..., 0)$ is a polynomial operation of $\mathbb{A}$ 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 1.9.7. An algebraic structure $\mathbb{A}$ is called *abelian* if there is a congruence Θ on $\mathbb{A} \times \mathbb{A}$ such that the diagonal $\Delta_{\mathbb{A}} = \{(a, a) \mid a \in \mathbb{A}\}$ is one of the congruence classes of Θ.

Proposition 1.9.8. *A group is abelian iff it is commutative.*

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

$$(1, b)(a, a)(1, b)^{-1} = (a, bab^{-1}),$$

the normality of $\Delta_{\mathbb{G}}$ is equivalent to the identity $a \approx bab^{-1}$, which is equivalent to $ab \approx ba$.

Alternatively, we can argue as follows. The group $\mathbb{G}$ is commutative iff the map $\mathbb{G} \to \mathbb{G}$ given by $x \mapsto x^{-1}$ is a homomorphism, and if this occurs then there is a homomorphism $\mathbb{G} \times \mathbb{G} \to \mathbb{G}$ such that the restriction $\mathbb{G} \times \{1\} \to \mathbb{G}$ is the identity, and such that the diagonal maps to $\{1\}$. Conversely, if the diagonal is a normal subgroup, then every coset intersects $\mathbb{G} \times \{1\}$ and $\{1\} \times \mathbb{G}$ exactly once, so the quotient $\mathbb{G} \times \mathbb{G}/\Delta_{\mathbb{G}}$ is isomorphic to $\mathbb{G}$ in two different ways, and composing these isomorphisms we obtain the map $x \mapsto x^{-1}$, so $\mathbb{G}$ 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 1.9.9. 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

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

Proposition 1.9.10. *An algebra $\mathbb{A}$ is abelian iff it satisfies the term condition.*

Proof. We think of congruences on $\mathbb{A}^2$ as subalgebras of $\mathbb{A}^{2\times2}$, the set of 2×2 matrices with entries in $\mathbb{A}$ (here elements of $\mathbb{A}^2$ are visualized as column vectors, and an element of $\mathbb{A}^{2\times2}$ 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 $\mathbb{M} \le \mathbb{A}^{2\times2}$ generated by matrices of the form

$$\begin{bmatrix} u & v \\ u & v \end{bmatrix}, \quad \begin{bmatrix} a & a \\ b & b \end{bmatrix},$$

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 $\mathbb{M}$ as a binary relation on $\mathbb{A}^2$, the transitive closure of $\mathbb{M}$ is a congruence Θ on $\mathbb{A}^2$, and it is clearly as small as possible given that $\Delta_{\mathbb{A}}$ is contained in a congruence class of Θ.

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

$$t(u_1, ..., u_m, a_1, ..., a_n) = t(u_1, ..., u_m, b_1, ..., b_n)$$

but

$$t(v_1, ..., v_m, a_1, ..., a_n) \ne t(v_1, ..., v_m, b_1, ..., b_n).$$

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

$$t(u_1, ..., u_m, a_1, ..., a_n) = t(u_1, ..., u_m, b_1, ..., b_n) \implies t(v_1, ..., v_{m-1}, u_m, a_1, ..., a_n) = t(v_1, ..., v_{m-1}, u_m, b_1, ..., b_n),$$

and then by the term condition applied to a version of t with variables permuted so that the mth variable becomes the first, this implies that

$$t(v_1, ..., v_m, a_1, ..., a_n) = t(v_1, ..., v_m, b_1, ..., b_n). \qquad \square$$

Proposition 1.9.11. *Every quasi-affine algebra satisfies the term condition and is therefore abelian.*

Proof. If t is an $n + 1$-ary term of a quasi-affine algebra, then we can write t in the form

$$t(x_0, ..., x_n) = \phi_0(x_0) + \cdots + \phi_n(x_n) + c,$$

where the ϕ_i are unary and c is a constant. Then for any $u \in \mathbb{A}, a_i, b_i \in \mathbb{A}$, we have

$$t(u, a_1, ..., a_n) = t(u, b_1, ..., b_n) \iff \phi_1(a_1) + \cdots + \phi_n(a_n) = \phi_1(b_1) + \cdots + \phi_n(b_n),$$

and this is a condition which does not depend on the value of u. $\qquad \square$

Example 1.9.1. 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 $t(x, y) = yxy^{-1}$, then the group is commutative, since we have $t(1, 1) = t(1, y) \iff t(x, 1) = t(x, y)$, that is, $1 = yy^{-1} \iff x = yxy^{-1}$.

Example 1.9.2. A ring is abelian in the sense of universal algebra iff it is a *zero ring*, that is, a ring satisfying the identity $xy \approx 0$. To see the necessity, we apply the term condition with the term $t(x, y) = xy$ and the pairs $(u, v) = (0, x)$ and $(a, b) = (0, y)$, to see that $0 \cdot 0 = 0 \cdot y \iff x \cdot 0 = x \cdot y$. To see the sufficiency, note that every zero ring is affine.

Example 1.9.3. The quasigroup with multiplication table

$\cdot$	0	1	2	3
0	3	2	0	1
1	2	3	1	0
2	1	0	2	3
3	0	1	3	2

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 ϕ is the transposition (2 3). This example is from [62].

In terms of congruence lattices, the main important feature of an affine algebra $\mathbb{A}$ is that $\mathrm{Con}(\mathbb{A} \times \mathbb{A})$ contains the following five element sublattice.
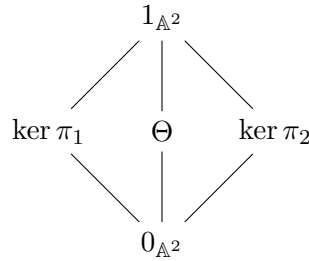
The abstract five element lattice corresponding to this picture is known as the diamond lattice $\mathcal{M}_3$. The lattice $\mathcal{M}_3$ has a special role in lattice theory: every modular lattice which isn't distributive contains a sublattice which is isomorphic to $\mathcal{M}_3$ (see Proposition A.4.2 in the appendix).

Theorem 1.9.12. *If $\mathbb{A}$ is an abelian Mal'cev algebra, and if Θ 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 $\mathrm{Con}(\mathbb{A}^2)$ isomorphic to $\mathcal{M}_3$.*

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

$$\begin{bmatrix} b \\ d \end{bmatrix} = p\left(\begin{bmatrix} b \\ b \end{bmatrix}, \begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} a \\ d \end{bmatrix} \right) \equiv_\Theta p\left(\begin{bmatrix} b \\ b \end{bmatrix}, \begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} a \\ b \end{bmatrix} \right) = \begin{bmatrix} b \\ b \end{bmatrix} \in \Delta_{\mathbb{A}},$$

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 $\Theta \wedge \ker \pi_1 = 0_{\mathbb{A}^2}$. $\square$

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 1.9.13. 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 Θ_i which contains p, and

- if l_i, l_j are lines of Θ_i, Θ_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 Θ_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 1.9.14. *If $\mathbb{A}$ is an abelian Mal'cev algebra and Θ is any congruence of $\mathbb{A}^2$ with the diagonal as a congruence class, then $(\mathbb{A}^2, \ker \pi_1, \ker \pi_2, \Theta)$ 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 [118]. 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 1.9.15. *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 $\mathbb{L} = (L, \cdot, 1)$ 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*

$$
\begin{aligned}
(x, y) \equiv_{\Theta_1} (x', y') &\iff x = x', \\
(x, y) \equiv_{\Theta_2} (x', y') &\iff y = y', \\
(x, y) \equiv_{\Theta_3} (x', y') &\iff x \cdot y = x' \cdot y',
\end{aligned}
$$

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 Θ_1, and take $1 = e$. Let l_2 be the line through e in the parallel class Θ_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 Θ_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.
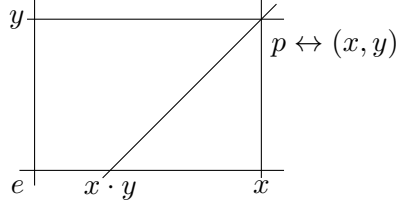
Finally, to define the multiplication $\cdot$ on L, note that for every $x, y \in L$ there is a point $p \in S$ corresponding to (x, y), and this point p is in a unique line $l_3 \in \Theta_3$. We then define $x \cdot y$ to be the element of L corresponding to the point $l_3 \cap l_1$, or alternatively 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). $\square$

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

Lemma 1.9.16. *If an S-3-system* $\mathbf{S} = (S, \Theta_1, \Theta_2, \Theta_3)$ *has a Mal'cev polymorphism* p, *then* p *is completely determined by* $\mathbf{S}$. *In fact,* $p(x, y, z)$ *can be "geometrically constructed" from the points* x, y, z.
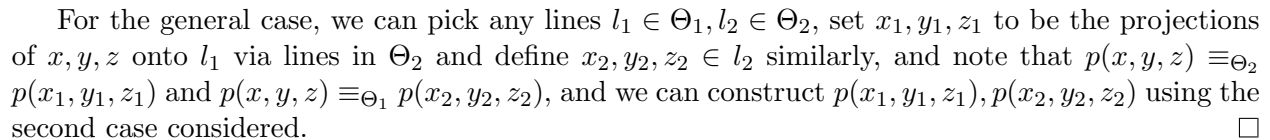
Proof. First consider the special case where x, y lie on a line l_1 and y, z lie on a different line l_2. Suppose that $l_1 \in \Theta_1$ and $l_2 \in \Theta_2$. Then $p(x, y, z) \equiv_{\Theta_1} p(y, y, z) = z$ and $p(x, y, z) \equiv_{\Theta_2} p(x, y, y) = x$, so if we draw the line $l_2' \in \Theta_2$ through x and the line $l_1' \in \Theta_1$ through z, we see that $p(x, y, z)$ is the intersection point $l_1' \cap l_2'$.
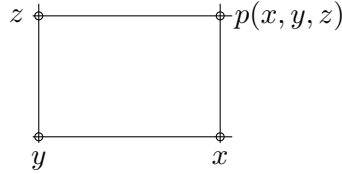
Next consider the special case where x, y, z lie on a line l_1, and suppose $l_1 \in \Theta_1$. Draw the line $l_2 \in \Theta_2$ through y and the line $l_3 \in \Theta_3$ through x, and let $y' \in l_2 \cap l_3$ be their point of intersection. Draw the line l_1' through y' parallel to l_1, draw the line l_2' through z 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'.
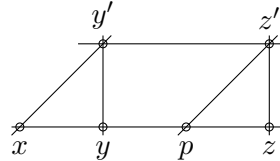
We claim that $p = p(x, y, z)$. To see this, note that $x \equiv_{\Theta_3} y'$, so $p(x, y, z) \equiv_{\Theta_3} p(y', y, z)$, and $p(y', y, z) = z'$ by the first case we considered. Thus $p(x, y, z) \equiv_{\Theta_3} z'$, i.e. $p(x, y, z) \in l_3'$, and since $x \equiv_{\Theta_1} y \equiv_{\Theta_1} z$, 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.)

74

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 Θ_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. $\square$

Corollary 1.9.17. *If p is a Mal'cev polymorphism of an S-3-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 $p(x, y, z)$. $\square$

Corollary 1.9.18. *If p is a Mal'cev polymorphism of an S-3-system $(S, \Theta_1, \Theta_2, \Theta_3)$, then the graph Γ_p of p, considered as a 4-ary relation on S, is primitively positively definable from $\Theta_1, \Theta_2, \Theta_3$.*

Corollary 1.9.18 can also be interpreted as saying that the map $p : \mathbb{S}^3 \to \mathbb{S}$ is a homomorphism of the *algebraic* structure $\mathbb{S}$ consisting of all polymorphisms of the relational structure $\mathbf{S}$. In particular, p "commutes with itself", that is, the two ways of computing $p * p$ on a 3×3 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 1.9.19. An n-ary term t of an algebraic structure $\mathbb{A}$ is called *central* if the map $t : \mathbb{A}^n \to \mathbb{A}$ is a homomorphism.

Now we relate the Mal'cev polymorphism to the coordinate loop $\mathbb{L}$. First we will show that $\mathbb{L}$ is associative.

Lemma 1.9.20. *If $\mathbf{S} = (S, \Theta_1, \Theta_2, \Theta_3)$ is an S-3-system with a Mal'cev polymorphism p, and if $\mathbb{L}$ is a coordinate loop of $\mathbf{S}$, then $\mathbb{L}$ satisfies*

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

In particular, $\mathbb{L}$ is associative, that is, $\mathbb{L}$ is a group.

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

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

To see that this implies the associativity of $\mathbb{L}$, 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

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

For a geometric way to visualize the proof, note that the stated property of $\mathbb{L}$ corresponds to the existence of the dashed line in the following picture.
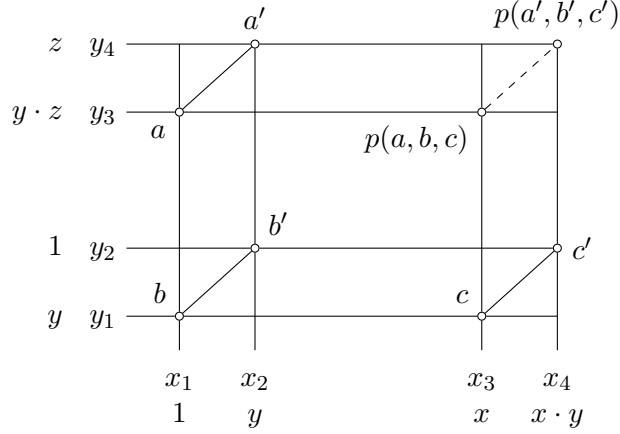
If we set $a = (x_1, y_3)$, etc. as in the picture, then the existence of the dashed line follows from the fact that p preserves the congruence Θ_3 and the fact that $p(a, b, c)$ completes the parallelogram through a, b, c and $p(a', b', c')$ completes the parallelogram through a', b', c'. $\qquad\square$

Lemma 1.9.21. *If* $\mathbf{S} = (S, \Theta_1, \Theta_2, \Theta_3)$ *is an S-3-system with a Mal'cev polymorphism* p, *and if* $\mathbb{L}$ *is a coordinate group of* $\mathbf{S}$, *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 Θ_1-coordinate 1. Consider the diagram
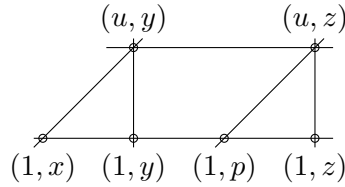
which we used to construct $p(x, y, z)$. Then from $(1, x) \equiv_{\Theta_3} (u, y)$ we have $1 \cdot x = u \cdot y$, and from $(1, p) \equiv_{\Theta_3} (u, z)$ we have $1 \cdot p = u \cdot z$. Solving for u we get $u = xy^{-1}$, and solving for p we get $p = xy^{-1}z$. $\qquad\square$

Corollary 1.9.22. *If* $\mathbf{S} = (S, \Theta_1, \Theta_2, \Theta_3)$ *is an S-3-system with a Mal'cev polymorphism* p, *and if* $\mathbb{L}$ *is a coordinate group of* $\mathbf{S}$, *then* $\mathbb{L}$ *is commutative.*

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

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

Theorem 1.9.23. *Any abelian Mal'cev algebra* $\mathbb{A}$ *is affine.*

Proof. By Theorem 1.9.12 and its corollary, $\mathbf{S} = (\mathbb{A}^2, \ker \pi_1, \ker \pi_2, \Theta)$ is an S-3-system with Mal'cev polymorphism p, where p is the Mal'cev term of $\mathbb{A}$ and Θ is a congruence on $\mathbb{A}^2$ with the diagonal as a congruence class. By Lemma 1.9.15, there is a loop structure $\mathbb{L}$ on the underlying set of $\mathbb{A}$ which describes $\mathbf{S}$. By Lemma 1.9.20, Lemma 1.9.21, and its corollary, $\mathbb{L}$ is an abelian group and p is given by $p(x, y, z) = x - y + z$ (writing the abelian group operation additively).

By Corollary 1.9.18, 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. $\qquad\qquad\square$

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

(1) $\mathbb{A}$ is affine,

(2) $\mathbb{A}$ is abelian and has a Mal'cev polynomial,

(3) $\mathbb{A}$ 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 $\mathbb{A}$ preserves every congruence of $\mathbb{A}^2$, so the polynomial clone of $\mathbb{A}$ 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 $\mathbb{A}$ is quasi-affine.

So suppose that p is a Mal'cev operation which commutes with itself, and pick any element to call 0 in $\mathbb{A}$. We define addition and negation on $\mathbb{A}$ 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 $p(x - y, 0, z) = x - y + z$. $\qquad\square$

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" [68], 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 [74], 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 [88] show that any abelian Taylor algebra is quasi-affine - the example $(\mathbb{R}, \frac{x+y}{2})$ 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 [21].

Remark 1.9.1. 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 1.9.1. 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 [123] 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 $x_3 \cdot x_2 \neq x_7 \cdot x_6$. Another example with just five elements is given in Example B.3.1.

Kearnes [87] 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 $\mathbb{A}^n$ 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 [85], Kearnes shows that any abelian algebra with a central binary polynomial which is cancellative is quasi-affine, and in [133] this is extended to the result that any abelian algebra with a commutative cancellative polynomial is quasi-affine. In [79], it is shown that abelian quandles are quasi-affine.

1.9.1 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 1.9.25. If $\alpha, \beta, \delta \in \text{Con}(\mathbb{A})$, we say that α *centralizes* β modulo δ, written $C(\alpha, \beta; \delta)$ (or $C(\alpha, \beta)$ if $\delta = 0_{\mathbb{A}}$), if for every $n+1$-ary term $t \in \text{Clo}_{n+1}(\mathbb{A})$, 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 δ which satisfies $C(\alpha, \beta; \delta)$ is called the *commutator* of α, β, 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 1.9.26. *If α centralizes β modulo δ, then for every $m + n$-ary term $t \in \text{Clo}_{m+n}(\mathbb{A})$, for any $(u_1, v_1), ..., (u_m, v_m) \in \alpha$, and for any $(a_1, b_1), ..., (a_n, b_n) \in \beta$, we have*

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

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

Proposition 1.9.27. *If $\mathbb{M}, \mathbb{N}$ are normal subgroups of a group $\mathbb{G}$, $[\mathbb{M}, \mathbb{N}]$ 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 $\mathbb{G}$ such that every element of $\mathbb{M}/\mathbb{K}$ commutes with every element of $\mathbb{N}/\mathbb{K}$ in $\mathbb{G}/\mathbb{K}$.

First suppose that $\theta_{\mathbb{M}}$ centralizes $\theta_{\mathbb{N}}$. Let t be the binary term $t(x, y) = yxy^{-1}$, 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 $t \in \text{Clo}_{n+1}(\mathbb{G})$ and any $(u, v) \in \theta_{\mathbb{M}}, (a_1, b_1), ..., (a_n, b_n) \in \theta_{\mathbb{N}}$ with

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

Thinking of $t(ux, a_1y_1, ..., a_ny_n)$ as a function of $x, y_1, ..., y_n$ with parameters $u, a_1, ..., a_n$, we may rearrange it into the form

$$t(ux, a_1y_1, ..., a_ny_n) = t'(x, y_1, ..., y_n)t(u, a_1, ..., a_n)$$

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

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

To show that

$$t(v, a_1, ..., a_n) = t(v, b_1, ..., b_n),$$

we just need to show that

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

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

Example 1.9.4. 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 1.9.5. In a majority algebra, the commutator is given by $[\alpha, \beta] = \alpha \wedge \beta$. To see this, suppose $(a, b) \in \alpha \wedge \beta$, and apply the term condition to the majority operation to see that

$$m(\boxed{a}, a, a) = m(\boxed{a}, a, b) \implies m(\boxed{b}, a, a) \ [\alpha, \beta] \ m(\boxed{b}, a, b),$$

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 1.9.6. In a semilattice, the commutator is given by $[\alpha, \beta] = \alpha \wedge \beta$. 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 $(a, b) \in \alpha \wedge \beta$, we have

$$s_3(\boxed{a}, a, b) = s_3(\boxed{a}, b, b) \implies s_3(\boxed{b}, a, b) \ [\alpha, \beta] \ s_3(\boxed{b}, b, b),$$

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 2×2 matrices.

Definition 1.9.28. For $\alpha, \beta \in \mathrm{Con}(\mathbb{A})$, we define the algebra $\mathbb{M}(\alpha, \beta) \leq \mathbb{A}^{2\times2}$ to be the subalgebra of 2×2 matrices which is generated by the matrices of the form

$$\begin{bmatrix} u & u \\ v & v \end{bmatrix} \text{ with } (u, v) \in \alpha, \quad \begin{bmatrix} a & b \\ a & b \end{bmatrix} \text{ with } (a, b) \in \beta.$$
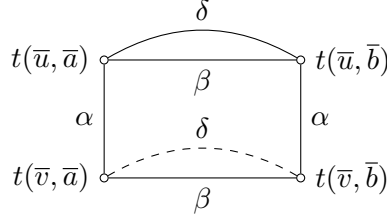
Proposition 1.9.29. *If $\alpha, \beta, \delta \in \mathrm{Con}(\mathbb{A})$, then α centralizes β modulo δ iff for all*

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathbb{M}(\alpha, \beta)$$

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:

$$\begin{array}{c}
t(\overline{u},\overline{a}) \overset{\delta}{\underset{\beta}{\rule{0pt}{0pt}}} t(\overline{u},\overline{b}) \\
\alpha \qquad \delta \qquad \alpha \\
t(\overline{v},\overline{a}) \underset{\beta}{\rule{0pt}{0pt}} t(\overline{v},\overline{b})
\end{array}$$

where the positioning of the four corners matches with the way we have laid out the 2×2 matrices in $\mathbb{M}(\alpha,\beta)$. A mnemonic for remembering where the δ edges go is that in the term condition $C(\alpha,\beta;\delta)$, "δ is next to β".

We now list a few elementary properties of the commutator which hold in general, which are given as exercises in Hobby and McKenzie's book [74].

Proposition 1.9.30. *For $\alpha,\beta,\delta \in \mathrm{Con}(\mathbb{A})$, 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 $(\alpha \vee (\beta \wedge \delta)) \wedge \beta \leq \delta$ then $C(\alpha,\beta;\delta)$ holds, so $[\alpha,\beta] \leq \alpha \wedge \beta$,*

(c) *if $\alpha' \leq \alpha, \beta' \leq \beta$, then $C(\alpha,\beta;\delta) \implies C(\alpha',\beta';\delta)$, so $[\alpha',\beta'] \leq [\alpha,\beta]$,*

(d) *for any γ we have $C(\alpha,\beta;\delta) \implies C(\alpha \wedge \gamma, \beta; \delta \wedge \gamma)$,*

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

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

(g) *if $\mathbb{B} \leq \mathbb{A}$ then $C(\alpha,\beta;\delta) \implies C(\alpha|_\mathbb{B}, \beta|_\mathbb{B}; \delta|_\mathbb{B})$, so $[\alpha|_\mathbb{B}, \beta|_\mathbb{B}] \leq [\alpha,\beta]|_\mathbb{B}$,*

(h) *if $[\alpha,\alpha] = 0_\mathbb{A}$, then any congruence class of α 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 $\begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathbb{M}(\alpha,\beta)$ with $a \equiv_\delta b$, we have $c \equiv_\alpha a \equiv_{\beta \wedge \delta} b \equiv_\alpha d$ and $c \equiv_\beta d$, so $(c,d) \in (\alpha \circ (\beta \wedge \delta) \circ \alpha) \wedge \beta$, which is a subset of δ by assumption.

For (e), we string together several instances of the term condition: if $(u,v) \in \bigvee_i \alpha_i$, $(a_i,b_i) \in \beta$, 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 $C(\alpha_{j_i}, \beta; \delta)$ we have

$$t(u_i, \overline{a}) \equiv_\delta t(u_i, \overline{b}) \implies t(u_{i+1}, \overline{a}) \equiv_\delta t(u_{i+1}, \overline{b}),$$

so by inducting on i we get $t(v, \overline{a}) \equiv_\delta t(v, \overline{b})$. $\qquad\square$

Corollary 1.9.31. *If an idempotent algebra $\mathbb{A}$ has any congruences $\alpha, \beta \in \mathrm{Con}(\mathbb{A})$ with $[\alpha,\beta] \neq \alpha \wedge \beta$, then some subalgebra of some quotient of $\mathbb{A}$ is a nontrivial abelian algebra.*

Proof. Let $\delta = \alpha \wedge \beta$, then from $\delta \leq \alpha, \beta$ we have $[\delta,\delta] \leq [\alpha,\beta] < \alpha \wedge \beta = \delta$. 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 $\mathbb{B}$ of δ' and $\mathbb{B}$ is an abelian subalgebra of $\mathbb{A}/[\delta,\delta]$. $\qquad\square$

Proposition 1.9.32. *If $[\alpha, \beta] = \alpha \wedge \beta$ for all $\alpha, \beta \in \mathrm{Con}(\mathbb{A})$, then $\mathrm{Con}(\mathbb{A})$ satisfies the meet-semidistributive law:*

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

Proof. If $\alpha, \beta, \gamma \in \mathrm{Con}(\mathbb{A})$ satisfy $\alpha \wedge \beta = \alpha \wedge \gamma$, then $C(\beta, \alpha; \alpha \wedge \beta)$ and $C(\gamma, \alpha; \alpha \wedge \beta)$ hold, so $C(\beta \vee \gamma, \alpha; \alpha \wedge \beta)$ holds, so $\alpha \wedge (\beta \vee \gamma) = [\beta \vee \gamma, \alpha] \leq \alpha \wedge \beta$. $\square$

Definition 1.9.33. An algebra $\mathbb{A}$ is *congruence meet-semidistributive*, written $\mathrm{SD}(\wedge)$ for short, if for all $\alpha, \beta, \gamma \in \mathrm{Con}(\mathbb{A})$ with $\alpha \wedge \beta = \alpha \wedge \gamma$, we have $\alpha \wedge (\beta \vee \gamma) = \alpha \wedge \beta$. A variety $\mathcal{V}$ is $\mathrm{SD}(\wedge)$ if every algebra $\mathbb{A} \in \mathcal{V}$ is $\mathrm{SD}(\wedge)$.

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 $\mathrm{SD}(\wedge)$ variety if and only if the associated CSP has bounded width.

Corollary 1.9.34. *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 $\mathbb{A}$ is affine, then $\mathrm{Con}(\mathbb{A}^2)$ contains a copy of the diamond lattice $\mathcal{M}_3$, and $\mathcal{M}_3$ doesn't satisfy the meet-semidistributive law. $\square$

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 1.9.35. Suppose that $\alpha \leq \beta \in \mathrm{Con}(\mathbb{A})$. We say that β is *abelian* over α if the term condition $C(\beta, \beta; \alpha)$ holds. We say that β is *solvable* over α if there is a chain of congruences $\alpha = \alpha_0 \leq \cdots \leq \alpha_n = \beta$ such that α_{i+1} is abelian over α_i for each i.

A congruence α is called abelian if it is abelian over $0_\mathbb{A}$ (equivalently $[\alpha, \alpha] = 0_\mathbb{A}$), and similarly α is called solvable if α is solvable over $0_\mathbb{A}$. An algebra $\mathbb{A}$ is called solvable if $1_\mathbb{A}$ is solvable.

The *center* of an algebra $\mathbb{A}$ is defined to be the largest ζ such that $C(\zeta, 1_\mathbb{A})$ holds (equivalently, the largest ζ with $[\zeta, 1_\mathbb{A}] = 0_\mathbb{A}$). For β a congruence, we define the *centralizer* of β, written $(0 : \beta)$, to be the largest congruence α such that $[\alpha, \beta] = 0$, and more generally for any δ we define $(\delta : \beta)$ to be the largest α such that $C(\alpha, \beta; \delta)$ holds.

Proposition 1.9.36. *For congruences on $\mathbb{A}$, we have the following:*

(a) *for any β, δ there exists a largest α such that $C(\alpha, \beta; \delta)$ holds, so $(\delta : \beta)$ (and, in particular, the center of $\mathbb{A}$) is well-defined,*

(b) *if γ is solvable over β and β is solvable over α, then γ is solvable over α,*

(c) *if β is solvable (abelian) over α, then $\beta \wedge \gamma$ is solvable (abelian) over $\alpha \wedge \gamma$ for any γ,*

(d) *if $\theta \leq \alpha \leq \beta$, then β is solvable (abelian) over α iff β/θ is solvable (abelian) over α/θ,*

(e) *$\mathbb{A}/\theta$ is solvable (abelian) iff $1_\mathbb{A}$ is solvable (abelian) over θ.*

Proof. Part (a) follows from Proposition 1.9.30(e), part (b) is obvious, part(c) follows from Proposition 1.9.30(d), part (d) follows from Proposition 1.9.30(f), and part (e) is part (d) specialized to the case $\beta = 1_\mathbb{A}, \alpha = \theta$. $\square$

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 1.9.37. 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 θ a congruence, we always have $p(x, y, y) \equiv_{[\theta, \theta]} x$.

Example 1.9.7. 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 [88].

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 1.9.38. For any $n \geq 1$, we define the *n-cycle commutator* $[\alpha, \beta]_n$ to be the least congruence δ such that for any cycle of n matrices

$$\begin{bmatrix} a_1 & b_1 \\ c_1 & d_1 \end{bmatrix}, \begin{bmatrix} a_2 & b_2 \\ c_2 & d_2 \end{bmatrix}, ..., \begin{bmatrix} a_n & b_n \\ c_n & d_n \end{bmatrix} \in \mathbb{M}(\alpha, \beta)$$

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 $\mathbb{A}$ is affine, then it is easy to check that $[1_\mathbb{A}, 1_\mathbb{A}]_n = 0_\mathbb{A}$ for every n. Note that for $n = 1$, we have $[\alpha, \beta]_1 = [\alpha, \beta]$. Additionally, since we can take the nth 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 [123] is an example of an algebra where $[1_\mathbb{A}, 1_\mathbb{A}]_1 = 0_\mathbb{A}$ but $[1_\mathbb{A}, 1_\mathbb{A}]_2 \neq 0_\mathbb{A}$.

For $n = 2$, the 2-cycle commutator is clearly symmetric: $[\alpha, \beta]_2 = [\beta, \alpha]_2$. Since it is defined via two matrices in $\mathbb{M}(\alpha, \beta)$, 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.

83
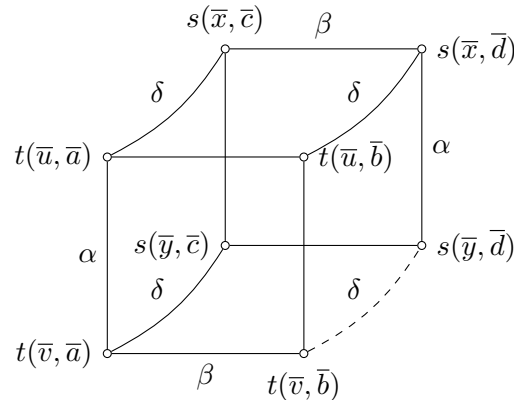
If we have a difference term, then all of the n-cycle commutators turn out to be equal.

Theorem 1.9.39 (Lipparini [102])**.** *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 $\begin{bmatrix} a_i & b_i \\ c_i & d_i \end{bmatrix} \in \mathbb{M}(\alpha, \beta)$ for $i \leq n$ are as in the definition of the n-cycle condition for $\delta = [\alpha, \beta]$. Applying the difference term, we have

$$p\left(\begin{bmatrix} a_i & b_i \\ c_i & d_i \end{bmatrix}, \begin{bmatrix} b_1 & b_1 \\ d_1 & d_1 \end{bmatrix}, \begin{bmatrix} a_1 & a_1 \\ c_1 & c_1 \end{bmatrix} \right) = \begin{bmatrix} p(a_i, b_1, a_1) & p(b_i, b_1, a_1) \\ p(c_i, d_1, c_1) & p(d_i, d_1, c_1) \end{bmatrix} \in \mathbb{M}(\alpha, \beta)$$

for $2 \leq i \leq n - 1$, and

$$p\left(\begin{bmatrix} a_n & b_n \\ c_n & d_n \end{bmatrix}, \begin{bmatrix} b_1 & a_1 \\ d_1 & c_1 \end{bmatrix}, \begin{bmatrix} a_1 & a_1 \\ c_1 & c_1 \end{bmatrix} \right) = \begin{bmatrix} p(a_n, b_1, a_1) & p(b_n, a_1, a_1) \\ p(c_n, d_1, c_1) & p(d_n, c_1, c_1) \end{bmatrix} \in \mathbb{M}(\alpha, \beta).$$

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 congruenct to c_1 modulo $[\alpha, \beta]$. From the fact that $c_1 \equiv_\beta d_n$ and $(c_1, d_n) \in \alpha \circ [\alpha, \beta] \circ \alpha = \alpha$, we have $(c_1, d_n) \in \alpha \wedge \beta$, 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 $c_1 \equiv_{[\alpha, \beta]} d_n$. $\qquad \square$

In fact, substantially more is true in varieties with a difference term. Kearnes [86] shows that almost all properties of the commutator which hold in congruence modular varieties generalize to varieties with a difference term, other than $[\alpha_1 \vee \alpha_2, \beta] = [\alpha_1, \beta] \vee [\alpha_2, \beta]$. This property must be weakened, but it is at least true that if $[\alpha_1, \beta] = [\alpha_2, \beta]$ then $[\alpha_1 \vee \alpha_2, \beta] = [\alpha_1, \beta]$ in varieties with difference terms.

If we go beyond varieties with a difference term, the commutator may no longer be symmetric. For instance, in the algebra $\mathbb{A} = (\{0, 1, 2, *\}, \cdot)$ with $\cdot$ given by

$\cdot$	0	1	2	$*$
0	0	2	1	$*$
1	2	1	0	$*$
2	1	0	2	$*$
$*$	$*$	$*$	$*$	$*$

,

if we let $\theta \in \mathrm{Con}(\mathbb{A})$ be the congruence corresponding to the partition $\{0, 1, 2\}, \{*\}$, then we have

$$[\theta, 1_\mathbb{A}] = 0_\mathbb{A}, \quad [1_\mathbb{A}, \theta] = \theta.$$

The simplest way to fix this asymmetry is to make the following definition from [88].

Definition 1.9.40. If $\alpha, \beta \in \mathrm{Con}(\mathbb{A})$, then we define their *symmetric commutator*, written $[\alpha, \beta]_s$, to be the least congruence δ such that both $C(\alpha, \beta; \delta)$ and $C(\beta, \alpha; \delta)$ hold.

84

To see that $[\alpha, \beta]_s$ is well-defined, we use Proposition 1.9.30(a) to see that the intersection of any collection of congruences that simultaneously satisfy $C(\alpha, \beta; \delta)$ and $C(\beta, \alpha; \delta)$ will also satisfy these two term conditions. Since the two-term commutator satisfies $C(\alpha, \beta; [\alpha, \beta]_2)$ and $C(\beta, \alpha; [\alpha, \beta]_2)$, we always have $[\alpha, \beta]_s \leq [\alpha, \beta]_2$.

We can also go in the other direction, and define a more general commutator by trying to directly think about what an algebra needs to satisfy to be quasi-affine. This leads to the following definition.

Definition 1.9.41. If $\alpha, \beta \in \mathrm{Con}(\mathbb{A})$, then we define their *linear commutator*, written $[\alpha, \beta]_\ell$, as follows. Define a group $\mathbb{G}$ with the following presentation: the generators of $\mathbb{G}$ are the elements of $\mathbb{A}$, and the relations are given by

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathbb{M}(\alpha, \beta) \quad \implies \quad a + d = b + c \text{ in } \mathbb{G}.$$

Then we define the equivalence relation $[\alpha, \beta]_\ell$ to be the kernel of the natural map $\mathbb{A} \to \mathbb{G}$.

Proposition 1.9.42. *The linear commutator $[\alpha, \beta]_\ell$ always defines a congruence of $\mathbb{A}$, we have $[\alpha, \beta]_n \leq [\alpha, \beta]_\ell$ for each n and $[\alpha, \beta]_\ell \leq \alpha \wedge \beta$, and $[1_\mathbb{A}, 1_\mathbb{A}]_\ell = 0_\mathbb{A}$ iff $\mathbb{A}$ is quasi-affine.*

Proof. The linear commutator can be defined combinatorially as follows. For each matrix $M = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathbb{M}(\alpha, \beta)$, call a, d the "positive" corners of the matrix M, and b, c the "negative" corners of M. Then $(x, y) \in [\alpha, \beta]_\ell$ iff there is some collection of matrices $M_i \in \mathbb{M}(\alpha, \beta)$ and a way to pair off values in the positive corners of the matrices M_i to equal values in negative corners of the matrices, so that the only unpaired values are x and y, with one occuring in a positive corner of some matrix and the other occuring in a negative corner of some matrix. This defines the linear commutator $[\alpha, \beta]_\ell$ as the union of a directed limit of relations defined by primitive positive formulas in $\mathbb{M}(\alpha, \beta)$, so the equivalence relation $[\alpha, \beta]_\ell$ is compatible with the operations of $\mathbb{A}$.

The inequality $[\alpha, \beta]_n \leq [\alpha, \beta]_\ell$ follows from the combinatorial description of $[\alpha, \beta]_n$ above (the n-cycle condition is a special case of the general setup of matching corners of matrices together). To prove that $[\alpha, \beta]_\ell \leq \alpha \wedge \beta$, we just need to check that $[\alpha, \beta]_\ell \leq \alpha$ by symmetry, and this follows by chasing equalities and congruences through the matrices in the combinatorial description of $[\alpha, \beta]_\ell$.

If $\mathbb{A}$ is quasi-affine, then it is easy to see that $[1_\mathbb{A}, 1_\mathbb{A}]_\ell = 0_\mathbb{A}$. Finally, if $[1_\mathbb{A}, 1_\mathbb{A}]_\ell = 0_\mathbb{A}$, then $\mathbb{A}$ embeds injectively into the group $\mathbb{G}$ from the definition of $[1_\mathbb{A}, 1_\mathbb{A}]_\ell$, and we can generalize the combinatorial description of $[\alpha, \beta]_\ell$ to get a combinatorial description of the restriction of the 4-ary relation $a + d = b + c$ to $\mathbb{A}$, so this 4-ary relation is preserved by the operations of $\mathbb{A}$. $\square$

We have the following relationship between the various commutators which have been defined so far:

$$[\alpha, \beta] \leq [\alpha, \beta]_s \leq [\alpha, \beta]_2 \leq [\alpha, \beta]_3 \leq \cdots \leq [\alpha, \beta]_\ell \leq \alpha \wedge \beta,$$

and among these, the commutators $[\alpha, \beta]_s, [\alpha, \beta]_2$, and $[\alpha, \beta]_\ell$ are symmetric by construction. In [88], Kearnes and Szendrei prove that in every Taylor variety we always have $[\alpha, \beta]_s = [\alpha, \beta]_\ell$, so almost all of the commutators collapse into a single concept in Taylor varieties (and they *all* collapse in varieties with difference terms). They also give an alternative characterization of the linear commutator by showing that it is equivalent to the commutator obtained by first "freely" extending your variety to make the basic operations multilinear inside some larger abelian group, and then computing commutators in the multilinear setting, which has a Mal'cev operation $x - y + z$.

85

Chapter 2

Compact Representations and algebras with Few Subpowers

2.1 Generalized Majority-Minority operations (motivating Few Subpowers)

The Few Subpowers algorithm was heavily influenced by Dalmau's paper on generalized majority-minority operations [51]. Dalmau's motivation was that in both near-unanimity algebras and Mal'cev algebras, every subalgebra of $\mathbb{A}^n$ 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 2.1.1. An operation φ is a *generalized majority-minority* operation (abbreviated as *gmm* operation) if for each pair a, b we either have

$$\varphi(x, y, ..., y) = \varphi(y, x, ..., y) = \cdots = \varphi(y, y, ..., x) = y \quad \text{for all } x, y \in \{a, b\},$$

or

$$\varphi(x, y, ..., y) = \varphi(y, y, ..., x) = x \quad \text{for all } x, y \in \{a, b\}.$$

In the second case we say that a, b is a *minority pair* for φ.

Definition 2.1.2. If $R \subseteq \mathbb{A}_1 \times \cdots \times \mathbb{A}_n$, then we define the *signature* of R, written $\mathrm{Sig}(R)$, to be the set of triples (i, a, b) with $i \in \{1, ..., n\}$, a, b a minority pair in $\mathbb{A}_i$, such that there are some $t_a, t_b \in R$ with $\pi_{1,...,i-1}(t_a) = \pi_{1,...,i-1}(t_b)$ 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 2.1.3. *If $\mathbb{R} \leq \mathbb{A}_1 \times \cdots \times \mathbb{A}_n$ is preserved by an $l+1$-ary gmm operation φ and $S \subseteq \mathbb{R}$ has $\mathrm{Sig}(S) = \mathrm{Sig}(\mathbb{R})$ and $\pi_I(S) = \pi_I(\mathbb{R})$ for all $I \subseteq \{1, ..., n\}$ with $|I| \leq l$, then $\mathbb{R}$ is generated by S (using only φ).*

Proof. We prove this by induction on the arity n of $\mathbb{R}$. Suppose that $a = (a_1, ..., a_n) \in \mathbb{R}$, by the induction hypothesis there is some b_n with $(a_1, ..., a_{n-1}, b_n)$ 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 $n \notin I$. 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 $b_{i_1}, ...$ such that

$$(b_{i_1}, a_{i_2}, ..., a_n), (a_{i_1}, b_{i_2}, ..., a_n), ..., (a_{i_1}, a_{i_2}, ..., b_n) \in \mathrm{Sg}_\varphi(S).$$

If some $b_i = a_i$ then we are done. If some pair a_i, b_i is minority then - assuming WLOG that a_{i_1}, b_{i_1} is minority - we have

$$\varphi\left(\begin{bmatrix} b_{i_1} & \cdots & b_{i_1} & a_{i_1} \\ a_{i_2} & \cdots & a_{i_2} & a_{i_2} \\ \vdots & \ddots & \vdots & \vdots \\ a_n & \cdots & a_n & b_n \end{bmatrix}\right) = \begin{bmatrix} a_{i_1} \\ a_{i_2} \\ \vdots \\ a_n \end{bmatrix} \in \mathrm{Sg}_\varphi(\pi_I S),$$

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

$$\varphi\left(\begin{bmatrix} b_{i_1} & a_{i_1} & \cdots & a_{i_1} \\ a_{i_2} & b_{i_2} & \cdots & a_{i_2} \\ \vdots & \vdots & \ddots & \vdots \\ a_n & a_n & \cdots & b_n \end{bmatrix}\right) = \begin{bmatrix} a_{i_1} \\ a_{i_2} \\ \vdots \\ a_n \end{bmatrix} \in \mathrm{Sg}_\varphi(\pi_I S),$$

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 $\mathrm{Sig}(S) = \mathrm{Sig}(\mathbb{R})$, 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

$$\varphi(b_n, ..., b_n, \varphi(b_n, d_n, ..., d_n, c_n)) = \varphi(b_n, ..., b_n, \varphi(b_n, ..., b_n, a_n)) = a_n,$$

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. $\quad\square$

Definition 2.1.4. 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 $\mathrm{Sig}(S) = \mathrm{Sig}(\mathbb{R})$, $\pi_I(S) = \pi_I(\mathbb{R})$ for every I with $|I| \leq l$, and $|S| \leq 2|\mathrm{Sig}(\mathbb{R})| + \sum_{|I| \leq l} |\pi_I(\mathbb{R})|$.

In order to manipulate compact representations of relations, we again define subroutines `Nonempty`, `Fix-values`, `Next-beta`, and `Intersect`:

- `Nonempty`$(R, i_1, ..., i_k, \mathbb{S})$ takes R a compact representation of $\mathbb{R} \leq \mathbb{A}_1 \times \cdots \times \mathbb{A}_n$, $\mathbb{S} \leq \mathbb{A}_{i_1} \times \cdots \times \mathbb{A}_{i_k}$, computes the subalgebra generated by $\pi_{i_1, ..., i_k}(R)$ under φ, and if this intersects with $\mathbb{S}$, then it returns an element of $\mathbb{R}$ which maps to an element of the intersection,

- `Fix-values`$(R, a_1, ..., a_m)$ 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 $x \in \mathbb{R} \wedge (x_1 = a_1) \wedge \cdots \wedge (x_m = a_m)$ 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 2.1.3,

- `Next-beta`$(R, i_1, ..., i_k, \mathbb{S})$ takes R a compact representation of $\mathbb{R} \leq \mathbb{A}_1 \times \cdots \times \mathbb{A}_n$, $\mathbb{S} \leq \mathbb{A}_{i_1} \times \cdots \times \mathbb{A}_{i_k}$, and returns a compact representation of $\mathbb{R} \cap \mathbb{S}$ 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`$(R, i_1, ..., i_k, S)$ takes R a compact representation of $\mathbb{R} \leq \mathbb{A}_1 \times \cdots \times \mathbb{A}_n$, S a compact representation of $\mathbb{S} \leq \mathbb{A}_{i_1} \times \cdots \times \mathbb{A}_{i_k}$, and computes a compact representation for $\mathbb{R} \cap \mathbb{S}$ by first making a compact representation of $\mathbb{R} \times \mathbb{S}$ 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`$(R, a_1, ..., a_m)$, φ 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** $I = \{i_1, ...\} \subseteq \{1, ..., n\}$ with $|I| \leq l$ and $(b_{i_1}, ...) \in \pi_I(R_{j-1})$ **do**
5:          Set $R_j \leftarrow R_j \cup$ `Nonempty`$(R_{j-1}, j, i_1, ..., i_{|I|}, \{(a_j, b_{i_1}, ..., b_{i_{|I|}})\})$.
6:      **for all** $(i, a, b) \in \mathrm{Sig}(R_{j-1})$ 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 $t \leftarrow$ `Nonempty`$(R_{j-1}, j, i, \{(a_j, a)\})$.
9:          **if** $t \neq \emptyset$ **then**
10:              Set $R_j \leftarrow R_j \cup \{t, \varphi(t, t, ..., t, \varphi(t, t_a, ..., t_a, t_b))\}$.
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 2.1.5. *Suppose that* $\mathbf{A} = (A, \Gamma)$ *is a relational structure where* Γ *is a set of relations which contains the swap relation* $S_{ab} = \{(a, b), (b, a)\}$ *for every pair* $a, b \in \mathbf{A}$. *Then either* $\mathrm{CSP}(\Gamma)$ *is NP-complete, or* $\mathbf{A}$ *has a ternary generalized majority-minority polymorphism. In the second case,* $\mathrm{CSP}(\Gamma)$ *can be solved in polynomial time by Dalmau's algorithm.*

Proof. Note that Γ is automatically core, since any unary polymorphism of S_{ab} must send a, b to distinct values in $\{a, b\}$. Thus if CSP(Γ) is not NP-complete, then it must have a Taylor polymorphism t.

First we will show that this implies that for all $a, b \in \mathbf{A}$ 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 $t' \in \mathrm{Clo}(t)$ 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 [134] 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 $t'(x, y, y) = t'(y, y, x) = t'(x, y, x) = x$ 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 ith pair given by $\{a_i, b_i\}$. We inductively define functions $f_i \in \mathrm{Clo}(t)$ 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 $\mathbf{A}$. 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 $\{a_i, b_i\}$. $\square$

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 2.1.1. The first example is $\mathbb{A}_1 = (\{a, b, c\}, \varphi_1)$, where φ_1 is a ternary gmm such that $\{a, x\}$ is a pure minority subalgebra of $\mathbb{A}_1$ for all x, $\{b, c\}$ is a majority subalgebra of $\mathbb{A}_1$, and the equivalence relation corresponding to the partition $\{a\}, \{b, c\}$ is a congruence α on $\mathbb{A}_1$ such that the quotient $\mathbb{A}_1/\alpha$ is a pure minority algebra. Explicitly, φ_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 2.1.2. The second example is $\mathbb{A}_2 = (\{a, b, c\}, \varphi_2)$, where φ_2 is a ternary gmm such that $\{a, x\}$ is a majority subalgebra of $\mathbb{A}_2$ for all x, $\{b, c\}$ is a pure minority subalgebra of $\mathbb{A}_2$, the equivalence relation corresponding to the partition $\{a\}, \{b, c\}$ is a congruence α on $\mathbb{A}_2$ such that

the quotient $\mathbb{A}_2/\alpha$ is a majority algebra, and the permutation $(b \ c)$ is an automorphism of $\mathbb{A}_2$. Explicitly, φ_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 $\{(a,b),(b,a)\}$, $\{(a,a),(a,b),(b,b)\}$, $\{(a,a),(b,c),(c,b)\}$ 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 φ, throughout the algorithm `Fix-values` and the proof of Theorem 2.1.3?

The first thing to note is that we often set almost all of the entries of φ 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 φ? It is only called directly in the subroutine `Nonempty`. It is crucial that it is actually used there, because the full function φ was necessary for Case 1 of Theorem 2.1.3. 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 2.1.3, 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

$$s(x_0, x_1, ..., x_l) = e(x_1, x_0, x_1, ..., x_l) \text{ if all but one of the } x_i \text{ are equal,}$$

p is related to e by

$$p(x, y, z) = e(y, x, z, ..., z) \text{ if } x = y \text{ or } y = z,$$

and e satisfies the identities

$$e(y, y, x, x, ..., x) = \varphi(y, d(y, x), ..., d(y, x), x) = x,$$
$$e(y, x, y, x, ..., x) = \varphi(x, y, d(y, x), ..., d(y, x)) = x,$$
$$e(x, x, x, y, ..., x) = \varphi(x, x, d(x, y), ..., x) = x,$$
$$\vdots$$
$$e(x, x, x, x, ..., y) = \varphi(x, x, x, ..., d(x, y)) = x.$$

Can we use this system of identities to prove an analogue of Theorem 2.1.3? 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 $(a_1, ..., a_{i-1}, b_i, a_{i+1}, ..., a_n)$ for each i, as well as a pair $(c_1, ..., c_{n-1}, a_n), (c_1, ..., c_{n-1}, b_n)$ which witnesses the triple (n, a_n, b_n), then we have

$$t\left(\begin{bmatrix} c_1 & c_1 & a_1 & b_1 & a_1 & \cdots \\ c_2 & c_2 & a_2 & a_2 & b_2 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \\ a_n & b_n & b_n & a_n & a_n & \cdots \end{bmatrix}\right) = e\left(\begin{bmatrix} b_1 & a_1 & b_1 & a_1 & \cdots \\ a_2 & a_2 & a_2 & b_2 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \\ d_n & d_n & a_n & a_n & \cdots \end{bmatrix}\right) = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix},$$

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 2.1.6. *If every subpower $\mathbb{R} \leq \mathbb{A}^n$ has a compact representation S consisting of at most $p(n)$ tuples, then the number of different subalgebras of $\mathbb{A}^n$ is at most $|\mathbb{A}^n|^{p(n)} = |\mathbb{A}|^{np(n)}$.*

Corollary 2.1.7. *No analogue of compact representations can exist for subpowers of a nontrivial semilattice.*

Proof. It's enough to consider the case $\mathbb{A} = (\{0,1\}, \max)$, since every semilattice contains a sub-algebra isomorphic to it. The number of subpowers of $\mathbb{A}^n$ 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 $\mathbb{A}^n$, so the number of subalgebras of $\mathbb{A}^n$ is at least

$$2^{\binom{n}{n/2}} \geq 2^{2^n/n},$$

and $2^n/n$ clearly grows faster than any polynomial. $\qquad\square$

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 xs 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 xs 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 $(\mathbb{A}^{\mathbb{A}^2})^n$.

2.2 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 $\mathbb{A}^n$.

Definition 2.2.1. If $\mathbb{A}$ is an algebraic structure and $a_1, ..., a_k \in \mathbb{A}$, we say that $a_1, ..., a_k$ are *independent* if no a_i is in the subalgebra generated by the rest of the a_js. For every n, we define $i_{\mathbb{A}}(n)$ to be the size of the largest independent set in $\mathbb{A}^n$.

Proposition 2.2.2. *If $\mathbb{A}$ is a finite algebra, then any subalgebra of $\mathbb{A}^n$ can be generated by at most $i_{\mathbb{A}}(n)$ elements, so the number of subalgebras of $\mathbb{A}^n$ is bounded above by $|\mathbb{A}^n|^{i_{\mathbb{A}}(n)} = 2^{n\lg(|\mathbb{A}|)i_{\mathbb{A}}(n)}$. The number of subalgebras of $\mathbb{A}^n$ is also bounded below by $2^{i_{\mathbb{A}}(n)}$.*

Proof. Since $\mathbb{A}$ is finite, every subalgebra of $\mathbb{A}^n$ 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 $a_1, ..., a_k$ are independent in $\mathbb{A}^n$. Then every subset S of $\{a_1, ..., a_k\}$ generates a distinct subalgebra of $\mathbb{A}^n$, since $\mathrm{Sg}_{\mathbb{A}^n}(S) \cap \{a_1, ..., a_k\} = S$ by the definition of independence. Thus $\mathbb{A}^n$ has at least 2^k distinct subalgebras. $\qquad\square$

Proposition 2.2.3. *If $\mathbb{B} \in HSP(\mathbb{A})$ is also finite, then $i_{\mathbb{B}}(n) \leq i_{\mathbb{A}}(cn)$ for some constant c depending only on $\mathbb{B}$.*

Proof. If $\mathbb{A}, \mathbb{B}$ are both finite, then there is some finite number c such that $\mathbb{B} \in HS(\mathbb{A}^c)$, that is, there is a subalgebra $\mathbb{C} \leq \mathbb{A}^c$ and a surjective homomorphism $f : \mathbb{C} \to \mathbb{B}$. Then every independent set in $\mathbb{B}^n$ lifts to an independent set in $(\mathbb{A}^c)^n = \mathbb{A}^{cn}$ by choosing any section of f and applying it coordinate-wise. $\qquad\square$

We will apply the above result to the free algebra on two generators $\mathcal{F}_{\mathcal{V}(\mathbb{A})}(x,y) \leq \mathbb{A}^{\mathbb{A}^2}$ 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 2.2.4. For every subset $S \subseteq \{1,...,k\}$, we define the k-dimensional column vector v^S by

$$v_i^S = \begin{cases} y & i \in S, \\ x & i \notin S. \end{cases}$$

A k-*cube term* is a term t with variables indexed by nonempty subsets of $\{1,...,k\}$, such that if we fix an enumeration $S_1, ..., S_{2^k-1}$ of these subsets, we have the identity

$$t(v^{S_1},...,v^{S_{2^k-1}}) \approx v^{\emptyset}.$$

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

$$t\left(\begin{bmatrix} y & y & y & x & y & x & x \\ y & y & x & y & x & y & x \\ y & x & y & y & x & x & y \end{bmatrix}\right) \approx \begin{bmatrix} x \\ x \\ x \end{bmatrix}.$$

Note that a Mal'cev term is the same as a 2-cube term (up to reordering variables).

Theorem 2.2.5 (Few subpowers implies cube term [25])**.** *Let* $\mathbb{F} = \mathcal{F}_{\mathcal{V}(\mathbb{A})}(x,y) \leq \mathbb{A}^{\mathbb{A}^2}$ *be the free algebra on two generators in the variety generated by* $\mathbb{A}$.

- *If* $i_{\mathbb{F}}(k) < 2^k$ *for any* k, *then* $\mathbb{A}$ *has a* k-*cube term.*
- *If* $i_{\mathbb{F}}(m) < \binom{m}{k}$ *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 $i_{\mathbb{F}}(k) < 2^k$, then the vectors v^S for $S \subseteq \{1,...,k\}$ 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 $v^{\emptyset} \in \mathrm{Sg}_{\mathbb{F}^k}\{v^S \mid S \neq \emptyset\}$, we see that there is a term t such that $t(v^{S_1},...) = v^{\emptyset}$, and since $\mathbb{F}$ 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 $S \in \binom{\{1,...,m\}}{k}$. 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 $\mathbb{F}^S \cong \mathbb{F}^k$. $\qquad\square$

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 2.2.6. If $\Delta \subseteq \mathcal{P}(\{1,...,k\}) \setminus \{\emptyset\}$, then we say that t is a Δ-*cube* term if it has variables indexed by elements of Δ and satisfies the identity $t(v^{S_1},...) = v^{\emptyset}$, where $S_1,...$ is an enumeration of the elements of Δ.

If we set $\Delta^e = \{\{1,2\}, \{1\}, \{2\}, ..., \{k\}\}$, then a Δ^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

$$
e\left(\begin{bmatrix}
y & y & x & x & \cdots & x \\
y & x & y & x & \cdots & x \\
x & x & x & y & \cdots & x \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
x & x & x & x & \cdots & y
\end{bmatrix}\right) \approx \begin{bmatrix} x \\ x \\ x \\ \vdots \\ x \end{bmatrix}.
$$

Theorem 2.2.7 (Cube term implies edge term [25])**.** *If $\mathbb{A}$ 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 $\mathbb{A}$ has a Δ^*-cube term t^* then it has a k-edge term, where

$$
\Delta^* = \{\{1,2\}, ..., \{1,k\}, \{1\}, \{2\}, ..., \{k\}\}
$$

only has $2k-1$ elements. The Δ^*-cube term identities for t^* state that

$$
t^*\left(\begin{bmatrix}
y & y & \cdots & y & y & x & x & \cdots & x \\
y & x & \cdots & x & x & y & x & \cdots & x \\
x & y & \cdots & x & x & x & y & \cdots & x \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
x & x & \cdots & y & x & x & x & \cdots & y
\end{bmatrix}\right) \approx \begin{bmatrix} x \\ x \\ x \\ \vdots \\ x \end{bmatrix}.
$$

In order to show that there is a k-edge term, we just need to show that $v^\emptyset$ can be generated from $\{v^S \mid S \in \Delta^e\}$ using the Δ^*-cube term t^*.

Let $a = t^*(x, ..., x, y, x, ..., x)$, 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 indentities 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 ys in the first coordinate are replaced by as. If $S \in \Delta^*$ and $1 \notin S$, then S is already in Δ^e and $v^{S,a} = v^S$, so we don't have to worry about these. If $S = \{1\}$, then we use

$$
t^*\left(\begin{bmatrix}
x & x & \cdots & x & y & x & x & \cdots & x \\
y & x & \cdots & x & x & y & x & \cdots & x \\
x & y & \cdots & x & x & x & y & \cdots & x \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
x & x & \cdots & y & x & x & x & \cdots & y
\end{bmatrix}\right) = \begin{bmatrix} a \\ x \\ x \\ \vdots \\ x \end{bmatrix},
$$

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

$$
t^*\left(\begin{bmatrix}
x & x & \cdots & x & y & x & x & \cdots & x \\
y & y & \cdots & y & y & y & y & \cdots & y \\
x & x & \cdots & x & x & x & x & \cdots & x \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
x & x & \cdots & x & x & x & x & \cdots & x
\end{bmatrix}\right) = \begin{bmatrix} a \\ y \\ x \\ \vdots \\ x \end{bmatrix},
$$

again noting that every column corresponds to an element of Δ^e. Finally, if $S = \{1, i\}$, say $S = \{1, 3\}$ without loss of generality, then we use

$$
t^* \left(\begin{bmatrix} x & x & \cdots & x & y & x & x & \cdots & x \\ y & y & \cdots & y & y & x & x & \cdots & x \\ x & x & \cdots & x & x & y & y & \cdots & y \\ x & x & \cdots & x & x & x & x & \cdots & x \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x & x & \cdots & x & x & x & x & \cdots & x \end{bmatrix} \right) = \begin{bmatrix} a \\ x \\ y \\ x \\ \vdots \\ x \end{bmatrix},
$$

where every row other than the first three (or other than the first, second, and ith in the general case) is all xs, and again every column belongs to Δ^e.

Now that we've constructed the $v^{S,a}$s for all $S \in \Delta^*$, we use t^* to put them all together:

$$
t^* \left(\begin{bmatrix} a & a & \cdots & a & a & x & x & \cdots & x \\ y & x & \cdots & x & x & y & x & \cdots & x \\ x & y & \cdots & x & x & x & y & \cdots & x \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x & x & \cdots & y & x & x & x & \cdots & y \end{bmatrix} \right) = \begin{bmatrix} x \\ x \\ x \\ \vdots \\ x \end{bmatrix}.
$$

Thus if $\mathbb{A}$ has a Δ^*-cube term, then it has a k-edge term. Explicitly, the construction we just worked through corresponds to the formula

$$
e(x_0, x_1, ..., x_k) = t^*(t^*(x_2, ..., x_2, x_0, x_2, ..., x_2), t^*(x_2, ..., x_2, x_0, x_3, ..., x_3), ...,
$$
$$
t^*(x_2, ..., x_2, x_0, x_k, ..., x_k), t^*(x_2, ..., x_k, x_1, x_2, ..., x_k), x_2, ..., x_k).
$$

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 Δ^*-cube term. Let $\Delta^{\ell*} = \Delta^* \cup \mathcal{P}(\{1, ..., \ell\}) \setminus \emptyset$. Note that a k-cube term is the same as a Δ^{k*}-cube term, and a Δ^*-cube term is the same as a Δ^{0*}-cube term.

Claim: If $\mathbb{A}$ has a $\Delta^{\ell*}$-cube term t^ℓ, then it also has a $\Delta^{(\ell-1)*}$-cube term.

Proof of Claim: We argue as before, this time taking $a = t^\ell(x, ..., x, y, x, ..., x)$, 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 ℓth 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^ℓ which looks just like the matrix which gives the defining identities for t^ℓ, but has the ℓth row replaced by the sequence of xs and ys we used to define a. Explicitly, M is given by

$M_{i,T}$	$T \neq \{\ell\}$	$T = \{\ell\}$
$i \neq \ell$	v_i^T	$v_i^T = x$
$i = \ell$	x	y.

Then $t^\ell(M) = v^{\{\ell\},a}$, and the Tth 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 Tth column of M^S is $v^{S\setminus\{\ell\}}$, if $\ell \in T$ but $T \neq \{\ell\}$ then the Tth column is $v^{\{1\}}$, and if $T = \{\ell\}$ then the Tth column is $v^{\{1,\ell\}}$. Explicitly, M^S is given by

$M^S_{i,T}$	$\ell \notin T$	$\ell \in T \neq \{\ell\}$	$T = \{\ell\}$
$i = 1 \in S$	y	y	y
$i = 1 \notin S$	x	y	y
$i \neq 1, \ell, \ i \in S$	y	x	x
$i \neq 1, \ell, \ i \notin S$	x	x	x
$i = \ell$	x	x	y.

These choices ensure that $t^\ell(M^S) = v^{S,a}$.

To finish, we apply t^ℓ to the set of vectors $v^{S,a}$ for $S \in \Delta^{\ell *}$, and see that the defining identities for t^ℓ imply that the resulting vector is $v^\emptyset$. 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 $t^\ell * t^\ell$. $\qquad\square$

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 [25], for the sake of readability and for consistency with the notation used in Appendix A.

Theorem 2.2.8 (Edge terms imply terms s, p, d [25]). *If e is a k-edge term on a finite algebra $\mathbb{A}$, then there are terms $s, p, d \in \mathrm{Clo}(e)$ with s k-ary which satisfy the system of identities*

$$s(y, x, x, ..., x) \approx d(y, x),$$
$$s(x, y, x, ..., x) \approx x,$$

$$\vdots$$

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

Furthermore, these terms can be computed from e in time $O(|\mathbb{A}|^k)$. If $\mathbb{A}$ is infinite, then we can find terms $s, p, d \in \mathrm{Clo}(e)$ 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:

$$s_1(x_1, x_2, ..., x_k) = e(x_2, x_1, x_2, ..., x_k),$$
$$p_1(x, y, z) = e(y, x, z, ..., z),$$
$$d_1(x, y) = e(y, x, y, ..., y).$$

We can get the last identity by an iteration argument. For each i, we set

$$s_{i+1}(x_1, x_2, ..., x_k) = s_1(s_i(x_1, x_2, ..., x_k), x_2, ..., x_k),$$
$$p_{i+1}(x, y, z) = p_1(d_i(x, y), y, z),$$
$$d_{i+1}(x, y) = d_1(d_i(x, y), y).$$

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 $a_2, ..., a_k \in \mathbb{A}$ we find the induced unary polynomial $f_{a_2,...,a_k} : x_1 \mapsto s_1(x_1, a_2, ..., a_k)$. 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 will go over later, but which the reader may enjoy trying to discover now as an exercise. $\qquad\square$

Now we can use the binary term d to define minority pairs and signatures.

Definition 2.2.9. If s, p, d are terms as in the Theorem 2.2.8, 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:

- a, b are a minority pair, i.e. $d(b, a) = b$,

- the pair t_a, t_b agree up to coordinate i: $\pi_{1,...,i-1}(t_a) = \pi_{1,...,i-1}(t_b)$, and

- we have $\pi_i(t_a) = a, \pi_i(t_b) = b$.

We define the *signature* of R, written $\mathrm{Sig}(R)$, to be the set of minority indices which are witnessed by pairs in R.

Definition 2.2.10. If $\mathbb{R} \leq \mathbb{A}_1 \times \cdots \times \mathbb{A}_n$ 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:

- $\mathrm{Sig}(S) = \mathrm{Sig}(\mathbb{R})$,

- for every $I \subseteq \{1, ..., n\}$ with $|I| \leq k - 1$ we have $\pi_I(S) = \pi_I(\mathbb{R})$, and

- $|S| \leq 2|\mathrm{Sig}(\mathbb{R})| + \sum_{I \subseteq \{1,...,n\}, |I| \leq k-1} |\pi_I(\mathbb{R})|$.

Theorem 2.2.11 (Subpowers with edge terms are generated by compact representations [25]). *If $\mathbb{R} \leq \mathbb{A}_1 \times \cdots \times \mathbb{A}_n$ 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 $\mathbb{R} = \mathrm{Sg}_e(S)$.*

Proof. Let s, p, d be terms as in Theorem 2.2.8. We induct on n. Suppose $a = (a_1, ..., a_n) \in \mathbb{R}$, then by the induction hypothesis there is $b_n \in \mathbb{A}_n$ with $(a_1, ..., a_{n-1}, b_n) \in \mathrm{Sg}_e(S)$. Then if we let $d_n = d(b_n, a_n)$ then we see that a_n, d_n is a minority pair and $(a_1, ..., a_n, d_n) \in \mathbb{R}$, so $(n, a_n, d_n) \in \mathrm{Sig}(\mathbb{R})$, and from the definition of a compact representation we see that there must be some $c_1, ..., c_{n-1}$ such that

$$(c_1, ..., c_{n-1}, a_n), (c_1, ..., c_{n-1}, d_n) \in S.$$

We show by an inner induction on subsets $I \subseteq \{1, ..., n\}$ that for each I, we have $\pi_I(a) \in \pi_I(\mathrm{Sg}_e(S))$. If $|I| \leq k - 1$ this follows from the definition of a compact representation, while if $n \notin I$ then this follows from the outer inductive hypothesis. For the sake of notational simplicity

97

we will assume that $I = \{1, ..., n\}$. Then by the inductive hypothesis, there are $b_1, ..., b_{n-1}$ such that for each i, we have

$$(a_1, ..., a_{i-1}, b_i, a_{i+1}, ..., a_n) \in \text{Sg}_e(S).$$

Then we have

$$s\left(\begin{bmatrix} a_1 & b_1 & a_1 & \cdots \\ a_2 & a_2 & b_2 & \cdots \\ \vdots & \vdots & \vdots & \ddots \\ b_n & a_n & a_n & \cdots \end{bmatrix}\right) = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ d_n \end{bmatrix} \in \text{Sg}_e(S).$$

Additionally, we have

$$p\left(\begin{bmatrix} c_1 & c_1 & b_1 \\ c_2 & c_2 & a_2 \\ \vdots & \vdots & \vdots \\ d_n & a_n & a_n \end{bmatrix}\right) = \begin{bmatrix} b_1 \\ a_2 \\ \vdots \\ d_n \end{bmatrix} \in \text{Sg}_e(S).$$

Now we can apply the k-edge term e to see that

$$e\left(\begin{bmatrix} b_1 & a_1 & b_1 & a_1 & \cdots \\ a_2 & a_2 & a_2 & b_2 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \\ d_n & d_n & a_n & a_n & \cdots \end{bmatrix}\right) = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} \in \text{Sg}_e(S). \qquad \square$$

Corollary 2.2.12. *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 $a_1, ..., a_m \in \mathbb{A}^n$ are independent, and consider the relations $\mathbb{R}_i = \text{Sg}_{\mathbb{A}^n}\{a_1, ..., a_i\}$. We can easily find a sequence of compact representations $S_1, ..., S_m$ of $\mathbb{R}_1, ..., \mathbb{R}_m$ with $S_i \subseteq S_{i+1}$ for each i. From the independence of the a_is, we have $\mathbb{R}_i \neq \mathbb{R}_{i+1}$ 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

$$m \leq |S_m| \leq 2n|\mathbb{A}|^2 + \sum_{I \subseteq \{1,...,n\}, |I| \leq k-1} |\mathbb{A}|^{k-1} = O(n^{k-1}). \qquad \square$$

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 φ, and to modify the `Fix-values` subroutine to use the ternary term p from Theorem 2.2.8.

That the modified `Fix-values` subroutine works follows from the following Proposition.

Proposition 2.2.13. *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).*

Algorithm 9 Fix-values$(R, a_1, ..., a_m)$, p, d terms as in Theorem 2.2.8, R a compact representation of $\mathbb{R} \le \mathbb{A}_1 \times \cdots \times \mathbb{A}_n$.

1: Set $R_0 \leftarrow R$.
2: **for** j from 1 to m **do**
3:      Let $R_j \leftarrow \emptyset$.
4:      **for all** $I = \{i_1, ...\} \subseteq \{1, ..., n\}$ with $|I| < k$ and $(b_{i_1}, ...) \in \pi_I(R_{j-1})$ **do**
5:          Set $R_j \leftarrow R_j \cup \texttt{Nonempty}(R_{j-1}, j, i_1, ..., i_{|I|}, \{(a_j, b_{i_1}, ..., b_{i_{|I|}})\})$.
6:      **for all** $(i, a, b) \in \mathrm{Sig}(R_{j-1})$ 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 $t \leftarrow \texttt{Nonempty}(R_{j-1}, j, i, \{(a_j, a)\})$.
9:          **if** $t \ne \emptyset$ **then**
10:             Set $R_j \leftarrow R_j \cup \{t, p(t_b, t_a, t)\}$.
11: **return** R_m.

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

$$\pi_i(p(t_b, t_a, t)) = p(b, a, a) = d(b, a) = b. \qquad \square$$

Example 2.2.1. There is an example of an algebra $\mathbb{A} = (\{a, b, c\}, g)$ with g a ternary operation such that $\mathbb{A}$ has a 3-edge term, but is not in the variety generated by generalized majority-minority algebras of any arity (up to term equivalence). 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 bs (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 2.2.8, 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

$$p\left(\begin{bmatrix} b & c & c \\ c & c & b \end{bmatrix}\right) = \begin{bmatrix} d(b, c) \\ b \end{bmatrix} \overset{?}{=} \begin{bmatrix} b \\ b \end{bmatrix}.$$

But this is impossible: the subalgebra of $\mathbb{A}^2$ 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 $\mathbb{A}$ has a 3-edge term, we define an auxiliary 4-ary term f by

$$f(u, x, y, z) = g(g(u, x, z), g(u, y, z), g(u, z, z)),$$

99

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 2.2.8, then d is given by

$d(x, y)$	a	b	c
a	a	b	a
b	a	b	a
c	c	c	c

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 $\mathrm{Sg}_{\mathbb{A}^2}\{(b, c), (c, c), (c, b)\}$ 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 [142] (which is closely related to $\{a, c\}$ being a ternary absorbing subalgebra), and the affine ternary relation $\{(a, a, c), (a, c, a), (c, a, a), (c, c, c)\}$.

For an idempotent algebra $\mathbb{A}$ with a nontrivial congruence $\theta \in \mathrm{Con}(\mathbb{A})$, such as the previous example, we can test whether $\mathbb{A}$ has few subpowers by checking that $\mathbb{A}/\theta$ has few subpowers and that each congruence class of θ has few subpowers separately. This follows from the following easy results from [104].

Proposition 2.2.14. *Suppose $\mathbb{A}$ is an idempotent algebra, $\theta \in \mathrm{Con}(\mathbb{A})$, and that there are terms t_1, t_2 such that t_1 acts as a Δ_1-cube term on $\mathbb{A}/\theta$ and t_2 acts as a Δ_2-cube term on each congruence class of θ. Then $t_2 * t_1$ is a Δ-cube term for $\mathbb{A}$, where $\Delta = \{S \times T \mid S \in \Delta_2, T \in \Delta_1\}$.*

Corollary 2.2.15. *If $\mathbb{A}_1, ..., \mathbb{A}_n$ are idempotent algebras with the same signature such that each $\mathbb{A}_i$ has a Δ_i-cube term t_i, then $t_1 * \cdots * t_n$ is a Δ-cube term for $\mathbb{A}_1 \times \cdots \times \mathbb{A}_n$, where $\Delta = \{S_1 \times \cdots \times S_n \mid S_i \in \Delta_i\}$.*

Corollary 2.2.16. *Suppose $\mathbb{A}$ is a finite idempotent algebra and $\theta \in \mathrm{Con}(\mathbb{A})$. Then $\mathbb{A}$ has few subpowers iff $\mathbb{A}/\theta$ has few subpowers and each congruence class of θ has few subpowers.*

2.2.1 Some connections with congruence modularity

Theorem 2.2.17. *If an algebra has an edge term, then it generates a congruence modular variety.*

Proof. By Theorem A.4.8 from Appendix A, we just need to check that an algebra with an edge term has directed Gumm terms, that is, terms $f_1, ..., f_k, p$ satisfying the system of identities

$$\begin{aligned}
f_1(x, x, y) &\approx x, \\
f_i(x, y, x) &\approx x \text{ for all } i, \\
f_i(x, y, y) &\approx f_{i+1}(x, x, y) \text{ for all } i, \\
f_k(x, y, y) &\approx p(x, y, y), \\
p(x, x, y) &\approx y.
\end{aligned}$$

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.4.8 to see why the existence of directed Gumm terms implies the existence of Gumm terms, then read part of the proof of Theorem A.4.7 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

$$f_i(x, y, z) = e(x, ..., x, y, z, ..., z),$$

such that there are $i - 1$ zs, a single y, and $k + 1 - i$ xs. Then we have

$$f_1(x, x, y) = e(x, ..., x, x) = x,$$

and for $i < k$ we have

$$f_i(x, y, x) = e(x, ..., x, y, x, ..., x) = x.$$

From the construction of the f_is we have

$$f_i(x, y, y) = e(x, ..., x, y, y, ..., y) = f_{i+1}(x, x, y)$$

for $i + 1 < k$. Finally, if we define $f_k(x, y, z)$ by

$$f_k(x, y, z) = e(y, x, y, z, ..., z)$$

and $p(x, y, z)$ by

$$p(x, y, z) = e(y, x, z, z, ..., z),$$

then

$$f_{k-1}(x, y, y) = e(x, x, x, y, ..., y) = f_k(x, x, y)$$

and

$$f_k(x, y, x) = e(y, x, y, x, ..., x) = x$$

by the k-edge identities, while

$$f_k(x, y, y) = e(y, x, y, y, ..., y) = p(x, y, y)$$

and

$$p(x, x, y) = e(x, x, y, y, ..., y) = y$$

by the k-edge identities again. Thus $f_1, ..., f_k, p$ are a sequence of directed Gumm terms. □

Theorem 2.2.18. *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 $\mathbb{A}$ 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.4.8 shows that we can construct a sequence of Jónsson terms from t, and then Theorem A.4.4 shows that $\mathbb{A}$ generates a congruence distributive variety.

Now the harder direction: assume that $\mathbb{A}$ generates a congruence distributive variety and has a k-edge term e. By Theorem A.4.8, there is a sequence of directed Jónsson terms $f_1, ..., f_m$, that is, a sequence satisfying the system of identities

$$
\begin{aligned}
f_1(x, x, y) &\approx x, \\
f_i(x, y, x) &\approx x \text{ for all } i, \\
f_i(x, y, y) &\approx f_{i+1}(x, x, y) \text{ for all } i, \\
f_m(x, y, y) &\approx y.
\end{aligned}
$$

Let $\mathcal{F} = \mathcal{F}_{\mathcal{V}(\mathbb{A})}(x, y) \leq \mathbb{A}^{\mathbb{A}^2}$ be the free algebra on two generators in the variety generated by $\mathbb{A}$. Let $\mathbb{S} \leq \mathcal{F}^k$ be generated by the vectors $(x, ..., x, y, x, ..., x)$ with all but one entry equal to x and the remaining entry equal to y. Note that $\mathbb{S}$ is symmetric under permuting its coordinates. We just need to prove that $(x, ..., x) \in \mathbb{S}$.

Claim: For all i, we have $(f_i(y, x, x), x, ..., x) \in \mathbb{S}$.

Proof of Claim: We induct on i, taking $(y, x, ..., x) \in \mathbb{S}$ as our base case. By the induction hypothesis, we have

$$(f_i(y, y, x), x, ..., x) = (f_{i-1}(y, x, x), x, ..., x) \in \mathbb{S}.$$

Additionally, the tuples

$$
\begin{bmatrix} f_i(y, x, x) \\ f_i(x, x, y) \\ x \\ \vdots \\ x \end{bmatrix} = f_i \left(\begin{bmatrix} y & x & x \\ x & x & y \\ x & y & x \\ \vdots & \vdots & \vdots \\ x & x & x \end{bmatrix} \right)
$$

and

$$
\begin{bmatrix} f_i(y, y, x) \\ f_i(x, x, y) \\ x \\ \vdots \\ x \end{bmatrix} = f_i \left(\begin{bmatrix} y & y & x \\ x & x & y \\ x & x & x \\ \vdots & \vdots & \vdots \\ x & x & x \end{bmatrix} \right)
$$

are both in $\mathbb{S}$. Now we apply the k-edge term e:

$$
e \left(\begin{bmatrix} f_i(y, y, x) & f_i(y, y, x) & f_i(y, x, x) & f_i(y, x, x) & \cdots & f_i(y, x, x) \\ f_i(x, x, y) & x & f_i(x, x, y) & x & \cdots & x \\ x & x & x & f_i(x, x, y) & \cdots & x \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x & x & x & x & \cdots & f_i(x, x, y) \end{bmatrix} \right) = \begin{bmatrix} f_i(y, x, x) \\ x \\ x \\ \vdots \\ x \end{bmatrix}.
$$

To finish the proof, we apply the Claim with $i = m$ to see that $(x, x, ..., x) = (f_m(y, x, x), x, ..., x) \in \mathbb{S}$. $\qquad\square$

Example 2.2.2. We give an example of a congruence distributive algebra without few subpowers. Recall from Example 1.6.6 that for each n, the relational structure $(\{0, 1\}, \{0\}, \leq, \{0, 1\}^n \setminus \{(0, ..., 0)\})$

has strict width exactly n. The limiting relational clone on $\{0,1\}$ generated by the relations $\{0\}, \leq$, and $\{0,1\}^n \setminus \{(0,...,0)\}$ for all $n \in \mathbb{N}$ corresponds to the clone generated by the ternary operation

$$f(x,y,z) = x \vee (y \wedge z).$$

Since the n-ary critical relation $\{0,1\}^n \setminus \{(0,...,0)\}$ 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 2.3.4.

To check that the algebra $\mathbb{A} = (\{0,1\}, x \vee (y \wedge z))$ generates a congruence distributive variety, consider the sequence of ternary terms given by

$$f_1(x,y,z) = x \vee (y \wedge z), \quad f_2(x,y,z) = (x \wedge y) \vee z.$$

To see that this is a sequence of directed Jónsson terms, note that they satisfy $f_i(x,y,x) = x \vee (y \wedge x) = x$, are connected by

$$f_1(x,y,y) = x \vee (y \wedge y) = x \vee y = (x \wedge x) \vee y = f_2(x,x,y),$$

and have $f_1(x,x,y) = x, f_2(x,y,y) = y$. By Theorem A.4.4 and the discussion before Theorem A.4.8, this implies that $\mathbb{A}$ is congruence distributive.

Example 2.2.3. We've seen earlier that the two-element semilattice $\mathbb{A} = (\{0,1\}, \max)$ 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 $\mathrm{Con}(\mathbb{A}^2)$ already fails to be modular. It turns out that every congruence on $\mathbb{A}^2$ is generated (as a congruence) by just one pair of elements a, b of $\mathbb{A}^2$, so we can label the nontrivial congruences on $\mathbb{A}^2$ by pairs of elements $a, b \in \mathbb{A}^2$, yielding the following congruence lattice.
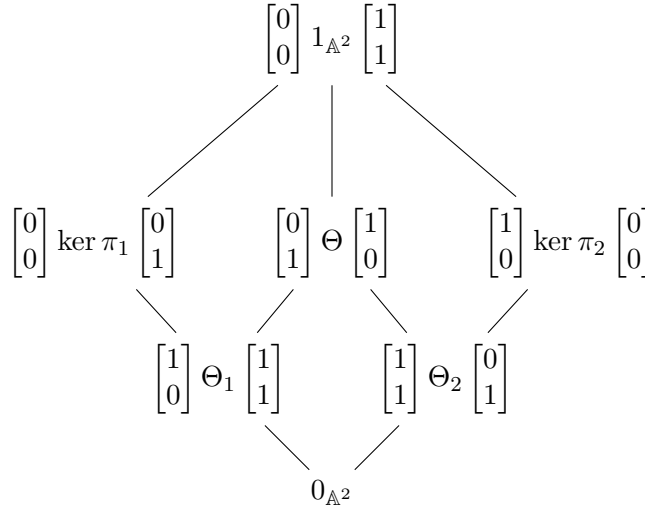
To see that this isn't modular, note that the sublattice generated by $\ker \pi_1, \ker \pi_2, \Theta_2$ is isomorphic to the pentagon lattice $\mathcal{N}_5$. Considered as an abstract lattice, $\mathrm{Con}(\mathbb{A}^2)$ is the standard example of a lattice which is meet-semidistributive (recall from Example 1.9.6 and Proposition 1.9.32 that the variety of semilattices is $\mathrm{SD}(\wedge)$) but not join-semidistributive (we have $\Theta \vee \ker \pi_1 = \Theta \vee \ker \pi_2 = 1_{\mathbb{A}^2}$, but $\Theta \vee (\ker \pi_1 \wedge \ker \pi_2) = \Theta \neq 1_{\mathbb{A}^2}$).

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 $\mathbb{A}$.

Definition 2.2.19. If $\mathbb{A}$ is an algebra, then we define the function $c_{\mathbb{A}}(n)$ to be the base-2 logarithm of the maximum size of $\mathrm{Con}(\mathbb{R})$ over all $\mathbb{R} \leq \mathbb{A}^n$.

Proposition 2.2.20. *A variety $\mathcal{V}$ is congruence distributive iff for all subdirect products $\mathbb{R} \leq_{sd}$ $\mathbb{A}_1 \times \cdots \times \mathbb{A}_n$ in $\mathcal{V}$, every congruence on $\mathbb{R}$ can be written as a product of congruences on the $\mathbb{A}_i s$.*

Proof. Suppose first that $\mathcal{V}$ is congruence distributive. Then for any congruence θ on $\mathbb{R}$, by distributivity and $\bigwedge_i \ker \pi_i = 0_{\mathbb{R}}$ we have

$$\bigwedge_i (\theta \vee \ker \pi_i) = \theta \vee \bigwedge_i \ker \pi_i = \theta \vee 0_{\mathbb{R}} = \theta,$$

so θ is the product of the congruences $\pi_i(\theta \vee \ker \pi_i) \in \mathrm{Con}(\mathbb{A}_i)$.

Conversely, suppose that $\mathbb{A} \in \mathcal{V}$, and suppose that $\alpha, \beta, \gamma \in \mathrm{Con}(\mathbb{A})$. Then $\mathbb{A}/(\beta \wedge \gamma)$ is a subdirect product of $\mathbb{A}/\beta$ and $\mathbb{A}/\gamma$, so the congruence

$$\alpha \vee (\beta \wedge \gamma),$$

considered as a congruence on $\mathbb{A}/(\beta \wedge \gamma)$, is a product congruence iff it is equal to

$$(\alpha \vee \beta) \wedge (\alpha \vee \gamma). \qquad \square$$

Corollary 2.2.21. *If $\mathcal{V}(\mathbb{A})$ is congruence distributive, then $c_{\mathbb{A}}(n) = n c_{\mathbb{A}}(1)$.*

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 $c_{\mathbb{A}}(n)$ for $\mathbb{A}$ a finite affine algebra, and since the congruence lattice only depends on the polynomial clone, we may assume that $\mathbb{A}$ is a module over a ring. In this case, there is a bijection between congruences on $\mathbb{A}^n$ and submodules of $\mathbb{A}^n$.

Proposition 2.2.22. *If $\mathbb{A}$ is a nontrivial finite module over a ring, then $c_{\mathbb{A}}(n) \geq \frac{n^2-1}{4}$.*

Proof. We may as well assume that $\mathbb{A}$ is simple. Let c be any nonzero element of $\mathbb{A}$. 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 $c_{\mathbb{A}}(2m) \geq m^2 \log_2(|\mathbb{A}|) \geq m^2$. $\qquad \square$

Corollary 2.2.23. *If $\mathbb{A}$ is finite and $\mathcal{V}(\mathbb{A})$ is congruence modular but not congruence distributive, then $c_{\mathbb{A}}(n) = \Omega(n^2)$.*

How can we get an upper bound on $c_{\mathbb{A}}(n)$ when $\mathbb{A}$ 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 2.2.24. *If $\mathbb{A}$ is a finite algebra such that $\mathrm{Con}(\mathbb{A})$ has height h, then*

$$|\mathrm{Con}(\mathbb{A})| \leq \sum_{i=0}^{h} \left(\tbinom{|\mathbb{A}|}{2}\right)^i \leq |\mathbb{A}|^{2h}.$$

Proof. Consider any congruence $\alpha \in \text{Con}(\mathbb{A})$. Since every cover of α is generated (as a congruence) by α together with some pair $(a, b) \notin \alpha$, the number of covers of α is bounded by $\binom{|\mathbb{A}|}{2}$. Since every element of $\text{Con}(\mathbb{A})$ can be reached from $0_{\mathbb{A}}$ by repeatedly choosing covers at most h times, we get the stated bound on $|\text{Con}(\mathbb{A})|$.

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 $\binom{\mathbb{A}}{2}$. Additionally, there is only one congruence at height h, since $\text{Con}(\mathbb{A})$ has a top element $1_{\mathbb{A}}$. So we have

$$|\text{Con}(\mathbb{A})| \leq 1 + \sum_{i=0}^{h-1} \binom{\binom{|\mathbb{A}|}{2}}{i}. \qquad \square$$

Corollary 2.2.25. *If $\mathbb{A}$ is finite and generates a congruence modular variety, then $c_{\mathbb{A}}(n) \leq n^2 \cdot 2|\mathbb{A}| \log_2(|\mathbb{A}|)$.*

Proof. Let c be the maximum height of $\text{Con}(\mathbb{B})$ over all subalgebras $\mathbb{B} \leq \mathbb{A}$ (c is automatically bounded by $|\mathbb{A}|$). We claim that for any $\mathbb{R} \leq \mathbb{A}^n$, the height of $\text{Con}(\mathbb{R})$ is bounded by cn. Since $\text{Con}(\mathbb{R})$ is modular, we can compute its height by looking at the size of *any* maximal chain in $\text{Con}(\mathbb{R})$.

We will choose our maximal chain to be any maximal extension of the chain

$$0_{\mathbb{R}} \leq \ker \pi_{[n-1]} \leq \cdots \leq \ker \pi_{[2]} \leq \ker \pi_1 \leq 1_{\mathbb{R}}.$$

By the Diamond Isomorphism Theorem A.2.5, the interval $[\![\ker \pi_{[i]}, \ker \pi_{[i-1]}]\!]$ is isomorphic to the interval $[\![\ker \pi_i, \ker \pi_{[i-1]} \vee \ker \pi_i]\!]$, so its height is bounded by the height of the interval $[\![\ker \pi_i, 1_{\mathbb{R}}]\!]$, which is isomorphic to $\text{Con}(\mathbb{R}/\ker \pi_i)$. Since $\mathbb{R}/\ker \pi_i \cong \pi_i(\mathbb{R}) \leq \mathbb{A}$, the height of $\text{Con}(\mathbb{R}/\ker \pi_i)$ is bounded by c, and putting these intervals together we see that the height of $\text{Con}(\mathbb{R})$ is bounded by cn.

Using the previous bound, we get

$$\log_2(|\text{Con}(\mathbb{R})|) \leq \log_2(|\mathbb{R}|^{2cn}) \leq 2cn \log_2(|\mathbb{A}|^n) = 2cn^2 \log_2(|\mathbb{A}|). \qquad \square$$

Theorem 2.2.26 (Few congruences on subpowers iff congruence modular [25]). *Let $\mathbb{A}$ be a finite algebra with at least two elements, and let $\mathcal{V}(\mathbb{A})$ be the variety it generates.*

- *If $\mathcal{V}(\mathbb{A})$ is congruence distributive, then $c_{\mathbb{A}}(n) = \Theta(n)$.*

- *If $\mathcal{V}(\mathbb{A})$ is congruence modular but not congruence distributive, then $c_{\mathbb{A}}(n) = \Theta(n^2)$.*

- *If $\mathcal{V}(\mathbb{A})$ is not congruence modular, then $c_{\mathbb{A}}(n) = 2^{\Theta(n)}$.*

Proof. By the previous results, all we need to check is that if $\mathcal{V}(\mathbb{A})$ is not congruence modular, then $c_{\mathbb{A}}(n) = 2^{\Omega(n)}$. Let $\mathbb{F} = \mathcal{F}_{\mathcal{V}(\mathbb{A})}(x, y, z, w) \leq \mathbb{A}^{\mathbb{A}^4}$ be the free algebra on four generators. We will show that if $c_{\mathbb{F}}(2n) < 2^n$ for any n, then $\mathbb{A}$ has Day terms, and is therefore congruence modular by Appendix A.1.

Define congruences on $\mathbb{F}$ as in Corollary A.1.2: let θ_{ab} be the congruence generated by the pair (a, b) for any pair of variables a, b, set $\alpha = \theta_{xy} \vee \theta_{zw}, \beta = \theta_{xz} \vee \theta_{yw}$, and $\gamma = (\alpha \wedge \beta) \vee \theta_{zw}$. This is the generic Shifting Lemma configuration:
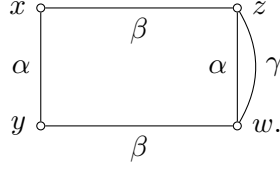
To show the existence of Day terms, we just need to show that $(x, y) \in \gamma$.

Pick an n such that $c_{\mathbb{F}}(2n) < 2^n$, and consider the subalgebra $\mathbb{R} \leq \mathbb{F}^{2n}$ consisting of tuples such that every pair of coordinates are related by β (it helps to imagine elements of $\mathbb{R}$ written out horizontally as row vectors, following the convention that variables which are related by β are laid out on horizontal lines). We will define a family of 2^n pairs of elements of $\mathbb{R}$ as follows.

First, we define elements $x^0, x^1, y^0, y^1 \in \mathbb{F}^2$ by $x^0 = (x, z), x^1 = (z, x)$ and similarly $y^0 = (y, w), y^1 = (w, y)$. Then, for any $i = (i_1, ..., i_n) \in \{0, 1\}^n$, we define $f_i, g_i \in \mathbb{R}$ by

$$f_i = (x^{i_1}, ..., x^{i_n}),$$
$$g_i = (y^{i_1}, ..., y^{i_n}).$$

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 $c_{\mathbb{F}}(2n) < 2^n$, 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 $i = (0, ..., 0)$. 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 $f_0' = (x, ..., x)$ and g_0 becomes the vector $g_0' = (y, ..., y)$, and defining elements f_j', g_j' by dropping half the coordinates of f_j, g_j, we see that

$$(f_0', g_0') \in \bigvee_{j \neq (0, ..., 0)} \Theta'(j),$$

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 $\alpha \wedge \beta$, then every pair of coordinates of b are also related by $\alpha \wedge \beta$. 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 $j \neq (0, ..., 0)$, $\Theta'(j)$ 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 ith coordinates of f_j', g_j' are z, w respectively, with $(z, w) \in \gamma$ by the definition of γ.

106
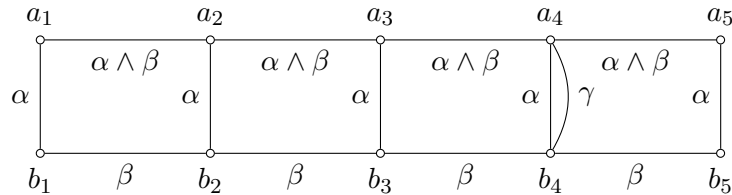
Putting the above properties together, and using $\alpha \wedge \beta \leq \gamma$, we see that $(f_0', b) \in \bigvee_{j \neq (0,\ldots,0)} \Theta'(j)$ implies that every coordinate of f_0' is congruent modulo γ to every coordinate of b, and taking $b = g_0'$ we see that $(x, y) \in \gamma$, which completes the proof. $\qquad \square$

Example 2.2.4. 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 θ_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.

2.3   Parallelogram terms

Examining the proof of Theorem 2.2.11, 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 2.3.1. If $k = m + n$, then an m, n-*parallelogram term* is a $k + 3$-ary term r which satisfies the identities

$$
r\left(\begin{bmatrix} y & y & x & z & \cdots & x & x & \cdots & x \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ y & y & x & x & \cdots & z & x & \cdots & x \\ x & y & y & x & \cdots & x & z & \cdots & x \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x & y & y & x & \cdots & x & x & \cdots & z \end{bmatrix}\right) = \begin{bmatrix} x \\ \vdots \\ x \\ x \\ \vdots \\ x \end{bmatrix},
$$

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 zs on the diagonal and xs elsewhere.

Theorem 2.3.2 (Edge term implies parallelogram terms [89])**.** *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 Δ-cube term for

$$
\Delta = \{\{1, \ldots, m\}, \{1, \ldots, k\}, \{m+1, \ldots, m+n\}, \{1\}, \ldots, \{k\}\},
$$

so by Theorem 2.2.7 if $\mathbb{A}$ 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 $\mathcal{F}(x, y, z)^k$ of all xs 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 xs and the last is a z, the vector of all ys, and the vectors $(x, y, \ldots, y), (y, x, \ldots, x)$.

Letting $d = d(y, x) = e(x, y, x, ..., x)$, we have

$$e\left(\begin{bmatrix} x & y & x & x & \cdots & x \\ z & x & z & x & \cdots & x \\ x & x & x & z & \cdots & x \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x & x & x & x & \cdots & z \end{bmatrix}\right) = \begin{bmatrix} d \\ x \\ x \\ \vdots \\ x \end{bmatrix}$$

and

$$e\left(\begin{bmatrix} x & y & x & x & \cdots & x \\ y & y & z & z & \cdots & z \\ y & y & x & x & \cdots & x \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ y & y & x & x & \cdots & x \end{bmatrix}\right) = \begin{bmatrix} d \\ z \\ x \\ \vdots \\ x \end{bmatrix},$$

so the vectors $(d, x, x, ..., x)$ and $(d, z, x, ..., x)$ 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

$$s(x_1, ..., x_k) = e(x_2, x_1, x_2, ..., x_k) \text{ and } p(x, y, z) = e(y, x, z, ..., z)$$

which act like near-unanimity and Mal'cev terms, respectively. To get the vector of all xs, we apply e one more time:

$$e\left(\begin{bmatrix} d & d & x & x & \cdots & x \\ z & x & z & x & \cdots & x \\ x & x & x & z & \cdots & x \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x & x & x & x & \cdots & z \end{bmatrix}\right) = \begin{bmatrix} x \\ x \\ x \\ \vdots \\ x \end{bmatrix}.$$

Explicitly, our $1, k-1$-parallelogram term r_1 is defined from the edge term e by

$$r_1(x, y, z, u_1, ..., u_k) = e(p(y, z, u_2), s(x, u_2, ..., u_k), u_2, ..., u_k)$$
$$= e(e(z, y, u_2, ..., u_2), e(u_2, x, u_2, ..., u_k), u_2, ..., u_k).$$

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 mth rows of our matrices. Let

$$a = r_{m-1}(y, y, x, x, ..., x, z, x, ..., x),$$

where the z occurs in the $m + 3$rd entry. We want to construct tuples $(x, ..., x, a, x, ..., x)$ and $(x, ..., x, a, y, ..., y)$ from the columns of the defining matrix for an $m, k-m$-parallelogram term. We construct these tuples via

$$r_{m-1}\left(\begin{bmatrix} y & y & x & z & \cdots & x & x & x & \cdots & x \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ y & y & x & x & \cdots & z & x & x & \cdots & x \\ y & y & x & x & \cdots & x & z & x & \cdots & x \\ x & y & y & x & \cdots & x & x & z & \cdots & x \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x & y & y & x & \cdots & x & x & x & \cdots & z \end{bmatrix}\right) = \begin{bmatrix} x \\ \vdots \\ x \\ a \\ x \\ \vdots \\ x \end{bmatrix}$$

and

$$
r_{m-1} \left(\begin{bmatrix} y & y & x & x & \cdots & x & x & x & \cdots & x \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ y & y & x & x & \cdots & x & x & x & \cdots & x \\ y & y & x & x & \cdots & x & z & x & \cdots & x \\ y & y & y & y & \cdots & y & x & y & \cdots & y \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ y & y & y & y & \cdots & y & x & y & \cdots & y \end{bmatrix} \right) = \begin{bmatrix} x \\ \vdots \\ x \\ a \\ y \\ \vdots \\ y \end{bmatrix}.
$$

To get to the vector of all xs, we use

$$
r_{m-1} \left(\begin{bmatrix} x & x & x & x & x & \cdots & z & x & \cdots & x \\ \vdots & \vdots & \vdots & \vdots & \vdots & \iddots & \vdots & \vdots & \ddots & \vdots \\ x & x & x & x & z & \cdots & x & x & \cdots & x \\ a & a & x & z & x & \cdots & x & x & \cdots & x \\ x & y & y & x & x & \cdots & x & z & \cdots & x \\ \vdots & \vdots & \vdots & \vdots & \vdots & \iddots & \vdots & \vdots & \ddots & \vdots \\ x & y & y & x & x & \cdots & x & x & \cdots & z \end{bmatrix} \right) = \begin{bmatrix} x \\ \vdots \\ x \\ x \\ x \\ \vdots \\ x \end{bmatrix},
$$

where the middle row works out because $m > 1$. Explicitly, r_m is defined in terms of r_{m-1} by

$$
r_m(x, y, z, u_1, ..., u_k) = t_{m-1}(t_{m-1}(x, y, z, u_1, ..., u_k), t_{m-1}(y, y, z, z, ..., u_m, ..., z), z, u_m, ..., u_1, u_{m+1}, ..., u_k).
$$
□

To understand what parallelogram terms tell us, it is necessary to restrict to certain special relations, known as *critical* relations.

Definition 2.3.3. A subalgebra $\mathbb{R} \in \mathbb{A}_1 \times \cdots \times \mathbb{A}_n$ 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.5.6 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 2.3.4 (Parallelogram terms constrain critical relations [89])**.** *A variety $\mathcal{V}$ has k-parallelogram terms iff for all critical $\mathbb{R} \leq \mathbb{A}_1 \times \cdots \times \mathbb{A}_n$ with $\mathbb{A}_i \in \mathcal{V}$, either $n < k$ or $\mathbb{R}$ has the parallelogram property.*

Proof. First suppose that $\mathcal{V}$ has k-parallelogram terms, and let $\mathbb{R} \leq \mathbb{A}_1 \times \cdots \times \mathbb{A}_n$ 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 [145], 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 $\mathbb{R}$.

109

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 $(\mathbb{A}_1 \times \cdots \times \mathbb{A}_i) \times (\mathbb{A}_{i+1} \times \cdots \times \mathbb{A}_n)$. 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

$$a \in \mathrm{Sg}(\mathbb{R} \cup \{x\}).$$

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

$$
\begin{bmatrix} a_1 \\ \vdots \\ a_i \\ a_{i+1} \\ \vdots \\ a_n \end{bmatrix} = r\left(\begin{bmatrix} d_1 & d_1 & a_1 & b_1 & \cdots & a_1 & a_1 & \cdots & a_1 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ d_i & d_i & a_i & a_i & \cdots & b_i & a_i & \cdots & a_i \\ a_{i+1} & d_{i+1} & d_{i+1} & a_{i+1} & \cdots & a_{i+1} & b_{i+1} & \cdots & a_{i+1} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ a_n & d_n & d_n & a_n & \cdots & a_n & a_n & \cdots & b_n \end{bmatrix} \right) \in \mathbb{R},
$$

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

$$
\begin{bmatrix} x \\ \vdots \\ x \\ x \\ \vdots \\ x \end{bmatrix} \notin \mathrm{Sg}_{\mathcal{F}^k} \left\{ \begin{bmatrix} y & y & x & z & \cdots & x & x & \cdots & x \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ y & y & x & x & \cdots & z & x & \cdots & x \\ x & y & y & x & \cdots & x & z & \cdots & x \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x & y & y & x & \cdots & x & x & \cdots & z \end{bmatrix} \right\},
$$

so by Zorn's Lemma there exists a maximal k-ary relation $\mathbb{R}$ on $\mathcal{F}$ which contains the right hand side but does not contain the tuple $(x, ..., x)$. The relation $\mathbb{R}$ is then a critical k-ary relation on $\mathcal{F}$, since every relation which properly contains $\mathbb{R}$ must contain $\mathbb{R}^* = \mathrm{Sg}(\mathbb{R} \cup \{(x, ..., x)\})$ and since every existential projection of $\mathbb{R}$ onto a proper subset of the coordinates contains a vector of all xs (by the last k columns of the matrix of generators above). However, $\mathbb{R}$ 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 $\mathcal{V}$. $\quad\square$

Corollary 2.3.5. *A variety $\mathcal{V}$ has k-parallelogram terms iff for every relation $\mathbb{R} \leq \mathbb{A}_1 \times \cdots \times \mathbb{A}_n$ with $\mathbb{A}_i \in \mathcal{V}$, there exists a relation $\mathbb{R}' \leq \mathbb{A}_1 \times \cdots \times \mathbb{A}_n$ such that $\mathbb{R}'$ has the parallelogram property and*

$$\mathbb{R} = \mathbb{R}' \cap \bigcap_{I \subseteq [n], |I| < k} \pi_I(\mathbb{R}).$$

The relation $\mathbb{R}'$ from the corollary need not be so mysterious: we can take it to be the *least* relation $\mathbb{R}'$ which contains $\mathbb{R}$ and has the parallelogram property, since any intersection of relations which have the parallelogram property also has the parallelogram property. This choice of $\mathbb{R}'$ 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 $\mathbb{R}^I$ which contains $\mathbb{R}$ and has the (binary) parallelogram property when considered as a subalgebra of

$$\left(\prod_{i \in I} \mathbb{A}_i\right) \times \left(\prod_{j \notin I} \mathbb{A}_j\right),$$

by finding the linking congruence of $\mathbb{R}$ when considered as a subalgebra of the above, which restricts to a congruence $\alpha_I \in \mathrm{Con}(\pi_I(\mathbb{R}))$, and taking $\mathbb{R}^I$ to be the relation $\alpha_I \circ \mathbb{R}$. We can then take

$$\mathbb{R}' = \bigcup_{I_1, I_2, \ldots \subseteq [n]} \mathbb{R}^{I_1 I_2 \cdots}.$$

In particular, if all of the algebras $\mathbb{A}_i$ are finite, then $\mathbb{R}'$ is contained in the (multisorted) relational clone generated by $\mathbb{R}$.

2.3.1 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 2.3.6. A relation $\mathbb{R} \leq \mathbb{A}_1 \times \cdots \times \mathbb{A}_k$ is said to have the $1, k-1$-*parallelogram property*, or alternatively is called *rectangular*, if for any $i \leq k$, when we regard $\mathbb{R}$ as a binary relation on

$$(\mathbb{A}_1 \times \cdots \times \mathbb{A}_{i-1} \times \mathbb{A}_{i+1} \times \cdots \times \mathbb{A}_k) \times \mathbb{A}_i,$$

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 θ_i on $\mathbb{A}_i$ from the linking congruence of $\mathbb{R}$ (considered as a binary relation on $(\cdots) \times \mathbb{A}_i$), then we have $x \in \mathbb{R}$ iff $x/\prod_i \theta_i \in \mathbb{R}/\prod_i \theta_i$. Thus we may as well study the relation

$$\mathbb{R}/\prod_i \theta_i \leq_{sd} \mathbb{A}_1/\theta_1 \times \cdots \times \mathbb{A}_k/\theta_k$$

instead of studying $\mathbb{R}$. The reduced relation is critical if the original $\mathbb{R}$ is critical, is still rectangular, and has trivial linking congruences on each $\mathbb{A}_i/\theta_i$, so it can be viewed as the graph of a surjective homomorphism

$$\pi_{[k]\setminus\{i\}}\left(\mathbb{R}/\prod_i \theta_i\right) \twoheadrightarrow \mathbb{A}_i/\theta_i$$

for each i.

Definition 2.3.7. A subdirect rectangular relation $\mathbb{R} \leq_{sd} \mathbb{A}_1 \times \cdots \times \mathbb{A}_k$ is called *reduced* if for each $i \leq k$, $\mathbb{R}$ is the graph of a surjective homomorphism

$$\pi_{[k]\setminus\{i\}}(\mathbb{R}) \twoheadrightarrow \mathbb{A}_i,$$

or equivalently, for each i the map

$$\pi_{[k]\setminus\{i\}} : \mathbb{R} \to \pi_{[k]\setminus\{i\}}(\mathbb{R})$$

is an isomorphism, i.e. $\ker \pi_{[k]\setminus\{i\}} = 0_{\mathbb{R}}$.

Proposition 2.3.8. *If $\mathbb{R} \leq_{sd} \mathbb{A}_1 \times \cdots \times \mathbb{A}_k$ is a reduced subdirect critical rectangular relation, then each $\mathbb{A}_i$ is subdirectly irreducible.*

Proof. Let $\mathbb{R}^*$ be the cover of $\mathbb{R}$ is the lattice of subalgebras of $\mathbb{A}_1 \times \cdots \times \mathbb{A}_k$, and let $a = (a_1, ..., a_k)$ 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 $b_i \in \mathbb{A}_i$ such that $(a_1, ..., a_{i-1}, b_i, a_{i+1}, ..., a_k) \in \mathbb{R}$ (and this b_i is unique, since $\mathbb{R}$ is reduced). The claim is that for each i, every nontrivial congruence on $\mathbb{A}_i$ contains the pair (a_i, b_i) - that is, each $\mathbb{A}_i$ is subdirectly irreducible with monolith equal to the congruence generated by the pair (a_i, b_i).

Let $\psi_i \in \mathrm{Con}(\mathbb{A}_i)$ be any nontrivial congruence. Then the relation

$$\exists y_i \; ((x_1, ..., x_{i-1}, y_i, x_{i+1}, ..., x_k) \in \mathbb{R}) \wedge (x_i \equiv_{\psi_i} y_i)$$

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 ψ_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 2.3.9. If $\mathbb{A}_1, ..., \mathbb{A}_k$ are subdirectly irreducible algebras, then we say that an algebra $\mathbb{R}$ is the *graph of a joint similarity* between the $\mathbb{A}_i$s if for each i, $\mathbb{R}$ has a (critical) congruence α_i with $\mathbb{R}/\alpha_i \cong \mathbb{A}_i$, and for each pair i, j there are congruences $\gamma_{ij}, \delta_{ij} \in \mathrm{Con}(\mathbb{R})$ such that

$$[\![\alpha_i, \alpha_i^*]\!] \searrow [\![\gamma_{ij}, \delta_{ij}]\!] \nearrow [\![\alpha_j, \alpha_j^*]\!].$$

More explicitly, this means that $\alpha_i \vee \delta_{ij} = \alpha_i^*$, $\alpha_j \vee \delta_{ij} = \alpha_j^*$, and $\alpha_i \wedge \delta_{ij} = \alpha_j \wedge \delta_{ij}$.

Note that by Proposition A.5.36, $\mathbb{R}/(\alpha_1 \wedge \cdots \wedge \alpha_k)$ 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 $\mathbb{A}_i$s, with $\alpha_i = \ker \pi_i$.

Theorem 2.3.10 (Kearnes, Szendrei [89])**.** *If $\mathbb{R} \leq_{sd} \mathbb{A}_1 \times \cdots \times \mathbb{A}_k$ is a reduced subdirect critical rectangular relation of arity $k \geq 3$ in a congruence modular variety, then*

112

(a) $\mathbb{R}$ *is the graph of a joint similarity between the* $\mathbb{A}_i s$,

(b) *for each* i, j, *the image of* $\pi_{i,j}(\mathbb{R})$ *in* $\mathbb{A}_i/(0_{\mathbb{A}_i} : 0^*_{\mathbb{A}_i}) \times \mathbb{A}_j/(0_{\mathbb{A}_j} : 0^*_{\mathbb{A}_j})$ *is the graph of an isomorphism*

$$\mathbb{A}_i/(0_{\mathbb{A}_i} : 0^*_{\mathbb{A}_i}) \xrightarrow{\sim} \mathbb{A}_j/(0_{\mathbb{A}_j} : 0^*_{\mathbb{A}_j}),$$

(c) *each monolith* $0^*_{\mathbb{A}_i}$ *is abelian, and*

(d) *the cover* $\mathbb{R}^*$ *is also rectangular, and the linking congruence of* $\mathbb{R}^*$ *on* $\mathbb{A}_i$ *is the monolith* $0^*_{\mathbb{A}_i}$.

If $\mathbb{R}$ *has the parallelogram property, then so does its cover* $\mathbb{R}^*$.

Proof. (a) Let $a = (a_1, ..., a_k) \in \mathbb{R}^* \setminus \mathbb{R}$ be a key tuple for $\mathbb{R}$, and for each i let $b_i \in \mathbb{A}_i$ such that $(a_1, ..., a_{i-1}, b_i, a_{i+1}, ..., a_k) \in \mathbb{R}$. Let $a^i = (a_1, ..., a_{i-1}, b_i, a_{i+1}, ..., a_k)$. Then for any $i \neq j$, if we let δ_{ij} be the congruence generated by the pair (a^i, a^j), we claim that

$$[\![\ker \pi_i, (\ker \pi_i)^*]\!] \searrow [\![0_{\mathbb{R}}, \delta_{ij}]\!].$$

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.2.5 - for details, see Proposition A.5.36.

(c) By Proposition A.5.36 again, if $\pi_{i,j}(\mathbb{R})$ is not the graph of an isomorphism for any pair i, j, then each monolith $0^*_{\mathbb{A}_i}$ 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

$$\exists y_i \ ((x_1, ..., x_{i-1}, y_i, x_{i+1}, ..., x_k) \in \mathbb{R}) \wedge (x_i \equiv_{0^*_{\mathbb{A}_i}} y_i)$$

and $\mathbb{R}$ is reduced, we have $(u_i, v_i) \in 0^*_{\mathbb{A}_i}$. Let $p(x, y, z)$ be a Gumm difference term as in Theorem A.3.1, i.e. a term such that $p(y, y, x) \approx x$, and such that for $(x, y) \in \theta$ and θ any congruence we have $p(x, y, y) [\theta, \theta] x$. Then taking $\theta = 0^*_{\mathbb{A}_i}$, we have $p(u_i, v_i, v_i) = u_i$ 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 $\mathbb{A}_I \times \mathbb{A}_{[n]\setminus I}$, where we set $\mathbb{A}_I = \prod_{i \in I} \mathbb{A}_i$, and that we have $(a, b), (c, b), (c, d) \in \mathbb{R}^*$. Pick some $i \in I$ and $j \notin I$. Then there is some a' such that $\pi_{I\setminus\{i\}}(a') = \pi_{I\setminus\{i\}}(a)$, $a'_i \equiv_{0^*_{\mathbb{A}_i}} a_i$, and $(a', b) \in \mathbb{R}$. Similarly find c' which only differs from c in the ith coordinate, has $c'_i \equiv_{0^*_{\mathbb{A}_i}} a_i$, 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 jth coordinate, has $d'_i \equiv_{0^*_{\mathbb{A}_j}} d_i$, 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 2.3.1. Consider the generalized majority-minority algebra $\mathbb{A} = (\{a, b, c\}, \varphi_2)$ from Example 2.1.2, which is subdirectly irreducible with abelian monolith $0_\mathbb{A}^*$ corresponding to the partition $\{a\}, \{b, c\}$ of its elements, and has $\mathbb{A}/0_\mathbb{A}^*$ isomorphic to a two element majority algebra. We can check that the monolith $0_\mathbb{A}^*$ of $\mathbb{A}$ is equal to its own centralizer by verifying that $[1_\mathbb{A} : 0_\mathbb{A}^*] = 0_\mathbb{A}^*$ and $[0_\mathbb{A}^*, 0_\mathbb{A}^*] = 0_\mathbb{A}$: to see this, note that

$$\varphi_2\left(\begin{bmatrix} a & a \\ b & b \end{bmatrix}, \begin{bmatrix} a & a \\ b & b \end{bmatrix}, \begin{bmatrix} b & c \\ b & c \end{bmatrix}\right) = \begin{bmatrix} a & a \\ b & c \end{bmatrix} \in \mathbb{M}(1_\mathbb{A}, 0_\mathbb{A}^*),$$

so $(b, c) \in [1_\mathbb{A}, 0_\mathbb{A}^*]$, while every element of $\mathbb{M}(0_\mathbb{A}^*, 0_\mathbb{A}^*)$ 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 $\mathbb{R} \leq_{sd} \mathbb{A}^3$ corresponding to the columns of the matrix
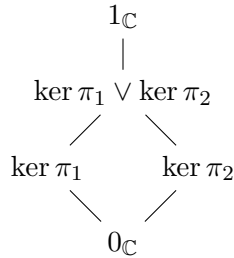
$$\begin{bmatrix} a & b & b & c & c \\ a & b & c & b & c \\ a & b & c & c & b \end{bmatrix}$$

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 $0_\mathbb{A}^* = (0_\mathbb{A} : 0_\mathbb{A}^*)$, and the cover $\mathbb{R}^*$ of $\mathbb{R}$ in $\mathrm{Inv}_3(\mathbb{A})$ is the relation $x \ 0_\mathbb{A}^* \ y \ 0_\mathbb{A}^* \ z$.

More generally, for any k we can define a relation $\mathbb{R}_k \leq_{sd} \mathbb{A}^k$ which contains the tuple $(a, ..., a)$ together with the 2^{k-1} tuples in $\{b, c\}^k$ such that the total number of cs 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 $\mathrm{Inv}_k(\mathbb{A})$: $\mathbb{R}_k$, $\mathbb{R}_k \setminus \{(a, ..., a)\}$, and the two relations we get from these by swapping bs and cs 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 $\mathbb{A}/0_\mathbb{A}^*$ 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 $\mathbb{A}/0_\mathbb{A}^* \to \mathbb{A}/0_\mathbb{A}^*$ which corresponds to swapping the equivalence classes $\{a\}$ and $\{b, c\}$ of $0_\mathbb{A}^*$. 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 $\mathrm{Con}(\mathbb{C})$ is given by the following picture.

$$
\begin{array}{c}
1_\mathbb{C} \\
| \\
\ker \pi_1 \vee \ker \pi_2 \\
\diagup \quad \diagdown \\
\ker \pi_1 \qquad \ker \pi_2 \\
\diagdown \quad \diagup \\
0_\mathbb{C}
\end{array}
$$

As the reader can see, there is no pair $\gamma, \delta \in \mathrm{Con}(\mathbb{C})$ such that $[\![\ker \pi_1, \ker \pi_1 \vee \ker \pi_2]\!] \searrow [\![\gamma, \delta]\!] \nearrow [\![\ker \pi_2, \ker \pi_1 \vee \ker \pi_2]\!]$, 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.5.38, since the corresponding congruence classes of $0^*_{\mathbb{A}}$ 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 $0^*_{\mathbb{A}}$, so $\mathbb{R}$ will consist of the tuple $(a, ..., a)$ together with some subalgebra of $\{b, c\}^k$. Since for any $\mathbb{S} \leq \{b, c\}^k$ the set $\mathbb{S} \cup \{(a, ..., a)\}$ will always be closed under φ_2, if $\mathbb{R}$ is critical then so is $\mathbb{R} \setminus \{(a, ..., a)\}$, and it's easy to check that there are only two critical relations $\mathbb{S} \leq_{sd} \{b, c\}^k$. This completes the classification of critical relations in $\text{Inv}_k(\mathbb{A})$ for $k > 2$.

Remark 2.3.1. Using the structure theorem 2.3.10 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.5.30), one can easily reduce the subpower membership problem 2.4.1 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.3.9). For details of the reduction, see [38].

Example 2.3.2. 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

$$
\begin{aligned}
g(a, a, b) &= a, \\
g(a, a, c) &= c, \\
g(a, a, d) &= c, \\
g(a, b, c) &= c.
\end{aligned}
$$

Then $\mathbb{A}$ has a unique nontrivial congruence $0^*_{\mathbb{A}}$ corresponding to the partition $\{a, c\}, \{b, d\}$, and $\mathbb{A}/0^*_{\mathbb{A}}$ is isomorphic to a two element majority algebra. The congruence classes of $0^*_{\mathbb{A}}$ are affine over $\mathbb{Z}/2$, and the algebra $\mathbb{S} = \text{Sg}_{\mathbb{A}^2}\{(a, b), (b, a)\}$ has a congruence ψ corresponding to the partition

$$
\left\{ \begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} b \\ c \end{bmatrix}, \begin{bmatrix} c \\ d \end{bmatrix}, \begin{bmatrix} d \\ a \end{bmatrix} \right\}, \left\{ \begin{bmatrix} a \\ d \end{bmatrix}, \begin{bmatrix} b \\ a \end{bmatrix}, \begin{bmatrix} c \\ b \end{bmatrix}, \begin{bmatrix} d \\ c \end{bmatrix} \right\},
$$

such that $\mathbb{S}/\psi$ is isomorphic to a two element affine algebra over $\mathbb{Z}/2$ (which is isomorphic to $\{a, c\}$). In fact, we have an isomorphism $\mathbb{S} \cong \mathbb{A} \times \{a, c\}$.

To see that $\mathbb{A}$ has few subpowers, let e be the term

$$
e(u, x, y, z) = g(x, g(u, y, y), g(y, g(x, y, z), g(x, y, z))).
$$

Then e acts like the majority operation $g(x, y, z)$ on $\mathbb{A}/0^*_{\mathbb{A}}$, acts like the minority operation $g(x, u, y)$ on $\{a, c\}$, and has

$$
e\left(\begin{bmatrix} b & b & a & a \\ b & a & b & a \\ a & a & a & b \end{bmatrix} \right) = g\left(\begin{bmatrix} b & a & a \\ a & b & a \\ a & a & a \end{bmatrix} \right) = \begin{bmatrix} a \\ a \\ a \end{bmatrix},
$$

$$
e\left(\begin{bmatrix} d & d & a & a \\ d & a & d & a \\ a & a & a & d \end{bmatrix} \right) = g\left(\begin{bmatrix} d & c & a \\ a & d & c \\ a & a & a \end{bmatrix} \right) = \begin{bmatrix} a \\ a \\ a \end{bmatrix}.
$$

Thus e is a 3-edge term.

Note that applying σ to the second coordinate of $\mathbb{S}$ turns it into $0^*_{\mathbb{A}}$, and under the isomorphism $(1, \sigma) : \mathbb{S} \xrightarrow{\sim} 0^*_{\mathbb{A}}$, one of the congruence classes of ψ becomes the diagonal $\{(x, x) \mid x \in \mathbb{A}\}$. Thus $0^*_{\mathbb{A}}$ is the center of $\mathbb{A}$, and $\mathbb{A}$ is similar to the idempotent reduct of $\mathbb{Z}/2$. Since $1_{\mathbb{A}} = (0_{\mathbb{A}} : 0^*_{\mathbb{A}})$ is not abelian, we see that $\mathbb{A}$ can't generate a residually small variety.

We can check that

$$\begin{bmatrix} a & b \\ a & b \\ a & b \end{bmatrix} \notin \mathrm{Sg}_{\mathbb{A}^{3 \times 2}} \left\{ \begin{bmatrix} a & b \\ a & b \\ b & a \end{bmatrix}, \begin{bmatrix} a & b \\ b & a \\ a & b \end{bmatrix}, \begin{bmatrix} b & a \\ a & b \\ a & b \end{bmatrix} \right\}$$

by taking the rows modulo ψ. Thus none of the subsets $\{a, b\}, \{b, c\}, \{c, d\}, \{d, a\}$ (which are taken to each other by powers of the automorphism σ) are closed under any term which acts nontrivially on $\mathbb{A}/0^*_{\mathbb{A}}$. Using this, one can show that $\mathrm{Clo}(g)$ does not contain any proper Taylor subclones.

What do critical relations on $\mathbb{A}$ look like? Suppose that $\mathbb{R} \leq_{sd} \mathbb{A}^m \times \{a, c\}^n$ is critical and subdirect for some m, n with $m + n \geq 3$. By Theorem 2.3.4, $\mathbb{R}$ has the parallelogram property. All we can conclude from Theorem 2.3.10 is that $\mathbb{R}^*$ has the parallelogram property and has linking congruence $(0^*_{\mathbb{A}})^m \times 1^n_{\{a,c\}}$, so the reduction $\mathbb{R}^*_{red}$ 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 $\mathbb{R}^*_{red}$ has $\pi_{ij}(\mathbb{R}^*_{red})$ 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 $\mathbb{R}^*_{red}$, we see that if we assume without loss of generality that $(a, ..., a) \in \mathbb{R}$ (by applying powers of σ 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 $0^*_{\mathbb{A}} \twoheadrightarrow \{a, c\}$ 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 σ, which is $\mathrm{Sg}_{\mathbb{A}^2}\{(a, b), (d, a)\}$, the critical binary relation

116

$\mathrm{Sg}_{\mathbb{A}^2}\{(a,a),(a,b),(b,b)\}$, which corresponds to a partial order on the majority algebra $\mathbb{A}/0^*_{\mathbb{A}}$, and the ternary relation $\mathrm{Sg}_{\mathbb{A}^3}\{(a,a,a),(a,c,c),(b,b,a)\}$, which is the graph of the homomorphism $0^*_{\mathbb{A}} \twoheadrightarrow \{a,c\}$.

2.4 Learnability of relations encoded by compact representations

We'll start off by reviewing some of the standard definitions of learning theory.

Definition 2.4.1. 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 2.4.2. Let $\mathcal{C}'$ be a concept class which contains $\mathcal{C}$, and call $\mathcal{C}'$ the *hypothesis* class. We define an *equivalence oracle* O_C with *target concept* $C \in \mathcal{C}$ to be the function which takes as input a hypothesis $C' \in \mathcal{C}'$, returns "true" if $C = C'$, and otherwise returns an (arbitrary) element of the symmetric difference $C\Delta C'$.

Definition 2.4.3. An algorithm which makes calls to an oracle O is said to *learn* the concept class $\mathcal{C}$ in the exact model with equivalence queries if, when the oracle O is the equivalence oracle O_C with target concept $C \in \mathcal{C}$, the algorithm makes finitely many calls to the oracle O with encodings of hypotheses $C' \in \mathcal{C}'$ before finally discovering the concept C. The learning algorithm is called *proper* if $\mathcal{C}' = \mathcal{C}$, and *improper* otherwise. If there is an algorithm which learns $\mathcal{C}$ in time polynomial in $\log|U|$, then we say that $\mathcal{C}$ 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 $\mathcal{C}$ consists of the set of subalgebras of $\mathbb{A}^n$, i.e. $\mathcal{C} = \mathrm{Inv}_n(\mathbb{A})$ (recall $\mathrm{Inv}_n(\mathbb{A})$ 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 $\mathcal{C}$ to exist, we need $\log|\mathcal{C}|$ 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 $\mathcal{C} = \mathrm{Inv}_n(\mathbb{A})$.

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 $\mathbb{R}$ 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 2.4.4. *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 $\mathbb{A}$ which can be built out of the terms e, s, p, d of Theorem 2.2.8 in time polynomial in n, such that $b^{[i]} = t^{[i]}(b, u_a, u_c, x^{I_1}, ...) \in \mathbb{R}$ satisfies $\pi_{[i]}(a) = \pi_{[i]}(b^{[i]})$.

Proof. The proof is a modification of the proof of Theorem 2.2.11, 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 $b^I = t^I(b, u_a, u_c, x^{I_1}, ...)$ 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 $\{j\} \cup J = \{l_1, ..., l_{k-1}, i\}$, and define sets $I_1, ..., I_{k-1}$ by deleting $l_1, ..., l_{k-1}$, 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

$$s(b, b^{I_1}, ..., b^{I_{k-1}}),$$

we see that if we restrict to the coordinates in $\{j\} \cup J$, we have

$$
s\left(\begin{bmatrix} a_{l_1} & b_{l_1}^{I_1} & a_{l_1} & \cdots & a_{l_1} \\ a_{l_2} & a_{l_2} & b_{l_2}^{I_2} & \cdots & a_{l_2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{l_{k-1}} & a_{l_{k-1}} & a_{l_{k-1}} & \cdots & b_{l_{k-1}}^{I_{k-1}} \\ b_i & a_i & a_i & \cdots & a_i \end{bmatrix}\right) = \begin{bmatrix} a_{l_1} \\ a_{l_2} \\ \vdots \\ a_{l_{k-1}} \\ c_i \end{bmatrix}.
$$

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

$$p\left(\begin{bmatrix} u_{l_1} & u_{l_1} & b^{I_1}_{l_1} \\ u_{l_2} & u_{l_2} & a_{l_2} \\ \vdots & \vdots & \vdots \\ u_{l_{k-1}} & u_{l_{k-1}} & a_{l_{k-1}} \\ c_i & a_i & a_i \end{bmatrix}\right) = \begin{bmatrix} b^{I_1}_{l_1} \\ a_{l_2} \\ \vdots \\ a_{l_{k-1}} \\ c_i \end{bmatrix}.$$

Thus, if we take t^I to be given by

$$t^I = e(p(u_c, u_a, t^{I_1}), s(b, t^{I_1}, ..., t^{I_{k-1}}), t^{I_1}, ..., t^{I_{k-1}}),$$

then when we restrict to the coordinates in $\{j\} \cup J$, we get

$$e\left(\begin{bmatrix} b^{I_1}_{l_1} & a_{l_1} & b^{I_1}_{l_1} & a_{l_1} & \cdots & a_{l_1} \\ a_{l_2} & a_{l_2} & a_{l_2} & b^{I_2}_{l_2} & \cdots & a_{l_2} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{l_{k-1}} & a_{l_{k-1}} & a_{l_{k-1}} & a_{l_{k-1}} & \cdots & b^{I_{k-1}}_{l_{k-1}} \\ c_i & c_i & a_i & a_i & \cdots & a_i \end{bmatrix}\right) = \begin{bmatrix} a_{l_1} \\ a_{l_2} \\ \vdots \\ a_{l_{k-1}} \\ a_i \end{bmatrix},$$

which completes the induction step. $\qquad\square$

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 $b^{[i]} \in \mathrm{Sg}_{\mathbb{A}^n}(R)$ iteratively with $\pi_{[i]}(b^{[i]}) = \pi_{[i]}(a)$. We start by taking $b^{[k-1]} = x^{[k-1]}$, and repeatedly invoke Lemma 2.4.4 to see that if $(i, a_i, c_i) \in \mathrm{Sig}(R)$, where $c_i = d(b^{[i-1]}_i, a_i)$, then we can construct $b^{[i]} \in \mathrm{Sg}_{\mathbb{A}^n}(R)$ in polynomial time. We formalize this procedure as a subroutine which I will call $\mathtt{Approximate}(R, a)$ (this is almost the same as the combination of the subroutines $\mathtt{Interpolate}$ and $\mathtt{New\text{-}Fix\text{-}values}$ from [76]).

The running time of $\mathtt{Approximate}$ 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^I takes $O(n)$ steps (since b^I has n coordinates). By only maintaining the values of $b^{[i-1]}$ and b^I with $i \in I$ in the ith 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 2.4.5. *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 $\mathtt{Approximate}(R, a)$ returns a and $a \in \mathrm{Sg}_{\mathbb{A}^n}(R)$, or $\mathtt{Approximate}(R, a)$ returns $b \neq a$ such that $b \in \mathrm{Sg}_{\mathbb{A}^n}(R)$, 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 $\mathbb{R} = \mathrm{Sg}_{\mathbb{A}^n}(R)$ 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 $\mathcal{C} = \mathrm{Inv}_n(\mathbb{A})$. While it's easy to test whether R and $\mathbb{R}$ have the same projections onto small subsets of the coordinates (just

Algorithm 10 Approximate(R, a), e, s, p, d terms as in Theorem 2.2.8, $R \subseteq \mathbb{A}^n$ 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:     Let x^I be the marked element of R with $\pi_I(x^I) = \pi_I(a)$.
3: Set $b^{[k-1] \cap [n]} = x^{[k-1] \cap [n]}$.
4: **for** i from k to n **do**
5:     Set $c_i \leftarrow d(b_i^{[i-1]}, a_i)$.
6:     **if** $(i, a_i, c_i) \notin \mathrm{Sig}(R)$ **then**
7:         **return** $b^{[i-1]}$.
8:     **else**
9:         Let (u_a^i, u_c^i) 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:         **for all** $l_1, ..., l_{k-1}$ with $j = l_1 < l_2 < \cdots < l_{k-1} < i$ **do**
12:             Set $I \leftarrow [j] \cup \{l_2, ..., l_{k-1}, i\}$.
13:             **for** m from 1 to $k - 1$ **do**
14:                 Set $I_m \leftarrow I \setminus \{l_m\}$.
15:             Set $b^I \leftarrow e(p(u_c^i, u_a^i, b^{I_1}), s(b^{[i-1]}, b^{I_1}, ..., b^{I_{k-1}}), b^{I_1}, ..., b^{I_{k-1}})$.
16: **return** $b^{[n]}$.

---

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 $\mathbb{R}$.

Let's think for a moment about the problem of checking whether R and $\mathrm{Sg}(R)$ witness the same minority indices. Since there are only $n|\mathbb{A}|^2$ possible minority indices, we may as well focus on one particular minority index (i, a, b). By replacing R with $\pi_{[i]}(R)$ and n with i, we may reduce to the case $i = n$.

Proposition 2.4.6. *Suppose that the minority index (i, a, b) is witnessed by some pair (u_a, u_b) in a relation $\mathbb{R} \leq \mathbb{A}^n$. Then for any tuple $t_a \in \mathbb{R}$ with $\pi_i(t_a) = a$, there is a tuple $t_b \in \mathbb{R}$ 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 $\pi_i(t_b) = p(b, a, a) = b$. $\square$

So we can check whether a minority index (i, a, b) is witnessed by $\mathrm{Sg}(R)$ 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 ith coordinate with a b. Finally, we check whether $\pi_{[i]}(t_b) \in \mathrm{Sg}(\pi_{[i]}(R))$. By the above results, we have $\pi_{[i]}(t_b) \in \mathrm{Sg}(\pi_{[i]}(R))$ if and only if $(i, a, b) \in \mathrm{Sig}(\mathrm{Sg}(R))$. We find ourselves naturally led to consider the *subpower membership problem*.

Problem 2.4.1 (Subpower Membership Problem)**.** Given a finite subset $S \subseteq \mathbb{A}^n$ and an element $x \in \mathbb{A}^n$, determine if x is in the subalgebra of $\mathbb{A}^n$ generated by S.

Theorem 2.4.7 (Bulatov, Mayr, Szendrei [38])**.** *For a fixed finite algebra $\mathbb{A}$ with few subpowers, the following problems are polynomial time reducible to eachother:*

120

- *The subpower membership problem for $\mathbb{A}$: determine if $x \in \mathrm{Sg}_{\mathbb{A}^n}(S)$, given $x \in \mathbb{A}^n$ and $S \subseteq \mathbb{A}^n$.*

- *Find a compact representation for $\mathrm{Sg}_{\mathbb{A}^n}(S)$, given a subset $S \subseteq \mathbb{A}^n$.*

- *The subpower intersection problem for $\mathbb{A}$: given subsets $R, S \subseteq \mathbb{A}^n$, find a set of generators for $\mathrm{Sg}_{\mathbb{A}^n}(R) \cap \mathrm{Sg}_{\mathbb{A}^n}(S)$.*

If $\mathbb{A}$ has few subpowers and has a finite number of basic operations, then the subpower membership problem for $\mathbb{A}$ 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 $a \in \mathbb{A}^n$ such that $\mathtt{Approximate}(R, a)$ returns a. If C is not closed under the basic operations of $\mathbb{A}$, 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. $\qquad\square$

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 $\mathcal{C} = \mathrm{Inv}_n(\mathbb{A})$ to a larger concept class $\mathcal{C}'$, where concepts in $\mathcal{C}'$ are encoded by "compact representations" $R \subseteq \mathbb{A}^n$, where we allow sets R which are not compact representations of the subalgebra $\mathbb{R} = \mathrm{Sg}_{\mathbb{A}^n}(R)$ which they generate. In order to be precise about exactly what concept C is encoded by R, we use the $\mathtt{Approximate}$ subroutine.

Definition 2.4.8. If $R \subseteq \mathbb{A}^n$ is a "compact representation", then the corresponding concept $C \subseteq \mathbb{A}^n$ encoded by R is defined by the following rule. An element $a \in \mathbb{A}^n$ 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 $\mathtt{Approximate}(R, a)$ returns a.

The penalty we will pay for this workaround is that since the new concept class $\mathcal{C}'$ is larger than $\mathcal{C} = \mathrm{Inv}_n(\mathbb{A})$, our learning algorithm will now be making *improper* equivalence queries. If the subpower membership problem for $\mathbb{A}$ 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.

Proposition 2.4.9. *For a fixed algebra $\mathbb{A}$ with few subpowers, the algorithm $\mathtt{Learn}(O)$ takes time polynomial in n to find an encoding R for the target concept $C \in \mathrm{Inv}_n(\mathbb{A})$.*

Proof. At every step of the algorithm, we have $\mathrm{Sg}(R) \subseteq C$: 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 $C' \subseteq \mathrm{Sg}(R) \subseteq C$, the value a returned by $O(R)$ will be contained in $C \Delta C' = C \setminus C' \subseteq C$, so $\mathrm{Sg}(R \cup \{a\}) \subseteq C$.

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. $\qquad\square$

121

Algorithm 11 Learn(O), O an equivalence oracle for an unknown target concept $C \in \text{Inv}_n(\mathbb{A})$.

1: Set $R \leftarrow \emptyset$.
2: **while** $O(R)$ does not return "true" **do**
3:     Set $a \leftarrow O(R)$.
4:     **for all** $I \subseteq [n]$ with $|I| < k$ such that $\pi_I(a)$ has no designated witness in R **do**
5:         Set $R \leftarrow R \cup \{a\}$.
6:         Mark a as the designated witness for $\pi_I(a)$.
7:     **while** Approximate(R, a) does not return a **do**
8:         Set $b \leftarrow$ Approximate(R, a).
9:         Let i be minimal such that $a_i \neq b_i$.
10:        Set $R \leftarrow R \cup \{a, d(b, a)\}$.
11:        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 $\text{Sg}(R)$.

Remark 2.4.1. 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 *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 [136]. In the PAC-learning model, the teacher (oracle) has access to both a target concept $C \in \mathcal{C}$ and a probability distribution μ 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 μ (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 2.4.10. If $C \in \mathcal{C}$ is a target concept and μ is a probability distribution on the universe U, then the *sampling oracle* for the pair C, μ is a randomized oracle which samples a random element $a \in U$ drawn from the distribution μ, 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 μ-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 2.4.11. 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* ϵ and *confidence* $1 - \delta$ if for any target concept $C \in \mathcal{C}$ and any probability distribution μ over the universe, the algorithm eventually returns a hypothesis $C' \in \mathcal{C}'$ such that

$$\mathbb{P}[\mu(C \Delta C') \leq \epsilon] \geq 1 - \delta.$$

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 μ*.

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" a the concept class $\mathcal{C}$, known as the VC-dimension.

Definition 2.4.12. 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 $\mathrm{VC}(\mathcal{C})$, 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 2.4.13 (Sauer-Shelah Lemma)**.** *If $\mathcal{C}$ is a collection of subsets of U with VC-dimension d, then*

$$|\mathcal{C}| \leq \sum_{i=0}^{d} \binom{|U|}{i}.$$

In fact, we have the stronger result that the number of sets $S \subseteq U$ which are shattered by $\mathcal{C}$ is at least $|\mathcal{C}|$.

Proof. We show that $\mathcal{C}$ shatters at least $|\mathcal{C}|$ sets by induction on $|\mathcal{C}|$. For the base case, note that the empty set is shattered by $\mathcal{C}$ as long as $|\mathcal{C}| \geq 1$. For the inductive step, let $x \in U$ be an element which is in some of the sets in $\mathcal{C}$ but not all of them, and let $\mathcal{C}_x$ be the collection of $C \in \mathcal{C}$ with $x \in C$ and $\mathcal{C}'_x = \mathcal{C} \setminus \mathcal{C}_x$. Inductively, $\mathcal{C}_x$ shatters at least $|\mathcal{C}_x|$ sets and $\mathcal{C}'_x$ shatters at least $|\mathcal{C}'_x|$ sets, and any set shattered by $\mathcal{C}_x$ or $\mathcal{C}'_x$ must not contain x.

To finish the induction, we just need to check that for any set S which is shattered by both $\mathcal{C}_x$ and $\mathcal{C}'_x$, the set $S \cup \{x\}$ is shattered by $\mathcal{C}$. $\qquad\square$

If the set S is shattered by $\mathcal{C}$, 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 2.4.14. *If an algorithm learns a concept class $\mathcal{C}$ with error ϵ and confidence $1 - \delta$ after requesting at most m classified examples, then*

$$m \geq (2(1 - \epsilon)(1 - \delta) - 1)\,\mathrm{VC}(\mathcal{C}).$$

Proof. Let S be a set with $|S| = \text{VC}(\mathcal{C})$ which is shattered by $\mathcal{C}$, and for each $X \subseteq S$ consider the sampling oracle O_X which samples from the uniform distribution μ on S, and has target concept some $C_X \in \mathcal{C}$ 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

$$\frac{1}{2} + \frac{m}{2|S|} \geq (1 - \epsilon)(1 - \delta). \qquad \square$$

Conversely, if the VC-dimension of $\mathcal{C}$ is small, then the standard learning algorithm in the PAC model performs well, so long as it can be implemented.

Theorem 2.4.15 (VC-dimension determines sample-complexity [30])**.** *If* $\text{VC}(\mathcal{C}') = d$, *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 $C' \in \mathcal{C}'$ *consistent with the data will learn* $\mathcal{C}$ *with error* ϵ *and confidence* $1 - \delta$.

Sketch. Consider the following process: pick $2m$ samples from the probability distribution μ, 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 $\mathcal{C}$ with error ϵ and confidence $1 - \delta$, then for some choice of target concept C and distribution μ, the process described will incorrectly classify at least $\frac{\epsilon m}{2}$ of the last m samples with probability at least $\frac{\delta}{2}$, by Chebyshev's inequality (at least for $m \geq \frac{8}{\epsilon}$). 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 2.4.13, the number of distinct subsets of X which can be written as $C' \cap X$ for some $C' \in \mathcal{C}'$ 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. $\qquad \square$

Note that if $\mathbb{A}$ is an algebraic structure and we take $U = \mathbb{A}^n$, $\mathcal{C} = \text{Inv}_n(\mathbb{A})$, then a set S is shattered by $\text{Inv}_n(\mathbb{A})$ iff S is an *independent* subset of $\mathbb{A}^n$. Thus the VC-dimension of $\text{Inv}_n(\mathbb{A})$ is exactly the same thing as the number $i_\mathbb{A}(n)$, so if the concept classes $\text{Inv}_n(\mathbb{A})$ are efficiently PAC-learnable for all n, then $\mathbb{A}$ 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 2.4.16 (Angluin [3])**.** *If a concept class $\mathcal{C}$ is efficiently learnable in the (improper) equivalence query model using a hypothesis class $\mathcal{C}'$ which has a polynomially evaluable encoding scheme, then $\mathcal{C}$ is also efficiently learnable in the PAC model.*

Proof. Given a sampling oracle O, we simulate an equivalence oracle as follows. The ith time the equivalence oracle is called by the learner, say to determine whether the target concept C is equivalent to a hypothesis $C' \in \mathcal{C}'$, 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

$$q_i \geq \frac{1}{\epsilon}(\ln(1/\delta) + i\ln(2)),$$

for instance, then we get

$$\sum_i (1 - \epsilon)^{q_i} \leq \sum_i e^{-\epsilon q_i} \leq \sum_i e^{\ln \delta - i \ln(2)} = \sum_i \frac{\delta}{2^i} \leq \delta. \qquad \square$$

Remark 2.4.2. Another learning model is the on-line learning model described by Littlestone [103]. 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 2.4.3. 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 [35], 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. The more recent paper [49] by Chen and Valeriote proves such a hardness result for algebraic structures which are not congruence modular, and for finitely related structures congruence modularity is equivalent to few subpowers (as we will see later).

2.5 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 $\mathrm{Pol}(R_1, ..., R_m)$ which preserve the relations $R_1, ..., R_m$. If $R_1, ..., R_m$ are all preserved by $\mathcal{O}$, then the clone $\mathrm{Pol}(R_1, ..., R_m)$ will certainly contain $\mathcal{O}$, but might end up being too large. In this case, $\mathrm{Pol}(R_1, ..., R_m)$ 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

$$\mathcal{O}_n = \mathcal{F}_\mathbb{A}(x_1, ..., x_n) \le \mathbb{A}^{\mathbb{A}^n}$$

generated by the elements $\pi_i : \mathbb{A}^n \to \mathbb{A}$ given by $\pi_i(a_1, ..., a_n) = a_i$, where $x_i \in \mathcal{F}_\mathbb{A}(x_1, ..., x_n)$ is identified with the element $\pi_i \in \mathbb{A}^{\mathbb{A}^n}$. Similarly, recall that the set of n-ary operations $f \in \mathrm{Pol}_n(R_1, ..., R_m)$, considered as a subalgebra of $\mathbb{A}^{\mathbb{A}^n}$, is given by the primitive positive formula

$$f \in \mathrm{Pol}_n(R_1, ..., R_m) \iff \bigwedge_{i \le m} \bigwedge_{M \in R_i^n} f(M) \in R_i.$$

To check that these two subalgebras of $\mathbb{A}^{\mathbb{A}^n}$ are equal, by Theorem 2.2.11 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 $\mathrm{Inv}(\mathbb{A})$ 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 $\mathbb{A}^{\mathbb{A}^n}$, we first need to choose an ordering on the coordinates of $\mathbb{A}^{\mathbb{A}^n}$, that is, an ordering on the elements of A^n. A natural choice is to first fix any total order $\le$ on the set A, and to extend this to the *lexicographic order* on A^n.

Definition 2.5.1. If $(A, \le)$ is a set with a total order, then we define the *lexicographic order* $\le_{lex}$ on A^n by $a \le_{lex} b$ iff either $a = b$ or there is some $i \le n$ such that $a_j = b_j$ for $j < i$ and $a_i < b_i$. In other words, $a <_{lex} b$ if $a_i < b_i$ at the first coordinate i where a and b differ.

Definition 2.5.2. If $(I, \le)$ 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 ith coordinate to be the set of pairs $(a, b) \in A^2$ given by

$$\mathrm{Forks}(R, i) := \{(a, b) \mid \exists t_a, t_b \in R, \ \pi_{<i}(t_a) = \pi_{<i}(t_b), \pi_i(t_a) = a, \pi_i(t_b) = b\}.$$

So in order to understand a clone $\mathcal{O}$ with a k-edge term, we need to understand the relations of arity less than k, together with the set of forks $\mathrm{Forks}(\mathcal{O}_n, a)$ for all $a \in A^n$ and all n. The issue is that while each set $\mathrm{Forks}(\mathcal{O}_n, a)$ 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 $\mathrm{Forks}(\mathcal{O}_n, a)$ to $\mathrm{Forks}(\mathcal{O}_m, b)$ for some choices of $a \in A^n, b \in A^m$.

Proposition 2.5.3. *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 $\phi^*(x_1, ..., x_n) = (x_{\phi(1)}, ..., x_{\phi(m)})$ satisfies the conditions*

- $\phi^*(a) = b$ *and*

- *for all $c <_{lex} a$ we have $\phi^*(c) <_{lex} b$,*

then for any clone $\mathcal{O}$ on A, we have $\mathrm{Forks}(\mathcal{O}_m, b) \subseteq \mathrm{Forks}(\mathcal{O}_n, a)$.

Proof. Letting $\mathbb{A} = (A, \mathcal{O})$, ϕ induces a natural map of free algebras $\mathcal{O}_m \to \mathcal{O}_n$ given by $x_i \mapsto x_{\phi(i)}$. We will write this natural map as $f \mapsto f_\phi$. Considering $\mathcal{O}_m, \mathcal{O}_n$ as subalgebras of $\mathbb{A}^{A^m}, \mathbb{A}^{A^n}$, respectively, we see that for $c \in A^n$ and $f \in \mathcal{O}_m$, the cth coordinate of the image f_ϕ of f under this map is given by

$$f_\phi(c) = f(\phi^*(c)).$$

In particular, if $t, t' \in \mathcal{O}_m$ with $\pi_{<_{lex}b}(t) = \pi_{<_{lex}b}(t')$, then

$$\pi_{<_{lex}a}(t_\phi) = \pi_{<_{lex}a}(t'_\phi),$$

and

$$\begin{bmatrix} t_\phi(a) \\ t'_\phi(a) \end{bmatrix} = \begin{bmatrix} t(\phi^*(a)) \\ t'(\phi^*(a)) \end{bmatrix} = \begin{bmatrix} t(b) \\ t'(b) \end{bmatrix},$$

so every fork in $\mathrm{Forks}(\mathcal{O}_m, b)$ is also a fork in $\mathrm{Forks}(\mathcal{O}_n, a)$. $\qquad\square$

Proposition 2.5.4. *A map ϕ 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 occurence of s in a is i, then $h(i)$ is the index of the first occurence of s in b.*

If no coordinate a_i of a is minimal or maximal with respect to the order $<$ on $\mathbb{A}$, then the converse is true: such a ϕ exists iff such an h exists.

Proof. Given such an h, we define ϕ 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 $c <_{lex} a$, 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 $\phi^*(c) <_{lex} \phi^*(a) = b$.

Now suppose that no coordinate a_i of a is minimal or maximal with respect to the order $<$ on $\mathbb{A}$. Then the map ϕ in the previous proposition must be surjective: if i is not in the image of ϕ, then we can define $c <_{lex} a$ which only differs from a on the ith coordinate, and $\phi^*(c) = \phi^*(a) \not<_{lex} b$, contradicting the choice of ϕ. Thus we can define $h : [n] \to [m]$ by

$$h(i) = \min\{j \in [m] \mid \phi(j) = i\},$$

so $h^*(b) = a$ and we see that a and b have the same set of symbols.

For any i, if we define $c <_{lex} a$ which matches a up to the ith 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 occurence of s in a, then $h(i)$ must be the index of the first occurence of s in b. $\qquad\square$

Definition 2.5.5. 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 occurences 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 $\text{Forks}(\mathcal{O}, a)$ 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 [73]. We can now simplify the description of the sets $\text{Forks}(\mathcal{O}, a)$ using the ordering $\leq_E$.

Definition 2.5.6. For any pair $(c, d) \in A^2$, we define the set $\lambda(\mathcal{O}, (c, d)) \subseteq A^+$ to be the set of $a \in A^+$ such that $(c, d) \notin \text{Forks}(\mathcal{O}, a)$.

Corollary 2.5.7. *For any clone $\mathcal{O}$ on a set A, the set $\lambda(\mathcal{O}, (c, d))$ is upwards closed in A^+ with respect to $\leq_E$, that is, if $a \in \lambda(\mathcal{O}, (c, d))$ and $a \leq_E b$, then $b \in \lambda(\mathcal{O}, (c, d))$.*

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 $\lambda(\mathcal{O}, (c, d))$ 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 2.5.8. A partial order $(X, \leq)$ is a *well partial order* if for every infinite sequence $x_1, x_2, ...$ 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 2.5.9. *A partial order $(X, \leq)$ is a well partial order iff it has no infinite descending chains and no infinite antichains.*

Proof. Let $x_1, x_2, ...$ 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 2.5.10. *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 $u_1, ..., u_k \in U$ 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 2.5.11. 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 2.5.12. *If B is the disjoint union of A with the two-element set of symbols $\{\#,'\}$, 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 $F : A^+ \to B^+$ to be the function which modifies $x \in A^+$ by inserting a $'$ after the first occurence of each symbol within x, inserting a $\#$ at the end of x, and then following that with a $'$ for each symbol which doesn't occur within x. For instance, if $A = \{a, b, c, d, e\}$, then

$$F(adaadca) = a'd'aadc'a\#'',$$

where the two $'$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 $'$. 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 occurence in $F(x)$, or inserted directly in front of a $'$, 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. $\qquad\square$

Theorem 2.5.13. *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 $i_1 < i_2 < \cdots$ such that every element x_{i_j} ends in the same element of A, say a. Let x'_{i_j} be the element of A^+ we obtain by deleting the a in the last coordinate of x_{i_j}, then from the definition of $\leq_e$ we see that $x_{i_j} \leq_e x_{i_k} \iff x'_{i_j} \leq_e x'_{i_k}$. Let j be minimal such that $x_j >_e x'_{i_k}$ for some k, and note that $j \leq i_1$ so j is well-defined. Then the sequence

$$x_1, x_2, \ldots, x_{j-1}, x'_{i_k}, x'_{i_{k+1}}, \ldots$$

is also an infinite antichain, and is lexicographically smaller than $x_1, x_2, \ldots, x_{j-1}, x_j, \ldots$, a contradiction. $\qquad\square$

Corollary 2.5.14. *If A is a finite set, then the partial order $\leq_E$ on A^+ is a well partial order.*

Theorem 2.5.15 (Few subpowers implies inherently finitely related [1])**.** *If a clone $\mathcal{O}$ contains a k-edge term, then it is finitely related. In fact, a set $\Gamma \subseteq \mathrm{Inv}(\mathcal{O})$ generates $\mathrm{Inv}(\mathcal{O})$ iff the following two conditions are satisfied:*

- *every relation of arity strictly less than k in $\mathrm{Inv}(\mathcal{O})$ is contained in $\langle \Gamma \rangle$, and*

- *for each minority pair $(c,d) \in A^2$ and each minimal element $a \in \lambda(\mathcal{O}, (c,d))$, if we set $n = |a|$, then the relation $\mathrm{Pol}_n(\Gamma)$ on $\mathbb{A}^{\mathbb{A}^n}$ defined by the primitive positive formula*

$$\bigwedge_{R \in \Gamma} \bigwedge_{M \in R^n} f(M) \in R$$

has $(c,d) \notin \mathrm{Forks}(\mathrm{Pol}_n(\Gamma), a)$.

Proof. By the fact that $\leq_E$ is a well partial order, we see that there is a finite set $\Gamma \subseteq \mathrm{Inv}(\mathcal{O})$ which satisfies the conditions given: for instance, we may take Γ to consist of the collection of all relations in $\mathrm{Inv}(\mathcal{O})$ of arity less than k, together with the relations $\mathcal{O}_n \leq \mathbb{A}^{\mathbb{A}^n}$ for every n such that some minimal element $a \in \lambda(\mathcal{O}, (c,d))$ has $|a| = n$ for some $(c,d) \in A^2$.

Now suppose that Γ satisfies the given conditions. Then for any $(c,d) \in A^2$ and any $b \in \lambda(\mathcal{O}, (c,d))$, there exists some minimal $a \in \lambda(\mathcal{O}, (c,d))$ with $a \leq_E b$. Thus if $|a| = n, |b| = m$, then $\mathrm{Forks}(\mathrm{Pol}_m(\Gamma), b) \subseteq \mathrm{Forks}(\mathrm{Pol}_n(\Gamma), a)$, and by the second condition on Γ we have $(c,d) \notin \mathrm{Forks}(\mathrm{Pol}_n(\Gamma), a)$. Thus for any $b \in A^+$ with $|b| = m$, we have

$$(c,d) \notin \mathrm{Forks}(\mathcal{O}_m, b) \implies (c,d) \notin \mathrm{Forks}(\mathrm{Pol}_m(\Gamma), b),$$

so $\mathrm{Forks}(\mathrm{Pol}_m(\Gamma), b) \subseteq \mathrm{Forks}(\mathcal{O}_m, b)$. Since $\mathrm{Pol}_m(\Gamma)$ contains $\mathcal{O}_m$ (by $\Gamma \subseteq \mathrm{Inv}(\mathcal{O})$), and every projection of $\mathrm{Pol}_m(\Gamma)$ onto fewer than k coordinates of $\mathbb{A}^{\mathbb{A}^m}$ is contained in the corresponding projection of $\mathcal{O}_m$ (by the first condition on Γ), we can apply Theorem 2.2.11 to see that $\mathrm{Pol}_m(\Gamma) = \mathcal{O}_m$. $\qquad\square$

Corollary 2.5.16. *The number of clones on a finite set which contain an edge term is countable.*

Remark 2.5.1. There is a converse to Theorem 2.5.15: if $\mathcal{O}$ is a clone on a finite set such that every clone $\mathcal{O}'$ with $\mathcal{O}' \supseteq \mathcal{O}$ is finitely related, then $\mathcal{O}$ has an edge term. The proof of this relies on the theory of *cube term blockers*, which roughly states that a clone $\mathcal{O}$ fails to contain a cube term iff there is an infinite sequence of invariant relations which look like the relations $\{0,1\}^n \setminus \{(0,...,0)\}$ - 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 1.1.3).

Example 2.5.1. Consider the algebra $\mathbb{A} = (\{a,b,c\}, g)$ from Example 2.2.1, which has $\{a,b\}$ a majority subalgebra and $\{a,c\}$ an absorbing minority subalgebra. Recall that the minority pairs of $\mathbb{A}$ were $(a,c), (c,a), (b,c)$. Since $a \in \mathrm{Sg}_{\mathbb{A}}\{b,c\}$, for any $s \in A^+$ we have

$$(b,c) \in \mathrm{Forks}(\langle g \rangle, s) \implies (a,c) \in \mathrm{Forks}(\langle g \rangle, s).$$

Take the standard alphabetical ordering $<$ on $\{a,b,c\}$. 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 $\mathcal{F}_{\mathbb{A}}(x,y)$ 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 aca come before it and that g preserves the affine ternary relation $\{(a,a,a), (a,c,c), (c,a,c), (c,c,a)\}$.

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 ith position and as elsewhere has $(a,c) \in \mathrm{Forks}(\langle g \rangle, s_{in})$. For this, we take the terms x_1 and $g(x_1, x_1, x_i)$ in the free algebra, and check that they make a fork at s_{in}. For $s' <_{lex} s_{in}$, we have $s'_1 = a$ and $s'_i < c$, so

$$g(s'_1, s'_1, s'_i) = g(a,a,a) \text{ or } g(a,a,b) = a = s_1,$$

130

so x_1 and $g(x_1, x_1, x_i)$ agree on tuples which come lexicographically before s_{in}. At s_{in}, we get the fork $(a, g(a, a, c)) = (a, c)$.

Example 2.5.2. Consider the gmm algebra $\mathbb{A}_2 = (\{a, b, c\}, \varphi_2)$ from Example 2.1.2, 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 $<$ on $\{a, b, c\}$, we find that $\lambda(\langle \varphi_2 \rangle, (b, c))$ contains the following 16 elements of A^+:

$$a, b, c, ab, ac, ba, ca, cb, acb, bcc, cab, cba, abcc, bacc, bcac, bcca.$$

Again, it is easy to check the strings of length 2 as $\mathcal{F}_{\mathbb{A}_2}(x, y)$ 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 φ_2 preserves order two the automorphism swapping b and c, and strings containing bcc such as $abcc$ don't work because the two strings where one of the cs is replaced by a b (i.e. $abbc$ and $abcb$ in this case) come before it and φ_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 $(b, c) \in \mathrm{Forks}(\langle \varphi_2 \rangle, s)$. 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 ith position for some $i \geq 2$. We need to show that there is some term $t \in \mathcal{F}_{\mathbb{A}_2}(x_1, ..., x_n)$ 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 $\mathrm{Inv}_k(\mathbb{A}_2)$ carried out in Example 2.3.1. By that analysis, we see that every relation $\mathbb{R} \leq \mathbb{A}_2^k$ is the intersection of some family of binary relations and some family of relations $\mathbb{R}_I \leq \mathbb{A}^I$ such that for each I and each $i, j \in I$, we have $\pi_{i,j}(\mathbb{R}_I) \subseteq 0^*_{\mathbb{A}_2}$, where $0^*_{\mathbb{A}_2}$ 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

$$\begin{bmatrix} s'_1 \\ c \end{bmatrix} \notin \mathrm{Sg}_{\mathbb{A}_2^2} \begin{bmatrix} s' \\ s \end{bmatrix},$$

or there is some family $s^1, ..., s^k <_{lex} s$ such that for each j, l, we have $(s^l_j, s_j) \in 0^*_{\mathbb{A}_2}$ but

$$\begin{bmatrix} s^1_1 \\ \vdots \\ s^k_1 \\ c \end{bmatrix} \notin \mathrm{Sg}_{\mathbb{A}_2^2} \begin{bmatrix} s^1 \\ \vdots \\ s^k \\ s \end{bmatrix}.$$

To rule out the first possibility, we note that if $s' <_{lex} s$ then $s'_1 \in \{a, b\}$, and if $(s'_1, s'_i) \neq (b, c)$, then (s'_1, s'_i) is a majority pair and taking $\varphi_2(x_1, x_1, x_i)$ does the trick, while if $(s'_1, s'_i) = (b, c)$, 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:

$$\varphi_2 \left(\begin{bmatrix} a & b & c \\ b & b & c \end{bmatrix} \right) = \begin{bmatrix} b \\ c \end{bmatrix}.$$

To rule out the second possibility, note that if $s^l <_{lex} s$ and $(s_j^l, s_j) \in 0_{\mathbb{A}_2}^*$ for all j, then the first coordinate where s^l and s can differ is at the coordinate i with $s_i = c$, so we must have $s_i^l = b$, $s_i = c$ and $s_1^l = s_1 = b$. Thus the term x_i rules out the second possibility.
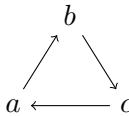
Chapter 3

Absorption and Bounded Width

3.1 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 $\mathbb{A} = (\{a, b, c\}, \cdot)$, where $\cdot$ is the binary, commutative, idempotent operation described by the following table.

The algebra $\mathbb{A}$ is not a semilattice, but every two-element subset of $\mathbb{A}$ is a semilattice. Thus, the binary operation $\cdot$ 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

$$R_{a,b}(x, y, z) := (x \in \{a, b\}) \wedge (x = a \implies y = z).$$

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 3.1.1. *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*

$$((x, y) \in \mathbb{R} \times \{a, b\}) \wedge (y = a \implies x \in \mathbb{S})$$

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:

$$\exists z \in \mathbb{A}^n \quad x \in \mathbb{R} \wedge z \in \mathbb{S} \wedge \bigwedge_{i \leq n} R_{a,b}(y, x_i, z_i). \qquad \square$$

Proposition 3.1.2. *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 $I \subseteq \{1, ..., n\}$ such that the projection π_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 $(a, ..., a) \notin \mathbb{R}$. Since $\mathbb{A}$ is idempotent, the set $\mathbb{R}'$ of triples (x, y, z) such that $(x, y, z, a, ..., a) \in \mathbb{R}$ 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. $\square$

Proposition 3.1.3. *If $\mathbb{R} \leq_{sd} \mathbb{A}^n \times \{a, b\}^k$ has full projection onto $\mathbb{A}^n$, then we have $\mathbb{A}^n \times \{(b, ..., b)\} \subseteq \mathbb{R}$.*

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 $(x_1, ..., x_n, b, ..., b) \in \mathbb{R}$. 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

$$(y, b) = ((x, b) \cdot (y^-, d)) \cdot (y, c) \in \mathbb{R}. \qquad \square$$

The previous two propositions are enough to describe an algorithm which solves $\text{CSP}(\mathbb{A})$. 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 3.1.4. An instance of a CSP is *cycle-consistent* if for every sequence of variables $v_1, ..., v_n$ and relations $R_1, ..., R_n$ 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 3.1.5. *Any cycle-consistent instance of* $\mathrm{CSP}(\mathbb{A})$ *has a solution.*

If we want to understand the complete structure of a general relation $\mathbb{R} \leq \mathbb{A}^n$, things become more complicated. A typical relation we need to consider has the form

$$x_1 \in \{a, b\} \wedge (x_1 = a \implies x_2 \in \{a, b\}) \wedge (x_1 = x_2 = a \implies x_3 \in \{a, b\})$$
$$\wedge \cdots \wedge (x_1 = \cdots = x_k = a \implies y = z).$$

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 3.1.6. *Suppose* $\mathbb{R} \leq \mathbb{A}^n$. *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 $I = \{i_1, ..., i_k\} \subseteq [n]$ be maximal such that, after applying cyclic automorphisms to coordinates in I, we have $x_{i_j} = a$ for all $j \leq k$, and such that the basic relation

$$y_{i_1} \in \{a, b\} \wedge (y_{i_1} = a \implies y_{i_2} \in \{a, b\}) \wedge (y_{i_1} = y_{i_2} = a \implies y_{i_3} \in \{a, b\})$$
$$\wedge \cdots \wedge (y_{i_1} = \cdots = y_{i_{k-1}} = a \implies y_{i_k} \in \{a, b\})$$

contains $\mathbb{R}$. Assume without loss of generality that the coordinates are ordered such that $I = \{n - k + 1, ..., n\}$ and such that the $n - k$-ary relation $\mathbb{R}'$ defined by

$$(y_1, ..., y_{n-k}) \in \mathbb{R}' \iff (y_1, ..., y_{n-k}, a, ..., a) \in \mathbb{R}$$

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 $x = (x_1, ..., x_m, b, ..., b, a, ..., a)$. By Propositions 3.1.2, 3.1.3, and our assumption that x satisfies all basic relations containing $\mathbb{R}$, we have $(x_1, ..., x_m) \in \pi_{[m]}(\mathbb{R}')$ and $\pi_{[m]}(\mathbb{R}') \times \{(b, ..., b)\} \subseteq \mathbb{R}'$, so $(x_1, ..., x_m, b, ..., b) \in \mathbb{R}'$, so $x \in \mathbb{R}$. $\square$

Remark 3.1.1. The intricate yet understandable structure of the basic relations considered above is at the heart of the uncountable region found by Zhuk [141] in the lattice of clones on a three-element domain. Each of the clones in Zhuk's uncountable region properly contains the clone of the rock-paper-scissors algebra, so the generating relations for the corresponding relational clones can be written in terms of the basic relations considered above.

Proposition 3.1.7. *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* $(x_1, ..., x_n) \in \{a, b\}^n \setminus \{(a, ..., a)\}$, *we have* $f(x_1, ..., x_n) = b$.

Proof. Suppose for contradiction that there is some $(x_1, ..., x_n) \in \{a, b\}^n \setminus \{(a, ..., a)\}$ with $f(x_1, ..., x_n) \neq b$. Since $\{a, b\} = \pi_1(R_{a,b})$ is preserved by f, we must then have $f(x_1, ..., x_n) = a$. We will show that for all i with $x_i = b$, f does not depend on its ith 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

$$\begin{bmatrix} a \\ f(y) \\ f(z) \end{bmatrix} = f \left(\begin{bmatrix} x_1 & x_2 & \cdots & x_n \\ y_1 & y_2 & \cdots & y_n \\ z_1 & z_2 & \cdots & z_n \end{bmatrix} \right) \in R_{a,b},$$

so $f(y) = f(z)$. $\qquad \square$

Theorem 3.1.8. *An n-ary operation f is contained in $\mathrm{Clo}_n(\mathbb{A})$ 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 $\mathcal{F} = \mathcal{F}_{\mathcal{V}(\mathbb{A})}(x_1, ..., x_n) \leq \mathbb{A}^{\mathbb{A}^n}$ be the subalgebra generated by $\pi_1, ..., \pi_n : \mathbb{A}^n \to \mathbb{A}$. The projection $\pi_x(\mathcal{F})$ of $\mathcal{F}$ onto the coordinate of $\mathbb{A}^{\mathbb{A}^n}$ corresponding to $x \in \mathbb{A}^n$ is the subalgebra of $\mathbb{A}$ generated by $\{\pi_1(x), ..., \pi_n(x)\} = \{x_1, ..., x_n\}$.

If x is a diagonal tuple, say $x = (a, ..., a)$, then $\pi_x(\mathcal{F}) = \{a\}$, corresponding to the fact that any $f \in \mathcal{F}$ must be idempotent, with $f(a, ..., a) = a$. 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 3.1.3 we have $f \in \mathcal{F} \iff \pi_I(f) \in \pi_I(\mathcal{F})$.

By Proposition 3.1.2, $\pi_I(\mathcal{F})$ 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 $\sigma \in \mathrm{Aut}(\mathbb{A})$ if $(\pi_i(x), \pi_i(y))$ is in the graph of σ 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 $f(y) = \sigma(f(x))$. $\qquad \square$

Note that one of the key steps behind the analysis of the rock-paper-scissors algebra was Proposition 3.1.2 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 3.1.9. An algebra is *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 3.1.10. *A finite idempotent algebra $\mathbb{A}$ is polynomially complete if every binary relation on $\mathbb{A}$ which contains the diagonal is either the equality relation or the full relation, and every ternary relation $\mathbb{R} \leq_{sd} \mathbb{A}^3$ such that every two variable projection of $\mathbb{R}$ is full is equal to the full relation $\mathbb{A}^3$.*

Proof. We will show by induction on n that every n-ary relation $\mathbb{R} \leq \mathbb{A}^n$ which contains the subalgebra of diagonal tuples $(x, ..., x)$, $x \in \mathbb{A}$ is given by a conjunction of equalities between pairs of coordinates. The base case $n = 2$ follows from our assumption on $\mathbb{A}$. By the inductive hypothesis, we may assume without loss of generality that $\pi_{[n] \setminus \{i\}} \mathbb{R} = \mathbb{A}^{n-1}$ for each i.

136

If $n = 3$, then our assumption on $\mathbb{A}$ implies that $\mathbb{R} = \mathbb{A}^3$. Otherwise, suppose for contradiction that $(x_1, ..., x_n) \notin \mathbb{A}^n$, and consider the ternary relation $\mathbb{R}'$ consisting of triples (u, v, w) such that $(u, v, w, x_4, ..., x_n) \in \mathbb{R}$. Since $\mathbb{A}$ is idempotent, $\mathbb{R}'$ is a subalgebra of $\mathbb{A}^3$, and every two-variable projection of $\mathbb{R}'$ is full, so by the $n = 3$ case we must have $(x_1, x_2, x_3) \in \mathbb{R}'$, a contradiction.

Note that we have shown that the relational clone corresponding to the polynomial clone of $\mathbb{A}$ is generated by the equality relation. The general $\text{Inv} - \text{Pol}$ Galois duality now shows that $\mathbb{A}$ is polynomially complete. To see this concretely, consider the subalgebra of $\mathbb{A}^{\mathbb{A}^n}$ generated by the functions π_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 $x \neq y \in \mathbb{A}^n$, then there is always some i such that $\pi_i(x) \neq \pi_i(y)$. Thus this subalgebra of $\mathbb{A}^{\mathbb{A}^n}$ must be the full set of operations $\mathbb{A}^n \to \mathbb{A}$. $\qquad \square$

Corollary 3.1.11. *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 $\mathbb{A}$ 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 3.1.12. The ternary *discriminator* function is the function t defined by

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

Proposition 3.1.13. *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 $\mathbb{A} = (A, t)$ 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 $\mathbb{R} \leq_{sd} \mathbb{A}^2$ is a relation properly containing the diagonal of $\mathbb{A}^2$, and assume without loss of generality that $(a, b) \in \mathbb{R}$ with $a \neq b$. Then for any $c \in \mathbb{A}$, we have

$$\begin{bmatrix} a \\ c \end{bmatrix} = t \left(\begin{bmatrix} a & b & c \\ b & b & c \end{bmatrix} \right) \in \mathbb{R},$$

and similarly $(d, b) \in \mathbb{R}$ for any $d \in \mathbb{A}$. Then for any $c, d \in \mathbb{A}$ we have

$$\begin{bmatrix} d \\ c \end{bmatrix} = t \left(\begin{bmatrix} a & a & d \\ c & b & b \end{bmatrix} \right) \in \mathbb{R},$$

so $\mathbb{R} = \mathbb{A}^2$.

To finish, we just need to show that any ternary relation $\mathbb{R} \leq_{sd} \mathbb{A}^3$ such that every two variable projection is full must be the full relation $\mathbb{A}^3$. Since $\mathbb{A}$ has full automorphism group, if

$\mathbb{R} \neq \mathbb{A}^3$ then we may assume without loss of generality that $(a, a, a) \notin \mathbb{R}$, while all three of $(a, a, b), (a, b, a), (b, a, a)$ are in $\mathbb{R}$. Then we have

$$\begin{bmatrix} b \\ a \\ b \end{bmatrix} = t \left(\begin{bmatrix} a & a & b \\ a & b & a \\ b & a & a \end{bmatrix} \right) \in \mathbb{R},$$

so

$$\begin{bmatrix} a \\ a \\ a \end{bmatrix} = t \left(\begin{bmatrix} a & b & b \\ a & a & a \\ b & b & a \end{bmatrix} \right) \in \mathbb{R},$$

contradicting the assumption $(a, a, a) \notin \mathbb{R}$. $\qquad \square$

Example 3.1.1. 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 $u(x, y, ?) = x$, so $t(x, y, z)$ is given by

$$((xx^-)x^+)^- = (xx^+)^- = (x^+)^- = x.$$

The ternary discriminator t satisfies the system of identities

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

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 1.6.5.

Theorem 3.1.14 (Pixley [120])**.** *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 $\mathcal{F} = \mathcal{F}_{\mathcal{V}(\mathbb{A})}(x, y, z)$ be the free algebra on three generators in this variety, and for $a, b \in \{x, y, z\}$ let θ_{ab} be the smallest congruence with $a \equiv_{\theta_{ab}} b$. Then $(x, z) \in \theta_{xz} \wedge (\theta_{xy} \circ \theta_{yz})$, so by congruence distributivity and permutability, we have

$$(x, z) \in \theta_{xz} \wedge (\theta_{xy} \vee \theta_{yz}) = (\theta_{xz} \wedge \theta_{xy}) \vee (\theta_{xz} \wedge \theta_{yz}) = (\theta_{xz} \wedge \theta_{yz}) \circ (\theta_{xz} \wedge \theta_{xy}).$$

Thus there is some $t \in \mathcal{F}$ such that

$$x \ (\theta_{xz} \wedge \theta_{yz}) \ t \ (\theta_{xz} \wedge \theta_{xy}) \ z.$$

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. $\qquad\square$

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.

3.2 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 3.2.1. If $f : A \to A$ is a unary function, we define $f^{\circ n}$ to be $f \circ \cdots \circ f$, with n copies of f. If (A, f) is either finite or profinite, we define f^{∞} by

$$f^{\infty}(x) := \lim_{n \to \infty} f^{\circ n!}(x).$$

Alternatively, we can define f^{∞} as the limit of $f^{\circ n}$ 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 3.2.2. *If (A, f) is profinite, then the limit defining f^∞ exists, and f^∞ satisfies the identity*

$$f^\infty(f^\infty(x)) \approx f^\infty(x).$$

Furthermore, if A is finite, then

$$f^\infty = f^{\circ \operatorname{lcm}\{1, \ldots, |A|\}},$$

and the graph of f^∞ 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 $\operatorname{lcm}\{1, \ldots, |A|\}$, we will show that $f^{\circ m} = f^{\circ m'}$: this will show that the limit is equal to $f^{\circ m}$, 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^∞ 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 $i < j \leq k$. 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^∞ and $f^{\infty-1}$ runs in time linear in $|A|$. $\qquad\square$

In the context of CSPs, the reduction to the case of core structures was based on the observation than any non-surjective unary polymorphism $f : \mathbf{A} \to \mathbf{A}$ 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 $f^\infty : \mathbf{A} \to \mathbf{A}$ will also be non-surjective, and in fact we have the guarantee that

$$f^\infty(A) \subseteq f^{\circ n}(A)$$

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 $e \in \operatorname{Clo}_1(\mathbb{A})$, we can define a reduct $\mathbb{A}_e$ of $\mathbb{A}$ as follows. For every n-ary operation $f \in \operatorname{Clo}_n(\mathbb{A})$, 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 $\mathbb{A}_e$ to be the algebraic structure $(A, \{f_e \mid f \in \mathrm{Clo}(\mathbb{A})\})$ having a basic operation f_e for each term f of $\mathbb{A}$.

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_e and f agree when they are restricted to $e(A)$. The reduct $\mathbb{A}_e$ 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 $\mathbb{A}_e$ and its subalgebra $e(A)$ are essentially interchangeable, and the subalgebra $e(A)$ of $\mathbb{A}_e$ has as its basic operations the terms of $\mathbb{A}$ 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 3.2.3. *If a system of height 1 identities is satisfied by terms $f^1, ..., f^k$ of $\mathbb{A}$, then the same system of height 1 identities is satisfied by the corresponding operations $f_e^1, ..., f_e^k$ of $\mathbb{A}_e$ (defined as above).*

Note that identities which involve nesting functions may not survive the process of passing from $\mathbb{A}$ to the reduct $\mathbb{A}_e$.

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 3.2.4. 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 π_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 3.2.5. If $t : \mathbb{A}^2 \to \mathbb{A}$ is a binary function and $\mathbb{A}$ is finite (or profinite), then we define t^∞ 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 3.2.6. *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^∞.

The function t^∞ now satisfies one of the two defining identities for a partial semilattice. Note that t^∞ 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^∞ into itself in a surprisingly counterintuitive way.

Proposition 3.2.7. *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

$$u(u(x, y), x) \approx f(u(x, y), f(x, u(x, y))) \approx f(u(x, y), u(x, y)) \approx u(x, y). \qquad \square$$

Finally, to get a term which satisfies *both* defining identities of a partial semilattice, we iterate the function u on its second variable.

Proposition 3.2.8. *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^∞. 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. $\qquad \square$

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 3.2.9. *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 $s \in \mathrm{Clo}(t)$ 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 3.2.10. *If $(b, b) \in \mathrm{Sg}_{\mathbb{A}^2}\{(a, b), (b, a)\}$, then there is a partial semilattice term $s \in \mathrm{Clo}(\mathbb{A})$ 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 $s_n \in \mathrm{Clo}(s)$ which act like the n-ary semilattice operation.

Proposition 3.2.11. *If s is a partial semilattice operation, then for all n there are terms $s_n \in \mathrm{Clo}(s)$ 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

$$s_n(x_1, ..., x_n) = s(s_{n-1}(x_1, ..., x_{n-1}), s(x_1, x_n)). \qquad \square$$

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 3.2.12. *Let $\mathbb{A} = (A, (f^i)_{i \in I})$ be a finite idempotent algebra, and let Σ 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 $(f_s^i)_{i \in I}$ of $\mathbb{A}_s$ will also satisfy the identities in Σ.*

Additionally, if $\mathbb{B}, \mathbb{C}$ are subalgebras of $\mathbb{A}$ such that for any $x \in \mathbb{B}$ and any $y \in \mathbb{C}$ we have $s(x, y), s(y, x) \in \mathbb{C}$, then for any n-ary term f of $\mathbb{A}$ and any $x_1, ..., x_n \in \mathbb{B} \cup \mathbb{C}$ such that at least one $x_i \in \mathbb{C}$, we have $f_s(x_1, ..., x_n) \in \mathbb{C}$.

Proof. Suppose we have an identity

$$f^i(a_1, ..., a_m) \approx f^j(b_1, ..., b_n),$$

with $\{a_1, ..., a_m\} = \{b_1, ..., b_n\} = \{x, y\}$. Define $a_1', ..., a_m'$ by $a_k' = s(x, y)$ if $a_k = x$ and $a_k' = s(y, x)$ if $a_k = y$, and define $b_1', ..., b_n'$ similarly. Then for each k, we have

$$s_m(a_k, ..., a_m, a_1, ..., a_{k-1}) \approx a_k',$$

and similarly for the b_l's, so

$$f_s^i(a_1, ..., a_m) \approx f^i(a_1', ..., a_m') \approx f^j(b_1', ..., b_n') \approx f_s^j(b_1, ..., b_n).$$

For the last statement, we just need to check that for any $x_1, ..., x_n \in \mathbb{B} \cup \mathbb{C}$ with at least one of the x_is in $\mathbb{C}$ we have $s_n(x_1, ..., x_n) \in \mathbb{C}$ (since $\mathbb{C}$ is closed under each term f of $\mathbb{A}$). This follows from the fact that s_n is defined from s in a way that involves all of its variables. $\qquad \square$

Since an algebra $\mathbb{A}$ is Taylor iff it satisfies a nontrivial system of two-variable height 1 identities, if $\mathbb{A}$ is Taylor then $\mathbb{A}_s$ 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. $\mathbb{A}$ has bounded width implies $\mathbb{A}_s$ 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 3.2.13. *If $\mathbb{A}$ has symmetric terms f_n of every arity, then it has symmetric terms f_n^s 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 $\sigma_1, ..., \sigma_{n!}$ be an enumeration of the permutations of $\{1, ..., n\}$, and define f_n^s by

$$f_n^s(x_1, ..., x_n) := f_{n!}(s_n(x_{\sigma_1(1)}, ..., x_{\sigma_1(n)}), ..., s_n(x_{\sigma_{n!}(1)}, ..., x_{\sigma_{n!}(n)})).$$

Then f_n^s is a symmetric term of arity n. $\qquad\square$

Proposition 3.2.14. *If $\mathbb{A}$ has totally symmetric terms f_n of every arity, then it has totally symmetric terms f_n^s 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 $\mathbb{A}$ such that there exist variables $y_1, ..., y_l$ with $\{y_1, ..., y_l\} = \{x_1, ..., x_n\}$ and such that for each i, the number of j with $y_j = x_i$ is at least m, and

$$t(x_1, ..., x_n) = s_l(y_1, ..., y_l).$$

Note that for $m' > m$ we have $S_{m'}^n \subseteq S_m^n$, 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 $a_1, ..., a_n \in \mathbb{A}$, the set of values

$$\{t(a_1, ..., a_n) \mid t \in S^n\}$$

depends only on the set $\{a_1, ..., a_n\}$. Thus we can take

$$f_n^s(x_1, ..., x_n) := f_{|S^n|}(\{t(x_1, ..., x_n) \mid t \in S^n\}). \qquad\square$$

Remark 3.2.1. The previous two propositions only used the fact that the restrictions of the s_ns to $\{a, b\}$ are symmetric and totally symmetric, respectively. So they can be generalized to show that if an algebra $\mathbb{A}$ has symmetric/totally symmetric operations of each arity, then for every subset X of $\mathbb{A}$ such that some collection of terms t_n of $\mathbb{A}$ preserve X and have symmetric/totally symmetric restrictions to X, we can find symmetric/totally symmetric operations of $\mathbb{A}$ 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 $\mathbb{A}$ iff the ternary relation $(x \in \{a, b\}) \wedge (x = a \implies y = z)$ defines a subalgebra of $\mathbb{A}^3$. We can generalize this somewhat.

Proposition 3.2.15. *If B, C are subsets of $\mathbb{A}$, then the ternary relation*

$$(x \in B \cup C) \wedge (x \notin C \implies y = z)$$

defines a subalgebra of $\mathbb{A}^3$ iff $B \cup C$ is a subalgebra of $\mathbb{A}$, and for any n, any n-ary term $f \in \mathrm{Clo}_n(\mathbb{A})$ 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 3.2.16. If $\mathbb{C} \leq \mathbb{B}$ are subalgebras of $\mathbb{A}$ such that there exists a term t with $t(\mathbb{B}, \mathbb{C}), t(\mathbb{C}, \mathbb{B}) \subseteq \mathbb{C}$, then we say that $\mathbb{C}$ *binary absorbs* $\mathbb{B}$, and write $\mathbb{C} \triangleleft_{bin} \mathbb{B}$. If for any n, any n-ary term $f \in \mathrm{Clo}_n(\mathbb{A})$ which depends on all of its inputs, and any $x_1, ..., x_n \in \mathbb{B}$ such that at least one $x_i \in \mathbb{C}$, we have $f(x_1, ..., x_n) \in \mathbb{C}$, then we say that $\mathbb{C}$ *strongly absorbs* $\mathbb{B}$, and write $\mathbb{C} \triangleleft_{str} \mathbb{B}$.

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 3.2.17. *If $\mathbb{C} \triangleleft_{bin} \mathbb{B}$, then there is a partial semilattice term s with $s(\mathbb{B}, \mathbb{C}), s(\mathbb{C}, \mathbb{B}) \subseteq \mathbb{C}$, and in the reduct $\mathbb{A}_s$ the subalgebras $\mathbb{B}_s, \mathbb{C}_s$ satisfy $\mathbb{C}_s \triangleleft_{str} \mathbb{B}_s$. Furthermore, $\mathbb{C} \triangleleft_{str} \mathbb{B}$ iff the ternary relation $(x \in \mathbb{B}) \wedge (x \notin \mathbb{C} \implies y = z)$ defines a subalgebra of $\mathbb{A}^3$ (and $\mathbb{C} \leq \mathbb{B}$).*

In general, a binary absorbing subalgebra of a binary absorbing subalgebra might not be binary absorbing (consider the 4 element lattice $(\{0,1\}^2, \wedge, \vee)$ and the sequence $\{(0,1)\} \triangleleft_{bin} \{(0,0), (0,1)\} \triangleleft_{bin} \{0,1\}^2$), and similarly for strongly absorbing subalgebras (consider the idempotent commutative groupoid $(\{a, b, c\}, \cdot)$ given by $ab = ac = b, bc = c$ and the sequence $\{c\} \triangleleft_{str} \{b, c\} \triangleleft_{str} \{a, b, c\}$). However, we can always chain together binary and strong absorption in one particular order.

Proposition 3.2.18. *If $\mathbb{C} \triangleleft_{bin} \mathbb{B} \triangleleft_{str} \mathbb{A}$, then $\mathbb{C} \triangleleft_{bin} \mathbb{A}$. Applying this repeatedly, we see that if*

$$\mathbb{C} \triangleleft_{bin} \mathbb{B}_n \triangleleft_{str} \cdots \triangleleft_{str} \mathbb{B}_1 \triangleleft_{str} \mathbb{A},$$

then $\mathbb{C} \triangleleft_{bin} \mathbb{A}$.

Proof. Suppose that $\mathbb{C}$ absorbs $\mathbb{B}$ with respect to the binary term t. Define a term u by

$$u(x, y) := t(t(x, t(x, y)), t(y, t(y, x))).$$

Then for any $a \in \mathbb{A}, c \in \mathbb{C}$, we have $t(a, c) \in \mathbb{B}$ and $t(a, t(a, c)) \in \mathbb{B}$ since $c \in \mathbb{B} \triangleleft_{str} \mathbb{A}$, so

$$u(a, c) \in t(\mathbb{B}, t(c, \mathbb{B})) \subseteq t(\mathbb{B}, \mathbb{C}) \subseteq \mathbb{C},$$

and similarly $u(c, a) \in \mathbb{C}$. $\qquad\square$

By iteratively replacing $\mathbb{A}$ with reducts $\mathbb{A}_s$ for partial semilattice terms s quadratically many times, we can reduce to the case where for all a, b, we have $(b, b) \in \mathrm{Sg}_{\mathbb{A}^2}\{(a, b), (b, a)\}$ iff $\{a, b\}$ is a semilattice subalgebra of $\mathbb{A}$ with absorbing element b.

Definition 3.2.19. We say that an idempotent algebra $\mathbb{A}$ has been *prepared* if for every pair a, b such that $(b, b) \in \mathrm{Sg}_{\mathbb{A}^2}\{(a, b), (b, a)\}$, the set $\{a, b\}$ is a semilattice subalgebra of $\mathbb{A}$.

For algebras which have been prepared, it makes sense to define a digraph whose edges correspond to semilattice subalgebras of $\mathbb{A}$.

Definition 3.2.20. 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 3.2.21. *Let s be a fixed nontrivial partial semilattice term of an idempotent algebra $\mathbb{A}$. If $\mathbb{A}$ 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) $\begin{bmatrix} b \\ b \end{bmatrix} \in \mathrm{Sg}_{\mathbb{A}^2} \left\{ \begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} b \\ a \end{bmatrix} \right\}$

(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 $f \in \mathrm{Clo}_n(\mathbb{A})$ 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 $(x \in \{a, b\}) \wedge (x = a \implies y = z)$ 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 3.2.22. *If $\mathbb{A}$ is prepared, then the following hold:*

(a) for $\mathbb{B} \lhd_{bin} \mathbb{A}$ and any $a \in \mathbb{A}$, there is some $b \in \mathbb{B}$ such that $a \to b$,

(b) if $\mathbb{B} \lhd_{bin} \mathbb{A}$ and $a \in \mathbb{A}$, $b \in \mathbb{B}$ have $b \to a$, then $a \in \mathbb{B}$,

(c) if $\mathbb{C} \lhd_{bin} \mathbb{B} \lhd_{bin} \mathbb{A}$, then $\mathbb{C} \lhd_{bin} \mathbb{A}$,

(d) if $\mathbb{B}_1, \mathbb{B}_2 \lhd_{bin} \mathbb{A}$, then $\mathbb{B}_1 \cap \mathbb{B}_2 \neq \emptyset$ and $\mathbb{B}_1 \cap \mathbb{B}_2 \lhd_{bin} \mathbb{A}$.

In particular, there is a unique minimal binary absorbing subalgebra $\mathbb{B} \lhd_{bin} \mathbb{A}$, and this $\mathbb{B}$ has no proper binary absorbing subalgebra.

Proof. Part (a) follows from the existence of a partial semilattice term s with $s(\mathbb{A}, \mathbb{B}) \subseteq \mathbb{B}$ and part (b) follows from part (g) of the previous proposition.

For part (c), choose a partial semilattice term s with $s(\mathbb{B}, \mathbb{C}), s(\mathbb{C}, \mathbb{B}) \subseteq \mathbb{C}$, and choose any binary term t with $t(\mathbb{A}, \mathbb{B}), t(\mathbb{B}, \mathbb{A}) \subseteq \mathbb{B}$. Define a binary term u by

$$u(x, y) := s(s(t(x, y), y), s(t(y, x), x)).$$

Then for $a \in \mathbb{A}, c \in \mathbb{C}$ we have $t(a, c), t(c, a) \in \mathbb{B}$, and we have $t(c, a) \to s(t(c, a), a)$, so by part (b) we have $s(t(c, a), a) \in \mathbb{B}$. Thus

$$u(a, c) \in s(s(\mathbb{B}, c), \mathbb{B}) \subseteq s(\mathbb{C}, \mathbb{B}) \subseteq \mathbb{C},$$

and similarly $u(c, a) \in \mathbb{C}$.

For part (d), pick $b_1 \in \mathbb{B}_1$, then by part (a) there is some $b_2 \in \mathbb{B}_2$ with $b_1 \to b_2$, and then by part (b) we have $b_2 \in \mathbb{B}_1$, so $b_2 \in \mathbb{B}_1 \cap \mathbb{B}_2$. Then from $\mathbb{B}_2 \lhd_{bin} \mathbb{A}$ we have $\mathbb{B}_1 \cap \mathbb{B}_2 \lhd_{bin} \mathbb{B}_1$, and we can apply part (c) to finish. $\square$

Proposition 3.2.23. *If $\mathbb{A}$ has been prepared and $a, b, c \in \mathbb{A}$ have $c \in \mathrm{Sg}\{a, b\}$ with $a \to c$, 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 $\{x, s(x, y)\}$. $\square$

In any digraph, the strongly connected components have a natural partial order.

Definition 3.2.24. We say that b is *reachable* from a if there is a sequence $a = a_0, a_1, ..., a_k = b$ such that $a_i \to a_{i+1}$ for $i = 0, ..., k - 1$.

Proposition 3.2.25. *If $\mathbb{A}$ is prepared and $s^1, ..., s^k$ are partial semilattice terms of $\mathbb{A}$, then for any n-ary term $f \in \langle s^1, ..., s^k \rangle$, $f(x_1, ..., x_n)$ is always reachable from at least one of the variables $x_1, ..., x_n$.*

Definition 3.2.26. 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 $a \to_s a'$, we also have $a' \in S$.

Definition 3.2.27. We say that a set A is *strongly connected* if for every subset $S \subset A$ with $S \neq \emptyset, A$ there is an $a \in S$ and a $b \in A \setminus S$ such that $a \to b$. 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 3.2.2. The digraph considered in this section is the same as the set of "thin red edges" of Andrei Bulatov's colored graph [43] 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 couterintuitive idea of taking a function f which satisfies $f(x, f(x, y)) \approx f(x, y)$ and plugging in $f(x, f(y, x))$.

3.3 Maximal strongly connected components and polynomial completeness

In this section we prove a few results of Andrei Bulatov [40] 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 3.3.1. *Fix a partial semilattice operation s. Suppose $\mathbb{R} \leq_{sd} \mathbb{A} \times \mathbb{B}$ is subdirect and A, B are maximal strongly connected subsets of $\mathbb{A}, \mathbb{B}$, respectively.*

(a) *The set of a such that $(\{a\} \times B) \cap \mathbb{R} \neq \emptyset$ is upwards closed. In particular, if $(A \times B) \cap \mathbb{R}$ is nonempty, then it is subdirect in $A \times B$.*

(b) *The set of a such that $\{a\} \times B \subseteq \mathbb{R}$ is upwards closed.*

(c) *If A is contained in a linked component of $\mathbb{R}$ (that is, a connected component of $\mathbb{R}$ considered as a bipartite graph on $\mathbb{A} \sqcup \mathbb{B}$), $(A \times B) \cap \mathbb{R} \neq \emptyset$, and A, B are finite, then $A \times B \subseteq \mathbb{R}$.*

Additionally, the product $A \times B$ is a maximal strongly connected subset of $\mathbb{A} \times \mathbb{B}$.

Proof. For part (a), suppose that $(a, b) \in \mathbb{R}$ and $b \in B$, and let $a \to a'$. Since $\mathbb{R}$ is subdirect, there is some b' with $(a', b') \in \mathbb{R}$. Then

$$\begin{bmatrix} a' \\ s(b, b') \end{bmatrix} = s\left(\begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} a' \\ b' \end{bmatrix} \right) \in \mathbb{R},$$

and $b \to s(b, b')$, so $s(b, b') \in B$.

For part (b), suppose that $\{a\} \times B \subseteq \mathbb{R}$ and $a \to a'$. Let S be the set of $b \in B$ such that $(a', b) \in \mathbb{R}$, that is, $S = \pi_2((\{a'\} \times B) \cap \mathbb{R})$. 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 $(a, b') \in \mathbb{R}$, so

$$\begin{bmatrix} a' \\ b' \end{bmatrix} = s\left(\begin{bmatrix} a' \\ b \end{bmatrix}, \begin{bmatrix} a \\ b' \end{bmatrix} \right) \in \mathbb{R}.$$

For part (c), suppose first that $A \times A \subseteq \mathbb{R} \circ \mathbb{R}^-$, where $\mathbb{R}^- \leq \mathbb{B} \times \mathbb{A}$ is the reverse of $\mathbb{R}$ (we will later reduce the general case to this case). Let a be any element of A, and let X be the set of $b \in \mathbb{B}$ such that $(a, b) \in \mathbb{R}$, that is, $X = \pi_2((\{a\} \times \mathbb{B}) \cap \mathbb{R})$. 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 $\mathbb{B}$, S is a maximal strongly connected component of X.

By the assumption $A \times A \subseteq \mathbb{R} \circ \mathbb{R}^-$ and the definition of X, we see that $(A \times X) \cap \mathbb{R}$ is subdirect in $A \times X$. Thus by part (b) and the fact that $\{a\} \times S \subseteq (A \times X) \cap \mathbb{R}$, we see that $A \times S \subseteq (A \times X) \cap \mathbb{R}$, so $A \times S \subseteq \mathbb{R}$. Then by part (b) applied to $\mathbb{R}^-$, we see that $A \times B \subseteq \mathbb{R}$.

Now suppose that $A \times A \not\subseteq \mathbb{R} \circ \mathbb{R}^-$. From the finiteness of A we see that there is some k such that $A \times A \subseteq (\mathbb{R} \circ \mathbb{R}^-)^{\circ k}$. Choose k minimal, and let $\mathbb{R}' = (\mathbb{R} \circ \mathbb{R}^-)^{\circ(k-1)} \leq_{sd} \mathbb{A}^2$. Then $\mathbb{R}'$ is equal to its own reverse $\mathbb{R}'^-$, and $A \times A \subseteq \mathbb{R}' \circ \mathbb{R}'$ since $2(k-1) \geq k$ for $k \geq 2$. Thus the previous paragraphs applied to $\mathbb{R}'$ (using $\mathbb{R}' = \mathbb{R}'^-$) show that $A \times A \subseteq \mathbb{R}'$, contradicting the minimality of k. $\qquad\square$

Corollary 3.3.2. *If $\pi : \mathbb{A} \twoheadrightarrow \mathbb{B}$ is a surjective homomorphism of finite algebras, then the subalgebra of $\mathbb{A}$ generated by the maximal elements of $\mathbb{A}$ maps surjectively onto the subalgebra of $\mathbb{B}$ generated by the maximal elements of $\mathbb{B}$.*

Corollary 3.3.3. *If we start with any arc-consistent instance of $\mathrm{CSP}(\mathbb{A}_1, ..., \mathbb{A}_n)$ 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 3.3.4. *Fix a partial semilattice operation s. Suppose that $\mathbb{R} \leq_{sd} \mathbb{A} \times \mathbb{B}$ is a subdirect product of finite algebras $\mathbb{A}, \mathbb{B}$, and that $\mathbb{B}$ is simple and $\mathbb{B} = \mathrm{Sg}(B)$, with B a maximal strongly connected component of $\mathbb{B}$. Then:*

(a) if $\mathbb{A}$ is also simple and $\mathbb{A} = \mathrm{Sg}(A)$ with A a maximal strongly connected component of $\mathbb{A}$, and if $\mathbb{R} \cap (A \times B) \neq \emptyset$, then $\mathbb{R}$ is either the graph of an isomorphism or is $\mathbb{A} \times \mathbb{B}$, and

(b) if $\mathbb{A}$ is arbitrary and $\mathbb{R}$ is not the graph of a homomorphism from $\mathbb{A}$ to $\mathbb{B}$, then there is an $a \in \mathbb{A}$ with $\{a\} \times B \subseteq \mathbb{R}$.

Proof. If $\mathbb{B}$ is simple, then the linking congruence of $\mathbb{R}$ on $\mathbb{B}$ must either be the trivial congruence $0_\mathbb{B}$, in which case $\mathbb{R}$ is the graph of a homomorphism from $\mathbb{A}$ to $\mathbb{B}$, or the full congruence $1_\mathbb{B}$, in which case $\mathbb{R}$ is linked. In the second case, the results follow from Theorem 3.3.1(c). $\qquad\square$

Theorem 3.3.5. *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 3.3.1(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 $(a, b, c) \in R$. Since $\pi_{13}(R) = A \times C$, there exists some $b' \in B$ such that $(a', b', c) \in R$. Since

$$
\begin{bmatrix} a' \\ s(b, b') \\ c \end{bmatrix} = s\left(\begin{bmatrix} a \\ b \\ c \end{bmatrix}, \begin{bmatrix} a' \\ b' \\ c \end{bmatrix} \right) \in R,
$$

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 $(a, b', c') \in R$. Since

$$
\begin{bmatrix} a \\ b' \\ s(c, c') \end{bmatrix} = s\left(\begin{bmatrix} a \\ b \\ c \end{bmatrix}, \begin{bmatrix} a \\ b' \\ c' \end{bmatrix} \right) \in R,
$$

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

$$
\begin{bmatrix} a' \\ b' \\ c' \end{bmatrix} = s\left(\begin{bmatrix} a' \\ b' \\ c \end{bmatrix}, \begin{bmatrix} a \\ b' \\ c' \end{bmatrix} \right) \in R.
$$

Thus both a and a' meet $(b', c') \in \pi_{23}(R)$. $\qquad\square$

Remark 3.3.1. The proof of Theorem 3.3.5 actually proves something slightly more general: if $R \subseteq A \times B \times C$ is closed under s, $\pi_{12}(R) = A \times B, \pi_{13}(R) = A \times C$, and A is weakly connected, then R is linked when considered as a subalgebra of $A \times \pi_{23}(R)$.

Corollary 3.3.6. *Fix a partial semilattice operation s. Suppose $R \subseteq A_1 \times \cdots \times A_n$ is closed under s, A_1 is strongly connected, $\pi_{[2,n]}(R)$ is strongly connected, $\pi_{1i}(R) = A_1 \times A_i$ for $i \in [2, n]$, and A_i are finite for all i. Then $R = A_1 \times \pi_{[2,n]}(R)$.*

149

Corollary 3.3.7. *Fix a partial semilattice operation s. Suppose $R \subseteq A_1 \times \cdots \times A_n$ is closed under s, all A_i are strongly connected, $\pi_{ij}(R) = A_i \times A_j$ for all $i \neq j$, and A_i are finite for all i. Then $R = A_1 \times \cdots \times A_n$.*

Corollary 3.3.8. *Fix a partial semilattice operation s. If $\mathbb{A}$ is simple and is generated by a finite maximal strongly connected component A, then $\mathbb{A}$ is polynomially complete.*

Proof. We just need to show that every relation $\mathbb{R} \leq \mathbb{A}^n$ which contains the set of constant tuples $\Delta_n = \{(a, ..., a) \mid a \in \mathbb{A}\}$ is an intersection of equality relations. First consider the case $n = 2$. From the assumption that $\mathbb{A}$ is simple we see that either $\mathbb{R}$ is the equality relation, or $\mathbb{R}$ is linked. If $\mathbb{R}$ is linked, then Theorem 3.3.1(c) and the fact that $\mathbb{R}$ contains Δ_2 implies that $A \times A \subseteq \mathbb{R}$, and from the assumption $\mathbb{A} = \mathrm{Sg}(A)$ we see that $\mathbb{R} = \mathbb{A} \times \mathbb{A}$.

Now consider the case $n \geq 3$. If any two-variable projection $\pi_{ij}(\mathbb{R})$ is the equality relation, then we can ignore one of the coordinates i, j, so we may assume without loss of generality that $\pi_{i,j}(\mathbb{R}) = \mathbb{A} \times \mathbb{A}$ for all i, j. Let a be any element of A, and let R be a maximal strongly connected component of $\mathbb{R}$ which is reachable from $(a, ..., a)$. Then for any i, j we must have $\pi_{i,j}(R) = A \times A$, so by the previous corollary we have $R = A^n$. Thus $A^n \subseteq \mathbb{R}$, and from the assumption $\mathbb{A} = \mathrm{Sg}(A)$ we see that $\mathbb{R} = \mathbb{A}^n$. $\square$

Example 3.3.1. The reader may be wondering whether we can weaken the assumption that $\pi_{23}(R)$ is strongly connected from Theorem 3.3.5 to the assumption that B, C are strongly connected. It seems plausible that if B, C are both strongly connected and $\pi_{23}(R)$ is a subdirect product of B and C, $\pi_{23}(R)$ might automatically be strongly connected.

However, there is an example of a strongly connected 2-semilattice $\mathbb{A}$ and a subdirect product $\mathbb{R} \leq_{sd} \mathbb{A}^2$ which is *not* strongly connected. The 2-semilattice $\mathbb{A}$ is pictured below.
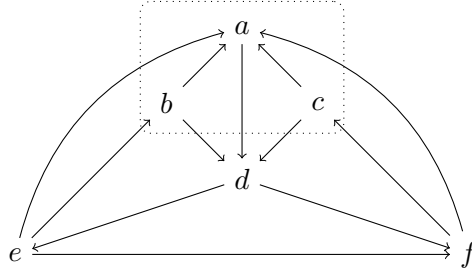

The missing values are given by $s(b, c) = s(b, f) = s(e, c) = a$.

If we let $\theta \leq_{sd} \mathbb{A}^2$ be the smallest congruence containing (b, c), then θ corresponds to the partition $\{a, b, c\}, \{d\}, \{e\}, \{f\}$, and $\mathbb{A}/\theta$ is a four element tournament. Considering θ as an algebra, we find that θ is *not* strongly connected: (b, c) and (c, b) are incomparable minimal elements of θ, and the remaining elements of θ form a maximal strongly connected component.

Example 3.3.2. 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 θ on a six-element algebra $\mathbb{A}$. This time $\mathbb{A}/\theta$ will be the three-element rock-paper-scissors algebra, and every congruence class of $\mathbb{A}$ will have two elements, with s acting as π_1 on the congruence class. As a digraph, $\mathbb{A}$ is just a directed six-cycle, pictured below.
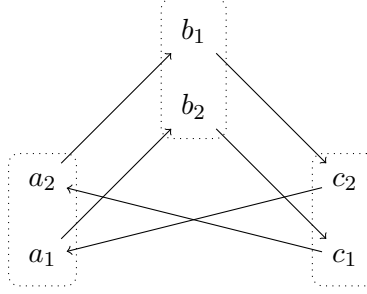
Given the above digraph structure and the assumption that there is a congruence θ corresponding to the partition $\{a_1, a_2\}, \{b_1, b_2\}, \{c_1, c_2\}$, there is only one way to fill in the values of the partial semilattice operation s. The reader can check that the congruence θ, considered as a subalgebra of $\mathbb{A}^2$, has two maximal strongly connected components which are both isomorphic to $\mathbb{A}$.

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 3.3.9. *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 $\Delta_A = \{(a,a) \mid a \in A\}$. Then R has a maximal strongly connected component which contains Δ_A.*

Proof. Since Δ_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 Δ_A is reachable from (a, b) in R. We will define a unary polynomial ϕ 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 ϕ, choose some sequence $(a_i, b_i) \in R$ such that $(a, a) = (a_0, b_0)$, $(a_i, b_i) \to (a_{i+1}, b_{i+1})$ for all i, and $(a_k, b_k) = (a, b)$ for some k. Then define ϕ by

$$\phi(x) = s\left(s\left(\cdots s\left(s\left(x, \begin{bmatrix} a_1 \\ b_1 \end{bmatrix}\right), \begin{bmatrix} a_2 \\ b_2 \end{bmatrix}\right), \cdots \right), \begin{bmatrix} a_k \\ b_k \end{bmatrix}\right).$$

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 ϕ_Δ of R by

$$\phi_\Delta(x) = s\left(s\left(\cdots s\left(s\left(x, \begin{bmatrix} b_1 \\ b_1 \end{bmatrix}\right), \begin{bmatrix} b_2 \\ b_2 \end{bmatrix}\right), \cdots \right), \begin{bmatrix} b_k \\ b_k \end{bmatrix}\right),$$

that is, by replacing each (a_i, b_i) in the definition of ϕ by (b_i, b_i). Then if $\phi^{\circ m}((a, a)) = (a, c)$, we have

$$\phi_\Delta^{\circ m}\left(\phi^{\circ(m-1)}\left(\begin{bmatrix} a \\ b \end{bmatrix}\right)\right) = \phi_\Delta^{\circ m}\left(\begin{bmatrix} a \\ c \end{bmatrix}\right) = \begin{bmatrix} c \\ c \end{bmatrix}.$$

Thus (c, c) is reachable from (a, b) in R. $\qquad\square$

Corollary 3.3.10. *If we start with any cycle-consistent instance of $\mathrm{CSP}(\mathbb{A}_1, ..., \mathbb{A}_n)$ 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 3.3.1(a), we just need to check this in the special case where our cycle-consistent instance is a cycle of binary relations $\mathbb{R}_i \leq_{sd} \mathbb{A}_i \times \mathbb{A}_{i+1}$ with indices taken modulo n. Let $\mathbb{R} \leq_{sd} \mathbb{A}_1 \times \cdots \times \mathbb{A}_n \times \mathbb{A}_1$ be the relation given by the formula

$$(x_1, x_2) \in \mathbb{R}_1 \wedge \cdots \wedge (x_n, x_{n+1}) \in \mathbb{R}_n.$$

The assumption that the instance is cycle-consistent implies that $\Delta_{\mathbb{A}_1} \subseteq \pi_{1,n+1}\mathbb{R}$. Set $\mathbb{R}_\Delta = \pi_{1,n+1}\mathbb{R}$.

For any algebra $\mathbb{A}$, let $\mathbb{A}^{\max}$ denote the subalgebra of $\mathbb{A}$ generated by the maximal elements of $\mathbb{A}$. We see from Theorem 3.3.1(a) that $\pi_{i,i+1}(\mathbb{R}^{\max}) = \mathbb{R}_i^{\max}$ for each i and that $\pi_{1,n+1}(\mathbb{R}^{\max}) = \mathbb{R}_\Delta^{\max}$. By Theorem 3.3.9 we have $\Delta_{\mathbb{A}_1^{\max}} \subseteq \mathbb{R}_\Delta^{\max}$, so the new instance is cycle-consistent at the first variable. $\square$

3.4   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 [34]. 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 preceeding 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 3.4.1. 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 3.4.2. *An algebra* $\mathbb{A} = (A, s)$ *is a 2-semilattice iff for all* $a, b \in \mathbb{A}$*, the subalgebra* $\mathrm{Sg}_{\mathbb{A}}\{a, b\}$ *is a semilattice under* s*.*

Proposition 3.4.3. *If* $\mathbb{A} = (A, s)$ *is a finite 2-semilattice, then* $\mathbb{A}$ *has a unique maximal strongly connected component.*

Proof. If a, b are any two maximal elements of $\mathbb{A}$, 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. $\square$

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 [34] 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 3.4.4. An idempotent algebra $\mathbb{A}$ with a fixed partial semilattice operation s is called *ancestral* if for all $a, b \in \mathbb{A}$, there is some $c \in \mathrm{Sg}_{\mathbb{A}}\{a, b\}$ which is reachable from both a and b. We call any such c a *common ancestor* of a and b.

Proposition 3.4.5. *A finite idempotent algebra* $\mathbb{A}$ *is ancestral iff every proper subalgebra of* $\mathbb{A}$ *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 3.4.6. An algebra $\mathbb{A} = (A, f)$ is a *spiral* if f is a commutative idempotent binary operation and every subalgebra of $\mathbb{A}$ which is generated by two elements either has size two or has a surjective homomorphism to the free semilattice on two generators.

Example 3.4.1. Here we give an example of a minimal spiral $\mathbb{A}_6$ which is not a 2-semilattice.
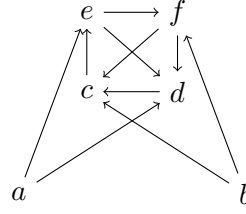
Every proper subalgebra of $\mathbb{A}_6$ is a 2-semilattice - in fact, every pair of elements other than $\{a, b\}$ generates a two or three element semilattice subalgebra of $\mathbb{A}_6$. The pair $\{a, b\}$ generates $\mathbb{A}_6$, and $\mathbb{A}_6$ has a congruence θ corresponding to the partition $\{a\}, \{b\}, \{c, d, e, f\}$ such that $\mathbb{A}_6/\theta$ is isomorphic to the free semilattice on two generators.

The reader may check that any nonempty subset S of $\mathbb{A}_6$ 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 $\text{Clo}(\mathbb{A}_6)$ contains no nontrivial proper subclones.

Theorem 3.4.7. *If $\mathbb{A} = (A, f)$ is a spiral, then for any partial semilattice term $s \in \text{Clo}(f)$ which is defined nontrivially in terms of f, the reduct $\mathbb{A}_s = (A, s)$ is ancestral.*

Proof. We prove this by induction on the size of A. Let a, b be any two elements of A. If $\text{Sg}_{\mathbb{A}}\{a, b\}$ 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 α from $\text{Sg}_{\mathbb{A}}\{a, b\}$ to the free semilattice on two generators. Clearly a and b must be sent to the two generators of the free semilattice by α, 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 α has congruence classes $\{a\}, \{b\}$, and $S = \text{Sg}_{\mathbb{A}}\{a, b\} \setminus \{a, b\}$, and S is a binary absorbing subalgebra of $\mathbb{A}$ with respect to f.

Since S is a binary absorbing subalgebra of $\mathbb{A}$ 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 $|S| \le |A| - 2$, we can apply the inductive hypothesis to see that $s(a, b), s(b, a)$ have a common ancestor in $\text{Sg}_{(S,s)}\{s(a, b), s(b, a)\} \subseteq \text{Sg}_{\mathbb{A}_s}\{a, b\}$. $\square$

Example 3.4.2. An example of an ancestral algebra which is not a 2-semilattice or a spiral is the algebra $\mathbb{A}_4 = (\{a, b, c, d\}, s)$, where s is the partial semilattice operation described below.

$$\begin{array}{c|cccc} s & a & b & c & d \\ \hline a & a & b & b & a \\ b & b & b & c & c \\ c & d & c & c & d \\ d & a & a & d & d \end{array}$$

The algebra $\mathbb{A}_4$ 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 $\mathbb{A}_4$, such that $s'(a,c) = a, s'(c,a) = c$. So the reduct $(\{a,b,c,d\}, s')$ of $\mathbb{A}_4$ 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 $\mathbb{A}_4$ is simple, and every proper subalgebra of $\mathbb{A}_4$ is a two element semilattice. By Corollary 3.3.8, $\mathbb{A}_4$ is polynomially complete, and in fact Theorem 3.3.1 and Theorem 3.3.5 imply that every subdirect relation $\mathbb{R} \leq_{sd} \mathbb{A}_4^n$ can be written as an intersection of two variable relations, each of which is the graph of an automorphism of $\mathbb{A}_4$. In particular, if we consider the ternary relation

$$\mathbb{R}_{ac} = \mathrm{Sg}_{\mathbb{A}^3}\left\{ \begin{bmatrix} a \\ a \\ c \end{bmatrix}, \begin{bmatrix} a \\ c \\ a \end{bmatrix}, \begin{bmatrix} c \\ a \\ a \end{bmatrix} \right\},$$

we find that $\mathbb{R}_{ac} = \mathbb{A}_4^3$. Since there is an automorphism of $\mathbb{A}_4$ which interchanges a and c, we see that there are ternary terms $g, g' \in \mathrm{Clo}(\mathbb{A}_4)$ 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

$$g'(x,x,y) \approx g'(x,y,x) \approx g'(y,x,x) \approx s'(y,x).$$

Proposition 3.4.8. *Every quotient of an ancestral algebra is ancestral.*

Theorem 3.4.9. *If $\mathbb{A}_1, ..., \mathbb{A}_n$ are ancestral algebras with partial semilattice operation s, then so is $\mathbb{A}_1 \times \cdots \times \mathbb{A}_n$.*

Proof. We prove this by induction on n. Let $a, b \in \mathbb{A}_1 \times \cdots \times \mathbb{A}_n$. Since $\mathbb{A}_1$ is ancestral, there is some $c_1 \in \mathrm{Sg}_{\mathbb{A}_1}\{a_1, b_1\}$ 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 $c' \in \mathrm{Sg}\{a,b\}$ with $c_1' = c_1$, and lifting the path from b_1 to c_1 to a path from b to some $c'' \in \mathrm{Sg}\{a,b\}$ 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 $\mathbb{A}_1$ is idempotent, we see that c', c'' have a common ancestor so long as $\mathbb{A}_2 \times \cdots \times \mathbb{A}_n$ is ancestral, which follows from the inductive hypothesis. $\square$

Corollary 3.4.10. *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 3.4.11. *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 $p_1, p_2, ..., p_m$ such that*

- *$a \to_s p_i(a, b)$ for all $a, b \in \mathbb{A}$ and all i,*

- *$a \to_s b$ 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 $\mathbb{F} = \mathcal{F}_{\mathbb{A}}(x, y) \leq \mathbb{A}^{\mathbb{A}^2}$ 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 $f_{i-1} \to_s f_i$, each $f_i \in \mathrm{Sg}_{\mathbb{F}}\{f_{i-1}(x, y), f_{i-1}(x, y)\}$, and such that the subset S of elements of the subalgebra $\mathbb{S} = \mathrm{Sg}_{\mathbb{F}}\{f_n(x, y), f_n(y, x)\}$ 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 $\mathbb{S} = \mathrm{Sg}_{\mathbb{F}}\{g(x, y), g(y, x)\}$. Thus we can extend our sequence $f_0, ..., f_n$ by $f_{n+1}, ..., f_m$ such that each $f_{i-1} \to_s f_i$, and $f_m(x, y) \approx f_n(y, x)$, and we will automatically have $f_i \in \mathrm{Sg}_{\mathbb{F}}\{f_{i-1}(x, y), f_{i-1}(x, y)\}$ for each i.

Note that $f_{i-1} \to_s f_i$ and $f_i \in \mathrm{Sg}_{\mathbb{F}}\{f_{i-1}(x, y), f_{i-1}(x, y)\}$ implies the existence of a binary term p_i such that $x \to_s p_i(x, y)$ and $f_i(x, y) = p_i(f_{i-1}(x, y), f_{i-1}(y, x))$, by the argument of Proposition 3.2.23. Note that the reduct with basic operations s, f_i is ancestral, and has the property that $a \to_s b$ 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 $\mathrm{Clo}(f_0, ..., f_m)$. Thus we may assume without loss of generality that $a \to_s b$ implies $p_i(a, b) = b$ for all i, and then the argument of Proposition 3.2.23 implies that each p_i is a partial semilattice term. $\square$

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 $(b, b) \in \mathrm{Sg}_{\mathbb{A}^2}\{(a, b), (b, a)\}$ iff $\{a, b\}$ is a semilattice subalgebra of $\mathbb{A}$ with $a \to b$.

Theorem 3.4.12. *Every finite ancestral algebra $\mathbb{A}$ has a reduct which is prepared and ancestral.*

Proof. Let s, f_i be as in Theorem 3.4.11, 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 $(b, b) \in \mathrm{Sg}_{\mathbb{A}^2}\{(a, b), (b, a)\}$ but $s(a, b) \neq b$. Let s' be a partial semilattice term with $s'(a, b) = b$. Then $c \to_s d$ implies $c \to_{s'} d$, 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 $\mathbb{A}$. $\square$

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 3.4.13. 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 3.4.14. *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. $\square$

Proposition 3.4.15. *If $\mathbb{A}$ is a minimal ancestral algebra and $\mathbb{B} \in HSP_{fin}(\mathbb{A})$, then $\mathbb{B}$ is also a minimal ancestral algebra.*

Proof. Let f_i be terms for $\mathbb{A}$ as in Theorem 3.4.11. 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 $f_0'(x,y) \approx x$, $f_i'(x,y) \to f_{i+1}'(x,y)$, 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

$$f_{m+i}'(x,y) := f_i(f_m'(x,y), f_n'(y,x)),$$

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 $\mathbb{A}$. $\square$

Theorem 3.4.16. *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 $\mathrm{Sg}_{\mathbb{A}}\{a,b\}$, then we have $\mathrm{Sg}_{\mathbb{A}}\{a,b\} = S \cup \{a,b\}$. If $\{a,b\} \not\subseteq S$, then $\mathrm{Sg}_{\mathbb{A}}\{a,b\}$ has a semilattice quotient with S as a congruence class which acts as the top element.*

Proof. Choose terms f_i as in Theorem 3.4.11. Let $\mathbb{F} = \mathcal{F}_{\mathbb{A}}(x,y)$ 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 $\mathbb{F}$, and note that since $g(x,y)$ is reachable from both x and y in $\mathbb{F}$, every term $t(x_1, ..., x_k) \in \mathrm{Clo}(g)$ which depends on all its inputs has the property that $t(x_1, ..., x_k)$ is reachable from each x_i in $\mathcal{F}_{\mathbb{A}}(x_1, ..., x_k)$.

Applying the semilattice iteration argument, we get a partial semilattice term $s'(x,y) \in \mathrm{Clo}(g)$, which is reachable from each of x, y, and $g(x,y)$ in $\mathbb{F}$. In particular, we see that $s'(x,y)$ is contained in the maximal strongly connected component of $\mathbb{F}$, 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 $\mathbb{F}$. Thus the clone generated by the f_i's must be equal to the clone of $\mathbb{A}$, and we see that every element of $\mathbb{F}$ is either equal to one of x, y or is contained in the maximal strongly connected component of $\mathbb{F}$. $\square$

Corollary 3.4.17. *If $\mathbb{A}$ is a minimal ancestral algebra, then the maximal strongly connected component of $\mathbb{A}$ is a strongly absorbing subalgebra of $\mathbb{A}$.*

There is a sense in which even the class of minimal ancestral algebras is unnecessarily large: it contains algebras such as the algebra $\mathbb{A}_4$ from Example 3.4.2 which have proper Taylor reducts with two element majority or affine subalgebras.

Theorem 3.4.18. *Suppose $\mathbb{A}$ is a minimal ancestral algebra which is generated by a and b, is strongly connected, and is simple. Then there are ternary terms $g, g' \in \mathrm{Clo}(\mathbb{A})$ 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 $\mathbb{S} = \mathrm{Sg}_{\mathbb{A}^2}\{(a, b), (b, a)\}$. If $\mathbb{S}$ is linked, then by Theorem 3.3.1(c) we must have $(b, b) \in \mathbb{S}$, so $a \to b$, a contradiction. Otherwise, $\mathbb{S}$ is the graph of an automorphism swapping a and b. In this case, the ternary relation $\mathbb{R} = \mathrm{Sg}_{\mathbb{A}^3}\{(a, a, b), (a, b, a), (b, a, a)\}$ has $(a, a), (a, b), (b, a) \in \pi_{i,j}(\mathbb{R})$ for each i, j, so by Theorem 3.3.1(c) we have $\pi_{i,j}(\mathbb{R}) = \mathbb{A}^2$, and then by Theorem 3.3.5 we have $\mathbb{R} = \mathbb{A}^3$. Thus $(a, a, a) \in \mathbb{R}$ and $(b, b, b) \in \mathbb{R}$, and we can take g, g' to be ternary terms of $\mathbb{A}$ which witness these facts. $\square$

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 3.4.19. *If $\mathbb{A}$ is a minimal ancestral algebra such that for all a, b the subalgebra $\mathrm{Sg}_{\mathbb{A}}\{a, b\}$ has no strongly connected quotient, then $\mathbb{A}$ is term equivalent to a spiral.*

Proof. Let s be a nontrivial partial semilattice operation on $\mathbb{A}$. 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 $|\mathbb{A}|$ that for each $a, b \in \mathbb{A}$, there is an n such that $f_n(a, b) = f_n(b, a)$. To see this, note that by Theorem 3.4.16, for any a, b the subalgebra generated by $s(a, b), s(b, a)$ is contained in the maximal strongly connected component S of $\mathrm{Sg}_{\mathbb{A}}\{a, b\}$, so as long as $S \neq \mathbb{A}$ we can apply the induction hypothesis to see that there is some i such that

$$f_i(s(a, b), s(b, a)) = f_i(s(b, a), s(a, b)),$$

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 $n = |\mathbb{A}|$). To finish, we need to show that if $\mathbb{A}$ is generated by two elements a, b with $|\mathbb{A}| > 2$, then the maximal strongly connected component S of $\mathbb{A}$ does not contain either of a, b. To this end, suppose for a contradiction that S contains b. Let $\mathbb{S} = \mathrm{Sg}_{\mathbb{A}^2}\{(a, b), (b, a)\}$. If S is contained in a linked component of $\mathbb{S}$, then by Theorem 3.3.1(c) we must have $(b, b) \in \mathbb{S}$, so $a \to b$, a contradiction. Otherwise, the linking congruence $\theta \in \mathrm{Con}(\mathbb{A})$ of $\mathbb{S}$ has $|S/\theta| > 1$ and $b/\theta \in S/\theta$, and so we may assume without loss of generality that θ is trivial. But if θ is trivial, then $\mathbb{A}$ has an automorphism which interchanges a and b, so S contains both a and b, so $\mathbb{A}$ is both strongly connected and generated by two elements, a contradiction. $\square$

3.5 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 [94] 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 $\mathbb{A}$ has a unique maximal strongly connected component $\mathbb{A}^{\max}$ (Proposition 3.4.5).

- If $\pi : \mathbb{A} \twoheadrightarrow \mathbb{B}$ is a surjective homomorphism, then $\pi(\mathbb{A}^{\max}) = \mathbb{B}^{\max}$ (Corollary 3.3.2 to Theorem 3.3.1(a)).

- If $\mathbb{R} \leq_{sd} \mathbb{A} \times \mathbb{B}$ and $\mathbb{A}^{\max}$ is contained in a linked component of $\mathbb{R}$, then $\mathbb{R}^{\max} = \mathbb{A}^{\max} \times \mathbb{B}^{\max}$ (Theorem 3.3.1(c)).

- In particular, if $\mathbb{R} \leq_{sd} \mathbb{A} \times \mathbb{B}$, $\mathbb{A}$ is generated by $\mathbb{A}^{\max}$, $\mathbb{B}$ is generated by $\mathbb{B}^{\max}$, and $\mathbb{B}$ is simple, then $\mathbb{R}$ is either the graph of a homomorphism $\mathbb{A} \twoheadrightarrow \mathbb{B}$ or $\mathbb{R} = \mathbb{A} \times \mathbb{B}$ (Corollary 3.3.4).

- If $\mathbb{R} \leq_{sd} \mathbb{A} \times \mathbb{B} \times \mathbb{C}$ has $\pi_{12}(\mathbb{R}) = \mathbb{A} \times \mathbb{B}$ and $\pi_{13}(\mathbb{R}) = \mathbb{A} \times \mathbb{C}$, then $\mathbb{R}^{\max} = \mathbb{A}^{\max} \times \pi_{23}(\mathbb{R})^{\max}$ (Theorem 3.3.5).

- Applying the above inductively, if $\mathbb{R} \leq_{sd} \mathbb{A}_1 \times \cdots \times \mathbb{A}_n$ has $\pi_{ij}(\mathbb{R}) = \mathbb{A}_i \times \mathbb{A}_j$ for all $i \neq j$, then $\mathbb{R}^{\max} = \mathbb{A}_1^{\max} \times \cdots \times \mathbb{A}_n^{\max}$ (Corollary 3.3.7).

- If $\mathbb{R} \leq_{sd} \mathbb{A} \times \mathbb{A}$ and $\mathbb{R}$ contains the diagonal $\Delta_{\mathbb{A}}$, then $\Delta_{\mathbb{A}^{\max}} \subseteq \mathbb{R}^{\max}$ (Theorem 3.3.9).

- If we start with any cycle-consistent instance of $\mathrm{CSP}(\mathbb{A}_1, ..., \mathbb{A}_n)$ 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 3.3.10).

If we assume that our algebras are minimal ancestral (rather than just ancestral), then each $\mathbb{A}^{\max}$ becomes a subalgebra (Corollary 3.4.17), 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 $\mathbb{A}_x$ is generated by $\mathbb{A}_x^{\max}$ (the last bullet point above).

To finish the argument, we need to find another strategy for reducing the variable domains when each $\mathbb{A}_x = \mathrm{Sg}(\mathbb{A}_x^{\max})$. The intuition is that if $\mathbb{A}_x = \mathrm{Sg}(\mathbb{A}_x^{\max})$, then there is some congruence $\theta_x \in \mathrm{Con}(\mathbb{A}_x)$ such that $\mathbb{A}_x/\theta_x$ is simple, and in fact $\mathbb{A}_x/\theta_x$ will be polynomially complete by Corollary 3.3.8. Since polynomially complete algebras should have few interesting subdirect relations, it's plausible that we can replace the domain $\mathbb{A}_x$ with an arbitrary congruence class of θ_x, and always obtain a cycle-consistent instance.

So fix a variable x with $|\mathbb{A}_x| > 1$, a maximal congruence θ_x in $\mathrm{Con}(\mathbb{A}_x)$, and a congruence class $\mathbb{A}'_x$ of θ_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 3.5.1. If $\mathbf{X}$ is an instance of a CSP and x, y are variables of $\mathbf{X}$, then a *path* p from x to y is defined as a sequence $x = v_0, (\mathbb{R}_1, i_1, j_1), v_1, ..., v_{n-1}, (\mathbb{R}_n, i_n, j_n), v_n = y$ such that each v_k is a variable, and each $\mathbb{R}_k$ is a relation such that one of the constraints of the instance $\mathbf{X}$ imposes the relation $\mathbb{R}_k$ 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 $\mathbb{P}_p \leq \mathbb{A}_x \times \mathbb{A}_y$ which is given by

$$\mathbb{P}_p := \pi_{i_1 j_1}(\mathbb{R}_1) \circ \cdots \circ \pi_{i_n j_n}(\mathbb{R}_n).$$

In other words, $\mathbb{P}_p$ is the set of pairs of values in $\mathbb{A}_x \times \mathbb{A}_y$ 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 $\mathbb{P}_{p+q} = \mathbb{P}_p \circ \mathbb{P}_q$, and $-p$ is a path from y to x with $\mathbb{P}_{-p} = \mathbb{P}_p^-$.

In particular, we see that an instance is arc-consistent iff for all paths p the associated binary relations $\mathbb{P}_p$ are subdirect, and it is cycle-consistent iff we additionally have $\Delta_{\mathbb{A}_v} \subseteq \mathbb{P}_p$ for every path p from a variable v back to itself.

Definition 3.5.2. Suppose that $\mathbf{X}$ is a cycle-consistent instance such that for all variable domains we have $\mathbb{A}_v = \mathrm{Sg}(\mathbb{A}_v^{\max})$, that x is any variable with $|\mathbb{A}_x| > 1$, that θ_x is any maximal congruence on $\mathbb{A}_x$, and that $\mathbb{A}'_x$ is any congruence class of $\mathbb{A}_x/\theta_x$.

For each variable y, we say that y is *proper* if there is a path p from y to x such that $\mathbb{P}_p/\theta_x \leq \mathbb{A}_y \times \mathbb{A}_x/\theta_x$ is the graph of a homomorphism $\iota_y : \mathbb{A}_y \twoheadrightarrow \mathbb{A}_x/\theta_x$. In this case, we define the congruence $\theta_y \in \mathrm{Con}(\mathbb{A}_y)$ to be the kernel of ι_y, and we define $\mathbb{A}'_y$ to be the preimage of $\mathbb{A}'_x$ under ι_y. If y is not proper, then we define $\mathbb{A}'_y$ to be $\mathbb{A}_y$.

We define the reduced instance $\mathbf{X}'$ by replacing the domain of each variable v by $\mathbb{A}'_v$, and replacing each constraint relation $\mathbb{R} \leq \mathbb{A}_{v_1} \times \cdots \times \mathbb{A}_{v_m}$ of $\mathbf{X}$ by $\mathbb{R}' = \mathbb{R} \cap (\mathbb{A}'_{v_1} \times \cdots \times \mathbb{A}'_{v_m})$.

The reason for the name "proper" is that a variable v is proper iff the reduced domain $\mathbb{A}'_v$ is a proper subalgebra of $\mathbb{A}_v$. First we need to check that the maps ι_y for the proper variables y are well-defined.

Lemma 3.5.3. *If y is a proper variable and p, q are two paths from y to x such that $\mathbb{P}_p/\theta_x, \mathbb{P}_q/\theta_x$ are graphs of homomorphisms $\iota_p, \iota_q : \mathbb{A}_y \twoheadrightarrow \mathbb{A}_x/\theta_x$, then in fact we have $\iota_p = \iota_q$. Thus ι_y, θ_y, and $\mathbb{A}'_y$ are all well-defined.*

Proof. The path $p - q$ connects y to itself, so by cycle-consistency we must have $\Delta_{\mathbb{A}_y} \subseteq \mathbb{P}_{p-q} = \mathbb{P}_p \circ \mathbb{P}_q^-$. Taking the quotient by θ_x, we see that $\Delta_{\mathbb{A}_y} \subseteq (\mathbb{P}_p/\theta_x) \circ (\mathbb{P}_q/\theta_x)^-$, so for every element $a \in \mathbb{A}_y$ we must have $\iota_p(a) = \iota_q(a)$. $\qquad\square$

We sometimes abuse notation, and think of ι_y as an isomorphism from $\mathbb{A}_y/\theta_y$ to $\mathbb{A}_x/\theta_x$.

Lemma 3.5.4. *Suppose p is a path from y to a proper variable z. Then one of the following is true:*

159

- $\mathbb{P}_p/\theta_z = \mathbb{A}_y \times \mathbb{A}_z/\theta_z$, *or*

- *y is also proper, and $\mathbb{P}_p/(\theta_y \times \theta_z)$ is the graph of an isomorphism $\iota_p : \mathbb{A}_y/\theta_y \xrightarrow{\sim} \mathbb{A}_z/\theta_z$ such that $\iota_y = \iota_z \circ \iota_p$.*

Proof. This follows from Corollary 3.3.4 and cycle-consistency (note that $\mathbb{A}_z/\theta_z$ is simple, since it is isomorphic to $\mathbb{A}_x/\theta_x$). $\qquad\square$

We have the ingredients necessary to check that the reduced instance $\mathbf{X}'$ is cycle-consistent. We start with arc-consistency.

Lemma 3.5.5. *Suppose $\mathbb{R} \leq_{sd} \mathbb{A}_{v_1} \times \cdots \times \mathbb{A}_{v_n}$ is a constraint of $\mathbf{X}$. Then the reduced constraint $\mathbb{R}' = \mathbb{R} \cap (\mathbb{A}'_{v_1} \times \cdots \times \mathbb{A}'_{v_n})$ is subdirect inside $\mathbb{A}'_{v_1} \times \cdots \times \mathbb{A}'_{v_n}$, that is, $\pi_i(\mathbb{R}') = \mathbb{A}'_{v_i}$ for each i.*

Proof. By symmetry, it's enough to prove that $\pi_1(\mathbb{R}') = \mathbb{A}'_{v_1}$. In other words, for each element $a \in \mathbb{A}'_{v_1}$, we want to find a tuple $s \in \mathbb{R}$ such that $s_i \in \mathbb{A}'_{v_i}$ 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 $\mathbb{A}_{v_i}$ to $\mathbb{A}'_{v_i} = \mathbb{A}_{v_i}$ has no effect. Similarly, for any two proper variables v_i, v_j such that $\pi_{ij}(\mathbb{R})$ induces an isomorphism between $\mathbb{A}_{v_i}/\theta_{v_i}$ and $\mathbb{A}_{v_j}/\theta_{v_j}$, we may ignore one of the two variables v_i, v_j, since any element $s \in \mathbb{R}$ which satisfies $s_i \in \mathbb{A}'_{v_i}$ will automatically also satisfy $s_j \in \mathbb{A}'_{v_j}$.

To formalize the process of ignoring variables, we define an equivalence relation $\sim$ on the set of indices of proper variables of $\mathbb{R}$, with $i \sim j$ when $\pi_{ij}(\mathbb{R})$ induces an isomorphism between $\mathbb{A}_{v_i}/\theta_{v_i}$ and $\mathbb{A}_{v_j}/\theta_{v_j}$ (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 $\mathbb{S} \leq \mathbb{A}_{v_1} \times \prod_{i \in I \setminus \{1\}} \mathbb{A}_{v_i}/\theta_{v_i}$ by

$$\mathbb{S} := \pi_I(\mathbb{R}) \Big/ \prod_{i \in I \setminus \{1\}} \theta_{v_i}.$$

We just need to show that for every $a \in \mathbb{A}'_{v_1}$ there is some $s \in \mathbb{S}$ with $s_1 = a$ and $s_i = \mathbb{A}'_{v_i}/\theta_{v_i}$ for each $i \in I \setminus \{1\}$. Note that by Lemma 3.5.4 and the construction of I, for every pair $i, j \in I$ the projection $\pi_{ij}(\mathbb{S})$ is full. Thus by Corollary 3.3.7, we in fact have

$$\mathbb{S}^{\max} = \mathbb{A}^{\max}_{v_1} \times \prod_{i \in I \setminus \{1\}} \mathbb{A}^{\max}_{v_i}/\theta_{v_i},$$

and since each $\mathbb{A}_{v_i}$ is generated by $\mathbb{A}^{\max}_{v_i}$, we have

$$\mathbb{S} = \mathbb{A}_{v_1} \times \prod_{i \in I \setminus \{1\}} \mathbb{A}_{v_i}/\theta_{v_i}. \qquad\square$$

Now we can check that cycle-consistency also holds for the reduced instance.

Lemma 3.5.6. *Suppose p is a path from v to v in the instance $\mathbf{X}$, and let p' be the corresponding path in $\mathbf{X}'$. If $\Delta_{\mathbb{A}_{v_1}} \subseteq \mathbb{P}_p$, then $\Delta_{\mathbb{A}'_{v_1}} \subseteq \mathbb{P}_{p'}$.*

Proof. Suppose that p is the path $v = v_0, (\mathbb{R}_1, i_1, j_1), v_1, ..., v_{n-1}, (\mathbb{R}_n, i_n, j_n), v_n = v$. Note that in the corresponding path p', we must replace each $\mathbb{R}_i$ with $\mathbb{R}'_i$, so we must also worry about the

proper variables which occur in $\mathbb{R}_i$ but do not lie along the path p. In order to do this cleanly, we consider the relation $\mathbb{R}$ defined by

$$\mathbb{R} := \left\{ (v_0, u^1, ..., u^n, v_n) \in \mathbb{A}_v \times \prod_{i \leq n} \mathbb{R}_i \times \mathbb{A}_v \;\middle|\; v_0 = u_{i_1}^1, u_{j_1}^1 = u_{i_2}^2, ..., u_{j_{n-1}}^{n-1} = u_{i_n}^n, u_{j_n}^n = v_n \right\}.$$

If each $\mathbb{R}_i$ has arity m_i, then $\mathbb{R}$ is thought of as a relation of arity $m = 2 + \sum_i m_i$, and the indices of $\mathbb{R}$ might contain several copies of variables of the instance $\mathbf{X}$. Let the ith index of $\mathbb{R}$ correspond to the variable y_i in $\mathbf{X}$, with $y_1 = v_0 = v$ and $y_m = v_n = v$, so

$$\mathbb{R} \leq_{sd} \mathbb{A}_{y_1} \times \cdots \times \mathbb{A}_{y_m}.$$

Note that by the arc-consistency of the instance $\mathbf{X}$, for any two indices i, j of the relation $\mathbb{R}$, the projection $\pi_{ij}(\mathbb{R})$ is the same as $\mathbb{P}_q$ for some path q from y_i to y_j formed out of the relations $\mathbb{R}_i$, and that $\pi_{1m}(\mathbb{R}) = \mathbb{P}_p$, so $\pi_{1m}(\mathbb{R}) \supseteq \Delta_{\mathbb{A}_v}$.

As in the argument for arc-consistency, we define an equivalence relation $\sim$ on the proper indices of $\mathbb{R}$ defined by $i \sim j$ when $\pi_{ij}(\mathbb{R})$ induces an isomorphism between $\mathbb{A}_{y_i}/\theta_{y_i}$ and $\mathbb{A}_{y_j}/\theta_{y_j}$. We let $I \subseteq [m]$ to be a set of indices of $\mathbb{R}$ with $1, m \in I$, such that I contains no indices of non-proper variables of $\mathbb{R}$ other than possibly 1 and m, such that $I \setminus \{m\}$ contains one representative from each $\sim$ class of $\{1, ..., m-1\}$, and such that $I \setminus \{1\}$ contains one representative from each $\sim$ class of $\{2, ..., m\}$. As before, we define a relation $\mathbb{S}$ by

$$\mathbb{S} := \pi_I(\mathbb{R}) \Big/ \prod_{i \in I \setminus \{1, m\}} \theta_{y_i}.$$

We just need to show that for every $a \in \mathbb{A}_v'$, there is some $s \in \mathbb{S}$ with $s_1 = s_m = a$ and $s_i = \mathbb{A}_{v_i}'/\theta_{v_i}$ for each $i \in I \setminus \{1, m\}$. By Lemma 3.5.4 and the construction of I, for every pair $i, j \in I$ with $\{i, j\} \neq \{1, m\}$ the projection $\pi_{ij}(\mathbb{S})$ is full. Thus by Corollary 3.3.7, we have

$$\pi_{I \setminus \{m\}}(\mathbb{S}) = \mathbb{A}_{y_1} \times \prod_{i \in I \setminus \{1, m\}} \mathbb{A}_{y_i}/\theta_{y_i}$$

and

$$\pi_{I \setminus \{1\}}(\mathbb{S}) = \mathbb{A}_{y_m} \times \prod_{i \in I \setminus \{1, m\}} \mathbb{A}_{y_i}/\theta_{y_i}.$$

Thus by Theorem 3.3.5, we have

$$\mathbb{S}^{\max} = \pi_{1m}(\mathbb{S})^{\max} \times \prod_{i \in I \setminus \{1, m\}} \mathbb{A}_{y_i}^{\max}/\theta_{y_i},$$

and by Theorem 3.3.9 and the assumption $\pi_{1m}(\mathbb{S}) = \pi_{1m}(\mathbb{R}) \supseteq \Delta_{\mathbb{A}_v}$, we have $\pi_{1m}(\mathbb{S})^{\max} \supseteq \Delta_{\mathbb{A}_v^{\max}}$. Since each $\mathbb{A}_y$ is generated by $\mathbb{A}_y^{\max}$, we have

$$\mathbb{S} \supseteq \Delta_{\mathbb{A}_v} \times \prod_{i \in I \setminus \{1, m\}} \mathbb{A}_{y_i}/\theta_{y_i},$$

so in particular for every $a \in \mathbb{A}_v'$ we have $\{a\} \times \prod_{i \in I \setminus \{1, m\}} \mathbb{A}_{y_i}'/\theta_{y_i} \times \{a\} \subseteq \mathbb{S}$, so $(a, a) \in \pi_{1m}(\mathbb{R}') = \mathbb{P}_{p'}$. $\qquad\square$

Thus the reduced instance $\mathbf{X}'$ is cycle-consistent. Since we can iteratively shrink our instance whenever some variable x has $\mathbb{A}_x \neq \mathrm{Sg}(\mathbb{A}_x^{\max})$ or has $\mathbb{A}_x = \mathrm{Sg}(\mathbb{A}_x^{\max})$ but $|\mathbb{A}_x| > 1$, we see that we eventually reach a situation where each $\mathbb{A}_x$ 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 3.5.7. *If $\mathbf{X}$ is a cycle-consistent instance of an ancestral CSP, then $\mathbf{X}$ has a solution.*

In fact, for any variable x of $\mathbf{X}$, and for any element $a \in \mathbb{A}_x$ such that there is a sequence of subalgebras $\mathbb{A}_x \supseteq \mathbb{A}_0 \supseteq \cdots \supseteq \mathbb{A}_n = \{a\}$ with $\mathbb{A}_0 = \mathrm{Sg}(\mathbb{A}_x^{\max})$ and such that for each i, there is a maximal congruence $\theta_i \in \mathrm{Con}(\mathbb{A}_i)$ and a congruence class $\mathbb{A}_i'$ of θ_i with $\mathbb{A}_{i+1} = \mathrm{Sg}(\mathbb{A}_i'^{\max})$, there is a solution to the instance $\mathbf{X}$ in which x is assigned the value a.

The simple construction of the reduced instance $\mathbf{X}'$ can be used to show that we can find a solution to any cycle-consistent instance of an ancestral CSP in linear time.

3.6 Cycle-consistency solves majority CSPs

The paper which prompted the study of cycle-consistency was a preliminary investigation by Chen, Dalmau, and Grußien [48], 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 [48] 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 $\mathbb{A}_x$ 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, [48] used the closely related concept of an *ideal* of a majority algebra.

Definition 3.6.1. If $\mathbb{A} = (A, m)$ is a majority algebra, then $\mathbb{B} \leq \mathbb{A}$ is called an *ideal* of $\mathbb{A}$ if $m(\mathbb{B}, \mathbb{A}, \mathbb{B}) \subseteq \mathbb{B}$.

The word "ideal" comes from the theory of median algebras - a subset $\mathbb{B}$ is an ideal of a median algebra $\mathbb{A}$ iff there is a congruence θ of $\mathbb{A}$ such that $\mathbb{B}$ is a congruence class of θ. The corresponding statement is not true of majority algebras in general: every subset of the dual discriminator algebra from Example 1.6.5 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 3.6.2. *Suppose that a relation $\mathbb{R}$ is defined by a primitive positive formula Φ involving the relations $\mathbb{R}_1, ..., \mathbb{R}_k$. If we replace each $\mathbb{R}_i$ with an ideal $\mathbb{R}_i'$ of $\mathbb{R}_i$ to make a primitive positive formula Φ', then the relation $\mathbb{R}'$ which is defined by Φ' is an ideal of $\mathbb{R}$.*

Proof. Let $\Phi(x) = \exists y \Psi(x, y)$, with Ψ quantifier-free, and let Ψ' be the corresponding formula with $\mathbb{R}_i$s replaced by $\mathbb{R}'_i$s. Then for any a, b, c with $a, c \in \mathbb{R}'$ and $b \in \mathbb{R}$, 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 $\mathbb{R}'_i$ is an ideal, so $\Phi'(m(a, b, c))$ holds. $\qquad\square$

Recall the definition of a path in an instance (Definition 3.5.1). It's notationally convenient to allow paths to act on subsets of the variable domains.

Definition 3.6.3. If p is a path connecting variables x, y of an instance $\mathbf{X}$, and if B is a subset of the variable domain $\mathbb{A}_x$, then we define $B + p$ to be the subset of $\mathbb{A}_y$ given by

$$B + p := \{c \in \mathbb{A}_y \mid \exists b \in B \text{ s.t. } (b, c) \in \mathbb{P}_p\} = \pi_2(\mathbb{P}_p \cap (B \times \mathbb{A}_y)).$$

Proposition 3.6.4. *If* $\mathbb{B} \leq \mathbb{A}_x$ *and* p *is a path from* x *to* y*, then* $\mathbb{B} + p$ *is a subalgebra of* $\mathbb{A}_y$*. If* $\mathbb{B}$ *is an ideal of* $\mathbb{A}_x$ *and the instance is arc-consistent, then* $\mathbb{B} + p$ *is an ideal of* $\mathbb{A}_y$*.*

Our overall strategy will be to start with a cycle-consistent instance $\mathbf{X}$, and find a collection of ideals $\mathbb{A}'_x$ of the variable domains $\mathbb{A}_x$ such that reducing each domain to $\mathbb{A}'_x$ produces a 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 $(x, \mathbb{B})$ where x is a variable and $\mathbb{B}$ is a proper ideal of $\mathbb{A}_x$. Note that $\mathcal{I}$ is nonempty as long as some x has $|\mathbb{A}_x| > 1$, since every singleton is an ideal.

Definition 3.6.5. Let $\mathcal{I}$ be the set of pairs $(x, \mathbb{B})$ where x is a variable and $\mathbb{B}$ is a proper ideal of $\mathbb{A}_x$. We define a quasiorder $\preceq$ on $\mathcal{I}$ by $(x, \mathbb{B}) \preceq (y, \mathbb{B} + p)$ for every path p from x to y with $\mathbb{B} + p \neq \mathbb{A}_y$.

Proposition 3.6.6. *If* $\mathbf{X}$ *is a cycle-consistent instance,* x *is a variable, and* $(x, \mathbb{B}) \preceq (x, \mathbb{C})$*, then* $\mathbb{B} \leq \mathbb{C}$*.*

Proof. Suppose p is a path from x to itself with $\mathbb{B} + p = \mathbb{C}$. By cycle-consistency we must have $\Delta_{\mathbb{A}_x} \subseteq \mathbb{P}_p$, so $\mathbb{B} \subseteq \mathbb{B} + p$. $\qquad\square$

Definition 3.6.7. 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 $(x, \mathbb{A}'_x)$ of $\mathcal{I}$ under the quasiorder $\preceq$.

Call a variable y *proper* if there is a path p from x to y such that $\mathbb{A}'_x + p \neq \mathbb{A}_y$, and in this case set $\mathbb{A}'_y = \mathbb{A}'_x + p$. If y is not proper, then set $\mathbb{A}'_y = \mathbb{A}_y$.

Define the reduced instance $\mathbf{X}'$ by replacing the domain of each variable v by $\mathbb{A}'_v$, and by replacing each constraint $\mathbb{R} \leq \mathbb{A}_u \times \mathbb{A}_v$ with $\mathbb{R}' = \mathbb{R} \cap (\mathbb{A}'_u \times \mathbb{A}'_v)$.

First we need to check that the sets $\mathbb{A}'_y$ are well-defined.

Lemma 3.6.8. *If there are paths* p, q *from* x *to* y *such that* $\mathbb{A}'_x + p \neq \mathbb{A}_y$ *and* $\mathbb{A}'_x + q \neq \mathbb{A}_y$*, then* $\mathbb{A}'_x + p = \mathbb{A}'_x + q$*.*

Proof. Since $(x, \mathbb{A}'_x)$ is maximal and $(x, \mathbb{A}'_x) \preceq (y, \mathbb{A}'_x + p)$, we must have $(y, \mathbb{A}'_x + p) \preceq (x, \mathbb{A}'_x) \preceq (y, \mathbb{A}'_x + q)$, so $\mathbb{A}'_x + p \leq \mathbb{A}'_x + q$. Similarly we have $\mathbb{A}'_x + q \leq \mathbb{A}'_x + p$, so $\mathbb{A}'_x + p = \mathbb{A}'_x + q$. $\qquad\square$

Next we check arc-consistency.

Lemma 3.6.9. *If p is a path from y to z and p' is the corresponding path in $\mathbf{X}'$, then $\mathbb{A}'_y + p' = \mathbb{A}'_z$.*

Proof. We just need to check this in the case when p has length 1, corresponding to a binary relation $\mathbb{R} \leq_{sd} \mathbb{A}_y \times \mathbb{A}_z$. If $\mathbb{A}'_y + p \neq \mathbb{A}_z$, then y, z must both be proper with $\mathbb{A}'_y + p = \mathbb{A}'_z$. Either way we see that $\mathbb{A}'_y + p \supseteq \mathbb{A}'_z$, and since $\mathbb{R}' = \mathbb{R} \cap (\mathbb{A}'_y \times \mathbb{A}'_z)$ we have $\mathbb{A}'_y + p' = \mathbb{A}'_z$ in the reduced instance. $\square$

Finally, we check that arc-consistency of $\mathbf{X}'$ and cycle-consistency of $\mathbf{X}$ implies cycle-consistency of $\mathbf{X}'$. For this, we note that if p is a path from v back to itself in $\mathbf{X}$, and if p' is the corresponding path in $\mathbf{X}'$, then $\mathbb{P}_{p'}$ is an ideal of $\mathbb{P}_p$. Since $\mathbb{P}_p \supseteq \Delta_{\mathbb{A}_v}$ we have

$$m(\mathbb{P}_{p'}, \Delta_{\mathbb{A}'_v}, \mathbb{P}_{p'}) \subseteq \mathbb{P}_{p'},$$

so the cycle-consistency of $\mathbf{X}'$ follows from the following result.

Theorem 3.6.10. *Suppose that $\mathbb{R} \leq_{sd} \mathbb{A} \times \mathbb{A}$ is subdirect with $m(\mathbb{R}, \Delta_{\mathbb{A}}, \mathbb{R}) \subseteq \mathbb{R}$, where m is a majority operation. Then $\Delta_{\mathbb{A}} \subseteq \mathbb{R}$.*
In fact, if $\mathbb{R} \leq_{sd} \mathbb{A}_1 \times \cdots \times \mathbb{A}_n$ is subdirect and satisfies $m(\mathbb{R}, S, \mathbb{R}) \subseteq \mathbb{R}$, where S is any subset of $\mathbb{A}_1 \times \cdots \times \mathbb{A}_n$, then $S \subseteq \mathbb{R}$.

Proof. First we prove the statement about binary relations, since this is all we will need. Let a be any element of $\mathbb{A}$. Since $\mathbb{R}$ is subdirect, there are $b, c \in \mathbb{A}$ such that $(a, b) \in \mathbb{R}$ and $(c, a) \in \mathbb{R}$. Then since $(a, a) \in \Delta_{\mathbb{A}}$, we have

$$\begin{bmatrix} a \\ a \end{bmatrix} = m\left(\begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} a \\ a \end{bmatrix}, \begin{bmatrix} c \\ a \end{bmatrix} \right) \in m(\mathbb{R}, \Delta_{\mathbb{A}}, \mathbb{R}) \subseteq \mathbb{R}.$$

For the more general statement, we show by induction on k that $\pi_{[k]}(S) \subseteq \pi_{[k]}(\mathbb{R})$ for each $k \leq n$. The base case $k = 1$ follows from the assumption that $\mathbb{R}$ is subdirect, and for the inductive step we may as well assume that we have already proven this for $k = n - 1$, and wish to show it for n. Let $(a_1, ..., a_n)$ be any element of S. Then by the inductive hypothesis there is some b such that $(a_1, ..., a_{n-1}, b) \in \mathbb{R}$, and by the assumption that $\mathbb{R}$ is subdirect there are $c_1, ..., c_{n-1}$ such that $(c_1, ..., c_{n-1}, a_n) \in \mathbb{R}$. Then we have

$$\begin{bmatrix} a_1 \\ \vdots \\ a_{n-1} \\ a_n \end{bmatrix} = m\left(\begin{bmatrix} a_1 \\ \vdots \\ a_{n-1} \\ b \end{bmatrix}, \begin{bmatrix} a_1 \\ \vdots \\ a_{n-1} \\ a_n \end{bmatrix}, \begin{bmatrix} c_1 \\ \vdots \\ c_{n-1} \\ a_n \end{bmatrix} \right) \in m(\mathbb{R}, S, \mathbb{R}) \subseteq \mathbb{R}. \qquad \square$$

Corollary 3.6.11. *The reduced instance $\mathbf{X}'$ is cycle-consistent.*

We have proved the main result of this section.

Theorem 3.6.12. *Every cycle-consistent instance $\mathbf{X}$ of a majority CSP has a solution.*
In fact, for any variable v of $\mathbf{X}$ and any value $a \in \mathbb{A}_v$, the instance $\mathbf{X}$ has a solution in which the variable v is assigned the value a.

Proof. For the second statement, we note that if $|\mathbb{A}_v| > 1$, then $(v, \{a\}) \in \mathcal{I}$, so there is some maximal element $(x, \mathbb{A}'_x) \in \mathcal{I}$ such that $(v, \{a\}) \preceq (x, \mathbb{A}'_x)$, and we define the reduction $\mathbf{X}'$ in terms of the maximal element $(x, \mathbb{A}'_x)$. If v is proper, then from $(v, \{a\}) \preceq (x, \mathbb{A}'_x) \preceq (v, \mathbb{A}'_v)$ we must have $a \in \mathbb{A}'_v$, and if v is not proper then we have $a \in \mathbb{A}_v = \mathbb{A}'_v$. Either way, we see by induction that the reduced instance $\mathbf{X}'$ has a solution in which the variable v is assigned the value a. $\square$

164

Corollary 3.6.13. *Suppose $\mathbb{A}$ is an algebra with a partial semilattice term s and a ternary term g such that for any subalgebra $\mathbb{B} \le \mathbb{A}$, the restriction of g to $\mathrm{Sg}(\mathbb{B}^{\max})$ is a majority operation. Then every cycle-consistent instance of $\mathrm{CSP}(\mathbb{A})$ has a solution.*

Proof. By Corollary 3.3.10, if we start with a cycle-consistent instance $\mathbf{X}$ and restrict all the variable domains $\mathbb{A}_i$ to $\mathrm{Sg}(\mathbb{A}_i^{\max})$ to create a new instance $\mathbf{X}'$, then $\mathbf{X}'$ will still be cycle-consistent, and by assumption $\mathbf{X}'$ will be preserved by the majority operation g. Then by the previous theorem, $\mathbf{X}'$ will have a solution. $\qquad\square$

Example 3.6.1. Consider $\mathbb{A} = (\{-, 0, +\}, g)$, where g is the idempotent cyclic ternary operaton with

$$g(0, 0, -) = g(0, -, -) = -,$$
$$g(0, -, +) = g(-, -, +) = -,$$
$$g(0, 0, +) = g(0, +, +) = +,$$
$$g(0, +, -) = g(-, +, +) = +.$$

This can be described more succinctly as follows: the permutation $(- \, +)$ is an automorphism of $\mathbb{A}$, $\{-, +\}$ is a majority subalgebra of $\mathbb{A}$, and $\{0, -\}, \{0, +\}$ are semilattice subalgebras of $\mathbb{A}$ with $0 \to -, +$. 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 $\mathrm{CSP}(\mathbb{A})$ has a solution. We give a table for s and draw the graph of two element subalgebras of $\mathbb{A}$ (with undirected edges for majority subalgebras and directed edges for semilattice subalgebras) below.

s	$-$	0	$+$
$-$	$-$	$-$	$-$
0	$-$	0	$+$
$+$	$+$	$+$	$+$

The relational clone $\mathrm{Inv}(g)$ is generated by the unary relation $x \ne 0$, the binary relation $x = -y$, the binary relation $x \le y$, 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 1.6.8, 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 [40] 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 3.6.1. It's tempting to try to generalize Theorem 3.6.10 to near-unanimity operations. We say that a subalgebra $\mathbb{B}$ *absorbs* $\mathbb{A}$ with respect to a near-unanimity operation t if

$$t(\mathbb{B}, ..., \mathbb{B}, \mathbb{A}, \mathbb{B}, ..., \mathbb{B}) \subseteq \mathbb{B}$$

for each possible location of $\mathbb{A}$. Suppose that $\mathbb{R} \le_{sd} \mathbb{A} \times \mathbb{A}$ absorbs $\Delta_{\mathbb{A}}$ with respect to t - can we conclude that $\mathbb{R}$ contains the diagonal?

Unfortunately the answer is no: even if $\mathbb{R}$ is subdirect and absorbs $\mathbb{A}^2$ with respect to a near-unanimity term, we might not have $\Delta_{\mathbb{A}} \subseteq \mathbb{R}$. Consider the threshold function t_2^n from Example 1.1.3 defined by

$$t_2^n(x_1, ..., x_n) = \begin{cases} 1 & \sum_i x_i \ge 2, \\ 0 & \sum_i x_i \le 1. \end{cases}$$

For $n \geq 4$, the relation

$$\mathbb{R} = \left\{ \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right\}$$

absorbs $\{0,1\}^2$ with respect to t_2^n, but does not contain the diagonal element $(0,0)$. However, $\mathbb{R}$ *does* intersect the diagonal at $(1,1)$. In the next section we will see that this weaker claim generalizes: if $\mathbb{R} \leq_{sd} \mathbb{A} \times \mathbb{A}$ absorbs $\Delta_\mathbb{A}$, then $\mathbb{R} \cap \Delta_\mathbb{A} \neq \emptyset$.

3.7 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 $\mathbb{R}'$ of binary relations $\mathbb{R}$ retain some of the connectivity properties of the original relations $\mathbb{R}$.

Definition 3.7.1. A subalgebra $\mathbb{B} \leq \mathbb{A}$ *absorbs* $\mathbb{A}$ with respect to an idempotent term t if

$$t(\mathbb{B}, ..., \mathbb{B}, \mathbb{A}, \mathbb{B}, ..., \mathbb{B}) \subseteq \mathbb{B}$$

for each possible location of $\mathbb{A}$. We just say that $\mathbb{B}$ absorbs $\mathbb{A}$, written $\mathbb{B} \lhd \mathbb{A}$, if there exists some idempotent term t such that $\mathbb{B}$ absorbs $\mathbb{A}$ 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 $\mathbb{B} \lhd \mathbb{A}$ is that there exists a common term t which witnesses all absorption within any finite collection of pairs $\mathbb{B}_i \lhd \mathbb{A}_i$.

Proposition 3.7.2. *If $\mathbb{B}_1 \lhd \mathbb{A}_1$ with respect to t_1 and $\mathbb{B}_2 \lhd \mathbb{A}_2$ with respect to t_2, then each $\mathbb{B}_i \lhd \mathbb{A}_i$ with respect to the star composition $t_1 * t_2$ (see Definition 1.5.3). If $\mathbb{A}_1 = \mathbb{B}_2$, then $\mathbb{B}_1 \lhd \mathbb{A}_2$ with respect to $t_1 * t_2$.*

Corollary 3.7.3. *A finite algebra $\mathbb{A}$ has a near-unanimity term iff for all $a \in \mathbb{A}$, the singleton $\{a\}$ absorbs $\mathbb{A}$.*

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

$$d_i(x, y, z) := t(\underbrace{x, ..., x}_{n-i}, y, \underbrace{z, ..., z}_{i-1}),$$

166

with $d_0(x, y, z) \coloneqq x$ and $d_{n+1}(x, y, z) \coloneqq z$, so that the d_i satisfy the system of identities

$$d_0(x, y, z) \approx x,$$
$$d_i(x, y, y) \approx d_{i+1}(x, x, y),$$
$$d_{n+1}(x, y, z) \approx z.$$

If $\mathbb{B}$ absorbs $\mathbb{A}$ with respect to the term t, then we will additionally have

$$d_i(\mathbb{B}, \mathbb{A}, \mathbb{B}) \subseteq \mathbb{B}$$

for all i.

Definition 3.7.4. A *Jónsson absorption chain* is a sequence of ternary terms $d_1, ..., d_n$ 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 $\mathbb{B}$ *Jónsson absorbs* $\mathbb{A}$ with respect to the Jónsson chain $d_1, ..., d_n$ if for each $i \in [n]$ we have

$$d_i(\mathbb{B}, \mathbb{A}, \mathbb{B}) \subseteq \mathbb{B}.$$

If $\mathbb{B}$ Jónsson absorbs $\mathbb{A}$ with respect to some Jónsson chain, then we write $\mathbb{B} \lhd_J \mathbb{A}$.

Proposition 3.7.5. *If $\mathbb{B} \lhd \mathbb{A}$, then $\mathbb{B} \lhd_J \mathbb{A}$.*

As with absorption, we can witness several instances of Jónsson absorption simultaneously with a single Jónsson absorption chain $d_1, ..., d_n$.

Proposition 3.7.6. *If $\mathbb{B}_1 \lhd_J \mathbb{A}_1$ with respect to $d_1, ..., d_m$ and $\mathbb{B}_2 \lhd_J \mathbb{A}_2$ with respect to $e_1, ..., e_n$, then the sequence of terms $f_1, ..., f_{mn}$ defined by*

$$f_{n(i-1)+j} \coloneqq d_i(x, e_j(x, y, z), z)$$

is a Jónsson absorption chain which witnesses both $\mathbb{B}_1 \lhd_J \mathbb{A}_1$ and $\mathbb{B}_2 \lhd_J \mathbb{A}_2$. If $\mathbb{A}_1 = \mathbb{B}_2$, then $\mathbb{B}_1 \lhd_J \mathbb{A}_2$ with respect to $f_1, ..., f_{mn}$.

Corollary 3.7.7. *A finite algebra $\mathbb{A}$ generates a congruence distributive variety iff for all $a \in \mathbb{A}$, the singleton $\{a\}$ Jónsson absorbs $\mathbb{A}$.*

Proof. A Jónsson absorbing chain which witnesses $\{a\} \lhd_J \mathbb{A}$ for all $a \in \mathbb{A}$ is the same as a sequence of terms $d_1, ..., d_m$ 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, $d_1, ..., d_m$ are a sequence of directed Jónsson terms. By Theorem A.4.8, a variety is congruence distributive iff it has directed Jónsson terms. $\square$

Example 3.7.1. If $\mathbb{A} = (A, s)$ is a 2-semilattice, then $\mathbb{B} \lhd_J \mathbb{A}$ iff $s(\mathbb{A}, \mathbb{B}) = s(\mathbb{B}, \mathbb{A}) \subseteq \mathbb{B}$, that is, iff $\mathbb{B} \lhd_{str} \mathbb{A}$.

Example 3.7.2. If $\mathbb{A}$ is abelian, then $\mathbb{A}$ has no Jónsson absorbing singleton subalgebras. To see this, note that if $\mathbb{A}$ is abelian, then for any Jónsson chain $d_1, ..., d_n$ witnessing $\{b\} \lhd_J \mathbb{A}$ and any $a \neq b$, we have $d_1(b, b, a) = b$, and then by induction we have

$$d_i(b, \boxed{b}, a) = b = d_i(b, \boxed{b}, b) \implies d_i(b, \boxed{a}, a) = d_i(b, \boxed{a}, b) = b \implies d_{i+1}(b, b, a) = d_i(b, a, a) = b,$$

so $a = d_n(b, a, a) = b$, a contradiction.

In particular, no affine algebra $\mathbb{A}$ has any proper Jónsson absorbing subalgebra $\mathbb{B}$, because we can apply the above argument to the quotient $\mathbb{A}/\theta_{\mathbb{B}}$, where $\theta_{\mathbb{B}}$ is the congruence of $\mathbb{A}$ which has $\mathbb{B}$ as a congruence class.

Example 3.7.3. Suppose $\mathbb{B}$ is an ideal of a majority algebra $\mathbb{A} = (A, m)$. Then $\mathbb{B} \lhd_J \mathbb{A}$ with respect to the Jónsson absorption chain $d_1(x, y, z) = m(x, y, z)$ (of length 1):

$$m(x, x, y) \approx x,$$
$$m(x, y, y) \approx y,$$
$$m(\mathbb{B}, \mathbb{A}, \mathbb{B}) \subseteq \mathbb{B}.$$

In fact, the converse holds: if $\mathbb{B} \lhd_J \mathbb{A}$, then there must be a majority term $m' \in \mathrm{Clo}(m)$ such that $m'(\mathbb{B}, \mathbb{A}, \mathbb{B}) \subseteq \mathbb{B}$. This follows from the fact that every ternary term in a majority algebra is either a projection or another majority operation.

If $\mathbb{A}$ generates a locally finite variety, then by applying the construction of Proposition 3.7.6 iteratively to all the majority operations in $\mathrm{Clo}(m)$, we can find a single majority term $\hat{m} \in \mathrm{Clo}(m)$ such that for any $\mathbb{C} \leq \mathbb{B} \in HSP(\mathbb{A})$ we have

$$\mathbb{C} \lhd_J \mathbb{B} \iff \hat{m}(\mathbb{C}, \mathbb{B}, \mathbb{C}) \subseteq \mathbb{C}.$$

As we will see later, for finite majority algebras $\mathbb{B} \lhd_J \mathbb{A}$ implies that $\mathbb{B} \lhd \mathbb{A}$ - possibly with respect to a term of very high arity (for instance, in the case where $\mathbb{A}$ is the dual discriminator algebra from Example 1.6.5 and $|\mathbb{B}| = |\mathbb{A}| - 1$, the minimal arity of a term t which witnesses $\mathbb{B} \lhd \mathbb{A}$ is $|\mathbb{A}| + 1$). So ideals of finite majority algebras are actually the same thing as absorbing subalgebras!

Remark 3.7.1. If we define a concept called *ideal absorption* by $\mathbb{B} \lhd_I \mathbb{A}$ when there is a ternary term d such that $d(x, x, y) \approx x \approx d(y, x, x)$ and $d(\mathbb{B}, \mathbb{A}, \mathbb{B}) \subseteq \mathbb{B}$, then all of the results about ideals of majority algebras generalize. I don't know any applications of this idea outside the context of majority algebras.

Like ideals of majority algebras, absorbing subalgebras play nice with primitive positive formulas.

Proposition 3.7.8. *Suppose that a relation $\mathbb{R}$ is defined by a primitive positive formula Φ involving the relations $\mathbb{R}_1, ..., \mathbb{R}_k$. If we replace each $\mathbb{R}_i$ with an absorbing subalgebra $\mathbb{R}_i' \lhd \mathbb{R}_i$ to make a primitive positive formula Φ', then the relation $\mathbb{R}'$ which is defined by Φ' is an absorbing subalgebra of $\mathbb{R}$. The same is true with "absorbing" replaced by "Jónsson absorbing".*

Proof. Since only finitely many relations $\mathbb{R}_i$ show up in Φ, we can find a single absorbing term (or Jónsson chain) which witnesses all absorptions $\mathbb{R}_i' \lhd \mathbb{R}_i$ (or $\mathbb{R}_i' \lhd_J \mathbb{R}_i$) simultaneously. From here the proof is similar to the proof in the case of ideals of majority algebras. $\square$

Now we will illustrate how Jónsson absorption is used, by proving a few connectivity results. Recall that every binary relation $\mathbb{R} \leq \mathbb{A} \times \mathbb{A}$ can be visualized as a graph in two different ways: we can either think of $\mathbb{R}$ as a bipartite graph on the disjoint union $\mathbb{A} \sqcup \mathbb{A}$, or we can think of $\mathbb{R}$ as a directed graph on $\mathbb{A}$. The next result is perhaps the most crucial.

Theorem 3.7.9 (Absorbing directed paths [15]). *If $\mathbb{S}, \mathbb{R} \leq \mathbb{A} \times \mathbb{A}$ are binary relations with $\mathbb{S} \lhd_J \mathbb{R}$, and $a, b \in \mathbb{A}$ satisfy*

- *$(a, a), (b, b) \in \mathbb{S}$, and*

- *$(a, b) \in \mathbb{R}$,*

then if we think of $\mathbb{S}$ as a directed graph on $\mathbb{A}$, there is a directed path from a to b in $\mathbb{S}$, that is, $(a, b) \in \mathbb{S}^{\circ n}$ for some n.

Proof. Suppose $\mathbb{S} \lhd_J \mathbb{R}$ with respect to the Jónsson chain $d_1, ..., d_n$. Then for each i we have

$$\begin{bmatrix} d_i(a, a, b) \\ d_{i+1}(a, a, b) \end{bmatrix} = \begin{bmatrix} d_i(a, a, b) \\ d_i(a, b, b) \end{bmatrix} = d_i \left(\begin{bmatrix} a \\ a \end{bmatrix}, \begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} b \\ b \end{bmatrix} \right) \in d_i(\mathbb{S}, \mathbb{R}, \mathbb{S}) \subseteq \mathbb{S}.$$

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 $(a, b) \in \mathbb{S}^{\circ n}$. $\qquad \square$

Applying the above to $\mathbb{S}^{\circ m} \lhd_J \mathbb{R}^{\circ m}$ for a sufficiently large m, we get the following stronger-looking corollary.

Corollary 3.7.10. *If $\mathbb{S}, \mathbb{R} \leq \mathbb{A} \times \mathbb{A}$ have $\mathbb{S} \lhd_J \mathbb{R}$, and $a, b \in \mathbb{A}$ satisfy*

- *each of a, b is contained in a directed cycle of the digraph $\mathbb{S}$, and*

- *there is a directed path from a to b in the digraph $\mathbb{R}$,*

then there is a directed path from a to b in the digraph $\mathbb{S}$.

For the sake of applying the previous result, it is useful to keep in mind the following basic fact about finite directed graphs.

Proposition 3.7.11. *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, $(a', a'), (a', a) \in R^{\circ n}$).*

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 $\varphi^{\circ 2n} = \varphi^{\circ n}$ by the finiteness of A: in fact, we may take $n = \mathrm{lcm}\{1, ..., |A|\}$. $\qquad \square$

In the next result, we think of binary relations as bipartite graphs. Recall that the *linked components* of a binary relation $\mathbb{R} \leq \mathbb{A} \times \mathbb{B}$ are the connected components of $\mathbb{R}$ 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 $\mathbb{R}$.

Theorem 3.7.12 (Absorbing linked components [15]). *If $\mathbb{S}, \mathbb{R} \leq \mathbb{A} \times \mathbb{B}$ are binary relations with $\mathbb{S} \lhd_J \mathbb{R}$, and $a, b \in \pi_1(\mathbb{S})$ are in the same linked component of $\mathbb{R}$, then a, b are in the same linked component of $\mathbb{S}$.*

Proof. If a, b are linked in $\mathbb{R}$, then there is some m such that $(a, b) \in (\mathbb{R} \circ \mathbb{R}^-)^{\circ m}$. Since $(\mathbb{S} \circ \mathbb{S}^-)^{\circ m} \lhd_J (\mathbb{R} \circ \mathbb{R}^-)^{\circ m}$ and $(a, a), (b, b) \in \mathbb{S} \circ \mathbb{S}^-$ by $a, b \in \pi_1(\mathbb{S})$, we can apply Theorem 3.7.9 to see that there is some n such that $(a, b) \in (\mathbb{S} \circ \mathbb{S}^-)^{\circ mn}$. Thus a, b are in the same linked component of $\mathbb{S}$. $\qquad\square$

The next result is an analogue of Theorem 3.3.9 and Theorem 3.6.10 for Jónsson absorption.

Theorem 3.7.13 (Loop Lemma, finite absorbing case [13]). *If $\mathbb{R} \leq_{sd} \mathbb{A} \times \mathbb{A}$ is subdirect, $\mathbb{A}$ is finite, and $\mathbb{R}$ Jónsson absorbs the diagonal $\Delta_{\mathbb{A}}$, then $\mathbb{R} \cap \Delta_{\mathbb{A}} \neq \emptyset$.*

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 $\mathbb{R} \cap (\mathbb{B} \times \mathbb{B})$ subdirect. Then $\mathbb{R} \cap (\mathbb{B} \times \mathbb{B})$ will Jónsson absorb $\Delta_{\mathbb{B}}$, and we can show by induction that $\mathbb{R} \cap \Delta_{\mathbb{B}} \neq \emptyset$.

Let b be any element of $\mathbb{A}$, and define a sequence of subalgebras $\mathbb{B}_i$ by $\mathbb{B}_0 = \{b\}$, $\mathbb{B}_{i+1} = \mathbb{B}_i + \mathbb{R}$, i.e. $\mathbb{B}_{i+1} = \pi_2(\mathbb{R} \cap (\mathbb{B}_i \times \mathbb{A}))$. If there is any i such that $\mathbb{B}_i \neq \mathbb{A}$ but $\mathbb{B}_{i+1} = \mathbb{A}$, then for every $\mathbb{C} \leq \mathbb{A}$ we have $(\mathbb{C} + \mathbb{R}^-) \cap \mathbb{B}_i \neq \emptyset$, so by the finiteness of $\mathbb{B}_i$ we may take

$$\mathbb{B} = \bigcap_{k \geq 0} \mathbb{B}_i + (\mathbb{R}^-)^{\circ k} = \{a \mid \exists a_0, a_1, \ldots \in \mathbb{B}_i \text{ s.t. } a_0 = a \text{ and } \forall j \ (a_j, a_{j+1}) \in \mathbb{R}\}.$$

Otherwise, each $\mathbb{B}_i \neq \mathbb{A}$, and by the finiteness of $\mathbb{A}$ there must be some m, n such that $\mathbb{B}_m = \mathbb{B}_{m+n}$. We will show that in this case we have $\mathbb{B}_{m+i} = \mathbb{B}_m$ for each i, so we may take $\mathbb{B} = \mathbb{B}_m$.

Consider any directed cycle $a_0, \ldots, a_{kn} = a_0$ 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 $(a_0, a_0), (a_i, a_i) \in \mathbb{R}^{\circ kn}$, that $\mathbb{R}^{\circ kn}$ Jónsson absorbs $(\mathbb{R} \cup \Delta_{\mathbb{A}})^{\circ kn}$, and that $(a_0, a_i) \in \mathbb{R}^{\circ i} \subseteq (\mathbb{R} \cup \Delta_{\mathbb{A}})^{\circ kn}$. Thus by Theorem 3.7.9 there is some l such that $(a_0, a_i) \in \mathbb{R}^{\circ ln}$, and since $\mathbb{B}_m + \mathbb{R}^{\circ ln} = \mathbb{B}_{m+ln} = \mathbb{B}_m$, we see that $a_i \in \mathbb{B}_m$.

Since $\mathbb{B}_{m+i} + \mathbb{R}^{\circ n} = \mathbb{B}_{m+i}$ 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 $\mathbb{R}^{\circ n}$ and a directed path of $\mathbb{R}^{\circ n}$ 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 $\mathbb{B}_m = \mathbb{B}_{m+i}$. $\qquad\square$

Corollary 3.7.14. *If $\mathbb{R} \leq_{sd} \mathbb{A} \times \mathbb{A}$ is subdirect, $\mathbb{A}$ is finite and has no proper absorbing subalgebra, and $\mathbb{R}$ absorbs the diagonal $\Delta_{\mathbb{A}}$, then $\Delta_{\mathbb{A}} \subseteq \mathbb{R}$.*

Proof. Since $\mathbb{R} \cap \Delta_{\mathbb{A}} \neq \emptyset$ is an absorbing subalgebra of $\Delta_{\mathbb{A}}$ and $\Delta_{\mathbb{A}} \cong \mathbb{A}$ has no proper absorbing subalgebra, we must have $\mathbb{R} \cap \Delta_{\mathbb{A}} = \Delta_{\mathbb{A}}$. $\qquad\square$

Definition 3.7.15. We say that $\mathbb{B}$ is a *minimal absorbing subalgebra* of $\mathbb{A}$, written $\mathbb{B} \lll \mathbb{A}$, if $\mathbb{B} \lhd \mathbb{A}$ and $\mathbb{B}$ has no proper absorbing subalgebra.

Proposition 3.7.16. *Every finite absorbing subalgebra of $\mathbb{A}$ contains a minimal absorbing subalgebra of $\mathbb{A}$, and any pair of distinct minimal absorbing subalgebras of $\mathbb{A}$ are disjoint.*

Proof. This follows from the fact that $\mathbb{C} \lhd \mathbb{B} \lhd \mathbb{A}$ implies $\mathbb{C} \lhd \mathbb{A}$, and the fact that the intersection of any pair of absorbing subalgebras is an absorbing subalgebra. $\qquad\square$

Theorem 3.7.17. *Suppose that $\mathbf{X}$ is an arc-consistent instance of a CSP, and suppose that for each variable domain $\mathbb{A}_v$ there is a minimal absorbing subalgebra $\mathbb{A}'_v \lll \mathbb{A}_v$ such that the reduced instance $\mathbf{X}'$ with variable domains replaced by $\mathbb{A}'_v$ and relations $\mathbb{R} \leq_{sd} \mathbb{A}_{v_1} \times \cdots \times \mathbb{A}_{v_n}$ replaced by $\mathbb{R}' = \mathbb{R} \cap (\mathbb{A}'_{v_1} \times \cdots \times \mathbb{A}'_{v_n})$ is arc-consistent.*

Then for any path p in $\mathbf{X}$ from a variable v to itself such that $\mathbb{P}_p \supseteq \Delta_{\mathbb{A}_v}$, the corresponding path p' of $\mathbf{X}'$ has $\mathbb{P}_{p'} \supseteq \Delta_{\mathbb{A}'_v}$. In particular, if $\mathbf{X}$ is cycle-consistent then so is $\mathbf{X}'$.

Proof. Note that $\mathbb{P}_{p'}$ is an absorbing subalgebra of $\mathbb{P}_p$, so $\mathbb{P}_{p'}$ absorbs $\Delta_{\mathbb{A}'_v}$. Since $\mathbb{P}_{p'}$ is subdirect in $\mathbb{A}'_v \times \mathbb{A}'_v$ by the arc-consistency of $\mathbf{X}'$ and $\mathbb{A}'_v$ has no proper absorbing subalgebra, we may apply Corollary 3.7.14 to see that $\mathbb{P}_{p'} \supseteq \Delta_{\mathbb{A}'_v}$. $\qquad\square$

Later we will show that any cycle-consistent instance $\mathbf{X}$ has an arc-consistent reduction $\mathbf{X}'$ where all variable domains are replaced by minimal absorbing subalgebras, which will set us up to apply Theorem 3.7.17. The argument strategy will be fairly generic, not using any specific properties of absorbing subalgebras other than Theorem 3.7.9 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 $\mathbb{P}_{j(p+q)+p} \supseteq \Delta_{\mathbb{A}_v}$.

3.7.1 Local criterion for Jónsson absorption

Since a finite algebra $\mathbb{A}$ has bounded strict width iff every singleton is an absorbing subalgebra of $\mathbb{A}$, we'd like to have a way to test whether a given subalgebra $\mathbb{B} \leq \mathbb{A}$ is an absorbing subalgebra. Since the arity of a potential absorbing term is unbounded, we'll start with the easier problem of testing whether $\mathbb{B}$ is a *Jónsson* absorbing subalgebra, since in this case there is an obvious algorithm which will at least eventually halt: list out every possible ternary term of $\mathbb{A}$ by brute force, and make a digraph of possible Jónsson chains.

The idea behind finding a better way to test whether $\mathbb{B} \lhd_J \mathbb{A}$ is to try to find a converse to the fundamental digraph connectivity result characterizing Jónsson absorption (Theorem 3.7.9). In order to formulate the converse, we need to consider generic pairs of digraphs $\mathbb{S} \leq \mathbb{R} \leq \mathbb{C} \times \mathbb{C}$ such that

$$\mathbb{B} \lhd_J \mathbb{A} \implies \mathbb{S} \lhd_J \mathbb{R}.$$

One natural way to do this is to write $\mathbb{R}$ as the projection to the last two coordinates of a ternary relation $\mathbb{X} \leq \mathbb{A} \times \mathbb{C} \times \mathbb{C}$, and to take $\mathbb{S}$ to be the corresponding projection of $\mathbb{X} \cap (\mathbb{B} \times \mathbb{C} \times \mathbb{C})$.

Definition 3.7.18. For $\mathbb{B} \leq \mathbb{A}$ and $\mathbb{C} \in \mathcal{V}(\mathbb{A})$ all idempotent, we say that $\mathbb{A}, \mathbb{B}, \mathbb{C}$ satisfy the condition $J(\mathbb{A}, \mathbb{B}; \mathbb{C})$ if for every $a \in \mathbb{A}$, $b, b' \in \mathbb{B}$, and $c, d \in \mathbb{C}$, if we set

$$\mathbb{S} = \pi_{23}\left(\mathrm{Sg}_{\mathbb{A} \times \mathbb{C} \times \mathbb{C}}\left\{\begin{bmatrix} b \\ c \\ c \end{bmatrix}, \begin{bmatrix} a \\ c \\ d \end{bmatrix}, \begin{bmatrix} b' \\ d \\ d \end{bmatrix}\right\} \cap \begin{bmatrix} \mathbb{B} \\ \mathbb{C} \\ \mathbb{C} \end{bmatrix}\right),$$

then there is some n such that $(c, d) \in \mathbb{S}^{\circ n}$.

We will show that the condition $J(\mathbb{A}, \mathbb{B}; \mathbb{A})$ is equivalent to $\mathbb{B} \lhd_J \mathbb{A}$, following the strategy of [15]. Note that Theorem 3.7.9 proves one direction of the equivalence, so we just need to prove that $J(\mathbb{A}, \mathbb{B}; \mathbb{A}) \implies \mathbb{B} \lhd_J \mathbb{A}$. The strategy will be to use induction to show that $J(\mathbb{A}^m, \mathbb{B}^m; \mathbb{A}^n)$ holds for all m, n, and then to take $m = |\mathbb{A}||\mathbb{B}|^2, n = |\mathbb{A}|^2$ to show that a certain directed path exists between binary terms in the free algebra on two generators, which will correspond to a Jónsson absorption chain. Before diving into the details, we will outline how this criterion could be used to test whether $\mathbb{B} \lhd_J \mathbb{A}$.

Note that if $\mathbb{A}$ is given in terms of tables for its basic operations, then the condition $J(\mathbb{A}, \mathbb{B}; \mathbb{A})$ can be tested in time polynomial in $|\mathbb{A}|, |\mathbb{B}|$ (with the degree of the polynomial depending on the arities of the basic operations), since the total number of tuples a, b, b', c, d is $|\mathbb{A}|^3|\mathbb{B}|^2$, computing $\mathbb{S}$

requires us to compute a ternary relation of size at most $|\mathbb{A}|^3$, and we only need to check whether $(c, d) \in \mathbb{S}^{\circ n}$ for $n \leq |\mathbb{A}|$.

If $\mathbb{A}$ is instead given in terms of a list of basic relations, then testing the condition $J(\mathbb{A}, \mathbb{B}; \mathbb{A})$ can be reduced to solving polynomially many polynomially large constraint satisfaction problems over the domain $\mathbb{A}$ - so in particular if $\mathrm{CSP}(\mathbb{A})$ can be solved in polynomial time, then we can test $J(\mathbb{A}, \mathbb{B}; \mathbb{A})$ in polynomial time. To see this, note that in order to test whether a given edge (e, f) is an element of $\mathbb{S}$, we just need to test whether $\mathbb{A}$ has a ternary polymorphism f such that

$$f(b, a, b') \in \mathbb{B},$$
$$f(c, c, d) = e,$$
$$f(c, d, d) = f,$$

and the set of ternary polymorphisms $f \in \mathcal{F}_{\mathbb{A}}(x, y, z) \leq \mathbb{A}^{\mathbb{A}^3}$ can be described by a primitive positive formula involving only $|\mathbb{A}|^3$ variables.

Lemma 3.7.19. *If $J(\mathbb{A}_1, \mathbb{B}_1; \mathbb{C})$ and $J(\mathbb{A}_2, \mathbb{B}_2; \mathbb{C})$ both hold, then so does $J(\mathbb{A}_1 \times \mathbb{A}_2, \mathbb{B}_1 \times \mathbb{B}_2; \mathbb{C})$.*

Proof. Suppose $a = (a_1, a_2) \in \mathbb{A}_1 \times \mathbb{A}_2$ and $b = (b_1, b_2), b' = (b'_1, b'_2) \in \mathbb{B}_1 \times \mathbb{B}_2$, $c, d \in \mathbb{C}$. Define $\mathbb{S}, \mathbb{R} \leq \mathbb{C} \times \mathbb{C}$ as usual, and define an intermediate digraph $\mathbb{S}_1$, where instead of restricting to $\mathbb{B}_1 \times \mathbb{B}_2$, we restrict to $\mathbb{B}_1 \times \mathbb{A}_2$ instead - so for the purposes of computing $\mathbb{S}_1$, we can ignore the $\mathbb{A}_2$ components. Then by $J(\mathbb{A}_1, \mathbb{B}_1; \mathbb{C})$, from $(c, d) \in \mathbb{R}$ we see that there is a directed path from c to d in $\mathbb{S}_1$.

To finish, we just need to check that for each $(e, f) \in \mathbb{S}_1$, there is a directed path from e to f in $\mathbb{S}$. Note that $(e, f) \in \mathbb{S}_1$ means that there are some $b''_1 \in \mathbb{B}_1, a''_2 \in \mathbb{A}_2$ such that

$$\begin{bmatrix} (b''_1, a''_2) \\ e \\ f \end{bmatrix} \in \mathrm{Sg} \left\{ \begin{bmatrix} (b_1, b_2) \\ c \\ c \end{bmatrix}, \begin{bmatrix} (a_1, a_2) \\ c \\ d \end{bmatrix}, \begin{bmatrix} (b'_1, b'_2) \\ d \\ d \end{bmatrix} \right\}.$$

Then from $e, f \in \mathrm{Sg}\{c, d\}$, there are some $(b'''_1, b'''_2) \in \mathbb{B}_1 \times \mathbb{B}_2$ with

$$\begin{bmatrix} (b'''_1, b'''_2) \\ e \\ e \end{bmatrix} \in \mathrm{Sg} \left\{ \begin{bmatrix} (b_1, b_2) \\ c \\ c \end{bmatrix}, \begin{bmatrix} (b'_1, b'_2) \\ d \\ d \end{bmatrix} \right\},$$

and similarly for (f, f), so we just need to check that

$$\pi_{23} \left(\mathrm{Sg} \left\{ \begin{bmatrix} (b'''_1, b'''_2) \\ e \\ e \end{bmatrix}, \begin{bmatrix} (b''_1, a''_2) \\ e \\ f \end{bmatrix}, \begin{bmatrix} (b''''_1, b''''_2) \\ f \\ f \end{bmatrix} \right\} \cap \begin{bmatrix} \mathbb{B}_1 \times \mathbb{B}_2 \\ \mathbb{C} \\ \mathbb{C} \end{bmatrix} \right)$$

contains a directed path from e to f. But now we can ignore the $\mathbb{B}_1$ component, so this follows from $J(\mathbb{A}_2, \mathbb{B}_2; \mathbb{C})$. $\square$

Lemma 3.7.20. *If $J(\mathbb{A}, \mathbb{B}; \mathbb{C}_1)$ and $J(\mathbb{A}, \mathbb{B}; \mathbb{C}_2)$ both hold and $\mathbb{C}_1, \mathbb{C}_2$ are finite and idempotent, then $J(\mathbb{A}, \mathbb{B}; \mathbb{C}_1 \times \mathbb{C}_2)$ holds as well.*

Proof. Suppose not. Choose $c = (c_1, c_2), d = (d_1, d_2) \in \mathbb{C}_1 \times \mathbb{C}_2$ such that $\mathrm{Sg}\{c, d\}$ is minimal among all pairs such that there exist $a \in \mathbb{A}, b, b' \in \mathbb{B}$ so that the associated digraph $\mathbb{S}$ has no directed path from c to d.

Ignoring the $\mathbb{C}_2$ components, we can apply $J(\mathbb{A}, \mathbb{B}; \mathbb{C}_1)$ to find a sequence of edges $(e^i, f^{i+1}) \in \mathbb{S}$ such that $f_1^i = e_1^i$ for each $i \leq n$, $c = f^1$, and $e^n = d$. Since we assumed that there is no directed path from c to $e^n = d$, we can consider the first i such that there is no directed path from c to e^i.

Since $e^i \in \mathrm{Sg}\{c, d\}$, we have

$$\begin{bmatrix} c \\ e^i \end{bmatrix} \in \mathrm{Sg} \left\{ \begin{bmatrix} c \\ c \end{bmatrix}, \begin{bmatrix} c \\ d \end{bmatrix}, \begin{bmatrix} d \\ d \end{bmatrix} \right\} = \mathbb{R},$$

and since there is no directed path from c to e^i in $\mathbb{S}$, we see that we must have $\mathrm{Sg}\{c, e^i\} = \mathrm{Sg}\{c, d\}$ by our minimality assumption, so in particular we have $f^i \in \mathrm{Sg}\{c, e^i\}$. Thus we have

$$\begin{bmatrix} f^i \\ e^i \end{bmatrix} \in \mathrm{Sg} \left\{ \begin{bmatrix} c \\ e^i \end{bmatrix}, \begin{bmatrix} e^i \\ e^i \end{bmatrix} \right\} \subseteq \mathbb{R}.$$

By the choice of i there is a path from c to f^i in $\mathbb{S}$ (passing through e^{i-1} if $i > 1$). To get a contradiction, we just need to show that there is a directed path from f^i to e^i in $\mathbb{S}$. Since $(e^i, e^i), (f^i, f^i) \in \mathbb{S}$, there are $a' \in \mathbb{A}, b'', b''' \in \mathbb{B}$ such that

$$\pi_{23} \left(\mathrm{Sg} \left\{ \begin{bmatrix} b'' \\ f^i \\ f^i \end{bmatrix}, \begin{bmatrix} a' \\ f^i \\ e^i \end{bmatrix}, \begin{bmatrix} b''' \\ e^i \\ e^i \end{bmatrix} \right\} \cap \begin{bmatrix} \mathbb{B} \\ \mathbb{C}_1 \times \mathbb{C}_2 \\ \mathbb{C}_1 \times \mathbb{C}_2 \end{bmatrix} \right) \subseteq \mathbb{S}.$$

Since $f_1^i = e_1^i$, we can ignore the $\mathbb{C}_1$ components in the above, so by $J(\mathbb{A}, \mathbb{B}; \mathbb{C}_2)$ there is a directed path from f^i to e^i in $\mathbb{S}$. $\square$

Theorem 3.7.21 (Local criterion for Jónsson absorption [15]). *If $\mathbb{B} \leq \mathbb{A}$ are finite and idempotent, then $\mathbb{B} \lhd_J \mathbb{A}$ if and only if $J(\mathbb{A}, \mathbb{B}; \mathbb{A})$ holds.*

Proof. By the previous two lemmas, $J(\mathbb{A}^m, \mathbb{B}^m; \mathbb{A}^n)$ holds for $m = \mathbb{B} \times \mathbb{A} \times \mathbb{B}$ and $n = \mathbb{A} \times \mathbb{A}$. There is a natural map $\Phi : \mathcal{F}_{\mathbb{A}}(x, y, z) \to \mathbb{A}^{\mathbb{B} \times \mathbb{A} \times \mathbb{B}}$ and a pair of natural maps $\Psi_1, \Psi_2 : \mathcal{F}_{\mathbb{A}}(x, y, z) \to \mathbb{A}^{\mathbb{A} \times \mathbb{A}}$: the first takes f to the restriction $f|_{\mathbb{B} \times \mathbb{A} \times \mathbb{B}}$, the other two take f to the functions $f(x, x, y), f(x, y, y)$.

Then we can apply $J(\mathbb{A}^m, \mathbb{B}^m; \mathbb{A}^n)$ with $a = \Phi(\pi_2)$, $b = \Phi(\pi_2), b' = \Phi(\pi_3)$, $c = \Psi_i(\pi_1) = \Psi_1(\pi_2), d = \Psi_i(\pi_3) = \Psi_2(\pi_2)$. If we set

$$\mathbb{S} = \pi_{23} \left(\mathrm{Sg} \left\{ \begin{bmatrix} x|_{x,z \in \mathbb{B}} \\ x \\ x \end{bmatrix}, \begin{bmatrix} y|_{x,z \in \mathbb{B}} \\ x \\ y \end{bmatrix}, \begin{bmatrix} z|_{x,z \in \mathbb{B}} \\ y \\ y \end{bmatrix} \right\} \cap \begin{bmatrix} \mathbb{B}^m \\ \mathbb{A}^n \\ \mathbb{A}^n \end{bmatrix} \right),$$

then the inner ternary subalgebra is exactly $\mathrm{Im}(\Phi, \Psi_1, \Psi_2)$, so $\mathbb{S}$ is exactly the digraph of pairs of binary terms $g(x, y), h(x, y)$ such that there is some ternary term $f(x, y, z)$ satisfying

$$f(\mathbb{B}, \mathbb{A}, \mathbb{B}) \subseteq \mathbb{B},$$
$$f(x, x, y) \approx g(x, y),$$
$$f(x, y, y) \approx h(x, y).$$

The condition $J(\mathbb{A}^m, \mathbb{B}^m; \mathbb{A}^n)$ says that this digraph contains a path from the term x to the term y, which is the same as a Jónsson absorption chain for $\mathbb{B} \lhd_J \mathbb{A}$. $\square$

173

Note that the same argument shows that it is enough to check $J(\mathbb{A}, \mathbb{B}; \mathbb{C}_i)$ for any collection of algebras $\mathbb{C}_1, ..., \mathbb{C}_n$ generating a variety $\mathcal{V}$ such that $\mathcal{F}_{\mathbb{A}}(x, y) = \mathcal{F}_{\mathcal{V}}(x, y)$. In cases where $\mathcal{F}_{\mathbb{A}}(x, y)$ is small the criterion becomes especially nice.

Corollary 3.7.22. *If $\mathbb{A} = (A, m)$ is a majority algebra, then $\mathbb{B} \lhd_J \mathbb{A}$ iff there do* not *exist $a \in \mathbb{A}$ and $b, c \in \mathbb{B}$ such that*

- *a, b, c are distinct,*

- *$\mathrm{Sg}_{\mathbb{A}}\{a, b, c\} \cap \mathbb{B} = \{b, c\}$,*

- *the partitions $\{\{b\}, \mathrm{Sg}\{a, b, c\} \setminus \{b\}\}$ and $\{\{c\}, \mathrm{Sg}\{a, b, c\} \setminus \{c\}\}$ of $\mathrm{Sg}\{a, b, c\}$ correspond to congruences θ_b, θ_c on $\mathrm{Sg}\{a, b, c\}$.*

The third bullet point can also be stated in the equivalent form: $\mathrm{Sg}\{a, b, c\}/(\theta_b \wedge \theta_c)$ is isomorphic to the three element median algebra, with median element $a/(\theta_b \wedge \theta_c) = \mathrm{Sg}\{a, b, c\} \setminus \{b, c\}$.

3.8 Absorption and $\mathbb{B}$-essential relations

In this section we'll give a relational description of absorption, as well as a first simplification via Ramsey theory. The relational description is a generalization of the way relations over near-unanimity algebras decompose.

Definition 3.8.1. Suppose $\mathbb{B} \le \mathbb{A}$. We say that a relation $\mathbb{R} \le \mathbb{A}^m$ is $\mathbb{B}$-*essential* if for every $1 \le i \le n$ we have

$$\mathbb{R} \cap (\mathbb{B}^{i-1} \times \mathbb{A} \times \mathbb{B}^{n-i}) \neq \emptyset,$$

but

$$\mathbb{R} \cap \mathbb{B}^n = \emptyset.$$

More generally, if $\mathbb{B}_i \le \mathbb{A}_i$ for all i, then we say that $\mathbb{R} \le \mathbb{A}_1 \times \cdots \times \mathbb{A}_m$ is $(\mathbb{B}_1, ..., \mathbb{B}_m)$-*essential* if

$$\mathbb{R} \cap (\mathbb{B}_1 \times \cdots \times \mathbb{B}_{i-1} \times \mathbb{A}_i \times \mathbb{B}_{i+1} \times \cdots \times \mathbb{B}_m) \neq \emptyset$$

for each i, but

$$\mathbb{R} \cap (\mathbb{B}_1 \times \cdots \times \mathbb{B}_m) = \emptyset.$$

Proposition 3.8.2. *If $\mathbb{R} \le \mathbb{A}^m$ is $\mathbb{B}$-essential, then so is*

$$\pi_{[m-1]}(\mathbb{R} \cap (\mathbb{A}^{m-1} \times \mathbb{B})).$$

In particular, if there is a $\mathbb{B}$-essential relation of some arity, then there are $\mathbb{B}$-essential relations of all smaller arities.

Proposition 3.8.3. *If $\mathbb{B}$ absorbs $\mathbb{A}$ with respect to a term t of arity m, then there are no $\mathbb{B}$-essential relations $\mathbb{R} \le \mathbb{A}^m$ of arity m.*

Proof. Suppose for contradiction that $\mathbb{R} \leq \mathbb{A}^m$ is $\mathbb{B}$-essential, and let $b_{ij} \in \mathbb{B}, a_i \in \mathbb{A}$ be such that

$$
\begin{bmatrix} a_1 \\ b_{21} \\ \vdots \\ b_{m1} \end{bmatrix}, \begin{bmatrix} b_{12} \\ a_2 \\ \vdots \\ b_{m2} \end{bmatrix}, ..., \begin{bmatrix} b_{1m} \\ b_{2m} \\ \vdots \\ a_m \end{bmatrix} \in \mathbb{R}.
$$

Then if we apply t, we have

$$
t\left(\begin{bmatrix} a_1 & b_{12} & \cdots & b_{1m} \\ b_{21} & a_2 & \cdots & b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m1} & b_{m2} & \cdots & a_m \end{bmatrix} \right) \in \mathbb{R} \cap \mathbb{B}^m
$$

since $\mathbb{B}$ absorbs $\mathbb{A}$ with respect to t, which is a contradiction. $\qquad\square$

Our main result is the converse to the above.

Theorem 3.8.4 (Relational description of absorption [15])**.** *If $\mathbb{A}$ is finite and idempotent, then $\mathbb{B}$ absorbs $\mathbb{A}$ with respect to a term of arity m if and only if there are no $\mathbb{B}$-essential relations of arity m. In particular, we have $\mathbb{B} \lhd \mathbb{A}$ if and only if there is a bound on the arity of $\mathbb{B}$-essential relations.*

The strategy of the proof is to show that if there are no m-ary terms t which absorb $\mathbb{B}$, then the projection of the free algebra $\mathcal{F}_{\mathbb{A}}(x_1, ..., x_m) \leq \mathbb{A}^{\mathbb{A}^m}$ onto the coordinates where all but one input x_i are in $\mathbb{B}$ looks like a $\mathbb{B}$-essential relation. The arity of this projection will be much higher than m, but the set of coordinates can be naturally grouped into m parts.

Lemma 3.8.5. *If $n_1, ..., n_m \geq 1$ and $\mathbb{R} \leq \mathbb{A}^{n_1} \times \cdots \times \mathbb{A}^{n_m}$ is $(\mathbb{B}^{n_1}, ..., \mathbb{B}^{n_m})$-essential, then there is a $\mathbb{B}$-essential relation $\mathbb{R}' \leq \mathbb{A}^m$ of arity m. In fact, $\mathbb{R}'$ can be chosen to have the form*

$$
\mathbb{R}' = \pi_I \left(\mathbb{R} \cap \left(\prod_i \mathbb{C}_i \right) \right)
$$

for some $I \subseteq [n_1 + \cdots + n_m]$ with $|I| = m$ and for some choice of $\mathbb{C}_i \in \{\mathbb{A}, \mathbb{B}\}$ for each i.

Proof. We prove this by induction on $n = n_1 + \cdots + n_m$. If all $n_i = 1$, then $\mathbb{R}$ is an m-ary $\mathbb{B}$-essential relation already. Otherwise, we may assume $n_m > 1$ without loss of generality. First consider the relation

$$
\mathbb{R}_1 = \pi_{[n-1]}(\mathbb{R} \cap (\mathbb{A}^{n-1} \times \mathbb{B})) \leq \mathbb{A}^{n_1} \times \cdots \times \mathbb{A}^{n_m - 1}.
$$

We have

$$
\mathbb{R}_1 \cap (\mathbb{B}^{n_1} \times \cdots \times \mathbb{A}^{n_i} \times \cdots \times \mathbb{B}^{n_{m-1}} \times \mathbb{B}^{n_m - 1}) \neq \emptyset
$$

for each $i \neq m$, and

$$
\mathbb{R}_1 \cap \mathbb{B}^{n-1} = \emptyset,
$$

so the only way for $\mathbb{R}_1$ to fail to be $(\mathbb{B}^{n_1}, ..., \mathbb{B}^{n_{m-1}}, \mathbb{B}^{n_m - 1})$-essential is if

$$
\pi_{[n-n_m] \cup \{n\}}(\mathbb{R}) \cap \mathbb{B}^{n - n_m + 1} = \emptyset.
$$

In this case, we see that

$$
\mathbb{R}_2 = \pi_{[n-n_m] \cup \{n\}}(\mathbb{R})
$$

is a $(\mathbb{B}^{n_1}, ..., \mathbb{B}^{n_{m-1}}, \mathbb{B})$-essential relation. $\qquad\square$

Proof of Theorem 3.8.4. We just need to prove that if there is no m-ary $\mathbb{B}$-essential relation, then $\mathbb{B}$ absorbs $\mathbb{A}$ with respect to some m-ary term t. For each i, let X_i be the set of tuples $(x_1, ..., x_m) \in \mathbb{A}^m$ such that $x_j \in \mathbb{B}$ for $j \neq i$, and $x_i \in \mathbb{A} \setminus \mathbb{B}$. Consider the relation

$$\mathbb{R} = \pi_{X_1 \cup \cdots \cup X_m}(\mathcal{F}_\mathbb{A}(x_1, ..., x_m)) \leq \mathbb{A}^{X_1} \times \cdots \times \mathbb{A}^{X_m}.$$

Since $\mathcal{F}_\mathbb{A}(x_1, ..., x_m)$ contains the projection functions $\pi_i : \mathbb{A}^m \to \mathbb{A}$, by the definition of the sets X_i we have

$$\mathbb{R} \cap (\mathbb{B}^{X_1} \times \cdots \times \mathbb{A}^{X_i} \times \cdots \times \mathbb{B}^{X_m}) \neq \emptyset$$

for all i. Since there is no $\mathbb{B}$-essential relation of arity m, we see that $\mathbb{R}$ can't be $(\mathbb{B}^{X_1}, ..., \mathbb{B}^{X_m})$-essential by Lemma 3.8.5, so we must have

$$\mathbb{R} \cap (\mathbb{B}^{X_1} \times \cdots \times \mathbb{B}^{X_m}) \neq \emptyset$$

as well. Then by the definition of $\mathbb{R}$, we see that there is a term $t \in \mathcal{F}_\mathbb{A}(x_1, ..., x_m)$ which absorbs $\mathbb{B}$. $\square$

We can simplify this slightly as follows.

Corollary 3.8.6. *We have $\mathbb{B} \lhd \mathbb{A}$ with respect to an m-ary term t iff for all $b_{ij} \in \mathbb{B}, a_i \in \mathbb{A}$ we have*

$$\mathrm{Sg}_{\mathbb{A}^m} \left\{ \begin{bmatrix} a_1 \\ b_{21} \\ \vdots \\ b_{m1} \end{bmatrix}, \begin{bmatrix} b_{12} \\ a_2 \\ \vdots \\ b_{m2} \end{bmatrix}, ..., \begin{bmatrix} b_{1m} \\ b_{2m} \\ \vdots \\ a_m \end{bmatrix} \right\} \cap \mathbb{B}^m \neq \emptyset.$$

We leave the following generalization as an exercise to the reader.

Theorem 3.8.7. *If $\mathbb{A}_1, ..., \mathbb{A}_k$ are finite and idempotent, and $\mathbb{B}_i \leq \mathbb{A}_i$ for each i are such that there is no $(\mathbb{B}_{i_1}, ..., \mathbb{B}_{i_m})$-essential relation $\mathbb{R} \leq \mathbb{A}_{i_1} \times \cdots \times \mathbb{A}_{i_m}$ for any choice of $i_1, ..., i_m \in [k]$, then there is an m-ary term t such that each $\mathbb{B}_i$ absorbs $\mathbb{A}_i$ with respect to t.*

Corollary 3.8.8. *A finite idempotent algebra $\mathbb{A}$ has a near-unanimity term of arity m iff for each choice of $a_i, b_i \in \mathbb{A}$, we have*

$$\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} \in \mathrm{Sg}_{\mathbb{A}^m} \left\{ \begin{bmatrix} a_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}, \begin{bmatrix} b_1 \\ a_2 \\ \vdots \\ b_m \end{bmatrix}, ..., \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ a_m \end{bmatrix} \right\}.$$

Now we move our focus to finding a simpler characterization of $\mathbb{B} \lhd \mathbb{A}$, without restricting to terms of a particular arity. We'll use the notation $r_k(m)$ for the multicolored Ramsey number $R(m, ..., m)$ (with k copies of m), defined as the least number n such that any edge coloring of K_n with k colors must have a monochromatic copy of K_m.

Theorem 3.8.9. *If $\mathbb{A}$ is finite and idempotent, then $\mathbb{B} \lhd \mathbb{A}$ iff there do not exist $a \in \mathbb{A}$ and $b, c \in \mathbb{B}$ such that for every m, we have*

$$
\mathrm{Sg}_{\mathbb{A}^m} \left\{ \begin{bmatrix} a & b & b & \cdots & b \\ c & a & b & \cdots & b \\ c & c & a & \cdots & b \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c & c & c & \cdots & a \end{bmatrix} \right\} \cap \mathbb{B}^m = \emptyset.
$$

Proof. We just need to show that if $\mathbb{B}$ does not absorb $\mathbb{A}$, then such a, b, c exist for every m. Let $n = |\mathbb{A}|(r_{|\mathbb{B}|^2}(m) - 1) + 1$. Then since $\mathbb{B}$ doesn't absorb $\mathbb{A}$ with respect to any term of arity n, there is some collection of $a_i \in \mathbb{A}, b_{ij} \in \mathbb{B}$ such that

$$
\mathrm{Sg}_{\mathbb{A}^n} \left\{ \begin{bmatrix} a_1 & b_{12} & \cdots & b_{1n} \\ b_{21} & a_2 & \cdots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & a_n \end{bmatrix} \right\} \cap \mathbb{B}^n = \emptyset.
$$

By the pigeonhole principle, there is some a which occurs at least $n' = r_{|\mathbb{B}|^2}(m)$ times among $a_1, ..., a_n$. Suppose without loss of generality that $a_1, ..., a_{n'}$ are all equal to a. If we restrict to the rows and columns with $a_i = a$, we find that

$$
\mathrm{Sg}_{\mathbb{A}^{n'}} \left\{ \begin{bmatrix} a & b_{12} & \cdots & b_{1n'} \\ b_{21} & a & \cdots & b_{2n'} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n'1} & b_{n'2} & \cdots & a \end{bmatrix} \right\} \cap \mathbb{B}^{n'} = \emptyset.
$$

Now we color the complete graph $K_{n'}$ with $|\mathbb{B}|^2$ colors, coloring the edge $\{i, j\}$ (with $i < j$) with the color corresponding to the ordered pair (b_{ij}, b_{ji}). Then by the definition of the Ramsey number $r_{|\mathbb{B}|^2}(m)$, there is a monochromatic copy of K_m, with all edges colored by the color corresponding to some pair $(b, c) \in \mathbb{B}^2$. By restricting to the rows and columns corresponding to the vertices of this monochromatic K_m, we see that

$$
\mathrm{Sg}_{\mathbb{A}^m} \left\{ \begin{bmatrix} a & b & \cdots & b \\ c & a & \cdots & b \\ \vdots & \vdots & \ddots & \vdots \\ c & c & \cdots & a \end{bmatrix} \right\} \cap \mathbb{B}^m = \emptyset. \qquad \square
$$

Corollary 3.8.10. *If $\mathbb{A} = (A, m)$ is a finite majority algebra, then $\mathbb{B} \lhd \mathbb{A}$ iff there is a majority term $m' \in \mathrm{Clo}(m)$ such that $m'(\mathbb{B}, \mathbb{A}, \mathbb{B}) \subseteq \mathbb{B}$. Equivalently, we have $\mathbb{B} \lhd \mathbb{A} \iff \mathbb{B} \lhd_J \mathbb{A}$.*

More precisely, if $\mathbb{B} \lhd_J \mathbb{A}$, then $\mathbb{B}$ absorbs $\mathbb{A}$ with respect to a term of arity at most $\lceil e \cdot |\mathbb{B}|! \rceil$, where e is Euler's constant $\sum_{n \geq 0} \frac{1}{n!} \approx 2.718$.

Proof. The weaker bound $\lceil e |\mathbb{A}| \cdot |\mathbb{B}|^2! \rceil$ on the arity of an absorbing term follows from the estimate $r_k(3) \leq \lceil e \cdot k! \rceil$ and the fact that

$$
\begin{bmatrix} b \\ m'(c, a, b) \\ c \end{bmatrix} \in \mathrm{Sg}_{\mathbb{A}^3} \left\{ \begin{bmatrix} a & b & b \\ c & a & b \\ c & c & a \end{bmatrix} \right\}.
$$

However, we don't need the exact setup above. It's enough to find n sufficiently large that for every $n \times n$ matrix

$$\begin{bmatrix} a_1 & b_{12} & \cdots & b_{1n} \\ b_{21} & a_2 & \cdots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & a_n \end{bmatrix}$$

with off-diagonal entries in $\mathbb{B}$, we can find i, j, k distinct such that $b_{ij} = b_{ik}$ and $b_{ki} = b_{kj}$. If we set $b = b_{ij} = b_{ik}$ and $c = b_{ki} = b_{kj}$, then this will give us a submatrix of the form

$$\begin{bmatrix} a_i & b & b \\ b_{ji} & a_j & b_{jk} \\ c & c & a_k \end{bmatrix},$$

and applying m' will give us an element of $\mathbb{B}^3$.

Taking $n = \lceil e \cdot |\mathbb{B}|! \rceil$ is good enough to find i, j, k with $b_{ij} = b_{ik}$ and $b_{ki} = b_{kj}$. The proof is a minor adaptation of the proof of the upper bound on $r_k(3)$, and is left as an exercise to the reader. $\qquad\square$

Remark 3.8.1. It's intriguing that in the case of majority algebras, the bound on the arity of the absorbing operation only depends on the size of $|\mathbb{B}|$.

Problem 3.8.1. Define $m(k)$ to be the least number such that whenever $\mathbb{A}$ is a finite majority algebra, $\mathbb{B} \lhd_J \mathbb{A}$, and $|\mathbb{B}| \leq k$, we can always find a term t of arity at most $m(k)$ such that $\mathbb{B}$ absorbs $\mathbb{A}$ with respect to t. How quickly does $m(k)$ grow?

The dual discriminator algebra from Example 1.6.5 shows that we always have $m(k) \geq k + 2$, while the previous result shows that $m(k) \leq \lceil e \cdot k! \rceil$. For $k = 1, 2$ we have $m(1) = 3, m(2) = 4$. Could it be that we have $m(k) = k + 2$ for every k?

3.9   Finding an arc-consistent absorbing subinstance

In this section we'll go over Marcin Kozik's proof from [95] (which refined the argument from [94]) of the fact that every cycle-consistent instance has a cycle-consistent subinstance such that every domain is absorption-free. In fact, Kozik proves something stronger, involving a weaker consistency notion known as *pq*-consistency. The technique for the proof can be viewed as a generalization of the argument for the case of majority algebras, but it is much more difficult because we can't assume that all the relations involved are binary. The main idea of the proof was originally developed in [22], for the sake of proving a technical lemma about absorption generalizing Theorem 3.7.13 and Theorem 3.6.10, which was needed to show that near-unanimity CSPs can be solved in NL (nondeterministic logspace).

First we define the weaker consistency notion known as *pq*-consistency (Kozik names it *jpq*-consistency in [95]). The basic idea behind this definition is that it is a consistency check which (aside from assuming arc-consistency) only involves pairwise projections of constraints, only computes compositions of these binary relations along cycles, and is strong enough to rule out the existence of a cycle such that each binary relation along it is the graph of a permutation and the composition of all these permutations is not the identity permutation.

Definition 3.9.1. A CSP instance $\mathbf{X}$ with domains $\mathbb{A}_{v_i}$ corresponding to variables v_i is called *pq-consistent* if

- it is arc-consistent, i.e. each relation $\mathbb{R} \leq \mathbb{A}_{v_1} \times \cdots \times \mathbb{A}_{v_k}$ imposed on the variables is subdirect, and

- for each variable v and each pair of cycles p, q of $\mathbf{X}$ which begin and end at v, there exists some $j \geq 0$ such that the binary relation $\mathbb{P}_{j(p+q)+p}$ corresponding to the path $j(p+q)+p$ (see Definition 3.5.1) contains the diagonal $\Delta_{\mathbb{A}_v}$, i.e. for each $a \in \mathbb{A}_v$, we have $a \in \{a\}+j(p+q)+p$ (see Definition 3.6.3).

The reader may find it interesting to check that in the proofs that we have already given for the fact that cycle-consistency solves ancestral CSPs and majority CSPs, we may substitute *pq*-consistency for cycle-consistency everywhere without significantly complicating the arguments. The reason for introducing the slightly more technical notion of *pq*-consistency is that the "standard semidefinite relaxation" of a CSP naturally produces a *pq*-consistent instance, but doesn't always produce a cycle-consistent instance - and the semidefinite relaxation is the tool used to "robustly" solve bounded width CSPs in [18].

The main step of the argument is the following technical result.

Theorem 3.9.2 (Kozik [95])**.** *If* $\mathbf{X}$ *is a pq-consistent instance and* $\mathbf{Y}$ *is an arc-consistent subinstance of* $\mathbf{X}$ *defined by restricting each domain and each relation of* $\mathbf{X}$ *to an absorbing subalgebra, and if any domain of* $\mathbf{Y}$ *has a proper absorbing subalgebra, then there is a proper arc-consistent subinstance* $\mathbf{Z}$ *of* $\mathbf{Y}$ *defined by restricting each domain to an absorbing subalgebra.*

Before diving into the proof of this result, we'll show how it can be used.

Theorem 3.9.3. *If* $\mathbf{X}$ *is a pq-consistent instance with domains* $\mathbb{A}_v$, *then there is a pq-consistent subinstance* $\mathbf{X}'$ *of* $\mathbf{X}$ *defined by restricting each domain* $\mathbb{A}_v$ *to a minimal absorbing subalgebra* $\mathbb{A}'_v$. *If* $\mathbf{X}$ *is cycle-consistent, then so is* $\mathbf{X}'$.

Proof. By repeatedly applying Theorem 3.9.2, we may find an arc-consistent subinstance $\mathbf{X}'$ of $\mathbf{X}$ such that each domain has no proper absorbing subalgebra. Then by Theorem 3.7.17, for every cycle r from v to v of $\mathbf{X}$ such that $\mathbb{P}_r \supseteq \Delta_{\mathbb{A}_v}$, if r' is the corresponding cycle in $\mathbf{X}'$, then we have $\mathbb{P}_{r'} \supseteq \Delta_{\mathbb{A}'_v}$. If $\mathbf{X}$ is *pq*-consistent, then for any cycles p, q from v to v there is some j such that $\mathbb{P}_{j(p+q)+p} \supseteq \Delta_{\mathbb{A}_v}$, so on taking $r = j(p+q)+p$ we see that the corresponding cycles p', q' have $\mathbb{P}_{j(p'+q')+p'} \supseteq \Delta_{\mathbb{A}'_v}$. □

The proof of Theorem 3.9.2 will only rely on three properties of absorption. Since there are several absorption-like concepts that have proven useful, and most of them satisfy these properties, we will consider an arbitrary "absorption concept" $\lhd_X$ which applies to certain pairs $\mathbb{B} \leq \mathbb{A}$, and which satisfies the following three properties.

- **Compatibility with pp-formulas.** If $\mathbb{S}_i \lhd_X \mathbb{R}_i$ are relations, and if a relation $\mathbb{R}$ is defined by a pp-formula Φ involving the relations $\mathbb{R}_1, ..., \mathbb{R}_k$ (and possibly some other relations), then if we define a relation $\mathbb{S}$ by the pp-formula Φ' defined by replacing each $\mathbb{R}_i$ by $\mathbb{S}_i$ in Φ, we have $\mathbb{S} \lhd_X \mathbb{R}$.

- **Transitive closure.** If $\mathbb{C} \lhd_X \mathbb{B} \lhd_X \mathbb{A}$, then $\mathbb{C} \lhd_X \mathbb{A}$.

179

- **Connectivity transfers.** If $\mathbb{S} \lhd_X \mathbb{R}$ and $\mathbb{R} \le \mathbb{A} \times \mathbb{A}$, and if $a, b \in \mathbb{A}$ are such that $(a,a), (b,b) \in \mathbb{S}$ and $(a,b) \in \mathbb{R}$, then there is some k such that $(a,b) \in \mathbb{S}^{\circ k}$.

Note that by the local criterion for Jónsson absorption (Theorem 3.7.21), if $\lhd_X$ is compatible with pp-formulas, then the connectivity transfer property of $\lhd_X$ is equivalent to the implication $\mathbb{B} \lhd_X \mathbb{A} \implies \mathbb{B} \lhd_J \mathbb{A}$. Also, a trivial case of compatibility with pp-formulas implies that for all $\mathbb{A}$, we have $\mathbb{A} \lhd_X \mathbb{A}$.

Throughout most of the proof, we will be focusing on the arc-consistent instance $\mathbf{Y}$. Therefore, for each variable v of $\mathbf{Y}$, we let $\mathbb{A}_v$ be the corresponding domain in $\mathbf{Y}$, and let $\mathbb{A}_v^{\mathbf{X}}$ be the corresponding domain in the original pq-consistent instance $\mathbf{X}$, so we have $\mathbb{A}_v \lhd_X \mathbb{A}_v^{\mathbf{X}}$. Similarly, if $\mathbb{R}$ refers to a relation in $\mathbf{Y}$, then we let $\mathbb{R}^{\mathbf{X}}$ refer to the corresponding relation in $\mathbf{X}$, with $\mathbb{R} \lhd_X \mathbb{R}^{\mathbf{X}}$.

The argument strategy generalizes the strategy used for majority algebras. We will consider the set $\mathcal{B}$ of ordered pairs $(x, \mathbb{B})$ such that x is a variable of $\mathbf{Y}$, $\mathbb{B} \lhd_X \mathbb{A}_x$, and $\mathbb{B} \ne \emptyset, \mathbb{A}_x$. We want to define a quasiorder $\preceq$ on $\mathcal{B}$, such that if restricting the domain of the variable x to $\mathbb{B}$ and imposing arc-consistency forces another variable y to have its domain restricted to $\mathbb{C}$, then we have $(x, \mathbb{B}) \preceq (y, \mathbb{C})$. Unfortunately, it is not enough to consider paths alone to define this partial order: general deductions involving arc-consistency involve reasoning about *trees*.

Definition 3.9.4. To every relational structure $\mathbf{A} = (A, R_1, ...)$ we associate the bipartite graph $\mathcal{G}_{\mathbf{A}}$ with vertex sets A and $R_1 \sqcup \cdots$, and edge set consisting of pairs (a, r) for every $a \in A$ and $r \in R_i$ such that some coordinate of r is equal to a (if a occurs as a coordinate of r multiple times, then we make multiple copies of the edge (a, r)).

We say that $\mathbf{A}$ is a *tree* if the associated bipartite graph $\mathcal{G}_{\mathbf{A}}$ is a tree (so in particular, no tuple r in any relation R_i can have any repeated coordinates).

Kozik [95] extends the concepts of paths and addition of paths to trees in order to define the partial order $\preceq$ on $\mathcal{B}$ properly.

Definition 3.9.5. If $\mathbf{Y}$ is a CSP instance, viewed as a relational structure, then we define a *tree pattern* p from x to y to consist of the following information:

- a relational structure $\mathbf{A} = (A, R_1, ...)$ which is a tree, with each relation of $\mathbf{A}$ corresponding to a relation of $\mathbf{Y}$,

- a homomorphism of relational structures $h : \mathbf{A} \to \mathbf{Y}$,

- a subset $I \subseteq A$ of the elements of A which we call the set of *inputs* to the pattern, such that for all $i \in I$ we have $h(i) = x$, and

- an element $o \in A$ which we call the *output* of the pattern, such that $h(o) = y$.

If p is a tree pattern from x to y, then we may view it as a CSP instance via the homomorphism $h : \mathbf{A} \to \mathbf{Y}$. If $\mathbb{B} \le \mathbb{A}_x$, then we define $\mathbb{B} + p$ to be the subalgebra of values $b \in \mathbb{A}_y$ such that the instance $\mathbf{A}$ has a solution with the variables from I assigned to values in $\mathbb{B}$, and with the variable o assigned to the value b.

If p is a tree pattern from x to y, and if q is a tree pattern from y to z, then we define the tree pattern $p + q$ by attaching a copy of p to each input of q, combining the output of each copy of p to the corresponding input of q. This definition is set up to ensure that $\mathbb{B} + (p + q) = (\mathbb{B} + p) + q$ for any $\mathbb{B} \le \mathbb{A}_x$.

Proposition 3.9.6. *If p is a tree pattern from x to y in an arc-consistent instance $\mathbf{Y}$ and $\mathbb{B} \lhd_X \mathbb{A}_x$, then $\mathbb{B} + p \lhd_X \mathbb{A}_y$.*

Proof. This follows from the fact that $\lhd_X$ is compatible with pp-formulas: we have $\mathbb{A}_x + p = \mathbb{A}_y$ if $\mathbf{Y}$ is arc-consistent, and so $\mathbb{B} + p \lhd_X \mathbb{A}_x + p = \mathbb{A}_y$. $\qquad\square$

Note that unlike the situation for path patterns, arc-consistency of the instance $\mathbf{Y}$ is no longer enough to ensure that $\mathbb{B} \neq \emptyset \implies \mathbb{B} + p \neq \emptyset$ for all tree patterns p. So we can no longer take as given that the subalgebras we construct will always be nonempty.

Definition 3.9.7. Define the quasiordered set $(\mathcal{B}, \preceq)$ to be the set of ordered pairs $(x, \mathbb{B})$ such that x is a variable of the instance $\mathbf{Y}$, $\mathbb{B} \lhd_X \mathbb{A}_x$, and $\mathbb{B} \neq \emptyset, \mathbb{A}_x$, with the quasiorder defined by $(x, \mathbb{B}) \preceq (y, \mathbb{C})$ if there exists a tree pattern p from x to y with $\mathbb{B} + p = \mathbb{C}$.

As in the argument for majority algebras, we now pick a maximal component $\mathcal{C}$ of the quasiordered set $(\mathcal{B}, \preceq)$ (since $\mathcal{B}$ is nonempty by assumption and is finite, such a maximal component exists). We would like to use $\mathcal{C}$ to define our reduced instance $\mathbf{Z}$, but we no longer have a guarantee that there is at most one set $\mathbb{B}$ with $(x, \mathbb{B}) \in \mathcal{C}$ for a given variable x.

A worst case scenario would be that there exist $\mathbb{B}_1, \mathbb{B}_2$ with $(x, \mathbb{B}_i) \in \mathcal{C}$ such that $\mathbb{B}_1 \cap \mathbb{B}_2 = \emptyset$: in this case, we would have no hope of using $\mathcal{C}$ to define an arc-consistent reduction, because no matter which $(y, \mathbb{C}) \in \mathcal{C}$ we pick, there exist tree patterns p_1, p_2 from y to x with $\mathbb{C} + p_i = \mathbb{B}_i$, so reducing the domain $\mathbb{A}_y$ to $\mathbb{C}$ and imposing arc-consistency would make it impossible to assign any value to x. The main step of the proof is ruling out this scenario.

Lemma 3.9.8. *If $\mathcal{C}$ is a maximal component of $(\mathcal{B}, \preceq)$, and if $(x, \mathbb{B}), (x, \mathbb{C}) \in \mathcal{C}$, then $\mathbb{B} \cap \mathbb{C} \neq \emptyset$.*

Before proving the lemma, we'll show how we can use it to finish the proof of Theorem 3.9.2. This step won't use the fact that the instance $\mathbf{X}$ is *pq*-consistent, or the fact that $\lhd_X$ transfers connectivity: the lemma is where these crucial facts are used.

Proof of Theorem 3.9.2, assuming the lemma. Note that if $(x, \mathbb{B}), (x, \mathbb{C}) \in \mathcal{C}$, then we can splice together tree patterns to show that $(x, \mathbb{B} \cap \mathbb{C}) \in \mathcal{C}$ as well (so long as $\mathbb{B} \cap \mathbb{C} \neq \emptyset$, which follows from the lemma). So for every x, we can define a subalgebra $\mathbb{B}_x \lhd_X \mathbb{A}_x$ by taking $\mathbb{B}_x$ to be the intersection of all $\mathbb{B}$ such that $(x, \mathbb{B}) \in \mathcal{C}$ (or taking $\mathbb{B}_x = \mathbb{A}_x$ if no such $\mathbb{B}$ exist). We define the absorbing subinstance $\mathbf{Z}$ by reducing the domains of $\mathbf{Y}$ from $\mathbb{A}_x$ to $\mathbb{B}_x$. We need to check that $\mathbf{Z}$ is arc-consistent.

Consider a single relation $\mathbb{R} \leq_{sd} \mathbb{A}_{x_1} \times \cdots \times \mathbb{A}_{x_k}$ of $\mathbf{Y}$. We wish to show that $\mathbb{R} \cap \prod_i \mathbb{B}_{x_i}$ is subdirect in $\prod_i \mathbb{B}_{x_i}$. We will show by induction on i that

$$\pi_i(\mathbb{R} \cap \prod_{j \leq i} \mathbb{B}_{x_j} \times \prod_{l > i} \mathbb{A}_{x_l}) = \mathbb{B}_{x_i}.$$

The base case $i = 1$ follows from the fact that $\mathbf{Y}$ is arc-consistent. For the inductive step, we pick any $(y, \mathbb{C}) \in \mathcal{C}$ and splice together tree patterns p_j from y to x_j with $\mathbb{C} + p_j = \mathbb{B}_{x_j}$ for $j < i$ such that $\mathbb{B}_{x_j} \neq \mathbb{A}_{x_j}$ together with the relation $\mathbb{R}$ to make a tree pattern p from y to x_i with

$$\mathbb{C} + p = \pi_i(\mathbb{R} \cap \prod_{j \leq i-1} \mathbb{B}_{x_j} \times \prod_{l > i-1} \mathbb{A}_{x_l}),$$

and note that by the induction hypothesis the right hand side is nonempty. Thus we either have $\mathbb{C} + p = \mathbb{A}_{x_i}$ or $(x_i, \mathbb{C} + p) \in \mathcal{C}$, and in either case we have $\mathbb{B}_{x_i} \subseteq \mathbb{C} + p$ (by the lemma), which completes the proof. $\qquad\square$

Now we finally prove the crucial lemma.

Proof of the lemma. Suppose for contradiction that the lemma is not true, and choose $\mathbb{C}$ maximal such that $(x, \mathbb{C}) \in \mathcal{C}$ and such that there exists $(x, \mathbb{B}) \in \mathcal{C}$ with $\mathbb{B} \cap \mathbb{C} = \emptyset$. Let $\mathbb{B}_1, ..., \mathbb{B}_k$ be the set of minimal $\mathbb{B}$s such that $(x, \mathbb{B}) \in \mathcal{C}$ and $\mathbb{B} \cap \mathbb{C} = \emptyset$. Note that since the set of $\mathbb{B}$s with $(x, \mathbb{B}) \in \mathcal{C}$ is closed under nonempty intersection, we must have $\mathbb{B}_i \cap \mathbb{B}_j = \emptyset$ for all $i \neq j$. Additionally, any $\mathbb{B}$ with $(x, \mathbb{B}) \in \mathcal{C}$ and $\mathbb{B} \cap \mathbb{C} = \emptyset$ must contain at least one $\mathbb{B}_i$.

Choose tree patterns p_i, q, r from x to x such that $\mathbb{B}_i + p_i = \mathbb{B}_{i+1}$, $\mathbb{C} + q = \mathbb{B}_1$, $\mathbb{B}_k + r = \mathbb{C}$. Define the tree pattern p by $p = q + p_1 + \cdots + p_{k-1} + r$, and note that $\mathbb{C} + p = \mathbb{C}$. We will mainly work inside the instance $\mathbf{A}$ corresponding to the tree pattern p.

First we prune the inputs of the tree pattern p a little bit to make a new tree pattern p' (with the same instance $\mathbf{A}$), removing variables of $\mathbf{A}$ from the input set one at a time as long as we can remove one while keeping $\mathbb{C} + p' = \mathbb{C}$. Now pick any remaining input variable $s \in \mathbf{A}$ of p' (at least one input variable remains at the end of the pruning process, by the arc-consistency of $\mathbf{Y}$), and let t be the output variable of p' (note that s, t are both mapped to x in $\mathbf{Y}$). Let p'' be p' with s removed from its input set. Consider the binary relation $\mathbb{S} \leq \mathbb{A}_x \times \mathbb{A}_x$ consisting of pairs (a, b) such that some solution of the instance $\mathbf{A}$ assigns the value a to s, assigns the value b to t, and assigns all input variables of p'' to values in $\mathbb{C}$.

Since $\mathbb{C} + p' = \mathbb{C}$, we have

$$\mathbb{C} + \mathbb{S} = \mathbb{C} + p' = \mathbb{C},$$

and because of the pruning process we have

$$\pi_2(\mathbb{S}) = \mathbb{C} + p'' \neq \mathbb{C},$$

so by the maximal choice of $\mathbb{C}$ we have $\pi_2(\mathbb{S}) \cap \mathbb{B}_i \neq \emptyset$ for all i. By splicing p'' together with a tree pattern q_i with $\mathbb{C} + q_i = \mathbb{B}_i$ (merging their outputs together), we see that $(x, \pi_2(\mathbb{S}) \cap \mathbb{B}_i) \in \mathcal{C}$, so by the minimality of $\mathbb{B}_i$ we have

$$\pi_2(\mathbb{S}) \supseteq \mathbb{B}_i$$

for all i. Thus the subalgebra

$$\mathbb{B}_i - \mathbb{S} = \pi_1(\mathbb{S} \cap \mathbb{A}_x \times \mathbb{B}_i)$$

is nonempty, has $(\mathbb{B}_i - \mathbb{S}) \cap \mathbb{C} = \emptyset$ since $(\mathbb{C} + \mathbb{S}) \cap \mathbb{B}_i = \emptyset$, and by splicing p'' with the same q_i and changing the output to s, we see that $(x, \mathbb{B}_i - \mathbb{S}) \in \mathcal{C}$. Thus there is some j_i such that $\mathbb{B}_i - \mathbb{S} \supseteq \mathbb{B}_{j_i}$. Then we have

$$(\mathbb{B}_{j_i} + \mathbb{S}) \cap \mathbb{B}_i \neq \emptyset,$$

and by another tree splice (this time splicing q_{j_i} into p'' by merging the output of q_{j_i} with s) we see that either $\mathbb{B}_{j_i} + \mathbb{S} = \mathbb{A}_x$ or $(x, \mathbb{B}_{j_i} + \mathbb{S}) \in \mathcal{C}$, so by the minimality of $\mathbb{B}_i$ we have

$$\mathbb{B}_{j_i} + \mathbb{S} \supseteq \mathbb{B}_i.$$

Thus we have

$$\cup_i \mathbb{B}_i + \mathbb{S} \supseteq \cup_i \mathbb{B}_i,$$

so if we consider $\mathbb{S}$ as a digraph on $\mathbb{A}_x$, we see that there is some directed cycle of $\mathbb{S}$ which is entirely contained in $\cup_i \mathbb{B}_i$. From $\mathbb{C} + \mathbb{S} = \mathbb{C}$, we also see that there is some directed cycle of $\mathbb{S}$ which is entirely contained in $\mathbb{C}$. The plan is to apply Corollary 3.7.10 to produce a directed path in $\mathbb{S}$ from

an element of $\mathbb{C}$ to an element of $\cup_i \mathbb{B}_i$, which will give us a contradiction since any directed path in $\mathbb{S}$ which starts in $\mathbb{C}$ must end up in $\mathbb{C}$.

In order to apply Corollary 3.7.10, we need to construct a binary relation $\mathbb{R}$ such that $\mathbb{S} \lhd_X \mathbb{R}$ and such that there is a directed path from $\mathbb{C}$ to $\cup_i \mathbb{B}_i$ in $\mathbb{R}$. This is where we will finally use the assumption that $\mathbf{Y}$ absorbs a bigger instance $\mathbf{X}$ which is pq-consistent. We define an instance $\mathbf{A}^{\mathbf{X}}$ similarly to $\mathbf{A}$, but with each domain replaced with the corresponding domain in $\mathbf{X}$ and similarly for the relations, and define $\mathbb{R}$ to be the projection of the solution set to $\mathbf{A}^{\mathbf{X}}$ onto the variables s, t. Then since $\lhd_X$ is compatible with pp-formulas and since every domain/relation restriction in sight is absorbing, we have $\mathbb{S} \lhd_X \mathbb{R}$.

Now pick any $b \in \cup_i \mathbb{B}_i$ which is contained in a directed cycle of $\mathbb{S}$. Suppose $b \in \mathbb{B}_i$. Consider the path from s to the output variable of $q + p_1 + \cdots + p_{i-1}$ in $\mathbf{A}$, call this path α, and let β be the path from that output variable to t in $\mathbf{A}$. The images of these paths in $\mathbf{X}$ are cycles $\alpha_{\mathbf{X}}, \beta_{\mathbf{X}}$ from x to x, so by the pq-consistency of $\mathbf{X}$ there must exist some $j \geq 0$ such that $b \in \{b\} + j(\beta_{\mathbf{X}} + \alpha_{\mathbf{X}}) + \beta_{\mathbf{X}}$. Note that by the arc-consistency of $\mathbf{X}$, $\mathbb{R}$ is the binary relation corresponding to the cycle $\alpha_{\mathbf{X}} + \beta_{\mathbf{X}}$. Additionally, since

$$\mathbb{C} + q + p_1 + \cdots + p_{i-1} = \mathbb{B}_i,$$

there is some $c \in \mathbb{C}$ such that $b \in \{c\} + \alpha_{\mathbf{X}}$. Thus we have

$$b \in \{c\} + \alpha_{\mathbf{X}} + j(\beta_{\mathbf{X}} + \alpha_{\mathbf{X}}) + \beta_{\mathbf{X}} = \{c\} + (j+1)(\alpha_{\mathbf{X}} + \beta_{\mathbf{X}}) = \{c\} + \mathbb{R}^{\circ(j+1)}.$$

Additionally, by following paths of $\mathbb{S}$ backwards sufficiently many times, we see that c is reachable from a directed cycle of $\mathbb{S}$ which is entirely contained in $\mathbb{C}$. Thus there is some m such that for some $a \in \mathbb{C}$, we have $(a, a), (b, b) \in \mathbb{S}^{\circ m}$ and $(a, b) \in \mathbb{R}^{\circ m}$, and since $\mathbb{S}^{\circ m} \lhd_X \mathbb{R}^{\circ m}$ we may apply the transfer of connectivity property to see that for some n we have $(a, b) \in \mathbb{S}^{\circ n}$, which gives us our contradiction. $\qquad \square$

To finish the analysis of bounded width algebras, we just need to understand the case where all the domains are absorption free. For this we need two main ingredients: first is that binary relations are forced to be boring unless some absorption occurs, and second is that if a simple algebra has an exciting ternary relation whose binary projections are boring, then the algebra must be affine and therefore does not have bounded width.

3.10   Zhuk's centers and ternary absorption

In this section we'll go over a very strong technique introduced by Zhuk in his proof of the dichotomy conjecture [142], which produces ternary absorption as soon as we have a certain type of binary relation on a pair of Taylor algebras. This technique allows us to both simplify and strengthen one of the key results needed for the study of general Taylor algebras, known as the "absorption theorem".

First, we'll go over the history of this idea, so the reader can understand where the definition comes from and why it is (somewhat) natural.

The main idea behind Zhuk's approach in [142] is to note that if an algebra is not polynomially complete, then its polynomial clone must be contained in a maximal proper subclone of the clone of all functions (that every proper subclone is contained in a *maximal* proper subclone follows from the fact that the clone of all functions is finitely generated: in fact, it's generated by the set of functions of arity 2). A maximal clone corresponds under the $\mathrm{Inv} - \mathrm{Pol}$ Galois connection to a

minimal relational clone, and every minimal relational clone can be generated by a single relation, of one of several special forms. Zhuk is very familiar with the theory of relational clones, so he was aware of Rosenberg's Completeness Theorem [128] (see [119] or chapter II.6 of [101] for alternate expositions), which completely classifies the special relations which correspond to maximal clones into six different types.

Zhuk then considered each of the types of relations from Rosenberg's classification, and investigated which of them might be preserved by the polynomial clone of a Taylor algebra, and what the existence of such a relation implies about the structure of the Taylor algebra. The most interesting case is the case of the relations known as *central relations*.

Definition 3.10.1. A relation $\mathbb{R} \leq \mathbb{A}^n$ is *central* if it has the following properties:

- $\mathbb{R}$ is symmetric under permuting its coordinates,

- $\mathbb{R}$ contains every tuple which has any pair of equal coordinates, and

- the set $\mathbb{C} \leq \mathbb{A}$ defined by

$$\mathbb{C} = \{c \in \mathbb{A} \mid \forall a_2, ..., a_n \in \mathbb{A}, (c, a_2, ..., a_n) \in \mathbb{R}\}$$

  is not empty and is not equal to $\mathbb{A}$.

The set $\mathbb{C}$ is known as the *center* of the central relation $\mathbb{R}$.

Since relations of high arity are hard to think about, Zhuk simplifies this to a special type of binary relation on $\mathbb{A} \times \mathbb{B}$, where $\mathbb{B}$ is secretly taken to be $\mathbb{A}^{n-1}$. To see that this step doesn't lose anything essential, we use the following fact about absorbing subalgebras of powers.

Proposition 3.10.2. *Suppose that $\mathbb{A}$ is idempotent and that $\mathbb{A}^k$ has a proper absorbing subalgebra for some k. Then $\mathbb{A}$ has a proper absorbing subalgebra.*

In fact, this holds for any absorption concept $\lhd_X$ which is compatible with pp-formulas.

Proof. We induct on k. Suppose that $\mathbb{B} \lhd_X \mathbb{A}^k$. If $\pi_1(\mathbb{B}) \neq \mathbb{A}$ then $\pi_1(\mathbb{B}) \lhd_X \mathbb{A}$ and we are done, otherwise since $\mathbb{B} \neq \mathbb{A}^k$ there must exist some $a \in \mathbb{A}$ such that $\pi_{[k]\setminus\{1\}}(\mathbb{B} \cap \{a\} \times \mathbb{A}^{k-1}) \neq \mathbb{A}^{k-1}$. Since $\lhd_X$ is compatible with pp-formulas and $\{a\} \leq \mathbb{A}$ by the idempotence of $\mathbb{A}$, we have

$$\pi_{[k]\setminus\{1\}}(\mathbb{B} \cap \{a\} \times \mathbb{A}^{k-1}) \lhd_X \mathbb{A}^{k-1},$$

so we can apply the induction hypothesis. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

With this in mind, it's natural to restrict our attention to binary relations $\mathbb{R} \leq \mathbb{A} \times \mathbb{B}$ which have a nontrivial proper "left center", and to try to use them to produce an absorbing subalgebra inside either $\mathbb{A}$ or $\mathbb{B}$.

Definition 3.10.3. If $\mathbb{R} \leq_{sd} \mathbb{A} \times \mathbb{B}$ is subdirect and $\mathbb{B}$ is finite and idempotent, then the *left center* of $\mathbb{R}$ is the subalgebra $\mathbb{C} \leq \mathbb{A}$ defined by

$$\mathbb{C} = \{c \in \mathbb{A} \mid \forall b \in \mathbb{B}, (c, b) \in \mathbb{R}\}.$$

The *right center* of a subdirect binary relation is defined similarly (so the right center of $\mathbb{R}$ is the left center of $\mathbb{R}^-$, and is a subalgebra of $\mathbb{B}$).

184

To see that the left center $\mathbb{C}$ is automatically a subalgebra of $\mathbb{A}$, note that it can be defined by the following pp-formula:

$$c \in \mathbb{C} \iff \bigwedge_{b \in \mathbb{B}} \exists x (x \in \{b\} \wedge (c, x) \in \mathbb{R}).$$

In order to do anything useful with such a binary relation, we will need to assume that $\mathbb{B}$ is Taylor. We will attempt to exploit the Taylor term to produce binary absorption on $\mathbb{B}$, using the following lemma.

Lemma 3.10.4. *Suppose $\mathbb{B} \leq \mathbb{A}$ and that there is an idempotent term $t \in \mathrm{Clo}_k(\mathbb{A})$ with the following two properties:*

- *t satisfies an identity of the form $t(x, u_2, ..., u_k) \approx t(y, v_2, ..., v_k)$, where each $u_i, v_i \in \{x, y\}$, and*

- *$t(\mathbb{B}, \mathbb{A}, ..., \mathbb{A}) \subseteq \mathbb{B}$.*

Then $\mathbb{B}$ absorbs $\mathbb{A}$ with respect to some idempotent binary operation f.

Proof. To make the notation more clear, we treat each u_i, v_i as a binary function, with $u_i = u_i(x, y)$ and $v_i = v_i(x, y)$. Define $f(x, y)$ by

$$f(x, y) := t(x, u_2(x, y), ..., u_k(x, y)) \approx t(y, v_2(x, y), ..., v_k(x, y)).$$

Then for any $a \in \mathbb{A}$ and $b \in \mathbb{B}$, we have

$$f(a, b) = t(b, v_2(a, b), ..., v_k(a, b)) \in t(\mathbb{B}, \mathbb{A}, ..., \mathbb{A}) \subseteq \mathbb{B},$$

and

$$f(b, a) = t(b, u_2(a, b), ..., u_k(a, b)) \in t(\mathbb{B}, \mathbb{A}, ..., \mathbb{A}) \subseteq \mathbb{B}. \qquad \square$$

Theorem 3.10.5 (Zhuk [142]). *Suppose that $\mathbb{A}, \mathbb{B}$ are finite idempotent algebras, and that there is a term t which is Taylor on $\mathbb{B}$. If $\mathbb{R} \leq_{sd} \mathbb{A} \times \mathbb{B}$ is subdirect and has a nontrivial left center $\mathbb{C}$, then either $\mathbb{B}$ has a proper binary absorbing subalgebra, or $\mathbb{C}$ absorbs $\mathbb{A}$ with respect to the term $t * \cdots * t$, with $|\mathbb{B}| - 1$ copies of t.*

Proof. Suppose t has arity k. We will show that if $\mathbb{B}$ has no proper absorbing subalgebra, then for any $a \in \mathbb{A} \setminus \mathbb{C}$ and for any $c_1, ..., c_k \in \mathbb{C}$ and any $i \leq k$, the value

$$t(c_1, ..., c_{i-1}, a, c_{i+1}, ..., c_k)$$

is "closer" to being in $\mathbb{C}$ than a is. To make this precise, we measure how close an element a is to being in $\mathbb{C}$ by looking at the size of the set

$$a + \mathbb{R} = \pi_2(\mathbb{R} \cap \{a\} \times \mathbb{B}).$$

By the definition of $\mathbb{C}$, we have $|a + \mathbb{R}| = |\mathbb{B}|$ if and only if $a \in \mathbb{C}$.

Since $\mathbb{R}$ is preserved by t, we have

$$t(c_1, ..., a, ..., c_k) + \mathbb{R} \supseteq t(c_1 + \mathbb{R}, ..., a + \mathbb{R}, ..., c_k + \mathbb{R}) = t(\mathbb{B}, ..., a + \mathbb{R}, ..., \mathbb{B}).$$

185

Since t is idempotent, the right hand side of the above must contain $a+\mathbb{R}$, and if it is equal to $a+\mathbb{R}$ then we can apply the previous lemma (since t is Taylor) to see that $a+\mathbb{R}$ is a binary absorbing subalgebra of $\mathbb{B}$. Thus if $a \notin \mathbb{C}$, then either $a+\mathbb{R}$ is a proper binary absorbing subalgebra of $\mathbb{B}$, or else

$$|t(c_1, ..., a, ..., c_k) + \mathbb{R}| > |a + \mathbb{R}|. \qquad \square$$

Keeping the same setup, the left center $\mathbb{C}$ has an additional nice property, which is much stronger than it looks.

Theorem 3.10.6 (Zhuk [142]). *Suppose $\mathbb{A}, \mathbb{B}$ are finite idempotent algebras. If $\mathbb{R} \leq_{sd} \mathbb{A} \times \mathbb{B}$ has a left center $\mathbb{C}$ and $\mathbb{B}$ has no proper binary absorbing subalgebras, then for any $a \in \mathbb{A}$ we have*

$$a \notin \mathbb{C} \implies \begin{bmatrix} a \\ a \end{bmatrix} \notin \mathrm{Sg}_{\mathbb{A}^2}\left\{ \begin{bmatrix} a \\ \mathbb{C} \end{bmatrix}, \begin{bmatrix} \mathbb{C} \\ \mathbb{C} \end{bmatrix}, \begin{bmatrix} \mathbb{C} \\ a \end{bmatrix} \right\}.$$

Proof. Suppose otherwise. Then there are i, j and $c_1, ..., c_i, c_j', ..., c_n' \in \mathbb{C}$ with $j \leq i+1$ and a term t of arity n such that

$$\begin{bmatrix} a \\ a \end{bmatrix} = t\left(\begin{bmatrix} a \\ c_1 \end{bmatrix}, \cdots, \begin{bmatrix} a \\ c_{j-1} \end{bmatrix}, \begin{bmatrix} c_j' \\ c_j \end{bmatrix}, \cdots, \begin{bmatrix} c_i' \\ c_i \end{bmatrix}, \begin{bmatrix} c_{i+1}' \\ a \end{bmatrix}, ..., \begin{bmatrix} c_n' \\ a \end{bmatrix} \right).$$

Looking at the neighbors via $\mathbb{R}$, we have

$$\begin{bmatrix} a + \mathbb{R} \\ a + \mathbb{R} \end{bmatrix} \supseteq t\left(\begin{bmatrix} a + \mathbb{R} & \cdots & a + \mathbb{R} & \mathbb{B} & \cdots & \mathbb{B} & \mathbb{B} & \cdots & \mathbb{B} \\ \mathbb{B} & \cdots & \mathbb{B} & \mathbb{B} & \cdots & \mathbb{B} & a + \mathbb{R} & \cdots & a + \mathbb{R} \end{bmatrix} \right).$$

Thus $a + \mathbb{R}$ absorbs $\mathbb{B}$ with respect to the binary term

$$f(x, y) := t(x, ..., x, y, ..., y)$$

as long as the number of xs is between $j - 1$ and i. $\qquad \square$

We can combine the previous two results about left centers to define a new type of absorption. We won't need the full power of the previous result, and instead will use a slightly weaker property.

Definition 3.10.7. We say that $\mathbb{C}$ *centrally absorbs* $\mathbb{A}$, written $\mathbb{C} \lhd_Z \mathbb{A}$, if the following two properties hold:

- $\mathbb{C} \lhd \mathbb{A}$, and

- for any $a \notin \mathbb{C}$, we have $\begin{bmatrix} a \\ a \end{bmatrix} \notin \mathrm{Sg}_{\mathbb{A}^2}\left\{ \begin{bmatrix} a \\ \mathbb{C} \end{bmatrix}, \begin{bmatrix} \mathbb{C} \\ a \end{bmatrix} \right\}.$

Corollary 3.10.8. *Suppose $\mathbb{A}, \mathbb{B}$ are finite and idempotent. If $\mathbb{R} \leq_{sd} \mathbb{A} \times \mathbb{B}$ has left center $\mathbb{C}$ and $\mathbb{B}$ is Taylor and binary absorption free, then $\mathbb{C} \lhd_Z \mathbb{A}$.*

There is an unfortunate naming collision between the centers considered here, and the centers considered in commutator theory. Generally it should be clear from context which sort of center is meant. (I have proposed the alternate name *stable absorption* instead of central absorption, but it seems unlikely to catch on.)

The key fact about central absorption that makes it so much more powerful than ordinary absorption is the following doubling trick due to Zhuk and Kozik.

Lemma 3.10.9 (Essential doubling trick [142])**.** *Suppose that* $\mathbb{R} \leq \mathbb{A}_0 \times \cdots \times \mathbb{A}_{n+1}$ *is* $(\mathbb{C}, \mathbb{B}_1, ..., \mathbb{B}_n, \mathbb{C}')$-*essential, with* $\mathbb{C}' \lhd_Z \mathbb{A}_{n+1}$ *and* $\mathbb{A}_{n+1}$ *finite and idempotent. Then there is a relation*

$$\mathbb{R}' \leq \mathbb{A}_0 \times \cdots \times \mathbb{A}_n \times \mathbb{A}_n \times \cdots \times \mathbb{A}_0$$

which is $(\mathbb{C}, \mathbb{B}_1, ..., \mathbb{B}_n, \mathbb{B}_n, ..., \mathbb{B}_1, \mathbb{C})$-*essential.*

Proof. Suppose $\mathbb{R}$ is chosen such that, subject to satisfying the assumptions of the lemma, the subalgebra $\mathbb{B}' \leq \mathbb{A}_{n+1}$ defined by

$$\mathbb{B}' = \pi_{n+1}(\mathbb{R} \cap \mathbb{C} \times \mathbb{B}_1 \times \cdots \times \mathbb{B}_n \times \mathbb{A}_{n+1})$$

is as small as possible. Note that $\mathbb{B}'$ is necessarily nonempty and disjoint from $\mathbb{C}'$ if $\mathbb{R}$ is $(\mathbb{C}, \mathbb{B}_1, ..., \mathbb{B}_n, \mathbb{C}')$-essential.

Since we may shrink $\mathbb{R}$ to the subalgebra generated by any collection of tuples witnessing $\mathbb{R} \cap (\mathbb{C} \times \cdots \times \mathbb{A}_i \times \cdots \times \mathbb{C}') \neq \emptyset$ for all i from 0 to $n+1$, we see that

$$b, b' \in \mathbb{B}' \implies b' \in \mathrm{Sg}_{\mathbb{A}_{n+1}}(\mathbb{C}' \cup \{b\}).$$

In particular, if we pick some $b \in \mathbb{B}'$ and define the symmetric binary relation $\mathbb{S} \leq \mathbb{A}_{n+1} \times \mathbb{A}_{n+1}$ by

$$\mathbb{S} = \mathrm{Sg}_{\mathbb{A}_{n+1}^2} \left\{ \begin{bmatrix} b \\ \mathbb{C}' \end{bmatrix}, \begin{bmatrix} \mathbb{C}' \\ b \end{bmatrix} \right\},$$

then $\pi_1(\mathbb{S}) \supseteq \mathbb{B}'$.

We now define the relation $\mathbb{R}'$ by

$$(x_0, ..., x_n, y_n, ..., y_0) \in \mathbb{R}' \iff \exists x_{n+1}, y_{n+1} \ (x_0, ..., x_{n+1}) \in \mathbb{R} \wedge (x_{n+1}, y_{n+1}) \in \mathbb{S} \wedge (y_0, ..., y_{n+1}) \in \mathbb{R}.$$

To see that

$$\mathbb{R}' \cap \mathbb{C} \times \mathbb{B}_1 \times \cdots \times \mathbb{A}_i \times \cdots \times \mathbb{B}_n \times \mathbb{B}_n \times \cdots \times \mathbb{B}_1 \times \mathbb{C} \neq \emptyset$$

for any $0 \leq i \leq n$, we choose $(x_0, ..., x_{n+1}) \in \mathbb{R} \cap (\mathbb{C} \times \cdots \times \mathbb{A}_i \times \cdots \times \mathbb{C}')$ and choose $(y_0, ..., y_{n+1}) \in \mathbb{R} \cap \mathbb{C} \times \mathbb{B}_1 \times \cdots \times \mathbb{B}_n \times \{b\}$, which is possible since $b \in \mathbb{B}'$ and $\mathbb{C}' \times \{b\} \subseteq \mathbb{S}$. We can check that

$$\mathbb{R}' \cap \mathbb{C} \times \mathbb{B}_1 \times \cdots \times \mathbb{B}_n \times \mathbb{B}_n \times \cdots \times \mathbb{A}_i \times \cdots \times \mathbb{B}_1 \times \mathbb{C} \neq \emptyset$$

for $0 \leq i \leq n$ similarly, by interchanging the roles of the x_is and y_is.

To finish, we just need to check that

$$\mathbb{R}' \cap \mathbb{C} \times \mathbb{B}_1 \times \cdots \times \mathbb{B}_n \times \mathbb{B}_n \times \cdots \times \mathbb{B}_1 \times \mathbb{C} = \emptyset,$$

or equivalently, that

$$\mathbb{S} \cap \mathbb{B}' \times \mathbb{B}' = \emptyset.$$

So suppose for contradiction that there are $b', b'' \in \mathbb{B}'$ with $(b', b'') \in \mathbb{S}$. Since

$$b \in \mathrm{Sg}(\mathbb{C}' \cup \{b'\}) \subseteq \mathbb{B}' - \mathbb{S},$$

we see that there is some $b''' \in \mathbb{B}'$ such that $(b, b''') \in \mathbb{S}$. But then we have

$$\{b\} + \mathbb{S} \supseteq \mathrm{Sg}(\mathbb{C}' \cup \{b'''\}) \supseteq \mathbb{B}',$$

so $(b, b) \in \mathbb{S}$, contradicting our assumption that $\mathbb{C}' \lhd_Z \mathbb{A}_{n+1}$. $\qquad\square$

Corollary 3.10.10. *If $\mathbb{C} \lhd_Z \mathbb{A}$ and $\mathbb{A}$ is finite and idempotent, then $\mathbb{C}$ absorbs $\mathbb{A}$ with respect to some ternary term.*

Proof. If $\mathbb{C}$ does not absorb $\mathbb{A}$ with respect to any ternary term, then by Theorem 3.8.4 there is some ternary $\mathbb{C}$-essential relation $\mathbb{R} \leq \mathbb{A}^3$. By repeatedly applying the doubling trick, we see that there exists some $\mathbb{C}$-essential relation of arity $2 + 2^k$ for every $k \geq 0$, so $\mathbb{C}$ can't absorb $\mathbb{A}$ with respect to a term of any arity, contradicting the assumption $\mathbb{C} \lhd_Z \mathbb{A}$. $\qquad\square$

Corollary 3.10.11. *If $\mathbb{C}_1 \lhd_Z \mathbb{A}_1, \mathbb{B}_2 \lhd \mathbb{A}_2$, and $\mathbb{C}_3 \lhd_Z \mathbb{A}_3$ with $\mathbb{A}_i$ finite and idempotent, then no $(\mathbb{C}_1, \mathbb{B}_2, \mathbb{C}_3)$-essential relation can exist.*

Proof. If a $(\mathbb{C}_1, \mathbb{B}_2, \mathbb{C}_3)$-essential relation exists, then by repeatedly applying the doubling trick we can find $(\mathbb{C}_1, \mathbb{B}_2, ..., \mathbb{B}_2, \mathbb{C}_1)$-essential relations of arbitrarily high arity. By forcing the first and last coordinates to be in $\mathbb{C}_1$ and existentially projecting, we see that there are $\mathbb{B}_2$-essential relations of arbitrarily high arity, which contradicts the assumption $\mathbb{B}_2 \lhd \mathbb{A}_2$. $\qquad\square$

Corollary 3.10.12. *If $\mathbb{A}_i$ are finite and idempotent, $\mathbb{C}_i \lhd \mathbb{A}_i$ for all i and for all but at most one i we have $\mathbb{C}_i \lhd_Z \mathbb{A}_i$, then for any relation $\mathbb{R} \leq \mathbb{A}_1 \times \cdots \times \mathbb{A}_n$ such that $\pi_{i,j}(\mathbb{R}) \cap \mathbb{C}_i \times \mathbb{C}_j \neq \emptyset$ for all i, j, we have*
$$\mathbb{R} \cap \mathbb{C}_1 \times \cdots \times \mathbb{C}_n \neq \emptyset.$$

Proof. We show by induction on $|I|$ that for all $I \subseteq [n]$ we have
$$\pi_I(\mathbb{R}) \cap \prod_{i \in I} \mathbb{C}_i \neq \emptyset.$$

The base case $|I| \leq 2$ is our assumption. For $|I| \geq 3$, pick $i, j, k \in I$ distinct. By the induction hypothesis, there are tuples $x_i, x_i, x_k \in \mathbb{R}$ such that $\pi_{I \setminus \{i\}}(x_i) \in \prod_{i' \in I \setminus \{i\}} \mathbb{C}_{i'}$, and similarly for x_j, x_k.

Now consider the subalgebra of $\mathbb{A}_i \times \mathbb{A}_j \times \mathbb{A}_k$ generated by $\pi_{i,j,k}(x_i), \pi_{i,j,k}(x_j), \pi_{i,j,k}(x_k)$. Since this subalgebra can't be a $(\mathbb{C}_i, \mathbb{C}_j, \mathbb{C}_k)$-essential relation (since at least two of $\mathbb{C}_i, \mathbb{C}_j, \mathbb{C}_k$ are centrally absorbing and the third is absorbing), it must contain an element of $\mathbb{C}_i \times \mathbb{C}_j \times \mathbb{C}_k$. Thus there is some $x \in \mathrm{Sg}\{x_i, x_j, x_k\}$ such that
$$\pi_{i,j,k}(x) \in \mathbb{C}_i \times \mathbb{C}_j \times \mathbb{C}_k,$$
and this x automatically satisfies
$$\pi_{I \setminus \{i,j,k\}}(x) \in \prod_{i' \in I \setminus \{i,j,k\}} \mathbb{C}_{i'}$$
since each of x_i, x_j, x_k do, which completes the inductive step. $\qquad\square$

Corollary 3.10.13. *If $\mathbb{A}$ is finite and idempotent, then there is a ternary term $t \in \mathrm{Clo}_3(\mathbb{A})$ such that for all finite $\mathbb{B} \in HSP(\mathbb{A})$ and each $\mathbb{C} \lhd_Z \mathbb{B}$, $\mathbb{C}$ absorbs $\mathbb{B}$ with respect to the term t.*

Proof. For any finite collection of pairs $\mathbb{C}_i \lhd_Z \mathbb{B}_i \in HSP(\mathbb{A})$, we can apply the previous corollary to find a term $t \in \mathrm{Clo}_3(\mathbb{A})$ which simultaneously witnesses all $\mathbb{C}_i \lhd \mathbb{B}_i$. Since there are only finitely many ternary terms t of $\mathbb{A}$, some t must work for all pairs $\mathbb{C} \lhd_Z \mathbb{B} \in HSP(\mathbb{A})$. $\qquad\square$

Central absorption turns out to be a good absorption concept (in the sense of the previous section), as long as we restrict ourselves to finite idempotent algebras. Unlike previous absorption concepts, in this case it is not so easy to see that $\lhd_Z$ is compatible with pp-formulas. For this, we need to consider the basic types of pp-formulas separately. The hardest case is the case of projections.

Proposition 3.10.14. *If* $\mathbb{C} \lhd_Z \mathbb{A}$ *with* $\mathbb{A}$ *finite and idempotent, and if there is a surjective homomorphism* $\pi : \mathbb{A} \twoheadrightarrow \mathbb{B}$, *then* $\pi(\mathbb{C}) \lhd_Z \mathbb{B}$.

Proof. Suppose there is some $b \in \mathbb{B} \setminus \pi(\mathbb{C})$ such that $(b, b) \in \mathrm{Sg}(\pi(\mathbb{C}) \times \{b\} \cup \{b\} \times \pi(\mathbb{C}))$. Choose $a \in \pi^{-1}(b)$ such that the subalgebra $\mathrm{Sg}(\mathbb{C} \cup \{a\})$ is as small as possible. Set

$$\mathbb{S} = \mathrm{Sg}_{\mathbb{A}^2} \left\{ \begin{bmatrix} a \\ \mathbb{C} \end{bmatrix}, \begin{bmatrix} \mathbb{C} \\ a \end{bmatrix} \right\}.$$

By the choice of b, there exist $a', a'' \in \mathbb{A}$ such that $(a', a'') \in \mathbb{S}$ and $\pi(a') = \pi(a'') = b$. By the choice of a, we have $a \in \mathrm{Sg}(\mathbb{C} \cup \{a''\})$. Thus we have

$$\mathrm{Sg}\{a, a'\} + \mathbb{S} \supseteq \mathrm{Sg}(\mathbb{C} \cup \{a''\}) \supseteq \{a\},$$

so there is some $a''' \in \mathrm{Sg}\{a, a'\}$ with $(a''', a) \in \mathbb{S}$, and by idempotence we have $\pi(a''') = b$, so $a \in \mathrm{Sg}(\mathbb{C} \cup \{a'''\})$. By idempotence we have $\{a\} \leq \mathbb{A}$, so

$$\{a\} - \mathbb{S} \supseteq \mathrm{Sg}(\mathbb{C} \cup \{a'''\}) \supseteq \{a\},$$

so $(a, a) \in \mathbb{S}$, which contradicts the assumption $\mathbb{C} \lhd_Z \mathbb{A}$. $\qquad\square$

Proposition 3.10.15. *If* $\mathbb{C} \lhd_Z \mathbb{B} \lhd_Z \mathbb{A}$, *then* $\mathbb{C} \lhd_Z \mathbb{A}$. *As a consequence, if* $\mathbb{C}_i \lhd_Z \mathbb{B}_i \leq \mathbb{A}$, *then* $\mathbb{C}_1 \cap \mathbb{C}_2 \lhd_Z \mathbb{B}_1 \cap \mathbb{B}_2$.

Proof. Suppose there is some $a \in \mathbb{A}$ such that $(a, a) \in \mathrm{Sg}(\mathbb{C} \times \{a\} \cup \{a\} \times \mathbb{C})$. Since $\mathbb{C} \leq \mathbb{B}$ and $\mathbb{B} \lhd_Z \mathbb{A}$, we must have $a \in \mathbb{B}$. Then since $\mathbb{C} \lhd_Z \mathbb{B}$, we must have $a \in \mathbb{C}$. Thus $\mathbb{C} \lhd_Z \mathbb{A}$.

For the second statement, note that $\mathbb{C}_2 \lhd_Z \mathbb{B}_2$ implies $\mathbb{C}_1 \cap \mathbb{C}_2 \lhd_Z \mathbb{C}_1 \cap \mathbb{B}_2$ and $\mathbb{C}_1 \lhd_Z \mathbb{B}_1$ implies $\mathbb{C}_1 \cap \mathbb{B}_2 \lhd_Z \mathbb{B}_1 \cap \mathbb{B}_2$. $\qquad\square$

Proposition 3.10.16. *If* $\mathbb{C}_1 \lhd_Z \mathbb{A}_1$, *then* $\mathbb{C}_1 \times \mathbb{A}_2 \lhd_Z \mathbb{A}_1 \times \mathbb{A}_2$.

Putting these three results together, we see that central absorption is a good absorption concept.

Proposition 3.10.17. *The absorption concept* $\lhd_Z$, *restricted to finite idempotent algebras, is compatible with pp-formulas, is transitively closed, and transfers connectivity.*

Remark 3.10.1. Annoyingly, binary absorption fails to be transitively closed or compatible with pp-formulas (the intersection of two binary absorbing subalgebras might not be binary absorbing). However, if we restrict ourselves to finite idempotent algebras which are *prepared*, that is, such that $(b, b) \in \mathrm{Sg}\{(a, b), (b, a)\}$ implies that $\{a, b\}$ is a semilattice subalgebra with absorbing element b, then binary absorption becomes compatible with pp-formulas and transitively closed (see Proposition 3.2.22).

In some cases central absorption implies binary absorption. To describe a criterion for when this happens, we will exploit partial semilattice operations.

Proposition 3.10.18. *Suppose that $\mathbb{C} \lhd_Z \mathbb{A}$ and that s is any partial semilattice operation. Then $s(\mathbb{C}, \mathbb{A}) \subseteq \mathbb{C}$.*

Proof. Suppose $c \in \mathbb{C}$ and $a \in \mathbb{A}$, and let $b = s(c, a)$. Then $s(c, b) = s(b, c) = b$ by the defining property of partial semilattice operations, so $(b, b) \in \mathrm{Sg}(\{b\} \times \mathbb{C} \cup \mathbb{C} \times \{b\})$. Thus by the definition of central absorption, we have $b \in \mathbb{C}$, that is, $s(c, a) \in \mathbb{C}$. $\qquad\square$

Proposition 3.10.19. *Suppose that $\mathbb{C} \lhd_Z \mathbb{A}$ in a finite idempotent algebra $\mathbb{A}$, or just that $s(\mathbb{C}, \mathbb{A}) \subseteq \mathbb{C}$ for all partial semilattice terms $s \in \mathrm{Clo}(\mathbb{A})$. Then the following are equivalent:*

(a) $\mathbb{C}$ *binary absorbs* $\mathbb{A}$,

(b) *for all $a \in \mathbb{A} \backslash \mathbb{C}$ and all $c \in \mathbb{C}$, the subalgebra $\mathrm{Sg}\{a, c\}$ has a proper binary absorbing subalgebra,*

(c) *for all $a \in \mathbb{A}$ and all $c \in \mathbb{C}$, there is a sequence of elements $a = a_0, a_1, ..., a_n \in \mathrm{Sg}\{a, c\}$ with $a_n \in \mathbb{C}$ such that $(a_i, a_i) \in \mathrm{Sg}\{(a_{i-1}, a_i), (a_i, a_{i-1})\}$ for all i.*

If $\mathbb{A}$ is prepared, then the third condition is equivalent to the assumption that for all a and for all $c \in \mathbb{C}$, the subalgebra $\mathrm{Sg}\{a, c\}$ contains a directed path from a to $\mathbb{C}$.

Proof. To see that (a) implies (b), note that $\mathbb{C} \lhd_{bin} \mathbb{A}$ implies that $\mathbb{C} \cap \mathrm{Sg}\{a, c\} \lhd_{bin} \mathrm{Sg}\{a, c\}$. To see that (b) implies (c), we induct on the size of $\mathrm{Sg}\{a, c\}$. Let $\mathbb{B}$ be a proper binary absorbing subalgebra of $\mathrm{Sg}\{a, c\}$, and let s be a partial semilattice term that witnesses this absorption (such an s exists by Proposition 3.2.17). Then for any $b \in \mathbb{B}$ we have $s(a, b) \in \mathbb{B}$, and if we take $a_1 = s(a, b)$ then $(a_1, a_1) \in \mathrm{Sg}\{(a, a_1), (a_1, a)\}$. Let $c_1 = s(c, b)$, then $c_1 \in \mathbb{B} \cap \mathbb{C}$, and so $\mathrm{Sg}\{a_1, c_1\} \subseteq \mathbb{B} < \mathrm{Sg}\{a, c\}$, so by the inductive hypothesis we can complete this to a sequence $a_1, ..., a_n \in \mathrm{Sg}\{a_1, c_1\}$ as in (c).

Now suppose that (c) holds. For each a, c with $c \in \mathbb{C}$, we will construct a binary function f_{ac} such that $f_{ac}(a, c) \in \mathbb{C}$ and $f_{ac}(\mathbb{C}, \mathbb{A}) \subseteq \mathbb{C}$. Then by cyclically composing the functions f_{ac} together, we can produce a binary term which absorbs $\mathbb{C}$. To construct f_{ac}, we pick a sequence of partial semilattice terms s_i such that $s_i(a_{i-1}, a_i) = a_i$ as well as binary terms t_i such that $t_i(a, c) = a_i$. We set

$$f_{ac}(x, y) := s_n(\cdots s_2(s_1(x, t_1(x, y)), t_2(x, y)) \cdots, t_n(x, y)).$$

Then we have

$$f_{ac}(a, c) = s_n(\cdots s_2(s_1(a, a_1), a_2) \cdots, a_n) = a_n \in \mathbb{C}$$

and

$$f_{ac}(\mathbb{C}, \mathbb{A}) \subseteq s_n(\cdots s_2(s_1(\mathbb{C}, \mathbb{A}), \mathbb{A}) \cdots, \mathbb{A}) \subseteq \mathbb{C},$$

as required. $\qquad\square$

3.11 Binary relations in Taylor algebras: the absorption theorem and the loop lemma

In this section we'll go over two of the main results from Barto and Kozik's paper [17] about absorption, known as the "absorption theorem" and the "loop lemma". The first of these results can be used to constrain the possible subdirect binary relations in simple absorption free algebras, while the second result makes no direct mention of absorption, but combines the theory of absorbing

subalgebras with an elementary argument in the absorption free case to give a criterion for a subdirect binary relation to intersect the diagonal.

The loop lemma was originally introduced in order to settle a special case of the dichotomy problem, where the template structure **A** consists of a set together with a single subdirect binary relation (considered as a directed graph). As a bonus, the loop lemma easily implies the existence of a Taylor term of a special form, known as a *Siggers* operation, named after the first person to notice that such special Taylor terms exist in the finite case [132] (this result was quickly refined, after the initial discovery: see [84] for the paper which introduced the 4-ary operations which are now commonly known as Siggers operations).

Here is a strong form of the absorption theorem, stated in terms of Zhuk's centers.

Theorem 3.11.1 (Absorption Theorem [17]). *If $\mathbb{R} \leq_{sd} \mathbb{A} \times \mathbb{B}$ is a subdirect binary relation and $\mathbb{A}, \mathbb{B}$ are finite idempotent Taylor algebras, and if $\mathbb{R}$ is* linked, *then either*

- $\mathbb{R} = \mathbb{A} \times \mathbb{B}$,

- $\mathbb{A}$ *has a proper binary absorbing or centrally absorbing subalgebra, or*

- $\mathbb{B}$ *has a proper subalgebra which is both binary absorbing and centrally absorbing.*

The absorption theorem can be viewed as a strengthening of Zhuk's results about central relations: as we will see, it actually follows from Zhuk's result by applying a few simple tricks. First we will bootstrap to the case of a subdirect relation $\mathbb{R}$ such that $\mathbb{R} \circ \mathbb{R}^- = \mathbb{A} \times \mathbb{A}$.

Lemma 3.11.2. *If $\mathbb{R} \leq_{sd} \mathbb{A} \times \mathbb{B}$ is a subdirect binary relation and $\mathbb{A}, \mathbb{B}$ are finite idempotent Taylor algebras, and if $\mathbb{R} \circ \mathbb{R}^- = \mathbb{A} \times \mathbb{A}$, then either*

- *there is some $b \in \mathbb{B}$ such that $\mathbb{A} \times \{b\} \subseteq \mathbb{R}$,*

- $\mathbb{A}$ *has a proper binary absorbing subalgebra, or*

- *every element of $\mathbb{A}$ is contained in a proper centrally absorbing subalgebra.*

Proof. Suppose that there is no $b \in \mathbb{B}$ with $\mathbb{A} \times \{b\} \subseteq \mathbb{R}$ and that $\mathbb{A}$ is binary absorption free, and choose any $a \in \mathbb{A}$. Choose a sequence of subalgebras $\{a\} + \mathbb{R} = \mathbb{D}_0 \geq \mathbb{D}_1 \geq \cdots \geq \mathbb{D}_n$ such that each $\mathbb{D}_{i+1}$ is a proper binary absorbing subalgebra of $\mathbb{D}_i$ and such that $\mathbb{D}_n$ has no proper binary absorbing subalgebras. We will first show that $\mathbb{D}_n - \mathbb{R} = \mathbb{A}$, and then we will apply Zhuk's result (Corollary 3.10.8) to the binary relation $\mathbb{R} \cap (\mathbb{A} \times \mathbb{D}_n)$.

We will show that $\mathbb{D}_i - \mathbb{R} = \mathbb{A}$ for each i, by induction on i. Note that $\mathbb{D}_0 - \mathbb{R} = \{a\} + \mathbb{R} - \mathbb{R} = \mathbb{A}$ by the assumption $\mathbb{R} \circ \mathbb{R}^- = \mathbb{A} \times \mathbb{A}$. For the inductive step, note that since $\mathbb{D}_{i+1} \lhd_{bin} \mathbb{D}_i$, we have

$$\mathbb{D}_{i+1} - \mathbb{R} \lhd_{bin} \mathbb{D}_i - \mathbb{R} = \mathbb{A},$$

so we must have $\mathbb{D}_{i+1} - \mathbb{R} = \mathbb{A}$ since $\mathbb{A}$ has no proper binary absorbing subalgebra.

If we set $\mathbb{R}' = \mathbb{R} \cap (\mathbb{A} \times \mathbb{D}_n)$, then we have

$$\{a\} \times \mathbb{D}_n \subseteq \mathbb{R}' \leq_{sd} \mathbb{A} \times \mathbb{D}_n.$$

Thus the left center $\mathbb{C}$ of $\mathbb{R}'$ contains a. Since $\mathbb{D}_n$ is binary absorption free, we see that $\mathbb{C}$ centrally absorbs $\mathbb{A}$ by Corollary 3.10.8. If $\mathbb{C} = \mathbb{A}$, then $\mathbb{A} \times \mathbb{D}_n \subseteq \mathbb{R}$, contradicting the assumption that there is no $b \in \mathbb{B}$ with $\mathbb{A} \times \{b\} \subseteq \mathbb{R}$. $\square$

In the case where $\mathbb{A}$ has no proper binary absorbing or centrally absorbing subalgebra and $\mathbb{R}$ has a nontrivial right center, we will use the criterion developed in Proposition 3.10.19 to show that the right center of $\mathbb{R}$ must actually be a binary absorbing subalgebra of $\mathbb{B}$.

Lemma 3.11.3. *If $\mathbb{R} \leq_{sd} \mathbb{A} \times \mathbb{B}$ is a subdirect binary relation and $\mathbb{A}, \mathbb{B}$ are finite idempotent Taylor algebras, and if there is some $b \in \mathbb{B}$ such that $\mathbb{A} \times \{b\} \subseteq \mathbb{R}$, then either*

- $\mathbb{R} = \mathbb{A} \times \mathbb{B}$,

- $\mathbb{A}$ *has a proper binary absorbing or centrally absorbing subalgebra, or*

- *the right center of $\mathbb{R}$ is a proper binary absorbing subalgebra of $\mathbb{B}$.*

Proof. Let $\mathbb{C} \leq \mathbb{B}$ be the right center of $\mathbb{R}$. By Corollary 3.10.8, if $\mathbb{A}$ has no proper binary absorbing subalgebra then we have $\mathbb{C} \triangleleft_Z \mathbb{B}$. If $\mathbb{C}$ is not a binary absorbing subalgebra of $\mathbb{B}$, then by Proposition 3.10.19 there must be some $b \in \mathbb{B} \setminus \mathbb{C}$ and $c \in \mathbb{C}$ such that $\mathrm{Sg}\{b, c\}$ has no proper binary absorbing subalgebra.

Since $\mathbb{R}$ is subdirect, there is some $a \in \mathbb{A}$ such that $(a, b) \in \mathbb{R}$. Since c is in the right center of $\mathbb{R}$, we also have $\mathbb{A} \times \{c\} \subseteq \mathbb{R}$. Thus if we set $\mathbb{R}' = \mathbb{R} \cap (\mathbb{A} \times \mathrm{Sg}\{b, c\})$, then we have

$$\{a\} \times \mathrm{Sg}\{b, c\} \subseteq \mathbb{R}' \leq_{sd} \mathbb{A} \times \mathrm{Sg}\{b, c\}$$

Since b is *not* in the right center of $\mathbb{R}$, the left center of $\mathbb{R}'$ is a proper subalgebra of $\mathbb{A}$. Then since $\mathrm{Sg}\{b, c\}$ has no proper binary absorbing subalgebra, Corollary 3.10.8 shows that the left center of $\mathbb{R}'$ is a proper centrally absorbing subalgebra of $\mathbb{A}$. $\square$

Proof of the Absorption Theorem. Let $\mathbb{S} = \mathbb{R} \circ \mathbb{R}^- \leq_{sd} \mathbb{A} \times \mathbb{A}$. If $\mathbb{S} = \mathbb{A} \times \mathbb{A}$, we may apply the lemmas to see that either $\mathbb{A}$ has a proper binary absorbing or centrally absorbing subalgebra, or that $\mathbb{B}$ has a proper subalgebra which is both binary absorbing and centrally absorbing. Otherwise, by the fact that $\mathbb{R}$ is linked and the finiteness of $\mathbb{A}$ there must be some minimal $k > 1$ such that $\mathbb{S}^{\circ k} = \mathbb{A} \times \mathbb{A}$. Then we can apply the lemmas to $\mathbb{S}^{\circ(k-1)}$ to see that $\mathbb{A}$ must have either a binary absorbing or centrally absorbing subalgebra. $\square$

Corollary 3.11.4. *Let $\mathbb{A}, \mathbb{B}$ be finite idempotent Taylor algebras with no proper binary or centrally absorbing subalgebras. If $\mathbb{B}$ is simple, then every subdirect binary relation $\mathbb{R} \leq_{sd} \mathbb{A} \times \mathbb{B}$ is either the full relation or the graph of a surjective homomorphism $\mathbb{A} \twoheadrightarrow \mathbb{B}$.*

Proof. Since $\mathbb{B}$ is simple, the linking congruence of $\mathbb{R}$ on $\mathbb{B}$ is either trivial or is full. If the linking congruence of $\mathbb{R}$ on $\mathbb{B}$ is trivial, then $\mathbb{R}$ must be the graph of a surjective homomorphism $\mathbb{A} \twoheadrightarrow \mathbb{B}$. Otherwise, $\mathbb{R}$ is linked, so we can apply the Absorption Theorem 3.11.1 to see that $\mathbb{R} = \mathbb{A} \times \mathbb{B}$. $\square$

Next we switch our focus to subdirect relations $\mathbb{R} \leq_{sd} \mathbb{A} \times \mathbb{A}$. In this case, it is often appropriate to think of $\mathbb{R}$ as a digraph on the vertex set $\mathbb{A}$, and we can ask questions about whether $\mathbb{R}$ (viewed as a digraph) is weakly connected, strongly connected, whether it contains any loops, etc. To be precise, the associated digraph is the relational structure $\mathbf{R} = (A, R)$, where A is the underlying set of $\mathbb{A}$ and $R \subseteq A \times A$ is the underlying set of $\mathbb{R}$ (often I abuse notation and write $\mathbf{R} = (\mathbb{A}, \mathbb{R})$ instead of explicitly replacing $\mathbb{A}, \mathbb{R}$ with their underlying sets).

Remark 3.11.1. Note that if $\mathbb{R} \le \mathbb{A} \times \mathbb{A}$ and $\mathbb{S} \le \mathbb{B} \times \mathbb{B}$ are subpowers of $\mathbb{A}, \mathbb{B}$, then a homomorphism $\mathbb{R} \to \mathbb{S}$ and a homomorphism $\mathbf{R} \to \mathbf{S}$ of the associated digraphs $\mathbf{R} = (\mathbb{A}, \mathbb{R}), \mathbf{S} = (\mathbb{B}, \mathbb{S})$ are completely different things! The first is a homomorphism of algebraic structures, and doesn't depend on how $\mathbb{R}, \mathbb{S}$ are represented as collections of ordered pairs of elements in $\mathbb{A}$ or $\mathbb{B}$ (but does depend on how the algebraic operations behave). The second is a digraph homomorphism, which ignores the algebraic structure, and is completely determined by a map $A \to B$ of the underlying sets of $\mathbb{A}, \mathbb{B}$ which is compatible with the digraph structures $\mathbb{R}, \mathbb{S}$.

In the context of digraphs, the case of a *subdirect* relation $\mathbb{R} \le_{sd} \mathbb{A} \times \mathbb{A}$ is actually rather special. The assumption $\pi_1(\mathbb{R}) = \mathbb{A}$ means that every vertex of the digraph $\mathbf{R}$ has outdegree at least one, and the assumption $\pi_2(\mathbb{R}) = \mathbb{A}$ means that every vertex of $\mathbb{R}$ has indegree at least one.

Definition 3.11.5. A digraph $\mathbf{D} = (V, E)$ is called *smooth* if every vertex of $\mathbf{D}$ has indegree at least one and outdegree at least one. Note that this is equivalent to the relation $E \subseteq V \times V$ being subdirect.

If a digraph is not smooth, it is often desirable to find a smooth digraph within it. The natural thing to do is to simply prune all of the vertice with indegree 0 or outdegree 0. Unfortunately, after this pruning step we may find ourselves with more vertices that need to be pruned, and so on - possibly ending up with no vertices at all! For instance, this actually occurs if our initial digraph is a finite directed path. Additionally, it may not be clear that these pruning operations are compatible with the algebraic structures which we started with. Luckily, there is a standard way to describe the result of this pruning process via a primitive positive formula, as well as a simple criterion for when the pruned digraph will be nonempty.

Proposition 3.11.6. *If* $\mathbf{D} = (V, E)$ *is a digraph, then the largest smooth digraph* $\mathbf{D}_{sm}$ *which is contained in* $\mathbf{D}$ *is exactly the set of vertices* v *of* $\mathbf{D}$ *such that there exists a bi-infinite directed walk through* v. *If* $\mathbf{D}$ *is finite, with n vertices, then the vertex set of* $\mathbf{D}_{sm}$ *may be defined by the pp-formula*

$$v \in \mathbf{D}_{sm} \iff \exists v_{-n}, ..., v_n \ (v_0 = v) \wedge \bigwedge_{-n \le i < n} (v_i, v_{i+1}) \in E.$$

The set $\mathbf{D}_{sm}$ *will be nonempty iff* $\mathbf{D}$ *contains a directed cycle (or a bi-infinite directed path, in the infinite case).*

Definition 3.11.7. If $\mathbf{D}$ is a digraph and $\mathbf{D}_{sm}$ is defined as in the previous proposition, then we call $\mathbf{D}_{sm}$ the *smooth part* of the digraph $\mathbf{D}$.

Note that the smooth part of a digraph may contain vertices which are not themselves part of any directed cycles: it may also contain intermediate vertices along directed paths connecting two directed cycles. In fact, the smooth part of a digraph enjoys the following convexity property.

Proposition 3.11.8. *If* $\mathbf{D}$ *is a digraph and* a, b *are in the smooth part of* $\mathbf{D}$, *then every vertex of* $\mathbf{D}$ *which can be found along any directed path from a to b is also contained in the smooth part of* $\mathbf{D}$.

One reason for introducing this terminology is that it lets us easily state results such as the following one.

Proposition 3.11.9. *If $\mathbb{S} \lhd \mathbb{R}$ and $\mathbb{R}, \mathbb{S} \le \mathbb{A} \times \mathbb{A}$ correspond to digraphs $\mathbf{R}, \mathbf{S}$ with vertex set $\mathbb{A}$, and if $\mathbf{R}$ is smooth, then the smooth part $\mathbf{S}_{sm}$ of the digraph $\mathbf{S}$ has vertex set equal to an absorbing subalgebra of $\mathbb{A}$, which will be nonempty as long as $\mathbf{S}$ contains some directed cycle.*

Of course, we will often abuse notation a little further, and talk about the "smooth part of the digraph $\mathbb{S}$" as long as this does not seem likely to cause confusion. It will be convenient to have the following criterion for the existence of a directed cycle contained in a subalgebra $\mathbb{B} \le \mathbb{A}$.

Proposition 3.11.10. *If $\mathbb{R} \le \mathbb{A} \times \mathbb{A}$, and if $\mathbb{B} \le \mathbb{A}$ is finite and satisfies either $\mathbb{B} \subseteq \mathbb{B} + \mathbb{R}$ or $\mathbb{B} \subseteq \mathbb{B} - \mathbb{R}$, then the restriction $\mathbb{R} \cap (\mathbb{B} \times \mathbb{B})$ of $\mathbb{R}$ to $\mathbb{B}$ has nonempty smooth part.*

Proof. Suppose that $\mathbb{B} \subseteq \mathbb{B} - \mathbb{R}$. Then every vertex in $\mathbb{B}$ has a an edge leaving it which lands in $\mathbb{B}$, so we can find an arbitrarily long directed walk of $\mathbb{R}$ which is entirely contained in $\mathbb{B}$. Since $\mathbb{B}$ is finite, this implies that there is some directed cycle which is entirely contained in $\mathbb{B}$. $\qquad\square$

As a warmup to the full loop lemma, we will first focus on the special case where the relation $\mathbb{R}$ is linked. This special case is usually enough to handle most applications.

Lemma 3.11.11 (Loop Lemma, linked case)**.** *Suppose that $\mathbb{A}$ is a finite Taylor algebra and that $\mathbb{R} \le_{sd} \mathbb{A} \times \mathbb{A}$ is a linked subdirect relation. Then $\mathbb{R}$ contains a loop, that is, $\mathbb{R} \cap \Delta_{\mathbb{A}} \ne \emptyset$.*

Proof. We prove this by induction on $|\mathbb{A}|$. We may assume that $\mathbb{A}$ is idempotent without loss of generality. If $\mathbb{R} \ne \mathbb{A} \times \mathbb{A}$, then $\mathbb{A}$ must have some proper absorbing subalgebra $\mathbb{B} \lhd \mathbb{A}$ by the Absorption Theorem 3.11.1. If we define a sequence of absorbing subalgebras $\mathbb{B} = \mathbb{B}_0, \mathbb{B}_1, ...$ of $\mathbb{A}$ by $\mathbb{B}_{i+1} = \mathbb{B}_i + \mathbb{R}$ for i even and $\mathbb{B}_{i+1} = \mathbb{B}_i - \mathbb{R}$ for i odd, then since $\mathbb{R}$ is linked and $\mathbb{A}$ is finite there must be some i such that $\mathbb{B}_{i+1} = \mathbb{A}$ but $\mathbb{B}_i \ne \mathbb{A}$. Since this $\mathbb{B}_i$ satisfies $\mathbb{B}_i \subseteq \mathbb{A} = \mathbb{B}_{i+1}$, we see that either $\mathbb{B}_i \subseteq \mathbb{B}_i + \mathbb{R}$ or $\mathbb{B}_i \subseteq \mathbb{B}_i - \mathbb{R}$, so by the previous proposition the relation $\mathbb{R} \cap (\mathbb{B}_i \times \mathbb{B}_i)$ has a nonempty smooth part $\mathbb{B}_{sm} \lhd \mathbb{B}_i$, with edge set $\mathbb{S} = \mathbb{R} \cap (\mathbb{B}_{sm} \times \mathbb{B}_{sm})$.

Since $\mathbb{B}_{sm} \lhd \mathbb{A}$, we have $\mathbb{S} \lhd \mathbb{R}$. Since $\mathbb{S}$ is smooth, we can transfer the linkedness of $\mathbb{R}$ to $\mathbb{S}$ using Theorem 3.7.12, to see that $\mathbb{S}$ must also be linked. By the inductive hypothesis applied to $\mathbb{B}_{sm}$, we see that $\mathbb{S}$ must contain a loop, and this loop will also be contained in $\mathbb{R}$ since $\mathbb{S} \le \mathbb{R}$. $\qquad\square$

To state the full loop lemma, we need another digraph concept.

Definition 3.11.12. The *algebraic length* of a weakly connected digraph $\mathbf{D}$ is the least common multiple of all integers k such that there is a digraph homomorphism from $\mathbf{D}$ to a directed cycle of length k.

Proposition 3.11.13. *The algebraic length of a weakly connected digraph $\mathbf{D} = (V, E)$ is the greatest common divisor of all integers k such that there exist $v \in V$ and $k_1, k_2, ..., k_m \in \mathbb{N}$ such that*

$$v \in \{v\} + k_1 E - k_2 E + \cdots \pm k_m E$$

and $k = k_1 - k_2 + \cdots \pm k_m$.

Furthermore, there exists a digraph homomorphism from $\mathbf{D}$ to a directed cycle $\mathbf{C}$ iff the algebraic length of $\mathbf{D}$ is a multiple of the length of the cycle $\mathbf{C}$.

Proposition 3.11.14. *If $\mathbf{D} = (V, E)$ is a smooth, weakly connected digraph of algebraic length k, then the digraph $\mathbf{D}^{\circ m} = (V, E^{\circ m})$ has $\gcd(k, m)$ weakly connected components, and each weakly connected component of $\mathbf{D}^{\circ m}$ has algebraic length $\frac{k}{\gcd(k,m)}$.*

194

Proposition 3.11.15. *If* $\mathbf{D} = (V, E)$ *is smooth and weakly connected, then* $\mathbf{D}$ *has algebraic length* 1 *if and only if there is some* $m \geq 0$ *such that the relation* $E^{\circ m}$ *is linked.*

Corollary 3.11.16. *If* $\mathbf{D} = (V, E)$ *is smooth and has a weakly connected component* $C \subseteq V$ *of algebraic length* 1*, and if* $v \in C$*, then the set* C *can be defined by a primitive positive formula using the singleton unary relation* $\{v\}$ *and the binary relation* E.

With these preliminaries out of the way, we can finally state the full version of the loop lemma for finite Taylor algebras.

Theorem 3.11.17 (Loop Lemma [17])**.** *If* $\mathbb{A}$ *is a finite Taylor algebra and* $\mathbb{R} \leq_{sd} \mathbb{A} \times \mathbb{A}$ *corresponds to a smooth digraph* $\mathbf{R} = (\mathbb{A}, \mathbb{R})$ *which has a weakly connected component of algebraic length* 1*, then* $\mathbb{R}$ *has a loop, i.e.* $\mathbb{R} \cap \Delta_{\mathbb{A}} \neq \emptyset$.

Proof. We prove this by induction on $|\mathbb{A}|$. We may assume that $\mathbb{A}$ is idempotent without loss of generality. We may also assume that $\mathbf{R}$ is weakly connected by restricting to a weakly connected component of algebraic length 1 (which forms a subalgebra of $\mathbb{A}$ by the results above). Let m be minimal such that $\mathbb{R}^{\circ m}$ is linked. We split into cases based on whether $\mathbb{R}^{\circ m} = \mathbb{A} \times \mathbb{A}$ or not.

If $\mathbb{R}^{\circ m} \neq \mathbb{A} \times \mathbb{A}$, then by the Absorption Theorem 3.11.1 we see that $\mathbb{A}$ must have some proper absorbing subalgebra. By a similar argument to the linked case (Lemma 3.11.11), we see that there is some proper absorbing $\mathbb{B} \lhd \mathbb{A}$ such that $\mathbb{S} = \mathbb{R} \cap (\mathbb{B} \times \mathbb{B})$ is subdirect in $\mathbb{B} \times \mathbb{B}$. Then since $\mathbb{S}^{\circ m} \lhd \mathbb{R}^{\circ m}$, we can apply Theorem 3.7.12 to see that $\mathbb{S}^{\circ m}$ is linked, so the smooth digraph $\mathbf{S} = (\mathbb{A}, \mathbb{S})$ has algebraic length 1 and $\mathbb{S}$ has a loop by the inductive hypothesis.

If $\mathbb{R}^{\circ m} = \mathbb{A} \times \mathbb{A}$, then we let $\mathbb{B}$ be any linked component of $\mathbb{R}^{\circ(m-1)}$ (note that $\mathbb{R}^{\circ(m-1)}$ is not linked by the choice of m, so $\mathbb{B}$ is a proper subalgebra of $\mathbb{A}$). First we will show that $\mathbb{B} \subseteq \mathbb{B} - \mathbb{R}$. To see this, let $b \in \mathbb{B}$ be arbitrary, pick any $c \in b + \mathbb{R}^{\circ(m-1)}$. Then since $\mathbb{R}^{\circ m} = \mathbb{A} \times \mathbb{A}$, we have

$$c \in b + \mathbb{R}^{\circ m},$$

and if we let d be the first element along a directed path of length m from b to c, then we have

$$d \in (b + \mathbb{R}) \cap (c - \mathbb{R}^{\circ(m-1)}) \subseteq (b + \mathbb{R}) \cap (b + \mathbb{R}^{\circ(m-1)} - \mathbb{R}^{\circ(m-1)}) \subseteq (b + \mathbb{R}) \cap \mathbb{B}.$$

Thus $b \in \mathbb{B} - \mathbb{R}$, and since b was an arbitrary element of $\mathbb{B}$ we see that $\mathbb{B} \subseteq \mathbb{B} - \mathbb{R}$. Thus the smooth part $\mathbb{B}_{sm}$ of $\mathbb{R} \cap (\mathbb{B} \times \mathbb{B})$ is nonempty.

To finish, we just need to check that the smooth digraph corresponding to $\mathbb{S} = \mathbb{R} \cap (\mathbb{B}_{sm} \times \mathbb{B}_{sm})$ has algebraic length 1. For this, we pick any $(b, c) \in \mathbb{S}^{\circ(m-1)}$, and pick any directed path $b = b_1, ..., b_m = c$ of length $m - 1$ with all $b_i \in \mathbb{B}_{sm}$. Since $(b, c) \in \mathbb{R}^{\circ m}$, we may also find directed path $b = c_0, ..., c_m = c$ from b to c of length m in $\mathbf{R}$. We will show that every c_i along this path is actually in $\mathbb{B}_{sm}$. For this, we just note that b_i, c_i are in the same linked component of $\mathbb{R}^{\circ(m-1)}$ for each $i \geq 1$ (since b_i and c_i can both reach c in exactly $m - i$ steps), so each c_i is at least in $\mathbb{B}$, and then since each c_i is along a directed path between two vertices of $\mathbb{B}_{sm}$ we see that each c_i belongs to the smooth part $\mathbb{B}_{sm}$ as well. Thus $b \in b + \mathbb{S}^{\circ(m-1)} - \mathbb{S}^{\circ m}$, so $\mathbb{S}$ has algebraic length 1 and we may apply the inductive hypothesis to see that $\mathbb{S}$ contains a loop. $\square$

**Corollary 3.11.18** (Siggers term [132], [84])**.** *If* $\mathbb{A}$ *is a finite Taylor algebra, then* $\mathbb{A}$ *has a 4-ary idempotent term* t *which satisfies the identity*

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

Proof. Assume without loss of generality that $\mathbb{A}$ is idempotent. Let $\mathbb{F} = \mathcal{F}_\mathbb{A}(x, y, z)$ be the free algebra on three generators in the variety generated by $\mathbb{A}$. Let $\mathbb{R}$ be the binary relation

$$\mathbb{R} = \mathrm{Sg}_{\mathbb{F}^2}\left\{ \begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} x \\ z \end{bmatrix}, \begin{bmatrix} y \\ z \end{bmatrix}, \begin{bmatrix} z \\ x \end{bmatrix} \right\}.$$

Then $\mathbb{R}$ is clearly subdirect, and the generating set of $\mathbb{R}$ forms the binary relation on $\{x, y, z\}$ pictured below, as both a bipartite graph and as a digraph.
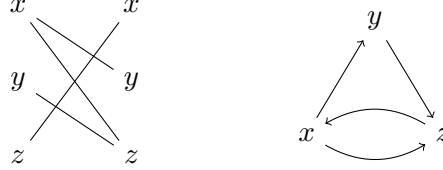
This digraph is smooth, strongly connected (in fact, it has $x + \mathbb{R}^{\circ 3} = \mathbb{F}$ and $\mathbb{R}^{\circ 5} = \mathbb{F} \times \mathbb{F}$), and has algebraic length 1 (since $x \in x + \mathbb{R}^{\circ 2} - \mathbb{R}^{\circ 1}$), so we can apply the Loop Lemma to see that $\mathbb{R}$ contains some loop (f, f) (we are using here the fact that $\mathbb{F} \le \mathbb{A}^{\mathbb{A}^3}$ is finite and Taylor). Then since $(f, f) \in \mathbb{R}$, there must be some 4-ary term t such that

$$t\left(\begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} x \\ z \end{bmatrix}, \begin{bmatrix} y \\ z \end{bmatrix}, \begin{bmatrix} z \\ x \end{bmatrix} \right) = \begin{bmatrix} f \\ f \end{bmatrix},$$

and this t then satisfies the identity

$$t(x, x, y, z) = f = t(y, z, z, x). \qquad \square$$

Remark 3.11.2. Suppose that t is a Siggers term, i.e. that $t(x, x, y, z) \approx t(y, z, z, x)$. If we substitute $y = z$ into the Siggers identity and rename variables, we see that

$$t(y, y, x, x) \approx t(x, x, x, y),$$

and if we substitute $x = y$ into the Siggers identity and rename variables, then we get

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

Thus there is some binary term $f(x, y)$ such that

$$t\left(\begin{bmatrix} y & y & x & x \\ x & y & y & x \\ x & x & x & y \end{bmatrix} \right) \approx \begin{bmatrix} f(x, y) \\ f(x, y) \\ f(x, y) \end{bmatrix}.$$

If we reorder the first and second inputs to t, the left hand side exactly becomes the left hand side of the equation for the 3-edge term. If $f(x, y)$ was equal to x, then t would become a 3-edge term (up to reordering inputs).

If $f(x, y)$ was instead equal to y, then $p(x, y, z) = t(x, x, y, z)$ would become a Mal'cev term, which is even better than a 3-edge term. However, if we allow for the possibility of semilattice subalgebras, then $f(x, y)$ must act as the semilattice operation on any two-element semilattice subalgebra, and of course in this case there couldn't possibly be any cube term of any arity. For this reason, the system of equations satisfied by t above are often summarized by calling such a t a "weak 3-edge term".

The fact that a Siggers term looks suspiciously similar to a 3-edge term is more than a coincidence: later we will see that every finite Taylor algebra either has a 3-edge term or has some pair of elements $a \ne b$ such that $(b, b) \in \mathrm{Sg}\{(a, b), (b, a)\}$.

3.12 Finite abelian Taylor algebras are affine, and Zhuk's four cases

First we recall the definition of an abelian algebra.

Definition 3.12.1. An algebraic structure $\mathbb{A}$ is called *abelian* if there is a congruence Θ on $\mathbb{A} \times \mathbb{A}$ such that the diagonal $\Delta_{\mathbb{A}} = \{(a,a) \mid a \in \mathbb{A}\}$ is one of the congruence classes of Θ.

The reader might be sceptical about how often such a congruence Θ actually shows up. After all, such a congruence is most naturally viewed as a 4-ary relation on $\mathbb{A}$, and for the most part we have only been able to prove interesting structural results about binary relations so far. The next result illustrates the most common situation which leads to the existence of such a congruence.

Proposition 3.12.2. *Suppose that* $\mathbb{R} \leq_{sd} \mathbb{A} \times \mathbb{A} \times \mathbb{A}$ *has the property that for each* $a \in \mathbb{A}$, *and for each permutation* (i,j,k) *of* $(1,2,3)$, *the binary relation*

$$\pi_{ij}(x \in \mathbb{R} \wedge x_k = a)$$

is the graph of an automorphism of $\mathbb{A}$. *Then* $\mathbb{A}$ *is abelian.*

Proof. Note that the assumption on $\mathbb{R}$ can be rephrased as saying that if we fix any pair of coordinates of a tuple in $\mathbb{R}$, then the last coordinate is uniquely determined. Therefore $\mathbb{R}$ can be viewed as the graph of a homomorphism

$$m : \mathbb{A} \times \mathbb{A} \twoheadrightarrow \mathbb{A}$$

such that the preimage $m^{-1}(a)$ is the graph of an automorphism of $\mathbb{A}$ for every $a \in \mathbb{A}$ (equivalently, m is the multiplication of some quasigroup which commutes with the operations of $\mathbb{A}$). In other words, every congruence class of the kernel $\ker m \in \mathrm{Con}(\mathbb{A} \times \mathbb{A})$ is the graph of an automorphism of $\mathbb{A}$. Twisting $\ker m$ by one of these automorphisms yields a congruence $\Theta \in \mathrm{Con}(\mathbb{A} \times \mathbb{A})$ such that one of its congruence classes is the graph of the identity permutation of $\mathbb{A}$. □

The proof we give in this section - following [21] - of the fact that finite abelian Taylor algebras are affine breaks into three steps:

- every finite abelian algebra is (hereditarily) absorption free,

- every finite, idempotent, Taylor, hereditarily absorption free algebra is Mal'cev, and

- every abelian Mal'cev algebra is affine.

We have already completed the third step in Section 1.9, Theorem 1.9.23. We will complete the remaining steps in reverse order as well.

Definition 3.12.3. We say that an algebra $\mathbb{A}$ is *hereditarily absorption free* if every subalgebra of $\mathbb{A}$ is absorption free, that is, if $\mathbb{C} \lhd \mathbb{B} \leq \mathbb{A}$ implies that $\mathbb{C} = \mathbb{B}$ or $\mathbb{C} = \emptyset$.

Proposition 3.12.4. *Suppose* $\mathbb{A}, \mathbb{B}$ *are idempotent and hereditarily absorption free. Then* $\mathbb{A} \times \mathbb{B}$ *is also hereditarily absorption free.*

Proof. Suppose that $\mathbb{S} \lhd \mathbb{R} \le \mathbb{A} \times \mathbb{B}$, with $\mathbb{S} \ne \emptyset$. Then since $\pi_1(\mathbb{S}) \lhd \pi_1(\mathbb{R}) \le \mathbb{A}$ and $\mathbb{A}$ is hereditarily absorption free, we see that $\pi_1(\mathbb{S}) = \pi_1(\mathbb{R})$. Thus for every $a \in \pi_1(\mathbb{R})$ we have $a + \mathbb{S} \ne \emptyset$, and since $\mathbb{A}$ is idempotent, we have

$$a + \mathbb{S} \lhd a + \mathbb{R} \le \mathbb{B}.$$

Then since $\mathbb{B}$ is hereditarily absorption free, we see that $a + \mathbb{S} = a + \mathbb{R}$. Since a was an arbitrary element of $\pi_1(\mathbb{R})$, we have $\mathbb{S} = \mathbb{R}$. $\qquad\square$

Theorem 3.12.5 (HAF implies Mal'cev [21])**.** *If $\mathbb{A}$ is finite, idempotent, Taylor, and hereditarily absorption free, then $\mathbb{A}$ is Mal'cev.*

Proof. By repeatedly applying the previous proposition, we see that the free algebra on two generators $\mathbb{F} = \mathcal{F}_\mathbb{A}(x, y) \le \mathbb{A}^{\mathbb{A}^2}$ is absorption free. Consider the binary relation $\mathbb{R} \le_{sd} \mathbb{F} \times \mathbb{F}$ defined by

$$\mathbb{R} = \mathrm{Sg}_{\mathbb{F}^2} \left\{ \begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} x \\ x \end{bmatrix}, \begin{bmatrix} y \\ x \end{bmatrix} \right\}.$$

Then $x + \mathbb{R} \supseteq \mathrm{Sg}_\mathbb{F}\{x, y\} = \mathbb{F}$, so x is contained in the left center of $\mathbb{R}$. Thus by the Absorption Theorem 3.11.1 (or just Zhuk's result Corollary 3.10.8) we must have $\mathbb{R} = \mathbb{F} \times \mathbb{F}$, and in particular $(y, y) \in \mathbb{R}$. Thus there is some ternary term p such that

$$p\left(\begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} x \\ x \end{bmatrix}, \begin{bmatrix} y \\ x \end{bmatrix} \right) = \begin{bmatrix} y \\ y \end{bmatrix}. \qquad\square$$

To finish the proof that finite abelian Taylor algebras are affine, we just need to check that every abelian algebra is absorption free. Note that every subalgebra of an abelian algebra is also abelian, so this will imply that abelian algebras are *hereditarily* absorption free as well. Additionally, every reduct of an abelian algebra is also abelian (since taking reducts can only increase the congruence lattice), so we see that the idempotent reduct of a finite abelian Taylor algebra will also be hereditarily absorption free, allowing us to apply the previous result to it.

It is not so easy to see how to use abelianness to rule out absorption. As a warmup, we will show that abelian algebras can't have any near-unanimity terms: this will give us the hint about how to show that finite abelian algebras are absorption free.

Proposition 3.12.6. *If an algebra $\mathbb{A}$ is abelian and has at least two elements, then $\mathbb{A}$ does not have a near-unanimity term.*

Proof. Let $\Theta \in \mathrm{Con}(\mathbb{A} \times \mathbb{A})$ be a congruence with the diagonal $\Delta_\mathbb{A}$ as a congruence class. Suppose for contradiction that t is a near-unanimity term of minimal arity n, and note that n must be at least 3 since $\mathbb{A}$ has at least two elements. Let a, b be any pair of elements of $\mathbb{A}$. Then we have

$$t\left(\begin{bmatrix} a & b & b & \cdots & b \\ b & b & b & \cdots & b \end{bmatrix} \right) = \begin{bmatrix} b \\ b \end{bmatrix} \in \Delta_\mathbb{A}.$$

Since the second column of inputs to t is $(b, b) \in \Delta_\mathbb{A}$, we can replace it with any other element of $\Delta_\mathbb{A}$ without changing the the result modulo Θ. Thus we have

$$t\left(\begin{bmatrix} a & a & b & \cdots & b \\ b & a & b & \cdots & b \end{bmatrix} \right) \equiv_\Theta \begin{bmatrix} b \\ b \end{bmatrix} \in \Delta_\mathbb{A}.$$

Since $t(b, a, b, ..., b) = b$, we see that we must have

$$t\left(\begin{bmatrix} a & a & b & \cdots & b \\ b & a & b & \cdots & b \end{bmatrix}\right) = \begin{bmatrix} b \\ b \end{bmatrix}.$$

Since a, b were arbitrary elements of $\mathbb{A}$, we see that

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

so the term $t(x, x, y_2, ..., y_{n-1})$ is a near-unanimity term of arity $n - 1$, contradicting the choice of t. $\square$

In order to mimic this argument to rule out absorption, we will need to assume finiteness of $\mathbb{A}$ and apply an iteration argument.

Theorem 3.12.7 (Abelian implies HAF [21])**.** *If a finite algebra $\mathbb{A}$ is abelian, then it is absorption free.*

Proof. Let $\Theta \in \mathrm{Con}(\mathbb{A} \times \mathbb{A})$ be a congruence with the diagonal $\Delta_{\mathbb{A}}$ as a congruence class. Suppose for contradiction that $\mathbb{B} \lhd \mathbb{A}$ is nonempty and proper, and let t be a term of minimal arity n among those which absorb $\mathbb{B}$. Note that $n \geq 2$ since $\mathbb{B}$ is a proper subalgebra of $\mathbb{A}$. Now iterate t on its first argument, i.e. define a sequence of terms t_i with $t_1 = t$ and

$$t_{i+1}(x, y_1, ..., y_{n-1}) := t(t_i(x, y_1, ..., y_{n-1}), y_1, ..., y_{n-1}).$$

By induction on i, each t_i absorbs $\mathbb{B}$. Since $\mathbb{A}$ is finite, there is some i such that $t_i = t_{2i}$, call this t_i t_∞. Then we have

$$t_\infty(t_\infty(x, y_1, ..., y_{n-1}), y_1, ..., y_{n-1}) \approx t_\infty(x, y_1, ..., y_{n-1}),$$

and t_∞ absorbs $\mathbb{B}$.

Now we argue as in the near-unanimity case: let $a \in \mathbb{A}$ and $b_1, b_2, ..., b_{n-1} \in \mathbb{B}$, and set

$$b = t_\infty(a, b_1, b_2, ..., b_{n-1}) \in \mathbb{B}.$$

Then we have

$$t_\infty\left(\begin{bmatrix} a & b_1 & b_2 & \cdots & b_{n-1} \\ b & b_1 & b_2 & \cdots & b_{n-1} \end{bmatrix}\right) = \begin{bmatrix} b \\ b \end{bmatrix} \in \Delta_{\mathbb{A}},$$

so since $(b_1, b_1) \equiv_\Theta (a, a)$, we have

$$t_\infty\left(\begin{bmatrix} a & a & b_2 & \cdots & b_{n-1} \\ b & a & b_2 & \cdots & b_{n-1} \end{bmatrix}\right) \equiv_\Theta \begin{bmatrix} b \\ b \end{bmatrix} \in \Delta_{\mathbb{A}}.$$

Thus since $\Delta_{\mathbb{A}}$ is a congruence class of Θ and $\mathbb{B}$ absorbs $\mathbb{A}$ with respect to t_∞, we have

$$t_\infty(a, a, b_2, ..., b_{n-1}) = t_\infty(b, a, b_2, ..., b_{n-1}) \in \mathbb{B}.$$

Since a was an arbitrary element of $\mathbb{A}$ and $b_2, ..., b_{n-1}$ were arbitrary elements of $\mathbb{B}$, we see that the term

$$t_\infty(x, x, y_2, ..., y_{n-1})$$

absorbs $\mathbb{B}$ and has arity $n - 1$, contradicting the choice of t. $\square$

Now we can put all the pieces together and get our main result.

Theorem 3.12.8 (Fundamental Theorem of Abelian Algebras, finite Taylor case [74], [21], [133], [145]). *If $\mathbb{A}$ is a finite abelian Taylor algebra, then $\mathbb{A}$ is affine.*

Proof. Let $\mathbb{A}^{id}$ be the idempotent reduct of $\mathbb{A}$, note that $\mathbb{A}^{id}$ is still abelian and Taylor (since Taylor terms are idempotent by definition). Then every subalgebra of $\mathbb{A}^{id}$ is also abelian, so by Theorem 3.12.7 $\mathbb{A}^{id}$ is hereditarily absorption free. Since $\mathbb{A}^{id}$ is finite, idempotent, Taylor, and hereditarily absorption free it has a Mal'cev term p by Theorem 3.12.5. Then p is also a Mal'cev term of $\mathbb{A}$, so we can apply Theorem 1.9.23 to see that $\mathbb{A}$ is affine. $\square$

Remark 3.12.1. It is not hard to generalize Theorem 3.12.7 to show that if a finite algebra $\mathbb{A}$ is solvable, then $\mathbb{A}$ is hereditarily absorption free. Thus finite solvable Taylor algebras are also Mal'cev by Theorem 3.12.5.

Now we can apply the fundamental theorem of abelian algebras to further constrain relations on absorption free algebras.

Theorem 3.12.9 (Zhuk [142]). *Suppose that $\mathbb{A}$ is finite, simple, idempotent, Taylor, absorption free, and* not *affine. Then every subdirect relation $\mathbb{R} \leq_{sd} \mathbb{A}^n$ is the intersection of its binary projections, each of which is either a full relation or the graph of an automorphism of $\mathbb{A}$.*

In fact, we can weaken the assumption that $\mathbb{A}$ is absorption free to the assumption that $\mathbb{A}$ has no binary or centrally absorbing subalgebras.

Proof. We call a subdirect relation $\mathbb{R} \leq \mathbb{A}^n$ *irredundant* if no $\pi_{ij}(\mathbb{R})$ is the graph of an automorphism of $\mathbb{A}$. We will prove by induction on n that every irredundant subdirect relation on $\mathbb{A}$ is the full relation.

The base cases of the induction are the cases $n = 1, 2, 3$. The case $n = 1$ is trivial (a unary subdirect relation must be full). The case $n = 2$ follows from the Absorption Theorem 3.11.1, since every subdirect binary relation on $\mathbb{A}$ is either the graph of an automorphism of $\mathbb{A}$, or is linked (since $\mathbb{A}$ is simple) and therefore is equal to the full relation (since $\mathbb{A}$ has no binary or centrally absorbing subalgebras). For the case $n = 3$, note that for any $a \in \mathbb{A}$, the binary relation

$$\mathbb{R}^a := \pi_{12}(\mathbb{R} \cap (\mathbb{A}^2 \times \{a\}))$$

is subdirect, so $\mathbb{R}^a$ is either the graph of an automorphism or is equal to $\mathbb{A}^2$. If there is any $a \in \mathbb{A}$ such that $\mathbb{R}^a = \mathbb{A}^2$, then a is contained in the right center of $\mathbb{R}$, considered as a binary relation on $(\mathbb{A}^2) \times \mathbb{A}$, so $\mathbb{R} = \mathbb{A}^3$ by the Absorption Theorem 3.11.1 (or just Corollary 3.10.8). Otherwise every $\mathbb{R}^a$ is the graph of an automorphism, and a similar argument applies if we permute the coordinates of $\mathbb{R}$, so we may apply Proposition 3.12.2 to see that $\mathbb{A}$ is abelian. But then by the fundamental theorem of abelian algebras 3.12.8 we see that $\mathbb{A}$ is affine, which contradicts our assumptions.

For the induction step, assume that $n > 3$. Then for every pair of distinct $i, j \leq n - 1$, the ternary relation $\pi_{ijn}(\mathbb{R})$ is full by the $n = 3$ case, so for every $a \in \mathbb{A}$, the binary relation

$$\pi_{ij}(\mathbb{R} \cap (\mathbb{A}^{n-1} \times \{a\}))$$

is the full relation $\mathbb{A}^2$. Thus the relation

$$\mathbb{R}^a := \pi_{[n-1]}(\mathbb{R} \cap (\mathbb{A}^{n-1} \times \{a\}))$$

is irredundant, so by the inductive hypothesis, $\mathbb{R}^a$ is the full relation $\mathbb{A}^{n-1}$ for every $a \in \mathbb{A}$. In other words, $\mathbb{R}$ is the full relation $\mathbb{A}^n$. $\square$

Since the conclusion of Theorem 3.12.9 is actually much stronger than merely being polynomially complete, we will give it a special name.

Definition 3.12.10. We say that an algebra $\mathbb{A}$ is *subdirectly complete* if every subdirect relation $\mathbb{R} \leq_{sd} \mathbb{A}^n$ is the intersection of its binary projections, each of which is either a full relation or the graph of an automorphism of $\mathbb{A}$.

Proposition 3.12.11. *Every subdirectly complete finite algebra is polynomially complete.*

Corollary 3.12.12 (Zhuk's four cases [142])**.** *If $\mathbb{A}$ is a nontrivial finite idempotent Taylor algebra, then at least one of the following is true.*

- *$\mathbb{A}$ has a proper binary absorbing subalgebra,*

- *$\mathbb{A}$ has a proper centrally absorbing subalgebra,*

- *$\mathbb{A}$ has a nontrivial affine quotient, or*

- *$\mathbb{A}$ has a nontrivial subdirectly complete quotient.*

Proof. Let $\theta \in \mathrm{Con}(\mathbb{A})$ be a maximal congruence on $\mathbb{A}$, so $\mathbb{A}/\theta$ is simple. If $\mathbb{A}/\theta$ has a proper binary or centrally absorbing subalgebra $\mathbb{B}$, then the preimage of $\mathbb{B}$ under the projection $\mathbb{A} \twoheadrightarrow \mathbb{A}/\theta$ is a proper binary or centrally absorbing subalgebra of $\mathbb{A}$. Otherwise, Theorem 3.12.9 shows that if $\mathbb{A}/\theta$ is not affine, then it is subdirectly complete. $\qquad\square$

Remark 3.12.2. For the sake of proving Theorem 3.12.9 and Corollary 3.12.12, we only need to show that if $\mathbb{A}$ is a finite idempotent Taylor algebra with a ternary relation $\mathbb{R} \leq_{sd} \mathbb{A}^3$ as in Proposition 3.12.2, then $\mathbb{A}$ is affine. It's possible to give a direct argument for this, as follows.

First, we reinterpret $\mathbb{R}$ as the graph of a quasigroup operation $\cdot : \mathbb{A} \times \mathbb{A} \to \mathbb{A}$. Using this quasigroup operation $\cdot$, we can define a Mal'cev operation $p : \mathbb{A}^3 \to \mathbb{A}$ which is centralized by the clone of $\mathbb{A}$, such that p is invertible in its first and last variables. We then pick any element $0 \in \mathbb{A}$, and define the binary operation $m : \mathbb{A}^2 \to \mathbb{A}$ by $m(x,y) \coloneqq p(x,0,y)$. Then we have $m(x,0) = p(x,0,0) = x$ and $m(0,x) = p(0,0,x) = x$ for all $x \in \mathbb{A}$, so we can apply the variant of the Eckmann-Hilton principle from Remark 1.5.3 to see that m must be commutative and associative. This m will also be cancellative by construction, so by the finiteness of $\mathbb{A}$ we see that m defines an abelian group structure on $\mathbb{A}$, which shows that $\mathbb{A}$ is quasiaffine. One then needs to check that any finite Taylor algebra which is quasiaffine has a Mal'cev polynomial to finish the argument.

3.13   Bounded width: affine-free CSPs are solved by cycle-consistency

Really, the title of this section should be referring to *pq*-consistency (see Definition 3.9.1), but I wanted to keep the table of contents understandable. We have already shown in Theorem 3.9.3 that if we have a *pq*-consistent instance of a CSP, then we can reduce some of the domains to find a *pq*-consistent instance in which every domain is absorption free. In this section, we will show that if every domain is absorption free and affine free, then we can reduce the instance further while preserving *pq*-consistency.

Definition 3.13.1. We say that a finite idempotent algebra $\mathbb{A}$ is *affine-free* if no quotient of any subalgebra of $\mathbb{A}$ is affine.

The argument strategy is very similar to the argument in the case of strongly connected algebras. We already have most of the pieces.

- If a binary subdirect relation $\mathbb{R} \leq_{sd} \mathbb{A} \times \mathbb{B}$ is linked and $\mathbb{A}, \mathbb{B}$ are absorption free and Taylor, then $\mathbb{R} = \mathbb{A} \times \mathbb{B}$ by the Absorption Theorem 3.11.1.

- If a binary relation $\mathbb{R} \leq \mathbb{A} \times \mathbb{A}$ absorbs the diagonal $\Delta_{\mathbb{A}}$ and $\mathbb{A}$ is absorption free, then $\Delta_{\mathbb{A}} \subseteq \mathbb{R}$ by Theorem 3.7.13.

- If a binary subdirect relation $\mathbb{R} \leq_{sd} \mathbb{A} \times \mathbb{A}$ is linked and $\mathbb{A}$ is Taylor, then $\mathbb{R} \cap \Delta_{\mathbb{A}} \neq \emptyset$ by the linked case of the Loop Lemma 3.11.11.

- If $\mathbb{A}$ is simple, idempotent, Taylor, absorption free, and not affine, then $\mathbb{A}$ is subdirectly complete, by Theorem 3.12.9.

The missing ingredient is an analogue of Theorem 3.3.5.

Theorem 3.13.2. *Suppose $\mathbb{R} \leq_{sd} \mathbb{A} \times \mathbb{B} \times \mathbb{C}$ is subdirect, $\mathbb{A}$ has no proper binary or centrally absorbing subalgebra and no affine quotient, $\pi_{23}(\mathbb{R})$ has no proper binary absorbing subalgebra, $\pi_{12}(\mathbb{R}) = \mathbb{A} \times \mathbb{B}$, $\pi_{13}(\mathbb{R}) = \mathbb{A} \times \mathbb{C}$, and $\mathbb{A}, \mathbb{B}, \mathbb{C}$ are finite idempotent Taylor algebras. Then $\mathbb{R} = \mathbb{A} \times \pi_{23}(\mathbb{R})$.*

Note that by the Absorption Theorem 3.11.1, we just need to prove that if we consider $\mathbb{R}$ as a subdirect binary relation $\mathbb{R} \leq_{sd} \mathbb{A} \times \pi_{23}(\mathbb{R})$, then $\mathbb{R}$ is linked. If not, then the linking congruence of $\mathbb{R}$ on $\mathbb{A}$ is contained in some maximal congruence $\theta \in \mathrm{Con}(\mathbb{A})$, and if we replace $\mathbb{R}$ by the quotient $\mathbb{R}/\theta \leq_{sd} \mathbb{A}/\theta \times \mathbb{B} \times \mathbb{C}$, then we have a smaller counterexample to Theorem 3.13.2 such that $\mathbb{A}$ is simple. So we just need to rule out the case where $\mathbb{A}$ is simple and $\mathbb{R}$ is the graph of a homomorphism $f : \pi_{23}(\mathbb{R}) \twoheadrightarrow \mathbb{A}$. For this, we will use a consequence of the linked case of the Loop Lemma 3.11.11.

Lemma 3.13.3. *If $\mathbb{R}, \mathbb{S} \leq_{sd} \mathbb{A} \times \mathbb{B}$ and the linking congruences of $\mathbb{R}$ on $\mathbb{A}$ and $\mathbb{B}$ contain the corresponding linking congruences of $\mathbb{S}$, then $\mathbb{R} \cap \mathbb{S} \neq \emptyset$.*

Proof. Let $\mathbb{A}' \leq \mathbb{A}, \mathbb{B}' \leq \mathbb{B}$ be corresponding linked components of $\mathbb{R}$, with $\mathbb{R} \cap (\mathbb{A}' \times \mathbb{B}') \neq \emptyset$. By replacing $\mathbb{A}, \mathbb{B}$ with $\mathbb{A}', \mathbb{B}'$ and shrinking $\mathbb{R}, \mathbb{S}$, we may assume without loss of generality that $\mathbb{R}$ is linked. Then $\mathbb{R} \circ \mathbb{S}^- \leq_{sd} \mathbb{A} \times \mathbb{A}$ is also linked, so by the linked case of the Loop Lemma 3.11.11, there is some $a \in \mathbb{A}$ such that $(a, a) \in \mathbb{R} \circ \mathbb{S}^-$. By the definition of $\mathbb{R} \circ \mathbb{S}^-$, this means that there is some $b \in \mathbb{B}$ such that $(a, b) \in \mathbb{R}$ and $(b, a) \in \mathbb{S}^-$, so $(a, b) \in \mathbb{R} \cap \mathbb{S}$. $\qquad \square$

Proof of Theorem 3.13.2. Write $\mathbb{S} = \pi_{23}(\mathbb{R}) \leq_{sd} \mathbb{B} \times \mathbb{C}$. Assume for the sake of contradiction that $\mathbb{A}$ is simple and that $\mathbb{R}$ is the graph of a homomorphism $f : \mathbb{S} \twoheadrightarrow \mathbb{A}$. Note that by the idempotence of $\mathbb{A}$, for each $a \in \mathbb{A}$ the set $f^{-1}(a) \subseteq \mathbb{S}$ is a subalgebra of $\mathbb{S}$, and let $\mathbb{S}_a := f^{-1}(a)$. The assumptions $\pi_{12}(\mathbb{R}) = \mathbb{A} \times \mathbb{B}, \pi_{13}(\mathbb{R}) = \mathbb{A} \times \mathbb{C}$ are equivalent to each $\mathbb{S}_a = f^{-1}(a)$ being a subdirect relation on $\mathbb{B} \times \mathbb{C}$.

If we can show that there are $a \neq a' \in \mathbb{A}$ such that $\mathbb{S}_a, \mathbb{S}_{a'} \leq_{sd} \mathbb{B} \times \mathbb{C}$ have the same linking congruences on $\mathbb{B}, \mathbb{C}$, then we can apply the lemma to see that $f^{-1}(a) \cap f^{-1}(a') = \mathbb{S}_a \cap \mathbb{S}_{a'} \neq \emptyset$, which will give us a contradiction. To accomplish this, we will show that each $\mathbb{S}_a$ has the same linking congruences on $\mathbb{B}, \mathbb{C}$ as $\mathbb{S}$. In fact, we will show that for every $a \in \mathbb{A}$, we have $\mathbb{S} \subseteq \mathbb{S}_a \circ \mathbb{S}_a^- \circ \mathbb{S}_a$.

Let (b, c) be any element of $\mathbb{S}$. Define a subalgebra $\mathbb{X}_{bc} \le \mathbb{A} \times \mathbb{A} \times \mathbb{A}$ by

$$\mathbb{X}_{bc} := \left\{ \begin{bmatrix} x \\ y \\ z \end{bmatrix} \;\middle|\; \exists b' \in \mathbb{B}, c' \in \mathbb{C} \text{ s.t. } \begin{bmatrix} x \\ b \\ c' \end{bmatrix} \in \mathbb{R} \wedge \begin{bmatrix} y \\ b' \\ c' \end{bmatrix} \in \mathbb{R} \wedge \begin{bmatrix} z \\ b' \\ c \end{bmatrix} \in \mathbb{R} \right\}.$$

Equivalently, we have

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} \in \mathbb{X}_{bc} \iff \begin{bmatrix} b \\ c \end{bmatrix} \in \mathbb{S}_x \circ \mathbb{S}_y^- \circ \mathbb{S}_z.$$

Since each $\mathbb{S}_a$ is subdirect, we have $(b, b) \in \mathbb{S}_a \circ \mathbb{S}_a^-$ and $(c, c) \in \mathbb{S}_a^- \circ \mathbb{S}_a$. Thus for each $a \in \mathbb{A}$, we have

$$\begin{bmatrix} a \\ a \\ f(b,c) \end{bmatrix}, \begin{bmatrix} f(b,c) \\ a \\ a \end{bmatrix} \in \mathbb{X}_{bc},$$

so $\mathbb{X}_{bc}$ is subdirect in $\mathbb{A}^3$, and for each $i \ne j \le 3$ the projection $\pi_{ij}(\mathbb{X}_{bc})$ is not the graph of an automorphism of $\mathbb{A}$. Thus by Theorem 3.12.9, we see that $\mathbb{X}_{bc} = \mathbb{A}^3$, so in particular we have $(a, a, a) \in \mathbb{X}_{bc}$ for all $a \in \mathbb{A}$. Since this holds for every $(b, c) \in \mathbb{S}$, we see that $\mathbb{S} \subseteq \mathbb{S}_a \circ \mathbb{S}_a^- \circ \mathbb{S}_a$ for all $a \in \mathbb{A}$, so each $\mathbb{S}_a$ has the same linking congruences on $\mathbb{B}, \mathbb{C}$ as $\mathbb{S}$, which completes the contradiction. $\qquad\square$

Corollary 3.13.4. *If $\mathbb{A}_1, ..., \mathbb{A}_n$ are finite idempotent Taylor algebras with no proper binary or centrally absorbing subalgebras such that all but at most two of the $\mathbb{A}_i$s have no affine quotients, and if $\mathbb{R} \le_{sd} \mathbb{A}_1 \times \cdots \times \mathbb{A}_n$ is a subdirect relation such that each $\pi_{ij}(\mathbb{R})$ is full, then $\mathbb{R} = \mathbb{A}_1 \times \cdots \times \mathbb{A}_n$.*

Corollary 3.13.5. *If $\mathbb{A}_1, ..., \mathbb{A}_n$ are finite idempotent Taylor algebras with no proper binary or centrally absorbing subalgebras and no affine quotients, and if $\mathbb{R} \le_{sd} \mathbb{A}_1 \times \cdots \times \mathbb{A}_n \times \mathbb{A}_1$ is a subdirect relation such that $\Delta_{\mathbb{A}_1} \subseteq \pi_{1,n+1}(\mathbb{R})$ and $\pi_{ij}(\mathbb{R})$ is full for all pairs (i, j) other than $(1, n+1)$, then $\mathbb{R}$ contains every tuple whose first and last coordinates are the same.*

Proof. Suppose first that $\mathbb{R}$ has a proper binary or centrally absorbing subalgebra $\mathbb{R}'$. Note that each $\pi_{ij}(\mathbb{R}')$ with $(i, j) \ne (1, n+1)$ is full since the $\mathbb{A}_i$ have no binary or centrally absorbing subalgebras. Additionally, $\pi_{1,n+1}(\mathbb{R}')$ absorbs $\Delta_{\mathbb{A}_1}$, so by Theorem 3.11.11 we see that $\Delta_{\mathbb{A}_1} \subseteq \pi_{1,n+1}(\mathbb{R}')$ as well. Thus we may replace $\mathbb{R}$ by $\mathbb{R}'$, until we eventually reach a situation where $\mathbb{R}$ has no proper binary or central absorption. In particular, we may assume that $\pi_{1,n+1}(\mathbb{R})$ has no proper binary or centrally absorbing subalgebras.

For any $2 \le i \le n$, we may apply Theorem 3.13.2 to $\pi_{i,1,n+1}(\mathbb{R})$ to see that $\pi_{i,1,n+1}(\mathbb{R}) = \mathbb{A}_i \times \pi_{1,n+1}(\mathbb{R})$. Now consider $\mathbb{R}$ as an n-ary relation

$$\mathbb{R} \le_{sd} \pi_{1,n+1}(\mathbb{R}) \times \mathbb{A}_2 \times \cdots \times \mathbb{A}_n,$$

and apply the previous corollary to see that $\mathbb{R} = \pi_{1,n+1}(\mathbb{R}) \times \mathbb{A}_2 \times \cdots \times \mathbb{A}_n$. In particular, since we have $\Delta_{\mathbb{A}_1} \subseteq \pi_{1,n+1}(\mathbb{R})$, we see that $\mathbb{R}$ contains every tuple whose first and last coordinates are equal. $\qquad\square$

Now that we've gathered up all the necessary ingredients, we argue as in the case of strongly connected algebras. We start by picking some variable x with $|\mathbb{A}_x| > 1$, pick a maximal congruence

$\theta_x \in \mathrm{Con}(\mathbb{A}_x)$, pick a congruence class $\mathbb{A}'_x \leq \mathbb{A}_x$ of θ_x. Then we refer back to Definition 3.5.2 to define the "proper" variables y to be the variables such that there exists a path p from y to x such that

$$\mathbb{P}_p/\theta_x \leq_{sd} \mathbb{A}_y \times \mathbb{A}_x/\theta_x$$

is the graph of a homomorphism $\iota_y : \mathbb{A}_y \twoheadrightarrow \mathbb{A}_x/\theta_x$, and define θ_y to be the kernel of ι_y and $\mathbb{A}'_y$ to be $\iota_y^{-1}(\mathbb{A}'_x)$.

As in the case of strongly connected algebras, we need to check that the homomorphism ι_y does not depend on the choice of path p. This time, we will check this using pq-consistency instead of cycle-consistency.

Lemma 3.13.6. *Suppose that the instance $\mathbf{X}$ is pq-consistent, and that x, θ_x are chosen as above. Suppose that y is a proper variable, and that p, q are two paths from y to x such that $\mathbb{P}_p/\theta_x, \mathbb{P}_q/\theta_x$ are the graphs of homomorphisms $\iota_p, \iota_q : \mathbb{A}_y \twoheadrightarrow \mathbb{A}_x/\theta_x$. Then $\iota_p = \iota_q$.*

Proof. Consider the cycles $p - q$ and $q - p$ from y to y, then by the definition of pq-consistency (Definition 3.9.1) we see that there must be some $j \geq 0$ such that for all $a \in \mathbb{A}_y$, we have

$$a \in \{a\} + j(p - q + q - p) + p - q.$$

For any $b \in \mathbb{A}_x$, we have $b/\theta_x - p + p = b/\theta_x$ and $b/\theta_x - q + q = b/\theta_x$ by the assumptions on $\mathbb{P}_p, \mathbb{P}_q$, so we see that

$$\{a\} + j(p - q + q - p) + p - q \subseteq \iota_p(a) - q = \iota_q^{-1}(\iota_p(a)),$$

so we must have $\iota_q(a) = \iota_p(a)$. □

As a consequence, we have the following analogue of Lemma 3.5.4.

Lemma 3.13.7. *Suppose that the instance $\mathbf{X}$ is pq-consistent, and that each domain has no proper binary or centrally absorbing subalgebra. Suppose p is a path from y to a proper variable z. Then one of the following is true:*

- $\mathbb{P}_p/\theta_z = \mathbb{A}_y \times \mathbb{A}_z/\theta_z$, *or*

- y *is also proper, and $\mathbb{P}_p/(\theta_y \times \theta_z)$ is the graph of an isomorphism $\iota_p : \mathbb{A}_y/\theta_y \xrightarrow{\sim} \mathbb{A}_z/\theta_z$ such that $\iota_y = \iota_z \circ \iota_p$.*

Proof. Since $\mathbb{A}_z/\theta_z$ is simple, the linking congruence of $\mathbb{P}_p/\theta_z$ must either be trivial or full. If the linking congruence of $\mathbb{P}_p/\theta_z$ is full, then by the Absorption Theorem 3.11.1 we see that $\mathbb{P}_p/\theta_z = \mathbb{A}_y \times \mathbb{A}_z/\theta_z$. Otherwise, $\mathbb{P}_p/\theta_z$ is the graph of a homomorphism from $\mathbb{A}_y$ to $\mathbb{A}_z/\theta_z$, so then by joining the path p with a path from z to x we see that y is proper and $\iota_y = \iota_z \circ \iota_p$. □

To finish, we just need to show that restricting each proper variable's domain $\mathbb{A}_x$ to $\mathbb{A}'_x$ gives us a pq-consistent instance $\mathbf{X}'$. To see that $\mathbf{X}'$ is arc-consistent, we apply Corollary 3.13.4 as in the proof of Lemma 3.5.5. To see that $\mathbf{X}'$ is pq-consistent, we apply Corollary 3.13.5 as in the proof of Lemma 3.5.6. We have proven our main result.

Theorem 3.13.8 (Kozik [95])**.** *If $\mathbf{X}$ is a pq-consistent instance of a CSP such that every domain is finite, idempotent, Taylor, and affine-free, then $\mathbf{X}$ has a solution.*

As a curiously roundabout consequence, we see that we can't build an affine (or even abelian) algebra out of affine-free algebras.

Corollary 3.13.9. *If $\mathbb{A}_1, ..., \mathbb{A}_n$ are finite, idempotent, Taylor, and affine-free, then the variety $\mathcal{V}(\mathbb{A}_1, ..., \mathbb{A}_n)$ which they generate does not contain any nontrivial abelian algebras.*

Proof. Since the variety $\mathcal{V}(\mathbb{A}_1, ..., \mathbb{A}_n)$ is finitely generated, it is locally finite, so any nontrivial abelian algebra in this variety must contain a finite abelian algebra $\mathbb{B}$ with $|\mathbb{B}| > 1$. Since $\mathbb{B}$ is finite, Taylor, and abelian, we see that $\mathbb{B}$ is affine by Theorem 3.12.8. But then $\mathbb{B}$ is a subquotient of some finite product of $\mathbb{A}_i$s, so $\mathrm{CSP}(\prod_i \mathbb{A}_i^k)$ fails to have bounded width for some finite k, which contradicts the fact that $\mathrm{CSP}(\mathbb{A}_1, ..., \mathbb{A}_n)$ is solved by *pq*-consistency. $\square$

Using commutator theory, we have the following consequence (see Corollary 1.9.34).

Corollary 3.13.10. *If $\mathbb{A}$ is a finite idempotent algebra, then $\mathbb{A}$ is Taylor and affine-free if and only if the variety $\mathcal{V}(\mathbb{A})$ is congruence meet-semidistributive.*

Using the language of pp-construcability (see Definition 1.4.14), we can rephrase Theorem 3.13.8 as follows.

Corollary 3.13.11. *A relational structure $\mathbf{A}$ with a finite domain has $\mathrm{CSP}(\mathbf{A})$ solved by pq-consistency if and only if $\mathbf{A}$ does not pp-construct any of the relational structures $(\mathbb{Z}/p, \{1\}, x + y = z)$, p prime.*

Proof. Since $\mathbf{A}$ pp-constructs its rigid core and vice-versa, we may assume without loss of generality that $\mathbf{A}$ is a rigid core. Then the associated algebra $\mathbb{A}$ is idempotent, so $\mathbb{A}$ is Taylor if and only if there is any relational structure which $\mathbf{A}$ does not pp-construct. To finish, we need to check that if $\mathbb{A}$ is not affine-free, then $\mathbf{A}$ pp-constructs some $(\mathbb{Z}/p, \{1\}, x + y = z)$. Since restricting to a subalgebra of $\mathbb{A}$ and taking a quotient can both be accomplished by pp-constructions, we may suppose that $\mathbb{A}$ is affine and nontrivial.

If $\mathbb{A}$ is affine, then by definition $\mathbb{A}$ is polynomially equivalent to some module $\mathbb{M}$. If $\mathbb{A}$ is also idempotent, then the relation $x + y = z$ is preserved by $\mathbb{A}$, as are all singleton unary relations, so $\mathbf{A}$ pp-constructs the relational structure $(\mathbb{M}, x + y = z)^{rig}$ (the superscript is shorthand for throwing in all unary singleton relations). Since $\mathbb{M}$ is finite, some element of $\mathbb{M}$ must have prime order, say order p. Then the set of all elements of $\mathbb{M}$ with order p is pp-definable, so we may suppose without loss of generality that every nonzero element of $\mathbb{M}$ has order exactly p. As an abelian group we then have $\mathbb{M} \cong (\mathbb{Z}/p)^k$ for some k. Letting c be any nonzero element of $\mathbb{M}$, we then see that $(\mathbb{M}, \{c\}, x + y = z)$ is homomorphically equivalent to $(\mathbb{Z}/p, \{1\}, x + y = z)$. $\square$

3.13.1 Weak Prague instances

The original proofs of the bounded width conjecture (i.e., that affine-free CSPs have bounded width) didn't use the concepts of *pq*-consistency or cycle-consistency. Bulatov's argument [40] used $(2, 3)$-consistency, and leveraged a local structure theory of bounded width algebras in terms of two element semilattice and majority subalgebras. The early arguments due to Barto and Kozik [12], [19] used simpler algebraic ingredients, but used a more complicated consistency condition satisfied by instances called *Prague instances*, which were then simplified to *weak Prague instances*. We won't go over the original Prague instance concept, but weak Prague instances have a nice definition.

Definition 3.13.12. An instance $\mathbf{X}$ of a CSP with variable domains $\mathbb{A}_x$ is called a *weak Prague instance* if it satisfies the following three conditions.

(P1) The instance $\mathbf{X}$ is arc-consistent, that is, each constraint relation $\mathbb{R} \leq \prod_{x_i} \mathbb{A}_{x_i}$ is subdirect.

(P2) For every variable x, every set $A \subseteq \mathbb{A}_x$, and every cycle p from x to x, we have the implication
$$A + p = A \implies A - p = A.$$

(P3) For every variable x, every set $A \subseteq \mathbb{A}_x$, and every pair of cycles p, q from x to x, we have the implication
$$A + p + q = A \implies A + p = A.$$

We can understand what condition (P2) says about an individual cycle p in terms of the digraph associated to the binary relation $\mathbb{P}_p \leq_{sd} \mathbb{A}_x \times \mathbb{A}_x$.

Proposition 3.13.13. *A subdirect binary relation $\mathbb{P} \leq_{sd} \mathbb{A} \times \mathbb{A}$ on a finite algebra $\mathbb{A}$ satisfies the implication*
$$A + \mathbb{P} = A \implies A - \mathbb{P} = A$$
for all $A \subseteq \mathbb{A}$ if and only if the digraph $\mathbf{P} = (\mathbb{A}, \mathbb{P})$ satisfies one of the following equivalent conditions:

- *every weakly connected component of $\mathbf{P}$ is strongly connected,*

- *every edge of $\mathbf{P}$ is contained in a directed cycle of $\mathbf{P}$,*

- *there is some $k \geq 0$ such that $\mathbb{P}^- \subseteq \mathbb{P}^{\circ k}$.*

An alternative form of condition (P2) is given in [18].

Proposition 3.13.14 (Barto, Kozik [18])**.** *If an instance satisfies condition (P1), then (P2) is equivalent to the following condition.*

(P2*) *For all variables x, sets $A \subseteq \mathbb{A}_x$, and cycles p from x to x such that $A + p = A$, if p_1 is the first step of the cycle p, then we have $A + p_1 - p_1 = A$.*

Note that $A + p_1 - p_1 = A$ if and only if A is a union of linked components of p_1.

Proof. It's easy to see that (P1) and (P2) imply (P2*), so we'll focus on proving the more difficult implication: that (P2*) implies (P2). Suppose that $A + p = A$, and write $p = p_1 + p_2 + \cdots + p_k$, where each p_i has length one. By the assumption $A + p = A$, we have
$$(A + p_1 + \cdots + p_i) + (p_{i+1} + \cdots + p_k + p_1 + \cdots + p_i) = (A + p) + p_1 + \cdots + p_i = A + p_1 + \cdots + p_i,$$
so we can apply (P2*) to see that
$$(A + p_1 + \cdots + p_i) + p_{i+1} - p_{i+1} = A + p_1 + \cdots + p_i.$$

Thus we have
$$
\begin{aligned}
A - p &= (A + p) - p \\
&= A + p_1 + \cdots + p_{k-1} + p_k - p_k - p_{k-1} - \cdots - p_1 \\
&= A + p_1 + \cdots + p_{k-1} - p_{k-1} - \cdots - p_1 \\
&= \cdots \\
&= A + p_1 - p_1 = A. \qquad \square
\end{aligned}
$$

Conditions (P1) and (P2) are closely related to the basic linear relaxation of a CSP, from subsection 1.6.1.

Theorem 3.13.15. *If $\mathbf{X}$ is an instance of a CSP such that the basic linear relaxation of $\mathbf{X}$ has a solution assigning probability vectors p_C to each constraint C of $\mathbf{X}$ and probability vectors p_x to each variable x, then the instance $\mathbf{X}'$ obtained by restricting each constraint relation of $\mathbf{X}$ to the support of the corresponding probability distribution p_C (and similarly for the variable domains) satisfies conditions (P1) and (P2).*

Proof. Assume for simplicity that $\mathbf{X} = \mathbf{X}'$, that is, that all of the probability vectors have full support. The compatibility of the probability vectors p_C with the probability vectors on the variable domains ensures that $\mathbf{X}$ is arc-consistent, so (P1) is satisfied. For (P2), it is easier to check condition (P2*) from Proposition 3.13.14. We attach to each set $A \subseteq \mathbb{A}_x$ a probability $P(A)$, given by

$$P(A) = \sum_{a \in A} p_{x,a}.$$

Now consider any step p_1 from a variable x to an adjacent variable y within a constraint C. Let $\mathbb{P} \subseteq \mathbb{A}_x \times \mathbb{A}_y$ be the binary projection of the corresponding constraint relation onto x and y, and let $p_{\mathbb{P}}$ be the corresponding marginal distribution of p_C. Then we have

$$P(A + \mathbb{P}) = \sum_{b \in A + \mathbb{P}} p_{y,b} \geq \sum_{b \in A + \mathbb{P}} \sum_{a \in A} p_{\mathbb{P},(a,b)} = \sum_{a \in A} p_{x,a} = P(A),$$

with equality when $A + \mathbb{P} - \mathbb{P} = A$. Thus if $A + p = A$, then we have

$$P(A) \leq P(A + p_1) \leq P(A + p) = P(A),$$

so $P(A + p_1) = P(A)$, and thus we have $A + p_1 - p_1 = A$. $\qquad\square$

In fact, Theorem 3.13.15 has a converse when we restrict our attention to a single cycle at a time.

Theorem 3.13.16. *If $\mathbf{X}$ is an instance of a CSP such that the associated hypergraph of variables and relations consists of a single cycle, then $\mathbf{X}$ has properties (P1) and (P2) if and only if the basic linear relaxation of $\mathbf{X}$ has a solution such that for each constraint C of $\mathbf{X}$, the support of the corresponding probability distribution p_C is exactly equal to the relation corresponding to C.*

Proof. Let $v_1, ..., v_n$ be the variables of $\mathbf{X}$ which occur in two constraints, in the order in which they appear around the cycle, and let the constraints $C_1, ..., C_n$ be numbered such that v_i and v_{i+1} are variables of C_i for each i (here we interpret the subscripts $i, i+1$ modulo n, so $v_{n+1} = v_1$).

Consider the following directed graph on the set of pairs (i, a) where $i \in \mathbb{Z}/n\mathbb{Z}$ and $a \in \mathbb{A}_{v_i}$: for every element r of the relation corresponding to constraint C_i, we make a directed edge from $(i, \pi_{v_i}(r))$ to $(i + 1, \pi_{v_{i+1}}(r))$. Then conditions (P1) and (P2) guarantee that every edge of this digraph is contained in a directed cycle. Choose some finite set of directed cycles $\mathcal{C}$ of this digraph which covers each edge at least once. Then for each constraint C_i, we let p_{C_i} be the probability distribution defined by first choosing a cycle from $\mathcal{C}$ uniformly at random, and then choosing uniformly among the elements r of the relation corresponding to the constraint C_i such that the edge from $(i, \pi_{v_i}(r))$ to $(i + 1, \pi_{v_{i+1}}(r))$ is contained in our chosen cycle. $\qquad\square$

The condition (P3) can be rephrased to look slightly more similar to the condition for pq-consistency.

Proposition 3.13.17. *An instance* $\mathbf{X}$ *with finite variable domains* $\mathbb{A}_x$ *satisfies condition (P3) if and only if it satisfies the following condition.*

(P3*) *For all variables* x, *for all pairs of cycles* p, q *from* x *to* x, *and for all* $a \in \mathbb{A}_x$, *there is some* $j \geq 0$ *such that*

$$\{a\} + j(p+q) = \{a\} + j(p+q) + p = \{a\} + j(p+q) + p + q.$$

Proof. First we show that (P3) implies (P3*). For this, note that if we define a sequence of subsets $A_i \subseteq \mathbb{A}_x$ by $A_i = \{a\} + i(p+q)$, then by the finiteness of $\mathbb{A}_x$ there must be some j, k with $k > 0$ such that $A_j = A_{j+k}$. But then (P3) implies that $A_j + p = A_j$ and similarly that $(A_j + p) + q = A_j + p$.

For the reverse direction, let $A \subseteq \mathbb{A}_x$ satisfy $A + p + q = A$. Then by the finiteness of A we can find j sufficiently large such that for each $a \in A$ we have $\{a\} + j(p+q) = \{a\} + j(p+q) + p$. For this choice of j, we then have

$$A = A + j(p+q) = A + j(p+q) + p = A + p. \qquad \square$$

There is also a natural way to certify that a given instance satisfies condition (P3), following a similar philosophy to the method we used to find absorbing reductions of cycle consistent majority CSPs.

Proposition 3.13.18. *An instance* $\mathbf{X}$ *satisfies condition (P3) at a variable* x *if and only if there is a partial order* $\preceq$ *on the power set* $\mathcal{P}(\mathbb{A}_x)$, *such that for every cycle* p *from* x *to* x *and every* $A \subseteq \mathbb{A}_x$, *we have*

$$A \preceq A + p.$$

The instance $\mathbf{X}$ *satisfies (P3) everywhere if and only if there is a quasiorder* $\preceq$ *on the set of ordered pairs* (x, A) *with* $A \subseteq \mathbb{A}_x$, *such that for each binary projection* $\mathbb{R}_{ij} \leq \mathbb{A}_x \times \mathbb{A}_y$ *of any constraint relation of* $\mathbf{X}$ *and for each* $A \subseteq \mathbb{A}_x$, *we have*

$$(x, A) \preceq (y, A + \mathbb{R}_{ij}),$$

and such that for each x, *the restriction of* $\preceq$ *to* $\{x\} \times \mathcal{P}(\mathbb{A}_x)$ *defines a partial order on* $\mathcal{P}(\mathbb{A}_x)$.

Weak Prague instances are closely related to pq-consistent instances, but they are not quite the same.

Theorem 3.13.19. *Every weak Prague instance is* pq-*consistent.*

Proof. Suppose $\mathbf{X}$ is a weak Prague instance, that x is a variable of $\mathbf{X}$, that p, q are cycles from x to x, and that $a \in \mathbb{A}_x$. We need to check that there is some $j \geq 0$ such that

$$a \in \{a\} + j(p+q) + p.$$

Since $\mathbb{A}_x$ is finite, there must be some $j > 0$ such that

$$\{a\} + j(p+q) = \{a\} + 2j(p+q).$$

Let $A = \{a\} + j(p+q)$ be the common value of both sides of the above equation (note that if $\mathbb{A}_x$ is idempotent, then A will actually be a subalgebra of $\mathbb{A}_x$). Then by (P2) we have

$$A = A + j(p+q) \implies A = A - j(p+q),$$

so

$$a \in \{a\} + j(p+q) - j(p+q) = A - j(p+q) = A.$$

Additionally, by (P3) we have

$$A = A + p + (q + (j-1)(p+q)) \implies A = A + p,$$

so

$$a \in A = A + p = \{a\} + j(p+q) + p. \qquad \square$$

Example 3.13.1. Here we give an example of a *pq*-consistent instance which is not a weak Prague instance. Consider the instance of 2-SAT with just one variable x, domain $\mathbb{A}_x = (\{0,1\}, \mathrm{maj})$, and a binary constraint relation $\mathbb{R} \leq_{sd} \mathbb{A}_x \times \mathbb{A}_x$ imposed on (x, x) given by $\mathbb{R} = \{(0,0), (0,1), (1,1)\}$ (that is, $\mathbb{R}$ is the binary relation $\leq$).

Since $\Delta_{\{0,1\}} \subseteq \mathbb{R}$, we see that this instance is *pq*-consistent. However, this instance does not satisfy property (P2) of a weak Prague instance: we have

$$\{1\} + \mathbb{R} = \{1\},$$

but

$$\{1\} - \mathbb{R} = \{0,1\} \neq \{1\}.$$

Alternatively, we can check that (P2) is not satisfied by noting that the digraph $(\{0,1\}, \leq)$ is weakly connected but not strongly conected.

Although not every *pq*-consistent instance satisfies (P2), we at least have the following implication.

Theorem 3.13.20 (Kozik [95])**.** *Every pq-consistent instance satisfies conditions (P1) and (P3).*

Proof. Suppose $\mathbf{X}$ is a *pq*-consistent instance, that x is a variable of $\mathbf{X}$, that p, q are cycles from x to x, and that $A \subseteq \mathbb{A}_x$ satisfies

$$A + p + q = A.$$

By *pq*-consistency, there is some $j \geq 0$ such that

$$A \subseteq A + j(p+q) + p = A + p.$$

Similarly, from

$$(A + p) + q + p = A + p,$$

we see that

$$A + p \subseteq (A + p) + q = A.$$

Thus we have $A = A + p$. $\qquad \square$

Example 3.13.2. There is an example of an instance which satisfies (P1) and (P3), but which is not *pq*-consistent. As in the previous example, this instance will have just a single variable x and a single binary constraint $\mathbb{R} \leq_{sd} \mathbb{A}_x \times \mathbb{A}_x$. We take the algebra $\mathbb{A}_x$ to be the three-element dual discriminator algebra $(\{0, 1, 2\}, d(x, y, z))$ from Example 1.6.5. The binary relation $\mathbb{R}$ is the 0/1/all constraint displayed below.
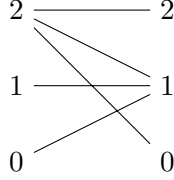
To see that this is not *pq*-consistent, note that there is no j such that $(0,0) \in \mathbb{R}^{\circ j}$. To see that this instance satisfies condition (P3), we use the following total ordering on $\mathcal{P}(\{0, 1, 2\})$:

$$\emptyset \preceq \{0\} \preceq \{0, 1\} \preceq \{0, 2\} \preceq \{1\} \preceq \{2\} \preceq \{1, 2\} \preceq \{0, 1, 2\}.$$

Libor Barto has raised the following question.

Problem 3.13.1. Is it true that every instance of an affine-free CSP which satisfies conditions (P1) and (P3) has a solution?

3.14 Terms for bounded width and the meta-problem

In this section we'll prove the existence of nice ternary terms characterizing bounded width algebras, which were first conjectured to exist by Jovanović [81] and later proved to exist using a Ramsey argument and the fact that bounded width CSPs are solved by $(2, 3)$-consistency [82]. Using *pq*-consistency instead of $(2, 3)$-consistency, it is possible to prove the existence of these terms directly, as noted by Kozik [95]. These nice ternary terms will allow us to efficiently solve the *meta-problem* for bounded width CSPs: given a core relational structure **A** as input, determine whether CSP(**A**) has bounded width.

Theorem 3.14.1 (Height 1 identities for bounded width [81], [82], [95])**.** *Suppose* **A** *is a relational structure on a finite domain. Then* CSP(**A**) *has bounded relational width iff there are ternary polymorphisms* $f, g \in \mathrm{Pol}_3(\mathbf{A})$ *satisfying the height* 1 *identities*

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

In this case, every pq-consistent instance of CSP(**A**) *has a solution.*

The identities in the statement of Theorem 3.14.1 may be interpreted as follows. If the common values $c(x, y)$ of $g(x, x, y)$, etc. are all equal to x, then g is a majority function, and f behaves as if it is first projection. If instead we have $c(x, y) = x \vee y$, then f, g both behave as if they are the three-element semilattice operation $x \vee y \vee z$. Finally, if $c(x, y) = y$, then f is a Pixley operation, so $f(x, f(x, y, z), z)$ is a majority operation, and additionally Theorem 3.1.14 applies.

Since having bounded relational width is preserved by homomorphic equivalence, we may reduce proving Theorem 3.14.1 to the special case where **A** is a core, and then we can use Theorem 1.4.7 to reduce to the case of a rigid core, so that the associated algebra $\mathbb{A}$ is idempotent. Since any

idempotent algebra $\mathbb{A}$ such that $\mathrm{CSP}(\mathbb{A})$ has bounded width must be Taylor and affine-free, we see from Theorem 3.13.8 that $\mathrm{CSP}(\mathbb{A})$ is solved by pq-consistency. Furthermore, by Corollary 3.13.9 we see that the free algebra $\mathbb{F} = \mathcal{F}_{\mathbb{A}}(x,y) \leq \mathbb{A}^{\mathbb{A}^2}$ is also affine-free, so $\mathrm{CSP}(\mathbb{F})$ is also solved by pq-consistency. The plan is to construct a pq-consistent instance of $\mathrm{CSP}(\mathbb{F})$ which encodes the existence of such ternary terms f, g, but before we do this we need a basic result about taking closures under algebraic operations.

Definition 3.14.2. Suppose that $\mathbf{X}$ is an instance of a CSP such that every variable domain is contained in $\mathbb{A}$, but possibly the variable domains and the relations of $\mathbf{X}$ are not closed under the operations of $\mathbb{A}$. Define $\mathrm{Sg}_{\mathbb{A}}(\mathbf{X})$ to be the instance of $\mathrm{CSP}(\mathbb{A})$ where every variable domain and every relation of $\mathbf{X}$ is replaced by the subalgebra it generates.

Proposition 3.14.3. *If $\mathbf{X}$ is a pq-consistent instance as above, then $\mathrm{Sg}_{\mathbb{A}}(\mathbf{X})$ is also pq-consistent.*

Proof. For arc-consistency, let $R \subseteq \mathbb{A}^n$ be any relation, and note that $\mathrm{Sg}_{\mathbb{A}}(\pi_1(R)) = \pi_1(\mathrm{Sg}_{\mathbb{A}}(R))$. For paths, let $R, S \subseteq \mathbb{A} \times \mathbb{A}$ be any binary relations, then we have $\mathrm{Sg}_{\mathbb{A}}(R \circ S) \subseteq \mathrm{Sg}_{\mathbb{A}}(R) \circ \mathrm{Sg}_{\mathbb{A}}(S)$. For cycles interacting well with the diagonal, note that for any $B \subseteq \mathbb{A}$ we have $\mathrm{Sg}_{\mathbb{A}^2}(\Delta_B) = \Delta_{\mathrm{Sg}_{\mathbb{A}}(B)}$. $\square$

We have a similar result for weak Prague instances (Definition 3.13.12), which we won't actually need.

Proposition 3.14.4. *If $\mathbf{X}$ is a weak Prague instance as above, then $\mathrm{Sg}_{\mathbb{A}}(\mathbf{X})$ is also a weak Prague instance.*

Proof. That $\mathrm{Sg}_{\mathbb{A}}(\mathbf{X})$ satisfies (P1) and (P3) follows from the fact that $\mathbf{X}$ is a pq-consistent instance (Theorem 3.13.19), which implies that $\mathrm{Sg}_{\mathbb{A}}(\mathbf{X})$ is also pq-consistent by the previous proposition, and this in turn implies that $\mathrm{Sg}_{\mathbb{A}}(\mathbf{X})$ satisfies (P1) and (P3) (Theorem 3.13.20). To check that $\mathrm{Sg}_{\mathbb{A}}(\mathbf{X})$ satisfies (P2), we use Proposition 3.13.13: note that if $P \subseteq \mathbb{A} \times \mathbb{A}$ satisfies $P^- \subseteq P^{\circ k}$, then $\mathrm{Sg}_{\mathbb{A}}(P)^- = \mathrm{Sg}_{\mathbb{A}}(P^-) \subseteq \mathrm{Sg}_{\mathbb{A}}(P^{\circ k}) \subseteq \mathrm{Sg}_{\mathbb{A}}(P)^{\circ k}$. $\square$

Lemma 3.14.5. *Suppose $\mathbf{X}$ is an instance of a CSP over the two-element domain $\{x, y\}$ with no unary relations, such that every binary projection $\pi_{i,j}(R)$ of every relation R is subdirect in $\{x, y\}^2$ and has $(x, x) \in \pi_{i,j}(R)$. Then $\mathbf{X}$ is pq-consistent.*

Proof. The assumptions on $\mathbf{X}$ directly imply that $\mathbf{X}$ is arc-consistent. Now consider any pair of cycles p, q from a variable v of $\mathbf{X}$ to itself. Note that the collection of binary relations on $\{x, y\}$ which are subdirect and contain (x, x) is closed under composition and reversal, so $\mathbb{P}_p, \mathbb{P}_q$ are both subdirect and contain (x, x). We just need to show that there is some j such that $y \in \{y\} + j(p + q) + p$.

If $(y, y) \in \mathbb{P}_p$, then we may take $j = 0$. Otherwise, we must have $\mathbb{P}_p = \{(x, x), (x, y), (y, x)\}$, and since $(x, x) \in \mathbb{P}_q$ this implies that $\mathbb{P}_p \circ \mathbb{P}_q \circ \mathbb{P}_p = \{x, y\}^2$, so we may take $j = 1$. $\square$

Proof of Theorem 3.14.1. First we prove the existence of such terms in any finite idempotent Taylor affine-free algebra $\mathbb{A}$. Consider the ternary relations $R, S \subseteq \{x, y\}^3$ given by

$$R = \left\{ \begin{bmatrix} x \\ x \\ y \end{bmatrix}, \begin{bmatrix} x \\ y \\ x \end{bmatrix}, \begin{bmatrix} y \\ x \\ x \end{bmatrix} \right\}$$

and

$$S = \left\{ \begin{bmatrix} x \\ x \\ x \end{bmatrix}, \begin{bmatrix} x \\ y \\ y \end{bmatrix}, \begin{bmatrix} y \\ x \\ y \end{bmatrix} \right\}.$$

It's easy to check that each binary projection of R and S is subdirect in $\{x,y\}^2$ and contains (x,x). Now consider the CSP instance $\mathbf{X}$ with just a single variable v, and the apply the constraints R and S to the triple (v,v,v) (if this makes you uncomfortable, you can instead use several different variables and impose equality constraints between them). By the lemma, $\mathbf{X}$ is a pq-consistent instance.

If we let $\mathbb{F} = \mathcal{F}_{\mathbb{A}}(x,y) \leq \mathbb{A}^{\mathbb{A}^2}$, then we may consider $\{x,y\}$ to be a subset of $\mathbb{F}$, and apply the proposition to see that $\mathrm{Sg}_{\mathbb{F}}(\mathbf{X})$ is also pq-consistent. Since $\mathbb{F}$ is finite, idempotent, Taylor, and affine-free, we can apply Theorem 3.13.8 to see that $\mathrm{Sg}_{\mathbb{F}}(\mathbf{X})$ has a solution. Suppose that this solution assigns the variable v to the value $c \in \mathbb{F}$. Then we have

$$\begin{bmatrix} c \\ c \\ c \end{bmatrix} \in \mathrm{Sg}_{\mathbb{F}}(R) \cap \mathrm{Sg}_{\mathbb{F}}(S) = \mathrm{Sg}_{\mathbb{F}} \left\{ \begin{bmatrix} x & x & y \\ x & y & x \\ y & x & x \end{bmatrix} \right\} \cap \mathrm{Sg}_{\mathbb{F}} \left\{ \begin{bmatrix} x & x & y \\ x & y & x \\ x & y & y \end{bmatrix} \right\}.$$

Thus there are ternary terms f, g of $\mathbb{A}$ such that

$$g\left(\begin{bmatrix} x & x & y \\ x & y & x \\ y & x & x \end{bmatrix} \right) = \begin{bmatrix} c \\ c \\ c \end{bmatrix} = f\left(\begin{bmatrix} x & x & y \\ x & y & x \\ x & y & y \end{bmatrix} \right),$$

and these f, g satisfy the required identities.

For the converse direction, we will suppose that such terms f, g exist for some idempotent algebra $\mathbb{A}$, and prove that $\mathbb{A}$ is Taylor and affine-free. It's easy to see that $\mathbb{A}$ must be Taylor, since the identities satisfied by g can't be satisfied by any projection. Since any identities which hold in $\mathbb{A}$ also hold in any subquotient of $\mathbb{A}$, we may suppose for contradiction that $\mathbb{A}$ is a nontrivial idempotent affine algebra. Then $\mathbb{A}$ is polynomially equivalent to some module $\mathbb{M}$ over some ring $\mathbb{R}$, and we may write

$$g(x,y,z) \approx \alpha x + \beta y + \gamma z$$

for some $\alpha, \beta, \gamma \in \mathbb{R}$ with $\alpha + \beta + \gamma = 1$. Plugging in $x = 0$ to the identities

$$g(x,x,y) \approx g(x,y,x) \approx g(y,x,x)$$

gives $\alpha y \approx \beta y \approx \gamma y$, so

$$g(x,y,z) \approx \alpha(x+y+z)$$

and $3\alpha x \approx x$. Then if we plug in $x = 0$ to the identities

$$2\alpha x + \alpha y \approx f(x,x,y) \approx f(x,y,x) \approx f(x,y,y),$$

we see that $\alpha y \approx 2\alpha y$, so $\alpha y \approx 0$. Multiplying by 3, we get $y \approx 0$, so in fact the algebra $\mathbb{A}$ must consist of just the single element 0, a contradiction. $\qquad\square$

The proof technique of Theorem 3.14.1 can be used to produce many further terms which mimic the monotone self-dual functions found in the clone of a two-element majority algebra.

Theorem 3.14.6. *Suppose* $\mathrm{CSP}(\mathbf{A})$ *has bounded relational width and* $\mathbf{A}$ *is finite. Then there is a binary polymorphism* $c(x,y)$, *and an infinite family of polymorphisms* $h_n^{\mathcal{F}} \in \mathrm{Pol}_n(\mathbf{A})$ *indexed by the collection of maximal intersecting families* $\mathcal{F}$ *of subsets of* $[n]$, *such that for each set* $S \in \mathcal{F}$ *with* $S \neq [n]$, *if we define* v_i^S *by*

$$v_i^S = \begin{cases} x & i \in S, \\ y & i \notin S, \end{cases}$$

we have the identity

$$h_n^{\mathcal{F}}(v_1^S, ..., v_n^S) \approx c(x,y).$$

Now we show how we can use the ternary terms f, g from Theorem 3.14.1 to solve the meta-problem.

Theorem 3.14.7. *Suppose we are given a finite relational structure* $\mathbf{A} = (A, R_1, ..., R_n)$, *where each relation* R_i *has arity* m_i *and is described by explicitly listing out its tuples, and suppose that we are promised that* $\mathbf{A}$ *is core. Then we can determine whether* $\mathrm{CSP}(\mathbf{A})$ *has bounded width in polynomial time, and in the case where* $\mathrm{CSP}(\mathbf{A})$ *has bounded width, we can explicitly find ternary functions* $f, g \in \mathrm{Pol}_3(\mathbf{A})$ *as in Theorem 3.14.1.*

Proof. We will define an instance $\mathbf{X}$ of $\mathrm{CSP}(\mathbf{A})$ such that every solution to $\mathbf{X}$ corresponds to a pair of terms f, g as in Theorem 3.14.1. The instance $\mathbf{X}$ will have two sets of $|A^3|$ variables, one variable for each value $f(a, b, c)$ for $a, b, c \in A$ and one variable for each value $g(a, b, c)$ for $a, b, c \in A$.

The relations of $\mathbf{X}$ will do two jobs: they will ensure that $f, g \in \mathrm{Pol}_3(\mathbf{A})$, and they will ensure that f, g satisfy the required identities. To ensure that $f \in \mathrm{Pol}_3(\mathbf{A})$, we consider every three tuples $a, b, c \in R_i$ (note that each of a, b, c is an m_i-tuple of values in A), and we impose the constraint

$$\begin{bmatrix} f(a_1, b_1, c_1) \\ f(a_2, b_2, c_2) \\ \vdots \\ f(a_{m_i}, b_{m_i}, c_{m_i}) \end{bmatrix} \in R_i$$

for each such tuple. The number of such constraints we need to impose to ensure that $f \in \mathrm{Pol}_3(\mathbf{A})$ is then

$$\sum_i |R_i|^3,$$

which is at most cubic in the size of the description of $\mathbf{A}$. We ensure that $g \in \mathrm{Pol}_3(\mathbf{A})$ with a similar collection of constraints.

To enforce the required identities between f, g, for every pair $a, b \in A$, we impose the equality constraints

$$g(a, a, b) = g(a, b, a) = g(b, a, a) = f(a, a, b) = f(a, b, a) = f(a, b, b).$$

This requires a total of $5|A|^2$ equality constraints. Thus, the instance $\mathbf{X}$ has overall size at most cubic in the size of the description of $\mathbf{A}$.

In order to solve $\mathbf{X}$, we view it as an instance of $\mathrm{CSP}(\mathbf{A}^{rig})$, where $\mathbf{A}^{rig}$ is the rigid core obtained from $\mathbf{A}$ by adding a singleton unary relation $\{a\}$ for each element $a \in A$. Note that if $\mathbf{A}$ is a core, then $\mathbf{A}$ has bounded width iff $\mathbf{A}^{rig}$ has bounded width (since each pp-constructs the other). We now attempt to solve the instance $\mathbf{X}$ by using the cycle-consistency algorithm, as follows. For each

variable v of $\mathbf{X}$, we go through the values $a \in A$ in order, and temporarily modify $\mathbf{X}$ by adding the extra constraint $v \in \{a\}$ to make an instance $\mathbf{X}_{v=a}$. Then we reduce $\mathbf{X}_{v=a}$ until it either becomes cycle-consistent or until we reach a contradiction. If there is any $a \in A$ such that $\mathbf{X}_{v=a}$ becomes cycle-consistent, then we replace $\mathbf{X}$ by $\mathbf{X}_{v=a}$ and move on to the next variable. If every choice of $a \in A$ leads to $\mathbf{X}_{v=a}$ reaching a contradiction, then we give up and report that $\mathrm{CSP}(\mathbf{A})$ does not have bounded width.

If the procedure ends without us giving up, then we have found f, g as in Theorem 3.14.1 and these terms prove that $\mathrm{CSP}(\mathbf{A})$ has bounded width. Conversely, if $\mathrm{CSP}(\mathbf{A})$ has bounded width, then the original instance $\mathbf{X}$ has a solution, and each time we replace $\mathbf{X}$ by $\mathbf{X}_{v=a}$, the fact that $\mathbf{X}_{v=a}$ can be reduced to a cycle-consistent instance implies that it has a solution, so the whole procedure will end by successfully finding a pair of functions f, g. Of course, if $\mathrm{CSP}(\mathbf{A})$ does not have bounded width, then we will fail to find a solution to $\mathbf{X}$. $\square$

A simple iteration argument allows us to give a criterion for bounded width involving just one ternary term and a binary term derived from it - however, the identities involved will not have height 1, so these new terms are unsuitable for the application to the meta-problem.

Theorem 3.14.8. *A finite relational structure $\mathbf{A}$ has bounded relational width if and only if it has a ternary polymorphism $g \in \mathrm{Pol}_3(\mathbf{A})$ such that, if f is the binary term $f(x, y) \coloneqq g(x, x, y)$, we have*

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

Proof. Suppose first that $\mathbf{A}$ has bounded relational width, and let $f_3, g_3 \in \mathrm{Pol}_3(\mathbf{A})$ be terms as in Theorem 3.14.1. Define a sequence of terms g^i by $g^1 \coloneqq g_3$ and

$$g^{i+1}(x, y, z) \coloneqq g^i(f_3(x, y, z), f_3(y, z, x), f_3(z, x, y)).$$

Define binary terms f^i by $f^i(x, y) \coloneqq g^i(x, x, y)$. Then we have

$$f^1(x, y) \approx g_3(x, x, y) \approx f_3(x, x, y) \approx f_3(x, y, x) \approx f_3(x, y, y),$$

and for each i we have

$$\begin{aligned}
f^{i+1}(x, y) \approx g^{i+1}(x, x, y) &\approx g^i(f_3(x, x, y), f_3(x, y, x), f_3(y, x, x)) \\
&\approx g^i(f^1(x, y), f^1(x, y), f^1(y, x)) \approx f^i(f^1(x, y), f^1(y, x)).
\end{aligned}$$

Thus the sequence $f^i(x, y)$ is generated by iterating the map $(x, y) \mapsto (f^1(x, y), f^1(y, x))$. Since $\mathbf{A}$ is finite, there is some N such that $g^N \approx g^{2N}$ and $f^N \approx f^{2N}$. Take $f \coloneqq f^N$ and $g \coloneqq g^N$ to finish the construction.

Now suppose that f, g satisfy the assumed identities. Let e be the unary operation $e(x) \coloneqq f(x, x) = g(x, x, x)$. The identity

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

implies that

$$e(e(x)) \approx e(x),$$

so $\mathbf{A}$ is homomorphically equivalent to $e(\mathbf{A})$, and the restrictions of f, g to $e(\mathbf{A})$ are idempotent. Let $\mathbb{A}_e$ be the idempotent algebra $(e(\mathbf{A}), f|_{e(\mathbf{A})}, g|_{e(\mathbf{A})})$. We will show that $\mathbb{A}_e$ is Taylor and affine-free.

214

That $\mathbb{A}_e$ is Taylor follows from the identity

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

For the sake of contradiction, assume that $\mathbb{B} \in HSP(\mathbb{A}_e)$ is a nontrivial affine algebra. Then we can write

$$g(x, y, z) \approx \alpha(x + y + z)$$

on $\mathbb{B}$, for some α with $3\alpha x \approx x$. Then we have

$$f(x, y) \approx 2\alpha x + \alpha y,$$

so

$$f(f(x, y), f(y, x)) \approx 2\alpha(2\alpha x + \alpha y) + \alpha(2\alpha y + \alpha x) \approx 5\alpha^2 x + 4\alpha^2 y.$$

Equating these and setting y to 0, we see that $2\alpha x \approx 5\alpha^2 x$. Multiplying by 9 and using $3\alpha x \approx x$, we get $6x \approx 5x$, so $x \approx 0$ on $\mathbb{B}$, a contradiction. $\qquad\square$

The identities satisfied by the term g of Theorem 3.14.8 have the following nice consequence.

Proposition 3.14.9. *Suppose that g is a ternary term as in Theorem 3.14.8, and that f is the associated binary term. Then for any a, b, either $f(a, b) = f(b, a)$, or the set $\{f(a, b), f(b, a)\}$ is closed under g, and $(\{f(a, b), f(b, a)\}, g)$ is isomorphic to a two-element majority algebra.*

For small examples of bounded width algebras $\mathbb{A}$ which do not contain large majority subalgebras, most of the structure of a bounded width algebra seems to be controlled by the binary term f from Theorem 3.14.8, with the exact values of the ternary term g only playing an important role on the majority subalgebras. I have also conjectured a very strong refinement of Theorem 3.14.8, which would give a much more explicit structure theory for bounded width algebras.

Conjecture 3.14.1. A finite idempotent algebra $\mathbb{A}$ has bounded relational width if and only if it has a ternary term m and an associated binary term $s(x, y) \coloneqq m(x, x, y)$, which satisfy the identities

$$m(x, x, y) \approx m(x, y, x) \approx m(y, x, x) \approx s(x, y)$$

and

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

3.15 Semidefinite Programming robustly solves bounded width CSPs

In this section we finally touch on a difficult topic: trying to maximize the number of satisfied constraints in a CSP instance which has no perfect solution. We consider only a very special case of this problem here: the problem of trying to approximately solve a CSP when we are promised that there exists a way to satisfy all but a tiny fraction of the constraints. This problem was considered by Guruswami and Zhou in [69].

Definition 3.15.1. We say that $\text{CSP}(\mathbf{A})$ is *robustly solvable* if there is a function $f : [0,1] \to [0,1]$ such that
$$\lim_{\epsilon \to 0} f(\epsilon) = 0,$$
and a polynomial time algorithm that takes as input an instance $\mathbf{X}$ of $\text{CSP}(\mathbf{A})$, and outputs an assignment to the variables of $\mathbf{X}$ such that if it is possible to satisfy a $1 - \epsilon$ fraction of the constraints of $\mathbf{X}$, then the assignment found by the algorithm satisfies at least a $1 - f(\epsilon)$ fraction of the constraints of $\mathbf{X}$.

Before we dive into our main topic, we first give evidence that certain CSPs are *not* robustly solvable. We won't prove the next result here.

Theorem 3.15.2 (Håstad [75])**.** *Let $\mathbf{A}$ be the affine CSP template with domain $\mathbb{A}$, where $\mathbb{A}$ is the idempotent reduct of any finite abelian group, and with relations given by $\mathbb{R}_c = \{(x, y, z) \mid x+y+z = c\} \leq_{sd} \mathbb{A}^3$ for every possible $c \in \mathbb{A}$.*

Then for every fixed $\epsilon > 0$, it is NP-hard to solve the following problem: given an instance $\mathbf{X}$ of $\text{CSP}(\mathbf{A})$ such that there exists an assignment satisfying at least a $1 - \epsilon$ fraction of the constraints, find an assignment which satisfies at least a $\frac{1}{|\mathbb{A}|} + \epsilon$ fraction of the constraints.

Note that for the affine CSP defined above, randomly guessing values for variables will produce an assignment which satisfies a $\frac{1}{|\mathbb{A}|}$ fraction of the constraints, on average. So Håstad's result tells us that it's NP-hard to find any improvement on randomly guessing, for affine CSPs which are not perfectly solvable.

Corollary 3.15.3. *If $\text{CSP}(\mathbf{A})$ is robustly solvable and $P \neq NP$, then $\mathbf{A}$ must be affine-free (and therefore $\mathbf{A}$ has bounded width).*

The best known approach to approximately solving CSPs, based on semidefinite programming, was laid out in Raghavendra's thesis [125] (see [124] for a short overview of the results). Under the Unique Games Conjecture, Raghavendra proved that this approach is actually optimal. The strategy is as follows.

As in the linear programming relaxation of a CSP, we imagine that we are looking for a probability distribution over solutions to the CSP. We do not give a full description of this unknown probability distribution: we only describe the marginal distribution over assignments to tuples of variables belonging to constraints of the CSP, as well as the marginal distribution over assignments to each pair of variables in the CSP. We impose compatibility conditions between the marginal distributions over each tuples of variables $(v_1, ..., v_m)$ belonging to some constraint and the marginal distribution over each pair (v_i, v_j) for $i, j \leq m$.

So far all the conditions given can be described by a system of linear inequalities. The semidefinite aspect comes from the following observation: every covariance matrix of any collection of random variables must be positive semidefinite.

To be more concrete, for each pair of variables x, y and each pair of values $a \in \mathbb{A}_x, b \in \mathbb{A}_y$, we have some variable $p_{(x,a),(y,b)}$ between 0 and 1, describing the probability that x is assigned the value a and y is assigned the value b. We create a matrix M_p with rows and columns indexed by ordered pairs (x, a) with $a \in \mathbb{A}_x$, and fill the $(x, a), (y, b)$ entry with $p_{(x,a),(y,b)}$ (I like to imagine M_p as a block matrix, with each block of rows or columns corresponding to a particular variable x). Then the matrix M_p must be positive semidefinite if these probabilities come from an actual probability distribution.

Before defining everything formally, we give an example.

216

Example 3.15.1. Consider the following instance of 2-SAT: we have three variables x, y, z, and each pair of variables has a $\neq$ constraint imposed between them. This instance has no perfect solution, but the standard linear programming relaxation is incapable of noticing this. Let's see how the semidefinite relaxation does.

The matrix M_p has six rows and six columns, corresponding to the pairs $(x,0), (x,1), (y,0), (y,1), (z,0), (z,1)$, in that order. If M_p comes from a probability distribution over perfect solutions to this instance of 2-SAT, then it must have the following shape:

$$M_p = \begin{pmatrix} * & 0 & 0 & * & 0 & * \\ 0 & * & * & 0 & * & 0 \\ 0 & * & * & 0 & 0 & * \\ * & 0 & 0 & * & * & 0 \\ 0 & * & 0 & * & * & 0 \\ * & 0 & * & 0 & 0 & * \end{pmatrix}.$$

Additionally, the entries in each block of M_p must sum to 1 (and be ≥ 0), and for each fixed row or column of M_p, the sum of the entries in the intersection of the row/column with any block must only depend on the row/column. Putting these linear constraints together, we quickly see that every nonzero entry of M_p must actually be equal to $\frac{1}{2}$. So far, this is exactly what the linear programming relaxation will guess.

The matrix M_p found above, with all nonzero entries equal to $\frac{1}{2}$, is *not* positive semidefinite. To see this, note that we have

$$\begin{pmatrix} 1 & -1 & | & 1 & -1 & | & 1 & -1 \end{pmatrix} \begin{pmatrix} \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \\ 1 \\ -1 \end{pmatrix} = -3 < 0.$$

So the semidefinite relaxation of the problem can detect that we can't perfectly solve this instance of 2-SAT.

Now suppose that we give up on finding a perfect solution, and instead look for an approximate solution. This means that some of the entries of M_p which were required to be 0 before are instead required to be *small*. One choice of M_p that works is

$$M_p = \frac{1}{8} \begin{pmatrix} 4 & 0 & 1 & 3 & 1 & 3 \\ 0 & 4 & 3 & 1 & 3 & 1 \\ 1 & 3 & 4 & 0 & 1 & 3 \\ 3 & 1 & 0 & 4 & 3 & 1 \\ 1 & 3 & 1 & 3 & 4 & 0 \\ 3 & 1 & 3 & 1 & 0 & 4 \end{pmatrix},$$

which the reader may verify is positive semidefinite. This seems to satisfy each particular constraint with a probability of $\frac{3}{4}$. So the semidefinite relaxation thinks it might be possible to satisfy a $\frac{3}{4}$ fraction of the constraints. An easy brute force search reveals that the best we can do in reality is to satisfy a $\frac{2}{3}$ fraction of the constraints.

There is one further step we will take to analyze the semidefinite relaxation, based on a standard fact from linear algebra about positive semidefinite matrices.

Proposition 3.15.4. *If M is an $n \times n$ positive semidefinite matrix, then there is a collection of vectors $x_1, ..., x_n \in \mathbb{R}^n$ such that $M_{ij} = x_i \cdot x_j$ for all $i, j \leq n$. Such a collection of vectors $x_1, ..., x_n$ can be computed from M in polynomial time.*

Proof. Perhaps the simplest approach is to compute a Cholesky decomposition of M, writing $M = LL^T$ for some lower triangular matrix L. The columns of L^T can then be used as the vectors $x_1, ..., x_n$. $\square$

Example 3.15.2. The matrix M_p from the end of the previous example is positive semidefinite, so there should exist vectors $x_0, x_1, y_0, y_1, z_0, z_1 \in \mathbb{R}^6$ whose matrix of dot products is equal to M_p. Since M_p has rank 3, we should even be able to find such vectors in $\mathbb{R}^3$. One particularly satisfying choice of vectors that works is

$$x_0 = \frac{1}{\sqrt{24}} \begin{bmatrix} \sqrt{2} - \sqrt{3} \\ \sqrt{2} \\ \sqrt{2} + \sqrt{3} \end{bmatrix}, x_1 = \frac{1}{\sqrt{24}} \begin{bmatrix} \sqrt{2} + \sqrt{3} \\ \sqrt{2} \\ \sqrt{2} - \sqrt{3} \end{bmatrix}, y_0 = \frac{1}{\sqrt{24}} \begin{bmatrix} \sqrt{2} \\ \sqrt{2} + \sqrt{3} \\ \sqrt{2} - \sqrt{3} \end{bmatrix}, y_1 = \frac{1}{\sqrt{24}} \begin{bmatrix} \sqrt{2} \\ \sqrt{2} - \sqrt{3} \\ \sqrt{2} + \sqrt{3} \end{bmatrix},$$

with z_0, z_1 defined similarly by cyclically shifting y_0, y_1.

Now we can give the definition of the basic semidefinite relaxation of a CSP (this is the LC relaxation from [125]).

Definition 3.15.5. Given an instance $\mathbf{X}$ of a CSP, with variable domains $\mathbb{A}_v$ and constraints C imposing relations $\mathbb{R}_C \leq \prod_{i \leq m_C} \mathbb{A}_{v_{C,i}}$ on the variables $v_{C,1}, ..., v_{C,m}$, the *basic semidefinite relaxation* of $\mathbf{X}$ is the following optimization problem. We wish to find a system of "probabilities" $p_{C,r}$ for $r \in \prod_{i \leq m_C} \mathbb{A}_{v_{C,i}}$, such that

$$\sum_r p_{C,r} = 1$$

for each constraint C and

$$p_{C,r} \geq 0$$

for each C, r, and to find vectors

$$x_a \in \mathbb{R}^N$$

for each variable x and value $a \in \mathbb{A}_x$, where $N = \sum_x |\mathbb{A}_x|$ is the number of pairs (x, a), such that for each C and each pair $i, j \leq m_C$ we satisfy the compatibility condition

$$(v_{C,i})_a \cdot (v_{C,j})_b = \sum_{r_i = a, r_j = b} p_{C,r}.$$

These vectors must also satisfy the following constraints, for each pair of variables x, y of $\mathbf{X}$.

- For all $a \neq b \in \mathbb{A}_x$, we have $x_a \cdot x_b = 0$.

- We have $\sum_{a \in \mathbb{A}_x} \|x_a\|^2 = \| \sum_{a \in \mathbb{A}_x} x_a \|^2 = 1$.

- For all $a \in \mathbb{A}_x, b \in \mathbb{A}_y$, we have $x_a \cdot y_b \geq 0$.

- We have $\sum_{a \in \mathbb{A}_x, b \in \mathbb{A}_y} x_a \cdot y_b = 1$.

Note that these constraints are slightly redundant - if every pair of variables shows up in the scope of some constraint C, then they are implied by the compatibility conditions between the probabilities $p_{C,r}$ and the dot products of the vectors.

Our goal is to find such a system of probabilities $p_{C,r}$ and vectors x_a such that the quantity

$$\frac{1}{\#C} \sum_C \sum_{r \in \mathbb{R}_C} p_{C,r}$$

is maximized. The maximum possible value of that sum is called the *value* of the semidefinite relaxation. If the value of the semidefinite relaxation is equal to 1, then we say that the system of probabilities $p_{C,r}$ and vectors x_a *perfectly solves* the semidefinite relaxation.

Note that the constraints we make on the vectors x_a only involve their dot products, and that they are always linear equalities/inequalities in terms of these dot products. We can deduce from these constraints a result which involves the vectors directly.

Proposition 3.15.6. *Suppose that a system of probabilities $p_{C,r}$ and vectors x_a are as in the basic semidefinite relaxation of a CSP instance $\mathbf{X}$. Then for any variables x, y of $\mathbf{X}$, we have*

$$\sum_{a \in \mathbb{A}_x} x_a = \sum_{b \in \mathbb{A}_y} y_b.$$

Proof. Let $x_{\mathbb{A}_x} = \sum_{a \in \mathbb{A}_x} x_a$ and similarly let $y_{\mathbb{A}_y} = \sum_{b \in \mathbb{A}_y} y_b$. Then we have $\|x_{\mathbb{A}_x}\|^2 = \|y_{\mathbb{A}_y}\|^2 = x_{\mathbb{A}_x} \cdot y_{\mathbb{A}_y} = 1$, so by the equality case of Cauchy-Schwarz we must have $x_{\mathbb{A}_x} = y_{\mathbb{A}_y}$. $\square$

A useful generalization of the notation for the vectors x_a was used heavily in [18].

Definition 3.15.7. Suppose we are in the setup of the basic semidefinite relaxation of a CSP instance $\mathbf{X}$. If x is a variable of $\mathbf{X}$ and $A \subseteq \mathbb{A}_x$, then we define the vector x_A by

$$x_A = \sum_{a \in A} x_a.$$

Note that since the vectors x_a are pairwise orthogonal for a fixed variable x, we have

$$\|x_A\|^2 = \sum_{a \in A} \|x_a\|^2 = x_A \cdot x_{\mathbb{A}_x}.$$

Additionally, for each pair of variables x, y, we have

$$x_A \cdot y_B = \sum_{a \in A, b \in B} x_a \cdot y_b$$

by the distributive law.

Before explaining how to use the basic semidefinite relaxation to robustly solve affine-free CSPs, we will first show that if an instance $\mathbf{X}$ of an affine-free CSP has a perfect solution to its basic semidefinite relaxation, then in fact $\mathbf{X}$ has a solution. The main idea is to prove that the set of values $a \in \mathbb{A}_x$ such that the vectors x_a are nonzero can be used to restrict the instance to get a weak Prague instance, which will then be *pq*-consistent by Theorem 3.13.19. The crucial computation is analyzing what happens to the vector x_A when we take a single step along a path.

Lemma 3.15.8. *Suppose we are in the setup of the basic semidefinite relaxation of a CSP instance* **X**. *Let x, y be variables of* **X**, *and define a binary relation $P \subseteq \mathbb{A}_x \times \mathbb{A}_y$ by*

$$(a, b) \in P \quad \Longleftrightarrow \quad x_a \cdot y_b > 0.$$

Then for any set $A \subseteq \mathbb{A}_x$, we have

$$\|x_A\|^2 \leq \|y_{A+P}\|^2,$$

with equality only when $x_A = y_{A+P}$. In fact, we have

$$x_A \cdot (y_{A+P} - x_A) = 0,$$

that is, y_{A+P} is the sum of x_A with a vector which is perpendicular to x_A.

Proof. We have $x_A \cdot x_A = x_A \cdot x_{\mathbb{A}_x}$, and by the definition of P we have

$$x_A \cdot y_{A+P} = \sum_{a \in A, b \in A+P} x_a \cdot y_b = \sum_{a \in A, b \in \mathbb{A}_y} x_a \cdot y_b = x_A \cdot y_{\mathbb{A}_y}.$$

Since $x_{\mathbb{A}_x} = y_{\mathbb{A}_y}$, we have

$$x_A \cdot x_A = x_A \cdot x_{\mathbb{A}_x} = x_A \cdot y_{\mathbb{A}_y} = x_A \cdot y_{A+P},$$

so x_A is orthogonal to $y_{A+P} - x_A$.

From the orthogonality of x_A with $y_{A+P} - x_A$, we have

$$\|y_{A+P}\|^2 = \|x_A + (y_{A+P} - x_A)\|^2 = \|x_A\|^2 + \|y_{A+P} - x_A\|^2 \geq \|x_A\|^2,$$

with equality exactly when $y_{A+P} = x_A$. $\qquad\square$

Theorem 3.15.9. *Suppose* **X** *is an instance of an affine-free CSP, and that there is a system of probabilities $p_{C,r}$ and vectors x_a which perfectly solves the basic semidefinite relaxation of* **X**. *Then* **X** *has a solution.*

Proof. We define a restriction $\mathbf{X}'$ of **X** by restricting each relation $\mathbb{R}_C$ to the support R'_C of the marginal distribution $p_{C,r}$, and restricting each variable domain $\mathbb{A}_x$ to the set A'_x of $a \in \mathbb{A}_x$ such that $\|x_a\|^2 \neq 0$. Note that each R'_C will be contained in the original relation $\mathbb{R}_C$ if we have a perfect solution to the basic semidefinite relaxation. Additionally, for each constraint C and each $i, j \leq m_C$, the binary projection $\pi_{i,j}(R'_C)$ will be equal to the set of ordered pairs $(a, b) \in A'_{v_{C,i}} \times A'_{v_{C,j}}$ such that $(v_{C,i})_a \cdot (v_{C,j})_b \neq 0$, by the compatibility between the probabilities $p_{C,r}$ and the vectors x_a.

We will check that $\mathbf{X}'$ is a weak Prague instance (see Definition 3.13.12). Arc-consistency (aka condition (P1)) of $\mathbf{X}'$ follows from the compatibility between the probabilities and the vectors. To check (P2) and (P3), we use Lemma 3.15.8. Let $A \subseteq A'_x$ and let p be a cycle from x to x in the instance $\mathbf{X}'$, with p_1 from x to y the first step of the cycle p. If we have

$$A + p = A,$$

then by Lemma 3.15.8 we have

$$\|x_A\|^2 \leq \|y_{A+p_1}\|^2 \leq \|x_{A+p}\|^2 = \|x_A\|^2,$$

so by the equality case of Lemma 3.15.8 we must have

$$x_A = y_{A+p_1}.$$

Thus for any $a' \notin A$, we must have $x_{a'} \cdot y_{A+p_1} = x_{a'} \cdot x_A = 0$, so we have

$$A + p_1 - p_1 = A.$$

Thus by Proposition 3.13.14 we see that $\mathbf{X}'$ satisfies condition (P2). We could have also checked condition (P2) using only the system of probabilities $p_{C,r}$, without mentioning the vectors x_a, by Theorem 3.13.15.

To check (P3), let $A \subseteq A'_x$, and let p, q be cycles from x to x in the instance $\mathbf{X}'$, with

$$A + p + q = A.$$

Then by Lemma 3.15.8 we have

$$\|x_A\|^2 \le \|x_{A+p}\|^2 \le \|x_{A+p+q}\|^2 = \|x_A\|^2,$$

so by the equality case of Lemma 3.15.8 we must have

$$x_A = x_{A+p}.$$

In particular, we must have $A = A + p$, which proves (P3).

To finish, we note that $\mathbf{X}'$ is pq-consistent by Theorem 3.13.19, so $\mathrm{Sg}(\mathbf{X}')$ is also pq-consistent by Proposition 3.14.3, and $\mathrm{Sg}(\mathbf{X}')$ is a restriction of the instance $\mathbf{X}$ since $\mathbf{X} = \mathrm{Sg}(\mathbf{X})$. Thus $\mathrm{Sg}(\mathbf{X}')$ has a solution by Theorem 3.13.8 and its corollaries, which is also a solution to the original instance $\mathbf{X}$. $\qquad\square$

In order to extend the previous result to an algorithm for *robustly* solving affine-free CSPs, we need to find some approximate analogue of Lemma 3.15.8. The plan is to start by arguing as in Theorem 1.6.17, using the probabilities $p_{C,r}$ to produce an arc-consistent instance, and then to randomly cut the unit ball of possible values for the vectors x_A into finitely many pieces, so that a version of Lemma 3.15.8 holds when we forget what the exact values of the vectors x_A are, and instead only keep track of which piece of the ball x_A is contained in. I plan to write up a detailed exposition of this at some point, but for now the interested reader can find the proof in [18].

Chapter 4

Finite Taylor Algebras

4.1 Cyclic terms

In this section we will prove that every finite Taylor algebra has a cyclic term.

Definition 4.1.1. An m-ary operation c is called *cyclic* if it satisfies the identity

$$c(x_1, x_2, ..., x_m) \approx c(x_2, ..., x_m, x_1).$$

Cyclic terms were first proved to exist for finite congruence modular algebras [20], and most of the basic facts about cyclic terms are developed in that paper. This was extended to finite congruence join-semidistributive algebras in [16], and then finally to all finite Taylor algebras in [17]. We'll start by showing that we only care about cyclic terms of prime arity.

Proposition 4.1.2 (Multiplicative property of cyclic terms [20])**.** *A variety $\mathcal{V}$ has a cyclic term c_{mn} of arity mn if and only if $\mathcal{V}$ has cyclic terms c_m, c_n of arity m and n, respectively.*

Proof. Suppose first that c_{mn} is a cyclic term of arity mn. Then we can define a cyclic term of arity m by plugging in

$$c_{mn}(\underbrace{x_1, ..., x_1}_{n}, \underbrace{x_2, ..., x_2}_{n}, ..., \underbrace{x_m, ..., x_m}_{n}),$$

and we can define a cyclic term of arity n similarly.

Conversely, suppose that c_m, c_n are cyclic terms of arity m and n. We define a cyclic term of arity mn by renumbering the inputs of the star composition $c_n * c_m$:

$$c_n \begin{pmatrix} c_m(x_1, & x_{n+1}, & ..., & x_{(m-1)n+1}), \\ c_m(x_2, & x_{n+2}, & ..., & x_{(m-1)n+2}), \\ \vdots & \vdots & \ddots & \vdots \\ c_m(x_n, & x_{2n}, & ..., & x_{mn}) \end{pmatrix} \approx c_n \begin{pmatrix} c_m(x_2, & x_{n+2}, & ..., & x_{(m-1)n+2}), \\ \vdots & \vdots & \ddots & \vdots \\ c_m(x_n, & x_{2n}, & ..., & x_{mn}), \\ c_m(x_{n+1}, & ..., & x_{(m-1)n+1}, & x_1) \end{pmatrix}. \quad \square$$

Corollary 4.1.3. *A variety $\mathcal{V}$ has a cyclic term of arity m if and only if $\mathcal{V}$ has a cyclic term of arity p for every prime p which divides m.*

Next we will describe the main obstruction to the existence of a cyclic term of a given arity.

Proposition 4.1.4 (Semantic meaning of cyclic terms [20]). *Suppose that $\mathcal{V}$ is a variety. Then for any $m \in \mathbb{N}$, the following are equivalent.*

    (a) $\mathcal{V}$ has no cyclic term of arity m.

    (b) There is some $\mathbb{A} \in \mathcal{V}$ and an automorphism $\sigma \in \operatorname{Aut}(\mathbb{A})$ such that $\sigma^m = 1$ and σ has no fixed point.

Proof. We start by showing that (b) implies (a). Suppose that $\mathbb{A}, \sigma$ are as in (b), and suppose for contradiction that $\mathbb{A}$ has some cyclic term c_m of arity m. Let a be any element of $\mathbb{A}$, and define a_i by $a_i = \sigma^i(a)$. Then we have

$$c_m \begin{pmatrix} a_1, & a_2, & ..., & a_m \\ a_2, & a_3, & ..., & a_1 \end{pmatrix} \in \sigma,$$

so $c_m(a_1, ..., a_m)$ is a fixed point of σ, contradicting the assumption in (b).

Now suppose that (a) holds. Let $\mathbb{A} = \mathcal{F}_{\mathcal{V}}(x_1, ..., x_m)$ be the free algebra on m generators, and let σ be the automorphism of $\mathbb{A}$ defined by cyclically permuting the generators $x_1, ..., x_m$. Then a fixed point of σ is precisely the same thing as a cyclic term of $\mathcal{V}$ of arity m, so if $\mathcal{V}$ has no cyclic term of arity m, then σ has no fixed points (and satisfies $\sigma^m = 1$). $\square$

For finite algebras, we can give a local criterion for the existence of a cyclic term.

Proposition 4.1.5 (Local criterion for cyclic terms [20]). *If $\mathbb{A}$ is a finite algebra, then $\mathbb{A}$ has an m-ary cyclic term if and only if it is the case that for all $a_1, ..., a_m \in \mathbb{A}$, there exists some m-ary term t such that*

$$t(a_1, a_2, ..., a_m) = t(a_2, ..., a_m, a_1) = \cdots = t(a_m, a_1, ..., a_{m-1}).$$

Proof. Say that an m-ary term t is cyclic for a tuple $(a_1, ..., a_m)$ if it satisfies the displayed equation from the statement of the proposition. Let c be an m-ary term which is cyclic for a maximal set of tuples (we are using finiteness of $\mathbb{A}$ here). Suppose for contradiction that c is not cyclic, and let $a = (a_1, ..., a_m)$ be any tuple such that c is not cyclic for a.

Define a tuple $a' = (a'_1, ..., a'_m)$ by

$$a'_i = c(a_i, a_{i+1}, ..., a_{i-1}),$$

with indices taken modulo m. By assumption, there is some m-ary term t which is cyclic for a'. But then the m-ary term

$$t(c(x_1, x_2..., x_m), c(x_2, ..., x_m, x_1), ..., c(x_m, x_1, ..., x_{m-1}))$$

is cyclic for a, and is also cyclic for every tuple which c was cyclic for, contradicting the maximality assumption on c. $\square$

For the sake of checking the local condition of Proposition 4.1.5 for a particular tuple $a_1, ..., a_m$, the natural approach is to compute the m-ary relation

$$\operatorname{Sg}_{\mathbb{A}^m} \left\{ \begin{bmatrix} a_1 & a_2 & \cdots & a_m \\ a_2 & a_3 & \cdots & a_1 \\ \vdots & \vdots & \ddots & \vdots \\ a_m & a_1 & \cdots & a_{m-1} \end{bmatrix} \right\},$$

223

and to check if it contains any constant tuples. This relation is invariant under cyclically permuting its coordinates, which leads us to make the following definition.

Definition 4.1.6. A relation $\mathbb{R} \leq \mathbb{A}^m$ is called *cyclic* if $\mathbb{R}$ is invariant under cyclically permuting its coordinates, that is,

$$(a_1, a_2, ..., a_m) \in \mathbb{R} \quad \Longleftrightarrow \quad (a_2, ..., a_m, a_1) \in \mathbb{R}.$$

Corollary 4.1.7 (Relational criterion for cyclic terms [20])**.** *If $\mathbb{A}$ is a finite algebra, then $\mathbb{A}$ has a cyclic term of arity m if and only if every m-ary cyclic relation $\mathbb{R} \leq \mathbb{A}^m$ contains a constant tuple.*

Now we are finally ready to prove one of the main results of [17], which states that every finite Taylor algebra $\mathbb{A}$ has cyclic terms of every prime arity $p > |\mathbb{A}|$. In fact, we will prove a stronger version of this result due to Zhuk (currently unpublished).

Theorem 4.1.8 (Finite Taylor algebras have cyclic terms [17], refined by Zhuk)**.** *Suppose $\mathbb{A}$ is a finite idempotent Taylor algebra and that p is prime. Then one of the following is true:*

(a) *either $\mathbb{A}$ has a cylic term of arity p, or*

(b) *there is some $\mathbb{B} \in HS(\mathbb{A})$ and some automorphism $\sigma \in \mathrm{Aut}(\mathbb{B})$ such that $\sigma^p = 1$ and σ has no fixed points.*

In particular, if $p > |\mathbb{A}|$ then $\mathbb{A}$ has a cyclic term of arity p.

Proof. We prove this by induction on $|\mathbb{A}|$. Suppose that there is no $\mathbb{B} \in HS(\mathbb{A}), \sigma \in \mathrm{Aut}(\mathbb{B})$ as in (b), and let $\mathbb{R} \leq \mathbb{A}^p$ be any p-ary cyclic relation. It's enough to show that $\mathbb{R}$ contains a constant tuple.

If $\pi_1(\mathbb{R}) \neq \mathbb{A}$, then since $\mathbb{R}$ is cyclic we have $\mathbb{R} \leq \pi_1(\mathbb{R})^p$, so we can apply the induction hypothesis to the algebra $\pi_1(\mathbb{R})$ to see that $\mathbb{R}$ has a constant tuple. Thus we may assume without loss of generality that $\mathbb{R}$ is subdirect in $\mathbb{A}^p$.

If $\mathbb{A}$ has a nontrivial congruence $\theta \in \mathrm{Con}(\mathbb{A})$, then $\mathbb{R}/\theta^p \leq (\mathbb{A}/\theta)^p$ is a cyclic relation on $\mathbb{A}/\theta$, so by the induction hypothesis applied to $\mathbb{A}/\theta$ there is some congruence class a/θ such that $\mathbb{R} \cap (a/\theta)^p \neq \emptyset$. Setting $\mathbb{R}' = \mathbb{R} \cap (a/\theta)^p$, we see that $\mathbb{R}'$ is a cyclic relation on a/θ, so by the induction hypothesis applied to a/θ we see that $\mathbb{R}'$ (and therefore also $\mathbb{R}$) has a constant tuple. Thus we may assume that $\mathbb{A}$ is simple.

If any $\pi_{ij}(\mathbb{R})$ is the graph of an automorphism σ of $\mathbb{A}$, then since $\mathbb{R}$ is cyclic, we see that $\pi_{j,2j-i}(\mathbb{R})$ is also the graph of σ, and similarly so is $\pi_{2j-i,3j-2i}(\mathbb{R})$, etc., and we must have $\sigma^p = 1$. Since p is prime, we see that in fact every $\pi_{kl}(\mathbb{R})$ is the graph of some power of the automorphism σ. In this case we see that $\mathbb{R}$ has a constant tuple if and only σ has a fixed point. Thus we may assume without loss of generality that every $\pi_{ij}(\mathbb{R})$ is linked.

By Zhuk's four cases (Corollary 3.12.12), we see that $\mathbb{A}$ is either affine, subdirectly complete, or has a proper ternary absorbing subalgebra.

If $\mathbb{A}$ is affine, with underlying abelian group $(A, +, -, 0)$, then since $x - y + z$ is a term of $\mathbb{A}$ (by the definition of an affine algebra), we see that $k_1 x_1 + \cdots + k_m x_m$ is a term of $\mathbb{A}$ for all $k_1, ..., k_m \in \mathbb{Z}$ such that $k_1 + \cdots + k_m \equiv 1 \pmod{|\mathbb{A}|}$. In particular, if $p \nmid |\mathbb{A}|$, then

$$p^{-1}(x_1 + \cdots + x_p)$$

is a p-ary cyclic term of $\mathbb{A}$. On the other hand, if $p \mid |\mathbb{A}|$, then by elementary group theory there must be some element $c \in \mathbb{A}$ of order p, and then by the idempotence of $\mathbb{A}$ the relation

$$\{(x,y) \mid x = y + c\}$$

is a subalgebra of $\mathbb{A}^2$, and it is then the graph of an automorphism σ of $\mathbb{A}$ which has order p and has no fixed points.

If $\mathbb{A}$ is subdirectly complete, then since $\mathbb{R} \leq_{sd} \mathbb{A}^p$ is subdirect and every $\pi_{ij}(\mathbb{R})$ is linked, we must have $\mathbb{R} = \mathbb{A}^p$. In this case $\mathbb{R}$ contains *every* constant tuple.

If $\mathbb{A}$ has a proper ternary absorbing subalgebra, then we define a directed graph $\mathbf{D}$ whose vertices are proper ternary absorbing subalgebras $\mathbb{B} \lhd \mathbb{A}$, and with a directed edge $(\mathbb{B}, \mathbb{B} + \pi_{ij}(\mathbb{R}))$ whenever $\mathbb{B} + \pi_{ij}(\mathbb{R}) \neq \mathbb{A}$ and $i \neq j$.

Claim: The digraph $\mathbf{D}$ has no directed cycles.

Proof of claim: Note first that since $\mathbb{R}$ is cyclic we have

$$\pi_{ij}(\mathbb{R})^- \subseteq \pi_{ij}(\mathbb{R})^{\circ(p-1)},$$

so if $\mathbb{B} + \pi_{ij}(\mathbb{R}) = \mathbb{B}$ then we must have

$$\mathbb{B} + \pi_{ij}(\mathbb{R}) - \pi_{ij}(\mathbb{R}) = \mathbb{B},$$

so $\mathbb{B}$ is a union of linked components of $\pi_{ij}(\mathbb{R})$. Since $\pi_{ij}(\mathbb{R})$ is linked and $\mathbb{B}$ is proper, this is impossible. Thus $\mathbf{D}$ has no directed cycles of length 1. Since $\mathbb{R}$ is cyclic, we also have

$$\pi_{ij}(\mathbb{R}) \circ \pi_{kl}(\mathbb{R}) \supseteq \pi_{i+k,j+l}(\mathbb{R}),$$

so if $\mathbf{D}$ has a directed cycle, then $\mathbf{D}$ has a directed cycle of length 2, of the form $\mathbb{B}+\pi_{ij}(\mathbb{R})+\pi_{kl}(\mathbb{R}) = \mathbb{B}$. If $i + k \neq j + l$ this gives us a directed cycle of length 1, while if $i + k = j + l$ then we have $\mathbb{B} + \pi_{ij}(\mathbb{R}) - \pi_{ij}(\mathbb{R}) = \mathbb{B}$, so once again $\mathbb{B}$ must be a union of linked components of $\pi_{ij}(\mathbb{R})$, which is impossible. The claim is proved.

Since the digraph $\mathbf{D}$ is finite, nonempty, and has no directed cycles, there must be a proper ternary absorbing subalgebra $\mathbb{B} \lhd \mathbb{A}$ such that $\mathbb{B} + \pi_{ij}(\mathbb{R}) = \mathbb{A}$ for all $i \neq j$. In particular, we see that $\pi_{ij}(\mathbb{R}) \cap \mathbb{B}^2 \neq \emptyset$ for all i, j. Since $\mathbb{B}$ is ternary absorbing, this implies that in fact $\mathbb{R} \cap \mathbb{B}^p \neq \emptyset$. Setting $\mathbb{R}' = \mathbb{R} \cap \mathbb{B}^p$, we can apply the induction hypothesis to $\mathbb{B}$ to see that $\mathbb{R}'$ contains a constant tuple. Thus $\mathbb{R}$ contains a constant tuple, and we are done. $\qquad\square$

Corollary 4.1.9. *If $\mathbb{A}$ is a finite Taylor algebra and m has no prime factors p which are less than or equal to $|\mathbb{A}|$, then $\mathbb{A}$ has an idempotent m-ary cyclic term.*

Example 4.1.1. Let $\mathbb{A}_n$ be the dual discriminator algebra from Example 1.6.5 on a domain of size n. Then every subset of $\mathbb{A}_n$ is a subalgebra with full automorphism group, so $\mathbb{A}_n$ does not have cyclic terms of any arity between 2 and n. By the previous results, we see that $\mathbb{A}_n$ has a cyclic term of arity m if and only if m has no prime factors which are less than or equal to n.

Corollary 4.1.10 (Siggers term from cyclic term)**.** *If $\mathbb{A}$ is a finite Taylor algebra, then $\mathbb{A}$ has an idempotent 4-ary Siggers term t, satisfying the identity $t(x,x,y,z) \approx t(y,z,z,x)$.*

Proof. Let c be an idempotent m-ary cyclic term for some $m > 1$. Then there are numbers $a, b \in \mathbb{N}$ such that $2a + 3b = m$, and we may define t by

$$t(x,y,z,w) := c(\underbrace{x,...,x}_{b}, \underbrace{y,...,y}_{a}, \underbrace{z,...,z}_{b}, \underbrace{w,...,w}_{a+b}). \qquad\square$$

Corollary 4.1.11 (Daisy chain terms). *If $\mathbb{A}$ is a finite Taylor algebra, then there are idempotent terms $w_i(x, y, z)$ for $i \in \mathbb{Z}$ such that for all i we have*

$$w_i(x, x, y) \approx w_i(y, x, x) \approx w_{i-1}(x, y, x),$$

and the sequence of terms w_i is periodic with some finite period.

Proof. Choose p to be an extremely huge prime, let c be an idempotent p-ary cyclic term, and let $a = \lfloor \frac{p}{3} \rfloor$. Define a long sequence of numbers $a_0, a_1, \dots$ by $a_0 = a$ and

$$a_{i+1} = p - 2a_i,$$

stopping as soon as we hit the first a_i with $a_i > \frac{p}{2}$. Define terms w_i' by

$$w_i'(x, y, z) := c(\underbrace{x, \dots, x}_{a_i}, \underbrace{y, \dots, y}_{a_{i+1}}, \underbrace{z, \dots, z}_{a_i}).$$

Since c is cyclic, these w_i's will satisfy the identities

$$w_i'(x, x, y) \approx w_i'(y, x, x) \approx w_{i-1}'(x, y, x).$$

If p is large enough, then there must be some $j < k$ such that $w_j' = w_k'$. Then we define w_i by picking some $i' \in [j, k]$ such that $i \equiv i' \pmod{k - j}$ and setting $w_i = w_{i'}'$. $\qquad \square$

Corollary 4.1.12. *A finite algebra $\mathbb{A}$ is Taylor if and only if it has a pair of idempotent ternary terms p, q satisfying the identities*

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

Proof. To see that such p, q must exist in a Taylor algebra, we can take p, q to be any pair of consecutive daisy chain terms from the previous corollary. To see that any such p, q define Taylor terms, note that if p is a projection then p must be second projection, but in this case q must be a Mal'cev term. $\qquad \square$

4.2 Minimal Taylor clones

Since our main aim in these notes is to understand the most general CSPs which can be solved in polynomial time, it makes sense to study (core) relational structures $\mathbf{A}$ such that $\mathrm{CSP}(\mathbf{A})$ is in P, but such that adding any additional relations to $\mathbf{A}$ makes the problem NP-complete. According to the CSP dichotomy theorem of Bulatov [44] and Zhuk [142], these maximal relational structures correspond under the $\mathrm{Inv} - \mathrm{Pol}$ Galois correspondence to *minimal* Taylor clones.

Definition 4.2.1. A clone $\mathcal{O}$ on a finite domain is called a *minimal Taylor clone* if $\mathcal{O}$ is Taylor and every proper subclone of $\mathcal{O}$ is not Taylor. A finite algebra $\mathbb{A}$ is called a *minimal Taylor algebra* if $\mathrm{Clo}(\mathbb{A})$ is a minimal Taylor clone.

At first it may not be clear that minimal Taylor clones even exist: perhaps every Taylor clone contains a proper Taylor subclone, with the relevant Taylor operations having higher and higher arity. We can rule this out by using the existence of a Siggers term (Corollary 3.11.18).

Proposition 4.2.2. *Every Taylor clone on a finite domain contains a minimal Taylor clone.*

Proof. By Corollary 3.11.18, every Taylor clone contains a 4-ary Siggers operation t satisfying the identity $t(x, x, y, z) \approx t(y, z, z, x)$. Since any such t is Taylor, and since there are only finitely many 4-ary operations on a given finite domain, at least one of the 4-ary Siggers operations $t \in \mathcal{O}$ generates a minimal Taylor clone. $\qquad\square$

Since every minimal Taylor clone is generated by a single 4-ary operation, we see that the number of minimal Taylor clones on a domain of size n is at most n^{n^4}. We can get a much better upper bound on the number of minimal Taylor clones by using the daisy chain terms from the previous section.

Proposition 4.2.3. *The number of minimal Taylor clones on a domain of size n is at most n^{2n^3}.*

Proof. By Corollary 4.1.12, every minimal Taylor clone $\mathcal{O}$ contains a pair of ternary idempotent operations p, q satisfying the identities

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

Since $\langle p, q \rangle$ generates a Taylor clone, we must have $\mathcal{O} = \langle p, q \rangle$. Since the number of ordered pairs of ternary operations p, q on a domain of size n is n^{2n^3}, we see that the number of minimal Taylor clones is at most n^{2n^3}. $\qquad\square$

Remark 4.2.1. Later we will see that the upper bound n^{2n^3} can be reduced to n^{n^3}, by showing that every minimal Taylor clone is generated by a *single* ternary operation. On a domain of size 2, it is easy to check that every minimal Taylor algebra is term equivalent to either a semilattice, a majority algebra, or to the idempotent reduct of $\mathbb{Z}/2$. On a domain of size 3, there turn out to be a total of 24 minimal Taylor algebras, up to term equivalence and isomorphism.

Unfortunately, the number of minimal Taylor algebras grows quite rapidly as the size of the domain increases: even if we only consider majority algebras, it turns out that the number of minimal majority algebras (up to term-equivalence) such that every three-element subset is a subalgebra is $7^{\binom{n}{3}}$, and identifying isomorphic algebras can only reduce this by a factor of at most $n!$, which makes little difference to the asymptotics.

The key fact that makes the theory of minimal Taylor algebras work is the following result, which essentially says that anything that "looks like" it "could be" a subalgebra or quotient of a minimal Taylor algebra actually *is* a subalgebra or quotient, and is also minimal Taylor as well.

Theorem 4.2.4. *If $\mathbb{A}$ is a minimal Taylor algebra and $\mathbb{B} \in HSP_{fin}(\mathbb{A})$, then $\mathbb{B}$ is also a minimal Taylor algebra.*

In fact, if $S \subseteq \mathbb{B}$ is a subset of $\mathbb{B}$ (not assumed to be a subalgebra), $t \in \mathrm{Clo}(\mathbb{A})$ is any term of $\mathbb{A}$, and θ is an equivalence relation on S such that

- *the set S is closed under t,*

- *every equivalence class of θ is a subalgebra of $\mathbb{B}$,*

- *the equivalence relation θ is a congruence of the algebraic structure (S, t), and*

- *the quotient $(S,t)/\theta$ is a Taylor algebra,*

then in fact the following must all be true:

- *the set S is actually the underlying set of a subalgebra $\mathbb{S}$ of $\mathbb{B}$,*

- *the equivalence relation θ is actually a congruence on the subalgebra $\mathbb{S}$, and*

- *the restriction of every term of $\mathbb{A}$ to the quotient $\mathbb{S}/\theta$ is in the clone generated by the restriction of t to $(S,t)/\theta$.*

Note that taking θ to be the trivial equivalence relation 0_S is always allowed, since every minimal Taylor algebra is automatically idempotent.

Proof. Let p be any prime such that $p > |\mathbb{A}|$ and $p > |(S,t)/\theta|$. By Theorem 4.1.8, there is a p-ary cyclic term $c \in \mathrm{Clo}(\mathbb{A})$, as well as a p-ary term $u \in \mathrm{Clo}(t)$ such that the restriction of u to $(S,t)/\theta$ is cyclic. Define a p-ary term c' by

$$c'(x_1,...,x_p) \coloneqq c(u(x_1,...,x_p), u(x_2,...,x_p,x_1), ..., u(x_p,x_1,...,x_{p-1})).$$

Then since c is cyclic, c' will automatically be cyclic as well. Since $\mathbb{A}$ is assumed to be minimal Taylor, we must have $\mathrm{Clo}(\mathbb{A}) = \langle c' \rangle$.

Suppose that $x_1,...,x_p \in S$. Then since $u \in \mathrm{Clo}(t)$ preserves S and acts cyclically on $(S,t)/\theta$, we must have

$$u(x_1,...,x_p) \equiv_\theta u(x_2,...,x_p,x_1) \equiv_\theta \cdots \equiv_\theta u(x_p,x_1,...,x_{p-1}) \in S,$$

and since equivalence classes of θ were assumed to be subalgebras of $\mathbb{B}$, we have

$$c'(x_1,...,x_p) \equiv_\theta u(x_1,...,x_p) \in S.$$

Thus c' preserves S as well as the equivalence relation θ, and the restriction of c' to $(S,t)/\theta$ is the same as the restriction of u to $(S,t)/\theta$. Since c' generates $\mathrm{Clo}(\mathbb{A})$, this finishes the proof. $\square$

An immediate consequence of Theorem 4.2.4 is that minimal Taylor algebras are *prepared* in the sense of Definition 3.2.19.

Proposition 4.2.5. *If $\mathbb{A}$ is a minimal Taylor algebra, then $a,b \in \mathbb{A}$ have*

$$\begin{bmatrix} b \\ b \end{bmatrix} \in \mathrm{Sg}_{\mathbb{A}^2}\left\{ \begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} b \\ a \end{bmatrix} \right\}$$

if and only if $\{a,b\}$ is a semilattice subalgebra of $\mathbb{A}$ with absorbing element b.

Proof. If $(b,b) \in \mathrm{Sg}\{(a,b),(b,a)\}$, then there must be some binary term t such that $t(a,b) = t(b,a) = b$. By idempotence, we automatically have $t(a,a) = a$ and $t(b,b) = b$, so the set $S = \{a,b\}$ is closed under t and (S,t) is a two-element semilattice. Thus we can apply Theorem 4.2.4 to see that $\{a,b\}$ must be a subalgebra of $\mathbb{A}$, and that the restriction of every term of $\mathbb{A}$ to $\{a,b\}$ is in the clone generated by the restriction of t to $\{a,b\}$. $\square$

Similarly, we can recognize two-element majority subalgebras and $\mathbb{Z}/2^{\mathrm{aff}}$ subalgebras. To simplify the statements of these results, it is convenient to assume the existence of an order two automorphism.

Proposition 4.2.6. *If $\mathbb{A}$ is a minimal Taylor algebra and $a, b \in \mathbb{A}$ are such that $\mathrm{Sg}_{\mathbb{A}^2} \left\{ \begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} b \\ a \end{bmatrix} \right\}$ is the graph of an automorphism of order two, then*

- *we have $\begin{bmatrix} a \\ a \\ a \end{bmatrix} \in \mathrm{Sg}_{\mathbb{A}^3} \left\{ \begin{bmatrix} a \\ a \\ b \end{bmatrix}, \begin{bmatrix} a \\ b \\ a \end{bmatrix}, \begin{bmatrix} b \\ a \\ a \end{bmatrix} \right\}$ iff $\{a, b\}$ is a majority subalgebra of $\mathbb{A}$, and*

- *we have $\begin{bmatrix} b \\ b \\ b \end{bmatrix} \in \mathrm{Sg}_{\mathbb{A}^3} \left\{ \begin{bmatrix} a \\ a \\ b \end{bmatrix}, \begin{bmatrix} a \\ b \\ a \end{bmatrix}, \begin{bmatrix} b \\ a \\ a \end{bmatrix} \right\}$ iff $\{a, b\}$ is a $\mathbb{Z}/2^{\mathrm{aff}}$ subalgebra of $\mathbb{A}$.*

Corollary 4.2.7. *If a minimal Taylor algebra $\mathbb{A}$ is generated by two elements a, b, then $\mathbb{A}$ is not subdirectly complete. As a consequence, either $\mathbb{A}$ has an affine quotient or $\mathbb{A}$ has a proper ternary absorbing subalgebra.*

Proof. Suppose for contradiction that $\mathbb{A}$ is subdirectly complete. Define a subdirect binary relation $\mathbb{S} \leq_{sd} \mathbb{A}^2$ by

$$\mathbb{S} = \mathrm{Sg}_{\mathbb{A}^2} \left\{ \begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} b \\ a \end{bmatrix} \right\}.$$

If (a, a) or (b, b) is in $\mathbb{S}$, then $\{a, b\}$ must be a two-element semilattice, which is not subdirectly complete. Otherwise, $\mathbb{S}$ must be the graph of an automorphism of order two by our assumption that $\mathbb{A}$ is subdirectly complete. Now define a subdirect ternary relation $\mathbb{R} \leq_{sd} \mathbb{A}^3$ by

$$\mathbb{R} = \mathrm{Sg}_{\mathbb{A}^3} \left\{ \begin{bmatrix} a \\ a \\ b \end{bmatrix}, \begin{bmatrix} a \\ b \\ a \end{bmatrix}, \begin{bmatrix} b \\ a \\ a \end{bmatrix} \right\}.$$

Since no $\pi_{ij}(\mathbb{R})$ can be the graph of an automorphism, we see that we must have $\mathbb{R} = \mathbb{A}^3$ by our assumption that $\mathbb{A}$ is subdirectly complete. Thus we have $(a, a, a) \in \mathbb{R}$, so $\{a, b\}$ must be a two-element majority algebra, which is not subdirectly complete. This contradiction proves that $\mathbb{A}$ must not be subdirectly complete.

For the last claim, we recall Zhuk's four cases (Corollary 3.12.12), and note that both binary absorption and central absorption imply ternary absorption. $\square$

The general recognition theorem for two-element majority subalgebras is as follows.

Proposition 4.2.8. *If $\mathbb{A}$ is a minimal Taylor algebra, then $a, b \in \mathbb{A}$ have*

$$\begin{bmatrix} a & b \\ a & b \\ a & b \end{bmatrix} \in \mathrm{Sg}_{\mathbb{A}^{3 \times 2}} \left\{ \begin{bmatrix} a & b \\ a & b \\ b & a \end{bmatrix}, \begin{bmatrix} a & b \\ b & a \\ a & b \end{bmatrix}, \begin{bmatrix} b & a \\ a & b \\ a & b \end{bmatrix} \right\}$$

if and only if $\{a, b\}$ is a majority subalgebra of $\mathbb{A}$.

It is also easy to recognize copies of the free semilattice on two generators.

229

Proposition 4.2.9. *If $\mathbb{A}$ is a minimal Taylor algebra and $a, b, c \in \mathbb{A}$ satisfy $a \to c$, $b \to c$ (i.e. $\{a, c\}$ and $\{b, c\}$ are semilattice subalgebras of $\mathbb{A}$ with absorbing element c), then we have*

$$\begin{bmatrix} c \\ c \end{bmatrix} \in \mathrm{Sg}_{\mathbb{A}^2} \left\{ \begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} b \\ a \end{bmatrix} \right\}$$

if and only if $\{a, b, c\}$ is isomorphic to the free semilattice on two generators.

We can also characterize binary absorbing subalgebras of minimal Taylor algebras, and show that they are always automatically strongly absorbing (and therefore are automatically centrally absorbing as well).

Proposition 4.2.10. *Suppose that $\mathbb{A}$ is a minimal Taylor algebra, and that $\mathbb{B} \lhd_{bin} \mathbb{A}$ is a binary absorbing subalgebra of $\mathbb{A}$. Then the following must hold.*

(a) *$\mathbb{B}$ is a strongly absorbing subalgebra of $\mathbb{A}$, that is, any term $f \in \mathrm{Clo}(\mathbb{A})$ which depends on its first input satisfies $f(\mathbb{B}, \mathbb{A}, ..., \mathbb{A}) \subseteq \mathbb{B}$.*

(b) *There is an equivalence relation $\theta_{\mathbb{B}} \in \mathrm{Con}(\mathbb{A})$ such that $\mathbb{B}$ is a congruence class of $\theta_{\mathbb{B}}$, and all other congruence classes of $\theta_{\mathbb{B}}$ are singletons.*

(c) *For every $a \notin \mathbb{B}$, $\mathbb{B} \cup \{a\}$ is a subalgebra of $\mathbb{A}$, and $(\mathbb{B} \cup \{a\})/\theta_{\mathbb{B}}$ is a two-element semilattice with absorbing element $\mathbb{B}/\theta_{\mathbb{B}}$.*

(d) *For every $a \notin \mathbb{B}$, there is some $b \in \mathbb{B}$ such that $\{a, b\}$ is a two-element semilattice with absorbing element b.*

(e) *For every $a, b \notin \mathbb{B}$ such that $\mathrm{Sg}_{\mathbb{A}}\{a, b\} \cap \mathbb{B} \neq \emptyset$, $\mathrm{Sg}_{\mathbb{A}}\{a, b\}/\theta_{\mathbb{B}}$ is isomorphic to the free semilattice on two generators.*

(f) *For every $a_1, ..., a_k \notin \mathbb{B}$ such that $\mathrm{Sg}_{\mathbb{A}}\{a_i, a_j\} \cap \mathbb{B} \neq \emptyset$ for all $i \neq j$, $\mathrm{Sg}_{\mathbb{A}}\{a_1, ..., a_k\}/\theta_{\mathbb{B}}$ is isomorphic to a semilattice of size $k + 1$.*

In particular, if $\mathbb{A}$ is generated by two elements and $\mathbb{B}$ is a proper binary absorbing subalgebra, then $\mathbb{A}/\theta_{\mathbb{B}}$ is either a two-element semilattice, or is isomorphic to the free semilattice on two generators.

For the sake of concretely writing down minimal Taylor algebras, we should pick convenient terms. My preference is to write them down in terms of the daisy chain terms from Corollary 4.1.11.

Definition 4.2.11. We say that a sequence of idempotent ternary terms w_i, defined for all $i \in \mathbb{Z}$, is a sequence of *daisy chain terms* if it satisfies the following properties:

- the sequence w_i is purely periodic in i with some finite period, and

- for all $i \in \mathbb{Z}$, we have $w_i(x, x, y) \approx w_i(y, x, x) \approx w_{i-1}(x, y, x)$.

It is useful to work out all possible sequences of daisy chain terms in our three basic examples of minimal Taylor algebras: semilattices, majority algebras, and affine algebras.

Proposition 4.2.12. *If $\mathbb{A} = (A, \vee)$ is a semilattice, then any sequence of daisy chain terms of $\mathbb{A}$ must have*

$$w_i(x, y, z) \approx x \vee y \vee z$$

for all $i \in \mathbb{Z}$.

Proof. It's enough to show that $w_i(x, x, y) \approx w_i(x, y, x) \approx w_i(y, x, x) \approx x \vee y$ for all i. Note that since $w_i(x, x, y) \approx w_i(y, x, x)$, we can't have $w_i(x, x, y) = y$, since semilattices have no Mal'cev terms. Additionally, if we had $w_i(x, x, y) = w_i(y, x, x) = x$, then w_i could not depend on its first or last coordinates, so we would have $w_{i+1}(x, x, y) \approx w_{i+1}(y, x, x) \approx w_i(x, y, x) = y$, which again contradicts the fact that semilattices have no Mal'cev terms.

Since the only binary terms of a semilattice are x, y, and $x \vee y$, we see by process of elimination that we must have $w_i(x, x, y) \approx w_i(y, x, x) \approx x \vee y$, and similar reasoning shows that $w_i(x, y, x) \approx w_{i+1}(x, x, y) \approx x \vee y$, so we are done. $\square$

Proposition 4.2.13. *If $\mathbb{A} = (A, m)$ is a majority algebra, then in any sequence of daisy chain terms of $\mathbb{A}$, each w_i must be a majority term, that is, we have*

$$w_i(x, x, y) \approx w_i(x, y, x) \approx w_i(y, x, x) \approx x$$

for all $i \in \mathbb{Z}$.

Proof. Note that every ternary term of a majority algebra is either a projection or a majority term (as is easily checked by induction on the construction of the term in terms of the majority operation m). If some w_i is a projection, then the identity $w_i(x, x, y) \approx w_i(y, x, x)$ implies that it must be second projection, but then the identity $w_{i+1}(x, x, y) \approx w_{i+1}(y, x, x) \approx w_i(x, y, x) = y$ implies that w_{i+1} is a Mal'cev term, which is impossible. Thus each w_i must be a majority term. $\square$

For the affine case, we have the following simplification in the setting of minimal Taylor algebras.

Proposition 4.2.14. *If $\mathbb{A}$ is minimal Taylor and affine, then there is an abelian group structure on the underlying set A such that $\mathbb{A}$ is term equivalent to $(A, x - y + z)$.*

Proof. Every affine algebra has the ternary function $x - y + z$ as a term, by Proposition 1.9.6. Since the ternary operation $x - y + z$ is Mal'cev, it generates a Taylor clone, so a minimal Taylor algebra is affine if and only if its clone is generated by $x - y + z$. $\square$

Because of this result, we don't need to think about the general case of a module over a (possibly noncommutative) ring if we are only interested in minimal Taylor algebras: we only need to think about algebras of the form $(A, x - y + z)$ for A an abelian group. By the classification of finite abelian groups, we can write such an algebra as a product of cyclic factors of prime power order. Recall that the *exponent* of a group is the least number n such that every cyclic subgroup has order dividing n.

Proposition 4.2.15. *If $\mathbb{A} = (A, x - y + z)$ is an affine algebra such that the abelian group A has exponent n, then for any sequence of daisy chain terms w_i of $\mathbb{A}$, there is a sequence of elements $a_i \in \mathbb{Z}/n$ such that*

$$w_i(x, y, z) \approx a_i x + (1 - 2a_i)y + a_i z$$

and

$$a_{i+1} \equiv 1 - 2a_i \pmod{n}$$

for all $i \in \mathbb{Z}$.

Proof. Every $m-ary$ term $t \in \mathrm{Clo}(x - y + z)$ can be written in the form

$$t(x_1, ..., x_m) \approx k_1 x_1 + \cdots + k_m x_m,$$

for some $k_i \in \mathbb{Z}$ satisfying

$$k_1 + \cdots + k_m = 1.$$

Of course, only the congruence classes of the values of the coefficients k_i modulo n matter, and the set of m-ary terms $t \in \mathrm{Clo}(\mathbb{A})$ is in bijection with the set of tuples of $k_i \in \mathbb{Z}/n$ such that $k_1 + \cdots + k_m \equiv 1 \pmod{n}$.

Thus we can write

$$w_i(x, y, z) \approx a_i x + b_i y + c_i z$$

for some $a_i, b_i, c_i \in \mathbb{Z}/n$ such that $a_i + b_i + c_i \equiv 1 \pmod{n}$. The identity $w_i(x, x, y) \approx w(y, x, x)$ then implies that $a_i \equiv c_i$, so $b_i \equiv 1 - 2a_i$, while the identity $w_{i+1}(y, x, x) \approx w_i(x, y, x)$ implies that $a_{i+1} \equiv b_i \equiv 1 - 2a_i$. $\qquad \square$

Proposition 4.2.16. *If* $\mathbb{A} = (A, x - y + z)$ *is an affine algebra such that* $|A|$ *is a power of* 2, *then any sequence of daisy chain terms of* $\mathbb{A}$ *must have*

$$w_i(x, y, z) \approx \frac{x + y + z}{3}$$

for all $i \in \mathbb{Z}$. *In particular, if the abelian group* A *has exponent* 2, *then each* w_i *is the Mal'cev operation* $x - y + z \approx x + y + z$.

Proof. Suppose the exponent of A is 2^k. Then if we let $a_i \in \mathbb{Z}/2^k$ be the sequence from Proposition 4.2.15, we see from $a_{i+1} \equiv 1 - 2a_i \pmod{2^k}$ that we have

$$a_{i+1} - 1/3 \equiv -2(a_i - 1/3) \pmod{2^k}$$

for all $i \in \mathbb{Z}$, so in fact we must have

$$a_i - 1/3 \equiv 0 \pmod{2^k}$$

for all $i \in \mathbb{Z}$. $\qquad \square$

Proposition 4.2.17. *If* $\mathbb{A} = (A, x - y + z)$ *is an affine algebra such that the abelian group* A *has exponent* 3^k, *then any sequence of daisy chain terms of* $\mathbb{A}$ *must have period equal to* 3^k, *and there must be some* $i \in \mathbb{Z}$ *such that*

$$w_{i-1}(x, y, z) \approx \frac{x + z}{2},$$
$$w_i(x, y, z) \approx y,$$
$$w_{i+1}(x, y, z) \approx x - y + z.$$

Proof. We just need to show that the map $a \mapsto 1 - 2a \pmod{3^k}$ defines a cyclic permutation of $\mathbb{Z}/3^k$ for all $k \geq 0$. To see this, it's enough to check that the cycle containing 0 has length exactly 3^k.

Lifting to the integers, if we define a sequence $a_i \in \mathbb{Z}$ by $a_0 = 0$ and $a_{i+1} = 1 - 2a_i$, then we can solve the recurrence to obtain

$$a_i = \frac{1 - (-2)^i}{3}.$$

Then we see that $a_i \equiv 0 \pmod{3^k}$ if and only if $3^{k+1} \mid 1 - (-2)^i$. By induction on k we may assume that 3^{k-1} divides i, and by the binomial theorem, we have

$$1 - (-2)^i = 1 - (1 - 3)^i = 3i - 9\binom{i}{2} + 27\binom{i}{3} - + \cdots \equiv 3i - 0 + 0 - + \cdots \pmod{3^{k+1}},$$

so $3^{k+1} \mid 1 - (-2)^i$ if and only if 3^k divides i. $\qquad\square$

By going back to the original construction of the daisy chain terms from a huge cyclic term, we can simplify the situation slightly for affine algebras of odd order.

Proposition 4.2.18. *If $\mathbb{A}$ is a minimal Taylor algebra, then it is possible to choose a sequence of daisy chain terms w_i of $\mathbb{A}$ such that for every affine $\mathbb{B} \in HSP_{fin}(\mathbb{A})$ of odd order, the restriction of w_1 to $\mathbb{B}$ is the Mal'cev operation $x - y + z$.*

Proof. Since there are only finitely many ternary terms $w_1 \in \mathrm{Clo}(\mathbb{A})$, it's enough to prove that for every finite k we can find a w_1 that is part of a sequence of daisy chain terms of $\mathbb{A}$, such that w_1 restricts to the Mal'cev operation $x - y + z$ on every affine $\mathbb{B} \in HSP(\mathbb{A})$ such that $|\mathbb{B}|$ is odd and $|\mathbb{B}| \le k$.

Note that for any large prime p, the restriction of a p-ary cyclic term c to $\mathbb{B}$ must be given by

$$c(x_1, ..., x_p) = \frac{x_1 + \cdots + x_p}{p}.$$

Thus, in the construction of the terms w_i' from Corollary 4.1.11 where we plugged in

$$w_i'(x, y, z) := c(\underbrace{x, ..., x}_{a_i}, \underbrace{y, ..., y}_{p - 2a_i}, \underbrace{z, ..., z}_{a_i}),$$

we will have

$$w_i'(x, y, z) = \frac{a_i x + (p - 2a_i)y + a_i z}{p}$$

on $\mathbb{B}$. So as long as we choose a_1 such that $a_1 \equiv p \pmod{k!}$ and $a_1 \approx \frac{p}{3}$ (which is possible as long as we take p much larger than $k!$), we will have $w_1'(x, y, z) = x - y + z$ on every affine algebra $\mathbb{B} \in HSP(\mathbb{A})$ of size at most k. For $|\mathbb{B}|$ odd, the restriction of the sequence of terms w_i' to $\mathbb{B}$ will be purely periodic, so the final sequence of daisy chain terms constructed will have $w_1 = w_1'$ on such $\mathbb{B}$. $\qquad\square$

Remark 4.2.2. A similar argument shows that we can instead choose daisy chain terms w_i such that $w_i(x, y, z) \approx \frac{x + y + z}{3}$ for all i on every affine algebra $\mathbb{B} \in HSP(\mathbb{A})$ such that $|\mathbb{B}|$ is not a multiple of 3. In fact, for any profinite integer $a \in \hat{\mathbb{Z}} = \varprojlim \mathbb{Z}/n$ such that $a \equiv \frac{1}{3} \pmod{2^k}$ for all k, we can choose daisy chain terms such that $w_0(x, y, z) \approx ax + (1 - 2a)y + az$ on every affine algebra $\mathbb{B} \in HSP(\mathbb{A})$.

We can also limit the collection of affine algebras which may show up in $HSP(\mathbb{A})$.

Proposition 4.2.19. *If $\mathbb{A}$ is a minimal Taylor algebra and $\mathbb{B} \in HSP(\mathbb{A})$ is affine, then the exponent n of $\mathbb{B}$ is finite, with $n \leq |\mathbb{A}|^{|\mathbb{A}|^2}$, and every prime p which divides n is bounded by $|\mathbb{A}|$.*

Proof. First we show that every n such that $\mathbb{Z}/n^{\mathrm{aff}} \in HSP(\mathbb{A})$ has $n \leq |\mathbb{A}|^{|\mathbb{A}|^2}$. To see this, note that $\mathbb{Z}/n^{\mathrm{aff}}$ is generated by two elements (to be more specific, it is generated by 0 and 1), so if $\mathbb{Z}/n^{\mathrm{aff}} \in HSP(\mathbb{A})$ then $\mathbb{Z}/n^{\mathrm{aff}}$ must be a quotient of the free algebra on two generators $\mathcal{F}_{\mathbb{A}}(x,y) \leq \mathbb{A}^{\mathbb{A}^2}$.

Next, note that if p is prime and $\mathbb{Z}/p^{\mathrm{aff}} \in HSP(\mathbb{A})$, then $\mathbb{A}$ can't have any cyclic term of arity p, since $\mathbb{Z}/p^{\mathrm{aff}}$ has an automorphism of order p with no fixed points. Thus by Theorem 4.1.8 there is no prime $p > |\mathbb{A}|$ such that $\mathbb{Z}/p^{\mathrm{aff}} \in HSP(\mathbb{A})$. $\qquad\square$

We will end this section by characterizing Zhuk's centrally absorbing subalgebras in the case of minimal Taylor algebras, and using them to naturally produce majority subquotients of minimal Taylor algebras.

Theorem 4.2.20. *If $\mathbb{A}$ is minimal Taylor, $\mathbb{C} \leq \mathbb{A}$, and $\mathbb{M} \in HSP(\mathbb{A})$ is the two element majority algebra on the domain $\{0, 1\}$, then the following are equivalent:*

(a) *$\mathbb{C}$ is a ternary absorbing subalgebra of $\mathbb{A}$,*

(b) *for every prime $p > |\mathbb{A}|$ there is a p-ary cyclic term c of $\mathbb{A}$ such that whenever $\#\{i \mid x_i \in \mathbb{C}\} > \frac{p}{2}$, we have*
$$c(x_1, ..., x_p) \in \mathbb{C},$$
and furthermore the restriction of c to $\mathbb{M}$ is the p-ary majority operation,

(c) *the binary relation $\mathbb{R} \subseteq \mathbb{A} \times \mathbb{M}$ given by*
$$\mathbb{R} = (\mathbb{A} \times \{0\}) \cup (\mathbb{C} \times \{0,1\})$$
is a subalgebra of $\mathbb{A} \times \mathbb{M}$,

(d) *$\mathbb{C}$ centrally absorbs $\mathbb{A}$,*

(e) *every daisy chain term $w_i(x, y, z)$ witnesses the fact that $\mathbb{C}$ ternary absorbs $\mathbb{A}$.*

Proof. For (a) implies (b): let t be a ternary term which witnesses $\mathbb{C} \lhd \mathbb{A}$. If m is a ternary term of $\mathbb{A}$ which acts as majority on $\mathbb{M}$, then the ternary term
$$t'(x, y, z) := m(t(x, y, z), t(y, z, x), t(z, x, y))$$
also witnesses the absorption $\mathbb{C} \lhd \mathbb{A}$, and the restriction of t' to $\mathbb{M}$ is the majority operation. Now let $p > |\mathbb{A}|$ be prime, and let $u \in \mathrm{Clo}(t')$ be any term such that the restriction of u to $\mathbb{M}$ is a p-ary majority operation. Any such u must have the property that whenever $\#\{i \mid x_i \in \mathbb{C}\} > \frac{p}{2}$, we have
$$u(x_1, ..., x_p) \in \mathbb{C}.$$

Now let c' be any p-ary cyclic term of $\mathbb{A}$, and define c by
$$c(x_1, ..., x_p) := c'(u(x_1, ..., x_p), u(x_2, ..., x_p, x_1), ..., u(x_p, x_1, ..., x_{p-1})).$$

For (b) implies (c), note that the cyclic term c must generate the clone of $\mathbb{A}$, so it's enough to check that the relation $\mathbb{R}$ is preserved by c, which is easy to prove directly.

That (c) implies (d) follows from Zhuk's Theorem 3.10.8, since the left center of $\mathbb{R}$ is $\mathbb{C}$ and the majority algebra $\mathbb{M}$ is binary absorption free.

That (c) implies (e) follows from a direct computation: we have

$$\begin{bmatrix} w_i(\mathbb{C}, \mathbb{A}, \mathbb{C}) \\ 1 \end{bmatrix} = w_i\left(\begin{bmatrix} \mathbb{C} \\ 1 \end{bmatrix}, \begin{bmatrix} \mathbb{A} \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbb{C} \\ 1 \end{bmatrix}\right) \subseteq \mathbb{R},$$

so $w_i(\mathbb{C}, \mathbb{A}, \mathbb{C}) \subseteq \mathbb{C}$, and similarly $w_i(\mathbb{A}, \mathbb{C}, \mathbb{C}), w_i(\mathbb{C}, \mathbb{C}, \mathbb{A}) \subseteq \mathbb{C}$.

That (d) implies (a) follows from Zhuk's Corollary 3.10.10, while (e) implies (a) is immediate.
$\square$

Corollary 4.2.21. *Suppose that $\mathbb{A}$ is minimal Taylor and that $\mathbb{C}, \mathbb{D} \lhd_Z \mathbb{A}$ are two ternary absorbing subalgebras of $\mathbb{A}$. Then $\mathbb{C} \cup \mathbb{D}$ is a subalgebra of $\mathbb{A}$.*

If $\mathbb{C} \cap \mathbb{D} = \emptyset$, then the equivalence relation θ on $\mathbb{C} \cup \mathbb{D}$ with parts $\mathbb{C}$ and $\mathbb{D}$ is a congruence on $\mathbb{C} \cup \mathbb{D}$, and $(\mathbb{C} \cup \mathbb{D})/\theta$ is isomorphic to the two element majority algebra.

Proof. Note that since $\mathbb{A}$ is minimal Taylor, the clone of $\mathbb{A}$ is generated by any pair of consecutive daisy chain terms, so we just need to check that $\mathbb{C} \cup \mathbb{D}$ is closed under each daisy chain term w_i. For any $a, b, c \in \mathbb{C} \cup \mathbb{D}$, we either have at least two of a, b, c in $\mathbb{C}$ or at least two of a, b, c in $\mathbb{D}$, so the fact that w_i witnesses both $\mathbb{C} \lhd \mathbb{A}$ and $\mathbb{D} \lhd \mathbb{A}$ implies that $w_i(a, b, c) \in \mathbb{C} \cup \mathbb{D}$.

If $\mathbb{C} \cap \mathbb{D} = \emptyset$, then the fact that w_i witnesses both $\mathbb{C} \lhd \mathbb{A}$ and $\mathbb{D} \lhd \mathbb{A}$ implies that w_i is compatible with θ, and that the restriction of w_i to the two-element algebra $(\mathbb{C} \cup \mathbb{D})/\theta$ is the majority operation.
$\square$

The following conjecture, if true, would wrap everything up quite neatly.

Conjecture 4.2.1. If $\mathbb{A}$ is a minimal Taylor algebra which is generated by two elements $a, b \in \mathbb{A}$, then at least one of the following is true:

- there is a congruence $\theta \in \text{Con}(\mathbb{A})$ such that $\mathbb{A}/\theta$ is an affine algebra of prime order,

- there is a congruence $\theta \in \text{Con}(\mathbb{A})$ such that $\mathbb{A}/\theta$ is a two element semilattice,

- there is a congruence $\theta \in \text{Con}(\mathbb{A})$ such that $\mathbb{A}/\theta$ is a two element majority algebra, or

- there are proper ternary absorbing subalgebras $\mathbb{C}, \mathbb{D} \lhd_Z \mathbb{A}$ such that $a \in \mathbb{C}, b \in \mathbb{D}, \mathbb{C} \cup \mathbb{D} = \mathbb{A}$, and $\mathbb{C} \cap \mathbb{D} \neq \emptyset$.

Embarassingly, we don't know the answer to the following basic question.

Problem 4.2.1. Is there any minimal Taylor algebra which is simple, is generated by two elements, has size at least 3, and is not affine?

4.3 Bulatov's colored graph

In Bulatov's approach to the CSP dichotomy conjecture [44], the theory of absorbing subalgebras isn't used. Instead, Bulatov introduces a colored graph in [34] and [43], and uses connectivity properties of this graph to analyze finite Taylor algebras.

Definition 4.3.1. Suppose $\mathbb{A}$ is a finite idempotent algebra, and a, b are any pair of distinct elements of $\mathbb{A}$.

- We say that (a, b) is a *semilattice edge* if there is a binary term t such that $t(a, b) = t(b, a) = b$.

- We say that $\{a, b\}$ is a *weak majority edge* if there is a congruence θ on $\text{Sg}\{a, b\}$ and a ternary term m such that $\{a/\theta, b/\theta\}$ is closed under m and $(\{a/\theta, b/\theta\}, m)$ is a two-element majority algebra.

- We say that $\{a, b\}$ is a *weak affine edge* if there is a congruence θ on $\text{Sg}\{a, b\}$ and a term p such that $(\text{Sg}\{a, b\}/\theta, p)$ is an affine algebra.

We drop the modifier "weak" on an edge if θ is a maximal congruence on $\text{Sg}\{a, b\}$, and for any $a', b' \in \text{Sg}\{a, b\}$ such that $a' \equiv_\theta a$ and $b' \equiv_\theta b$, we have $\text{Sg}\{a, b\} = \text{Sg}\{a', b'\}$. Note that semilattice edges are directed, while majority and affine edges are undirected.

Note that a semilattice edge might not be a subalgebra, and similarly the set $a/\theta \cup b/\theta$ might not be a subalgebra if $\{a, b\}$ is a majority edge. If $\mathbb{A}$ is a *minimal* Taylor algebra, however, then Theorem 4.2.4 shows that $a/\theta \cup b/\theta$ is a subalgebra if (a, b) is a weak majority edge, and similary for semilattice edges. In [44], Bulatov calls an algebra *sm-smooth* if this special case of Theorem 4.2.4 applies to it.

We could have also defined "weak semilattice edges" in a similar way to the way we defined weak majority edges, but this is unnecessary.

Proposition 4.3.2. *If there is a congruence θ on $\text{Sg}\{a, b\}$ and an idempotent binary term t such that $t(a, b) \equiv_\theta t(b, a) \equiv_\theta b$, then there is some $b' \in \text{Sg}\{a, b\}$ such that $b' \equiv_\theta b$ and a partial semilattice term $s \in \text{Clo}(t)$ such that $s(a, b') = s(b', a) = b'$.*

Bulatov defines his colored graph by coloring the semilattice edges red, coloring the majority edges yellow, and coloring the affine edges blue (I don't know why these particular colors were chosen).

Definition 4.3.3. We say that a finite idempotent algebra $\mathbb{A}$ has a *hereditarily connected colored graph* if for all $\mathbb{B} \leq \mathbb{A}$, the colored graph of $\mathbb{B}$ is connected (ignoring the directions on the semilattice edges).

For the purposes of checking if an algebra is hereditarily connected, weak edges are interchangeable with edges by the following result.

Proposition 4.3.4. *Let $\mathbb{A}$ be a finite idempotent algebra. If the colored graph of weak edges of $\mathbb{A}$ is hereditarily connected, then the colored graph of edges of $\mathbb{A}$ is also hereditarily connected.*

Proof. Suppose that $\{a, b\}$ is a weak edge of $\mathbb{A}$, with corresponding congruence θ. We will prove by induction on $|\operatorname{Sg}\{a, b\}|$ that a and b are connected in the colored graph of edges of $\operatorname{Sg}\{a, b\}$. We may enlarge θ to a maximal congruence on $\operatorname{Sg}\{a, b\}$ without loss of generality, since any congruence of $\operatorname{Sg}\{a, b\}$ which identifies a and b is the full congruence. If (a, b) is not an edge, then we may pick $a' \in a/\theta$ and $b' \in b/\theta$ such that $\operatorname{Sg}\{a', b'\}$ is strictly smaller than $\operatorname{Sg}\{a, b\}$, and by the inductive hypothesis we see that a', b' are connected in the colored graph of edges of $\operatorname{Sg}\{a, b\}$. Since $\operatorname{Sg}\{a, a'\} \subseteq a/\theta$ and $\operatorname{Sg}\{b, b'\} \subseteq b/\theta$, we also see by the inductive hypothesis that a is connected to a' and b is connected to b' in the colored graph of edges of $\operatorname{Sg}\{a, b\}$. $\square$

Algebras with hereditarily connected colored graphs are closed under the usual algebraic operations.

Proposition 4.3.5. *If $\mathbb{A}, \mathbb{B}$ are finite idempotent algebras (of the same signature) with hereditarily connected colored graphs, then so is $\mathbb{A} \times \mathbb{B}$. More generally, if $\mathbb{A}$ is a finite idempotent algebra and $\theta \in \operatorname{Con}(\mathbb{A})$ is such that $\mathbb{A}/\theta$ is hereditarily connected and every congruence class of θ is also hereditarily connected, then $\mathbb{A}$ is hereditarily connected.*

Proof. We prove the more general statement. Let $a, b \in \mathbb{A}$ be any pair of elements. We will show that a and b are connected by weak edges in $\operatorname{Sg}\{a, b\}$. If $a/\theta = b/\theta$, then $\operatorname{Sg}\{a, b\}$ is contained in a congruence class of θ, so a, b are connected by edges of $\operatorname{Sg}\{a, b\}$. Otherwise, since $a/\theta, b/\theta$ are connected by edges in $\operatorname{Sg}\{a, b\}/\theta$, we can find a sequence of elements $a = a_0, a_1, ..., a_n = b$ such that $(a_i/\theta, a_{i+1}/\theta)$ is an edge of $\mathbb{A}/\theta$ for all i. Then each (a_i, a_{i+1}) will be a weak edge of $\operatorname{Sg}\{a, b\}$, with the corresponding congruence containing θ. $\square$

Proposition 4.3.6. *If $\mathbb{A}$ is a finite idempotent algebra with a hereditarily connected colored graph and $\theta \in \operatorname{Con}(\mathbb{A})$, then $\mathbb{A}/\theta$ also has a hereditarily connected colored graph.*

Proof. We just need to show that if (a, b) is an edge of $\mathbb{A}$ with $a/\theta \neq b/\theta$, then a/θ is connected to b/θ within the subalgebra they generate. We will induct on the size of $|\operatorname{Sg}\{a, b\}|$. If (a, b) is a semilattice edge, then $(a/\theta, b/\theta)$ will automatically be a semilattice edge as well. Otherwise, let η be the maximal congruence on $\operatorname{Sg}\{a, b\}$ corresponding to the edge (a, b).

Since every congruence class of η is a proper subalgebra of $\operatorname{Sg}\{a, b\}$, we see by induction that if $c \equiv_\eta d$, then c/θ and d/θ are connected in the subalgebra they generate, which is contained in $\operatorname{Sg}\{a/\theta, b/\theta\}$. Thus if the restriction of θ to $\operatorname{Sg}\{a, b\}$ is not contained in the maximal congruence η, then a/θ must be connected to b/θ in the subalgebra $\operatorname{Sg}\{a/\theta, b/\theta\}$. Otherwise, if the restriction of θ to $\operatorname{Sg}\{a, b\}$ *is* contained in η, then $(a/\theta, b/\theta)$ is an edge, with witnessing congruence η/θ. $\square$

Corollary 4.3.7. *If $\mathbb{A}$ is a finite idempotent algebra with a hereditarily connected colored graph, then $\mathbb{A}$ is Taylor.*

Bulatov's main result in [34] and [43] is that the converse to the above corollary holds. Since Bulatov didn't have the theory of absorbing subalgebras available to him, he proved this by using tame congruence theory. We will give a different proof, using a pair of consecutive daisy chain terms (whose existence followed from the existence of a cyclic term), and the fact that abelian Taylor algebras are affine.

Theorem 4.3.8 (Bulatov [34], [43])**.** *A finite idempotent algebra $\mathbb{A}$ is Taylor if and only if it has a hereditarily connected colored graph.*

We will prove Theorem 4.3.8 by induction on $|\mathbb{A}|$. A minimal counterexample $\mathbb{A}$ must be simple by Proposition 4.3.5, and every proper subalgebra of a minimal counterexample must have a hereditarily connected colored graph.

Definition 4.3.9. If $\mathbb{A}$ is any algebra, we define the *hypergraph of proper subalgebras* of $\mathbb{A}$ to be the hypergraph with vertex set equal to the underlying set of $\mathbb{A}$, and with a hyperedge $\mathbb{B}$ for every proper subalgebra $\mathbb{B} \leq \mathbb{A}$. We say that $\mathbb{A}$ is *disconnected* if the hypergraph of proper subalgebras of $\mathbb{A}$ is not connected.

We define the connected component equivalence relation $\sim_{\mathbb{A}}$ (or just $\sim$ if $\mathbb{A}$ is clear from context) on $\mathbb{A}$ by $a \sim_{\mathbb{A}} b$ if a is connected to b by a sequence of proper subalgebras of $\mathbb{A}$ (note that in general, $\sim_{\mathbb{A}}$ will *not* be a congruence).

Proposition 4.3.10. *If $\mathbb{A}$ is a disconnected algebra, then for any $a \not\sim_{\mathbb{A}} b$ we have $\mathrm{Sg}\{a, b\} = \mathbb{A}$.*

Proposition 4.3.11. *Suppose that $\mathbb{A}$ is finite, idempotent, simple, and disconnected. For any binary relation $\mathbb{R} \leq \mathbb{A} \times \mathbb{A}$ with $\pi_2(\mathbb{R}) = \mathbb{A}$, either $\mathbb{R}$ is the graph of an automorphism of $\mathbb{A}$ or there is some $a \in \mathbb{A}$ such that $\{a\} \times \mathbb{A} \subseteq \mathbb{R}$.*

Proof. If $\mathbb{R}$ is not the graph of an automorphism, then the linking congruence must be nontrivial, hence full (since $\mathbb{A}$ is simple). Thus there is some $a \in \mathbb{A}$ such that $(a, b), (a, c) \in \mathbb{A}$ for some pair of elements b, c with $b \not\sim c$, and from $\mathrm{Sg}_{\mathbb{A}}\{b, c\} = \mathbb{A}$ and idempotence, we see that $\{a\} \times \mathbb{A} \subseteq \mathbb{R}$. $\square$

Proposition 4.3.12. *Suppose that $\mathbb{A}$ is finite, idempotent, simple, and disconnected, and that $a \not\sim_{\mathbb{A}} b$ are such that neither (a, b) nor (b, a) are semilattice edges. Then the binary relation*

$$\mathbb{S}_{ab} := \mathrm{Sg}_{\mathbb{A}^2}\left\{\begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} b \\ a \end{bmatrix}\right\}$$

is the graph of an automorphism of order two which interchanges a and b.

Proof. Assume not. Then by Proposition 4.3.11, there is some $c \in \mathbb{A}$ with $\{c\} \times \mathbb{A} \subseteq \mathbb{S}_{ab}$. Since a, b are in different connected components of $\mathbb{A}$, at least one of them is in a different component than c, so we may suppose that a and c are in different connected components of $\mathbb{A}$ without loss of generality. Then since $(b, a), (b, c) \in \mathbb{S}_{ab}$ and $b \in \mathrm{Sg}_{\mathbb{A}}\{a, c\} = \mathbb{A}$, we have $(b, b) \in \mathbb{S}_{ab}$, so (a, b) is a semilattice edge. $\square$

Definition 4.3.13. Define the equivalence relation $\sim_{\mathbb{A}}^s$ by $a \sim_{\mathbb{A}}^s b$ if a can be connected to b by a chain of proper subalgebras and semilattice edges.

Corollary 4.3.14. *If $\mathbb{A}$ is finite, idempotent, and simple, and if $\sim_{\mathbb{A}}^s$ is not the full equivalence relation $\mathbb{A} \times \mathbb{A}$, then $\mathrm{Aut}(\mathbb{A})$ acts transitively on $\mathbb{A}$.*

Proof. For any pair $a, b \in \mathbb{A}$, either $a \not\sim_{\mathbb{A}}^s b$, or there is some $c \in \mathbb{A}$ such that $a \not\sim_{\mathbb{A}}^s c$ and $c \not\sim_{\mathbb{A}}^s b$. $\square$

Proposition 4.3.15. *Suppose that $\mathbb{A}$ is finite, idempotent, and simple. For any $a \not\sim_{\mathbb{A}}^s b$, if the ternary relation*

$$\mathbb{R}_{ab} := \mathrm{Sg}_{\mathbb{A}^3}\left\{\begin{bmatrix} b \\ a \\ a \end{bmatrix}, \begin{bmatrix} a \\ b \\ a \end{bmatrix}, \begin{bmatrix} a \\ a \\ b \end{bmatrix}\right\}$$

contains (a, a, a) or (b, a, b), then $\{a, b\}$ is a majority edge.

Proof. Suppose that there is some ternary term t witnessing the presence of one of those tuples in $\mathbb{R}_{ab}$. By Proposition 4.3.12, we see that $\{a, b\}$ is closed under t, and the restriction of t to $\{a, b\}$ is either a majority operation or a Pixley operation. Either way, the ternary term $t(x, t(x, y, z), z)$ acts like a majority operation on $\{a, b\}$. $\qquad\square$

Proof of Theorem 4.3.8. We only need to show that if $\mathbb{A}$ is a finite idempotent Taylor algebra, then $\mathbb{A}$ has a connected colored graph. Suppose that $\mathbb{A}$ is a counterexample of minimal size, and note that $\mathbb{A}$ must be simple by Proposition 4.3.5.

Our aim is to show that for any $a \not\sim_{\mathbb{A}}^s b$ such that $\{a, b\}$ is not a majority edge, the ternary relation

$$\mathbb{R}_{ab} := \operatorname{Sg}_{\mathbb{A}^3} \left\{ \begin{bmatrix} b \\ a \\ a \end{bmatrix}, \begin{bmatrix} a \\ b \\ a \end{bmatrix}, \begin{bmatrix} a \\ a \\ b \end{bmatrix} \right\}$$

must satisfy the conditions of Proposition 3.12.2, which will imply that $\mathbb{A}$ is affine. Note that $a \not\sim_{\mathbb{A}} b$ implies that $\mathbb{R}_{ab}$ is subdirect.

The first step to proving that $\mathbb{R}_{ab}$ satisfies the conditions of Proposition 3.12.2 is showing that $\pi_{12}(\mathbb{R}_{ab}) = \mathbb{A} \times \mathbb{A}$. Since $\pi_{12}(\mathbb{R}_{ab})$ contains $(a, a), (a, b)$, and (b, a), we need to check that $(b, b) \in \pi_{12}(\mathbb{R}_{ab})$. So it is natural to study the set of tuples in $\mathbb{R}_{ab}$ such that two of the coordinates are equal.

The next claim is the main place where we will use the fact that $\mathbb{A}$ is Taylor.

Claim 1: If we define $\mathbb{D}_{ab} \le \mathbb{A} \times \mathbb{A}$ to be the set of pairs (c, d) such that $(c, d, c) \in \mathbb{R}_{ab}$, then $\pi_1(\mathbb{D}_{ab}) \cap \pi_2(\mathbb{D}_{ab}) \ne \emptyset$.

Proof of Claim 1: Let p, q be consecutive daisy chain terms, i.e. p and q are ternary terms satisfying the identities

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

If we set $c = p(a, a, b), d = p(a, b, a) = q(a, a, b)$, and $e = q(a, b, a)$, then we have

$$p\left(\begin{bmatrix} b & a & a \\ a & b & a \\ a & a & b \end{bmatrix} \right) = \begin{bmatrix} c \\ d \\ c \end{bmatrix}$$

and

$$q\left(\begin{bmatrix} b & a & a \\ a & b & a \\ a & a & b \end{bmatrix} \right) = \begin{bmatrix} d \\ e \\ d \end{bmatrix},$$

so $(c, d), (d, e) \in \mathbb{D}_{ab}$, and $d \in \pi_1(\mathbb{D}_{ab}) \cap \pi_2(\mathbb{D}_{ab})$.

Claim 2: The binary relation $\mathbb{D}_{ab}$ from Claim 1 has $\pi_1(\mathbb{D}_{ab}) = \mathbb{A}$.

Proof of Claim 2: Suppose not. First consider the case where neither $\pi_1(\mathbb{D}_{ab}), \pi_2(\mathbb{D}_{ab})$ are equal to $\mathbb{A}$. Then if $c \in \pi_1(\mathbb{D}_{ab}) \cap \pi_2(\mathbb{D}_{ab})$, we see that both $\{a, c\}$ and $\{b, c\}$ are contained in proper subalgebras of $\mathbb{A}$, so $a \sim c \sim b$, contradicting the assumption $a \not\sim b$.

Suppose now that $\pi_1(\mathbb{D}_{ab}) \ne \mathbb{A}$ but $\pi_2(\mathbb{D}_{ab}) = \mathbb{A}$. Then by Proposition 4.3.11, there is some $c \in \mathbb{A}$ such that $\{c\} \times \mathbb{A} \subseteq \mathbb{D}_{ab}$. By Corollary 4.3.14, there is an automorphism $\sigma \in \operatorname{Aut}(\mathbb{A})$ with $\sigma(a) = c$, and from $(\sigma(a), \sigma(b)) \in \{c\} \times \mathbb{A} \subseteq \mathbb{D}_{ab}$, we see that $(\sigma(a), \sigma(b), \sigma(a)) \in \mathbb{R}_{ab}$, so in fact

$\sigma(\mathbb{R}_{ab}) \subseteq \mathbb{R}_{ab}$, and so $\sigma(\mathbb{D}_{ab}) = \mathbb{D}_{ab}$. Thus $(a, a) = \sigma^{-1}(c, c) \in \mathbb{D}_{ab}$, so by Proposition 4.3.15 this contradicts the assumption that $\{a, b\}$ is not a majority edge.

Claim 3: We have $\pi_{1,2}(\mathbb{R}_{ab}) = \mathbb{A} \times \mathbb{A}$.

Proof of Claim 3: By Claim 2, there is some c such that $(b, c) \in \mathbb{D}_{ab}$. Thus $(a, a), (a, b), (b, a), (b, b) \in \pi_{1,2}(\mathbb{R}_{ab})$, and these four elements generate $\mathbb{A}^2$.

Claim 4: For any c, the binary relation $\mathbb{R}_{ab}^c \leq \mathbb{A}^2$ defined as the set of pairs (d, e) such that $(c, d, e) \in \mathbb{R}_{ab}$ is the graph of an automorphism of order two.

Proof of Claim 4: Suppose not. Note that by Claim 3, the relation $\mathbb{R}_{ab}^c$ is subdirect. Thus if $\mathbb{R}_{ab}^c$ is not the graph of an automorphism, then by Proposition 4.3.11 there is some d such that $\{d\} \times \mathbb{A} \subseteq \mathbb{R}_{ab}^c$.

First suppose that $c = a$, so $(a, b) \in \mathbb{R}_{ab}^a$. Then either $a \not\sim d$ or $b \not\sim d$. If $a \not\sim d$, then from $(a, b), (d, b) \in \mathbb{R}_{ab}^a$ we see that $(b, b) \in \mathbb{R}_{ab}^a$, contradicting Proposition 4.3.15. Similarly if $b \not\sim d$, then from $(a, b), (a, d) \in \mathbb{R}_{ab}^a$ we see that $(a, a) \in \mathbb{R}_{ab}^a$, contradicting Proposition 4.3.15.

Now suppose that $c \neq a$. There is some $e \neq a$ such that $d \not\sim e$ (if not, then $\sim$ has just two equivalence classes, which are interchanged by the automorphism $\mathbb{S}_{ab}$, and which both have size 1, reducing us to the case $|\mathbb{A}| = 2$). Then $\mathbb{S}_{de}$ is the graph of an automorphism of order two which interchanges d and e, and from $(d, e), (e, d) \in \mathbb{R}_{ab}^c$ we see $\mathbb{S}_{de} \subseteq \mathbb{R}_{ab}^c$. Then since $e \neq a$ there is some $f \neq d$ such that $(a, f) \in \mathbb{S}_{de} \subseteq \mathbb{R}_{ab}^c$. Then from $(a, d), (a, f) \in \mathbb{R}_{ab}^c$ we see that $(c, d), (c, f) \in \mathbb{R}_{ab}^a$, contradicting the fact that $\mathbb{R}_{ab}^a$ is the graph of an automorphism of order two.

To finish the proof, note that Claim 4 shows that the relation $\mathbb{R}_{ab}$ satisfies the assumptions of Proposition 3.12.2, so $\mathbb{A}$ must be abelian. Thus we can apply the Theorem 3.12.8 to see that $\mathbb{A}$ must be affine. $\square$

4.4 Conservative Taylor algebras

Bulatov's colored graph was originally inspired by the study of conservative Taylor algebras. These algebras are easy to classify, and they are a great toy case for testing conjectures about general Taylor algebras.

Definition 4.4.1. A k-ary operation $f : A^k \to A$ is *conservative* if for all $a_1, ..., a_k \in A$ we have

$$f(a_1, ..., a_k) \in \{a_1, ..., a_k\}.$$

An algebraic structure $\mathbb{A}$ is called *conservative* if every basic operation of $\mathbb{A}$ is conservative.

Note that conservative algebras are automatically idempotent.

Proposition 4.4.2. *An algebraic structure $\mathbb{A}$ is conservative if and only if every subset $S \subseteq \mathbb{A}$ is actually a subalgebra of $\mathbb{A}$.*

On the relational side, we define conservative relational structures as follows.

Definition 4.4.3. A relational clone Γ on a domain A is called *conservative* if every unary relation $U \subseteq A$ is an element of Γ, i.e. $\mathcal{P}(A) \subseteq \Gamma$. A relational structure $\mathbf{A}$ is called *conservative* if every unary relation U can be primitively positively defined using the basic relations of $\mathbf{A}$.

Proposition 4.4.4. *If a relational structure $\mathbf{A}$ and an algebraic structure $\mathbb{A}$ are related by the $\mathrm{Inv} - \mathrm{Pol}$ Galois correspondence, then $\mathbf{A}$ is conservative if and only if $\mathbb{A}$ is conservative.*

If we are handed a relational structure, then the next result can be useful to decrease the amount of work needed to verify that it is conservative.

Proposition 4.4.5. *A relational structure* $\mathbf{A}$ *with finite underlying set* A *is conservative if and only if, for every* $a \in A$, *the unary relation* $A \setminus \{a\}$ *is primitively positively definable from the basic relations of* $\mathbf{A}$.

Example 4.4.1. A natural example of a conservative CSP template (on an infinite domain) is the *list-coloring* problem for graphs: the domain A is an infinite set, and the relations consist of the binary $\neq$ relation and the collection of all possible subsets $U \subseteq A$ as unary relations.

Example 4.4.2. A conservative 2-semilattice is called a *tournament*. The rock-paper-scissors algebra is probably the most famous example of a tournament which is not totally ordered.

Example 4.4.3. If an affine CSP is conservative, then the domain must have size two: the only conservative affine algebra is $\mathbb{Z}/2^{\mathrm{aff}}$, up to term equivalence.

Sometimes we will want to use the following refinement of the concept of conservative algebras.

Definition 4.4.6. We say that a relational clone Γ is *k-conservative* if every unary relation $U \subseteq A$ with size $|U| \leq k$ is an element of Γ, and we define k-conservative clones, algebras, and relational structures similarly.

Example 4.4.4. An algebra is idempotent iff it is 1-conservative.

Example 4.4.5. The *k-list-coloring* problem for graphs corresponds to the relational structure with infinite domain A, and relations consisting of the binary $\neq$ relation and the collection of all possible subsets $U \subseteq A$ with $|U| \leq k$. This problem is equivalent to 2SAT for $k = 2$, and is NP-hard for $k \geq 3$.

Example 4.4.6. The only 2-conservative affine algebras are $(\mathbb{Z}/2^{\mathrm{aff}})^k$, up to term equivalence and isomorphism.

For the sake of understanding CSPs, we would like to focus on minimal Taylor algebras. The next result shows that we can reduce the study of conservative Taylor algebras to conservative minimal Taylor algebras without losing anything essential.

Proposition 4.4.7. *Every reduct of a conservative algebra is also conservative. In particular, every conservative Taylor clone contains a minimal Taylor clone which is also conservative.*

Since every minimal Taylor clone can be generated by a pair of ternary terms (for instance, we can take a pair of consecutive daisy chain terms), we only have to focus on understanding conservative Taylor algebras of size 3.

Proposition 4.4.8. *A minimal Taylor algebra is conservative if and only if it is 3-conservative.*

In fact, looking carefully at how a daisy chain term must act on a conservative Taylor algebra, we have the following simplification.

Proposition 4.4.9. *If* $\mathbb{A}$ *is a 2-conservative minimal Taylor algebra and* w_i *is any daisy chain term for* $\mathbb{A}$, *then we have*
$$w_i(x, x, y) \approx w_i(x, y, x) \approx w_i(y, x, x),$$

so in fact every w_i is a ternary weak near-unanimity operation. The binary function

$$f(x, y) := w_i(x, x, y)$$

is independent of i, and completely determines the colored graph of $\mathbb{A}$. In addition, the binary function

$$s(x, y) := f(x, f(y, x))$$

is a partial semilattice term of $\mathbb{A}$.

Proof. Note that every pair of distinct elements $a, b \in \mathbb{A}$ must form an edge of the colored graph of $\mathbb{A}$ if $\mathbb{A}$ is a 2-conservative Taylor algebra. By our analysis of daisy chain terms on the basic two-element minimal Taylor algebras, we see that:

- if (a, b) is a semilattice edge, then $w_i(x, x, y) = w_i(x, y, x) = w_i(y, x, x) = x \vee y$ for $x, y \in \{a, b\}$,

- if $\{a, b\}$ is a majority edge, then $w_i(x, x, y) = w_i(x, y, x) = w_i(y, x, x) = x$ for $x, y \in \{a, b\}$,

- if $\{a, b\}$ is a $\mathbb{Z}/2^{\text{aff}}$ edge, then $w_i(x, x, y) = w_i(x, y, x) = w_i(y, x, x) = y$ for $x, y \in \{a, b\}$.

Thus we can tell what sort of edge $\{a, b\}$ is (as well as how it is directed, in case it is a semilattice edge) by examining the restriction of $f(x, y)$ to the set $\{a, b\}$. The claim about $s(x, y)$ follows easily by considering each of the three possible types of edge individually. $\square$

Thus, from now on we imagine that all conservative minimal Taylor algebras live in a variety $\mathcal{V}$ having just one ternary basic operation w, which satisfies the weak near-unanimity identity

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

Proposition 4.4.10. *The free algebra $\mathcal{F}_{\mathcal{V}}(x, y)$ on two generators in the variety $\mathcal{V}$ generated by minimal Taylor algebras has size 6: its elements are $x, y, f(x, y), f(y, x), s(x, y), s(y, x)$ (defined as in the previous proposition). The colored graph of the algebra $\mathcal{F}_{\mathcal{V}}(x, y)$ is as follows.*
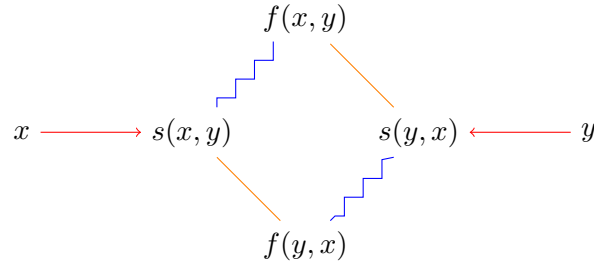
Here the semilattice edges are directed, the majority edges are straight and undirected, and the $\mathbb{Z}/2^{\text{aff}}$ edges are drawn as zigzags. This algebra has $\{f(x, y), f(y, x), s(x, y), s(y, x)\}$ as a binary absorbing subalgebra, corresponding to a semilattice quotient, and has $\{x, f(x, y), s(x, y)\}$ and $\{y, f(y, x), s(y, x)\}$ as ternary absorbing subalgebras, corresponding to a majority quotient.

In order to understand conservative minimal Taylor algebras, Proposition 4.4.8 implies that it's most important to understand the conservative algebras of size 3. Additionally, Proposition 4.4.9 implies that we just need to figure out which ternary weak near-unanimity operations on a three element set generate *minimal* Taylor clones. We get a further simplification by dividing into cases based on whether or not there is a ternary cyclic term. In the case where there is no cyclic term, the following result is useful.

Theorem 4.4.11. *If w is a ternary weak near-unanimity term of a finite algebra $\mathbb{A}$, then there is a ternary weak near-unanimity term $g \in \mathrm{Clo}(w)$ which also satisfies the identity*

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

If $|\mathbb{A}| = 3$ and $\mathbb{A}$ has no ternary cyclic term, then any such g satisfies $g(x,y,z) = x$ whenever x, y, z are all different.

Proof. Let $\gamma : \mathbb{A}^3 \to \mathbb{A}^3$ be the map given by

$$\gamma\left(\begin{bmatrix} x \\ y \\ z \end{bmatrix}\right) := \begin{bmatrix} w(x,y,z) \\ w(y,z,x) \\ w(z,x,y) \end{bmatrix}.$$

Then since $\mathbb{A}^3$ is finite, there is some k such that $\gamma^{\circ 2k} = \gamma^{\circ k}$. If we define g by

$$\gamma^{\circ k}\left(\begin{bmatrix} x \\ y \\ z \end{bmatrix}\right) = \begin{bmatrix} g(x,y,z) \\ g(y,z,x) \\ g(z,x,y) \end{bmatrix},$$

then $\gamma^{\circ k} \circ \gamma^{\circ k} = \gamma^{\circ k}$ implies that g satisfies the identity

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

Note that since w is a weak near-unanimity term, if any two of x, y, z are equal, then $\gamma(x,y,z)$ is a constant tuple, and then by idempotence so is $\gamma^{\circ k}(x,y,z)$. Therefore g is also a weak near-unanimity operation, and $\gamma^{\circ k}(x,y,z)$ can only avoid being a constant tuple if x, y, z are all different.

If $\mathbb{A}$ has no ternary cyclic term and has underlying set $\{a, b, c\}$, then $\gamma^{\circ k}(a,b,c)$ can't be a constant tuple by Proposition 4.1.5, and since $\gamma^{\circ 2k} = \gamma^{\circ k}$, $\gamma^{\circ k}(a,b,c)$ must not have any pair of coordinates equal, so $\gamma^{\circ k}(a,b,c)$ must be a permutation of (a,b,c). By Theorem 4.1.8, if $\mathbb{A}$ has no ternary cyclic term then it must have an automorphism of order three, so $\gamma(a,b,c)$ must be one of $(a,b,c), (b,c,a)$, or (c,a,b), and in each case we have $\gamma^{\circ k}(a,b,c) = (a,b,c)$. Similarly, we must also have $\gamma^{\circ k}(a,c,b) = (a,c,b)$, so we have $g(x,y,z) = x$ whenever x, y, z are all different. $\qquad\square$

Theorem 4.4.12. *If a minimal Taylor algebra has size 3 and has no ternary cyclic term, then (after renaming elements) it is term equivalent to one of the following four algebras:*

- *the affine algebra $\mathbb{Z}/3^{\mathrm{aff}}$,*

- *the rock-paper-scissors algebra from Section 3.1,*

- *the three element dual discriminator algebra from Example 1.6.5, or*

- *the three element simple nonabelian Mal'cev algebra from Example 1.7.2.*

All but the first are conservative, all but the second have a full automorphism group, and the first two have binary cyclic terms.

Proof. By Theorem 4.1.8, if a minimal Taylor algebra $\mathbb{A}$ with underlying set $\{a, b, c\}$ has size 3 and has no ternary cyclic term, then $\mathbb{A}$ must have an automorphism of order three with no fixed points, so the permutation $(a\ b\ c)$ is in $\mathrm{Aut}(\mathbb{A})$. By Theorem 4.3.8, either $\mathbb{A}$ is affine - in which case it must be term-equivalent to $\mathbb{Z}/3^{\mathrm{aff}}$ - or $\mathbb{A}$ has some proper subalgebra of size 2 (since $\mathbb{A}$ has an edge (a, b), and either $\mathrm{Sg}\{a, b\}$ has size 2, or $\mathrm{Sg}\{a, b\} = \mathbb{A}$ has a proper quotient, and one of the congruence classes is a subalgebra of size 2). Since $(a\ b\ c) \in \mathrm{Aut}(\mathbb{A})$, if any 2-element subset of $\mathbb{A}$ is a subalgebra, then *every* 2-element subset of $\mathbb{A}$ is a subalgebra, and all three 2-element subalgebras of $\mathbb{A}$ are isomorphic to $\{a, b\}$.

If $\{a, b\}$ is a semilattice, then any binary operation s that acts like the semilattice term on $\{a, b\}$ has $(\{a, b, c\}, s)$ isomorphic to the rock-paper-scissors algebra. If $\{a, b\}$ is a majority algebra and g is a ternary weak near-unanimity operation as in the previous theorem, then g is a majority operation which acts as first projection whenever all three of its inputs are distinct, so $(\{a, b, c\}, g)$ is isomorphic to the three-element dual discriminator algebra. If $\{a, b\}$ is an affine algebra and g is a ternary weak near-unanimity operation as in the previous theorem, then g is a Mal'cev operation which acts as a minority operation whenever two of its inputs are equal, and which acts as first projection whenever all three of its inputs are distinct, so $(\{a, b, c\}, g)$ is isomorphic to the three element simple nonabelian Mal'cev algebra from Example 1.7.2. $\qquad\square$

In most of the remaining cases, the colored graph already does not have any automorphisms of order 3. In these cases, it turns out to be relatively easy to pick out a specific ternary cyclic operation which is determined by the colored graph alone. In fact, we have the following slightly stronger statement.

Theorem 4.4.13. *Suppose that a minimal Taylor algebra $\mathbb{A}$ has the following properties:*

- *$\mathbb{A}$ is 2-conservative, that is, for all $a, b \in \mathbb{A}$ the subset $\{a, b\}$ is a subalgebra of $\mathbb{A}$,*

- *the colored graph of $\mathbb{A}$ does not contain any majority triangles, and*

- *the colored graph of $\mathbb{A}$ does not contain any affine triangles.*

Then $\mathbb{A}$ is conservative, and $\mathrm{Clo}(\mathbb{A})$ is determined by the colored graph of $\mathbb{A}$.

Proof. Let w be any daisy chain term for $\mathbb{A}$. Define a map $\gamma : \mathbb{A}^3 \to \mathbb{A}^3$ as in Theorem 4.4.11. We will make sure to only apply γ to triples where some pair of coordinates are equal, since the values γ takes on such triples is completely determined by the colored graph of $\mathbb{A}$ by Proposition 4.4.9. Define binary terms f, s as in Proposition 4.4.9, and note that f and s are uniquely determined by the colored graph of $\mathbb{A}$. Define maps $\alpha_f, \beta_f : \mathbb{A}^3 \to \mathbb{A}^3$ by

$$\alpha_f\left(\begin{bmatrix} x \\ y \\ z \end{bmatrix}\right) := \begin{bmatrix} f(x, y) \\ f(y, z) \\ f(z, x) \end{bmatrix}$$

and

$$\beta_f\left(\begin{bmatrix} x \\ y \\ z \end{bmatrix}\right) := \begin{bmatrix} f(x, z) \\ f(y, x) \\ f(z, y) \end{bmatrix},$$

and define maps $\alpha_s, \beta_s : \mathbb{A}^3 \to \mathbb{A}^3$ similarly, with f replaced by s. Note that α_f, β_f, etc. each have the property that if the input has two coordinates the same, then so does the output. As long as

244

a, b, c do not form a majority triangle, an affine triangle, or a rock-paper-scissors subalgebra, then the triple

$$\alpha_f \circ \beta_f \circ \alpha_s \circ \beta_s \left(\begin{bmatrix} x \\ y \\ z \end{bmatrix} \right)$$

has two of its three coordinates equal (to check this, consider the case where $\{a, b, c\}$ contains at least one semilattice edge separately from the case where it contains only majority and affine edges). Thus the ternary term

$$t := \pi_1 \circ \gamma \circ \alpha_f \circ \beta_f \circ \alpha_s \circ \beta_s : \mathbb{A}^3 \to \mathbb{A}$$

is cyclic on every such triple. Since we assumed that $\mathbb{A}$ has no majority triangles or affine triangles, the only possible triples of $\mathbb{A}$ such that the value of t is not uniquely determined by the colored graph of $\mathbb{A}$ are the rock-paper-scissors subsets of $\mathbb{A}$, which are necessarily subalgebras of $\mathbb{A}$ by Theorem 4.2.4. If we iterate t as in Theorem 4.4.11, then the resulting ternary function g has its values on rock-paper-scissors subalgebras fixed as well, so all of the values of g are determined purely by the colored graph of $\mathbb{A}$. Furthermore, this g is conservative and generates a Taylor clone, so $\mathrm{Clo}(\mathbb{A}) = \mathrm{Clo}(g)$ and $\mathbb{A}$ is conservative. $\square$

Finally, we need to understand the case of a majority triangle or affine triangle $\{a, b, c\}$ with a cyclic term. In these cases, it is helpful to keep track of the subalgebra

$$\pi_1 \left(\mathrm{Sg}_{\mathbb{A}^2} \left\{ \begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} b \\ c \end{bmatrix}, \begin{bmatrix} c \\ a \end{bmatrix} \right\} \cap \Delta_{\mathbb{A}} \right) \leq \mathbb{A},$$

since the set of possible outputs of a cyclic term applied to (a, b, c) must be contained in this subalgebra. This subalgebra is an invariant of $\mathrm{Clo}(\mathbb{A})$, and it shrinks when $\mathrm{Clo}(\mathbb{A})$ shrinks.

Definition 4.4.14. If $\mathbb{A}$ is a three element minimal Taylor algebra with underlying set $\{a, b, c\}$, then we will say that an element x of $\mathbb{A}$ is *circled* if $(x, x) \in \mathrm{Sg}\{(a, b), (b, c), (c, a)\}$. Note that the set of circled elements of $\mathbb{A}$ does not depend on the ordering of a, b, c.

Theorem 4.4.15. *Suppose that $\mathbb{A}$ is a conservative three element minimal Taylor algebra with a ternary cyclic term g, such that either all three of the edges of $\mathbb{A}$ are majority or all three are affine. Then (after renaming elements) $\mathbb{A}$ is term equivalent to one of the following three algebras:*

- *the three element solvable nonabelian Mal'cev algebra from Example 1.7.3, with $*$ as the unique circled element,*

- *the three element median algebra $\{0, 1, 2\}$, with the median element 1 as the unique circled element, or*

- *the three element minimal majority algebra $(\{a, b, c\}, m)$, with m a cyclic majority operation such that $m(a, b, c) = b$ and $m(a, c, b) = c$, with $\{b, c\}$ as the set of circled elements.*

In particular, every conservative three element minimal Taylor algebra is determined up to term equivalence by its colored graph and set of circled elements.

Furthermore, in any conservative minimal Taylor algebra, we can choose a ternary operation g as in Theorem 4.4.11 such that if we take g as the basic operation, then every three element majority subalgebra with two circled elements is isomorphic to the third algebra listed above (not just term-equivalent).

Proof. Let g be a ternary cyclic term for $\mathbb{A}$, and suppose that $\mathbb{A}$ has underlying set $\{a, b, c\}$. Once we know the types of the edges of $\mathbb{A}$, we only need to know the values of $g(a, b, c)$ and $g(a, c, b)$ to completely determine g. For each choice of edges, we have two cases: either $g(a, b, c) = g(a, c, b)$, or $g(a, b, c) \neq g(a, c, b)$. This gives us four cases total.

First consider the case where all three edges of $\mathbb{A}$ are affine (so g is Mal'cev), and $g(a, b, c) \neq g(a, c, b)$. Without loss of generality, we may assume that $g(a, b, c) = b$ and $g(a, c, b) = c$. We will show that this case does not occur, by constructing a term w which generates a strictly smaller Taylor clone. Note that the order two permutation which swaps b and c is an automorphism of $(\{a, b, c\}, g)$. Then if we define the ternary operation t by

$$t(x, y, z) := g(x, g(x, y, z), g(x, g(x, y, z), g(x, z, y))),$$

then t is also Mal'cev and satisfies

$$t(a, b, c) = a, t(b, a, c) = c, t(c, b, a) = c,$$

so if we define the ternary operation w by

$$w(x, y, z) := g(t(x, y, z), t(y, z, x), t(z, x, y)),$$

then w is a symmetric Mal'cev operation, with $w(a, b, c) = w(a, c, b) = a$. Then w generates a strictly smaller Taylor clone, since w preserves the equivalence relation with equivalence classes $\{a\}$ and $\{b, c\}$, while g does not. Thus this case does not occur.

In the remaining three cases, we get the three algebras described in the theorem statement. We need to check that these three algebras are really *minimal* Taylor. Note that in each case, there is a nontrivial congruence on $\mathbb{A}$ with quotient of size two and congruence classes of size at most two, so every Taylor reduct of $\mathbb{A}$ is forced to have a cyclic ternary term. We will show that the clone of each of these algebras contains only one or two ternary cyclic operations w. Note that the only values of $w(x, y, z)$ which are not determined by the types of the edges are the ones where x, y, z are all distinct.

In the case of the solvable Mal'cev algebra from Example 1.7.3 with underlying set $\{0, 1, *\}$, the congruence with congruence classes $\{*\}, \{0, 1\}$ forces the value of $w(0, 1, *)$ to be $*$, and similarly for other permutations of the inputs. Thus there is only one ternary cyclic operation w in the clone.

In the case of the three element median algebra $\{0, 1, 2\}$, the congruences corresponding to the partitions $\{0, 1\}, \{2\}$ and $\{0\}, \{1, 2\}$ force the value of $w(0, 1, 2)$ to be in $\{0, 1\} \cap \{1, 2\} = \{1\}$, and similarly for other permutations of the inputs. Thus there is only one ternary cyclic operation w in the clone.

In the last case, the congruence corresponding to the partition $\{a\}, \{b, c\}$ forces the value of $w(a, b, c)$ to be either b or c. Additionally, the order two automorphism which interchanges b and c forces us to have

$$w(a, b, c) = b \iff w(a, c, b) = c.$$

Thus we either have $w(x, y, z) \approx m(x, y, z)$, or $w(x, y, z) \approx m(x, z, y)$, so there are exactly two ternary cyclic operations w in the clone.

For the last statement, suppose that we have a minimal conservative algebra $\mathbb{A}$, with several majority subalgebras with two circled elements. Let g be any ternary operation as in Theorem 4.4.11. By the last case above, the restriction of g to any of these majority subalgebras either acts like $m(x, y, z)$ or like $m(x, z, y)$. Suppose for contradiction that two of these subalgebras are not

246

isomorphic, i.e., that g acts as $m(x, y, z)$ on one and acts as $m(x, z, y)$ on the other. We will produce a ternary weak near-unanimity term w which acts like $m(x, y, z)$ on both, which will generate a proper Taylor subclone. To this end, we define a ternary term t by

$$t(x, y, z) := g(x, g(x, y, z), g(x, z, y)),$$

and define w by

$$w(x, y, z) := g(t(x, y, z), t(y, z, x), t(z, x, y)).$$

Then w is cyclic on any subalgebra of $\mathbb{A}$ where g is cyclic, so in particular w is a weak near-unanimity operation. Note that if $\{a, b, c\}$ is a majority subalgebra of $\mathbb{A}$ with $\{b, c\}$ as the set of circled elements, then regardless of whether the restriction of g to $\{a, b, c\}$ is $m(x, y, z)$ or $m(x, z, y)$, we always have

$$t(a, b, c) = b, \quad t(b, c, a) = b, \quad t(c, a, b) = c,$$

so $w(a, b, c) = b$, and so the restriction of w to $\{a, b, c\}$ is $m(x, y, z)$. $\qquad\square$

Putting the proofs of the above theorems together, we get a procedure which puts the basic ternary weak near-unanimity operation g of any minimal conservative Taylor algebra into a standard form, such that the restriction of g to any three element subalgebra is completely determined by the edge types and the set of circled elements. In particular, we can exactly count the number of conservative minimal Taylor clones of a given size.

Corollary 4.4.16. *The number of conservative minimal Taylor clones on a set of size n is exactly*

$$\sum_{\textit{3-edge-colorings of } K_n} 2^{\#(\textit{semilattice})} 4^{\Delta(\textit{affine})} 7^{\Delta(\textit{majority})} = (1 + o(1)) \cdot 7^{\binom{n}{3}},$$

where $\Delta(c)$ is the number of monochromatic triangles of color c. In particular, for large n almost all conservative minimal Taylor clones are majority algebras.

If we only want to know the number of conservative minimal Taylor algebras of a given size up to term equivalence and *isomorphism*, then we can use the Burnside's counting theorem, together with the fact that the automorphism group of a conservative minimal Taylor algebra is determined by its colored graph and the choices of circled vertices on its three-element majority and Mal'cev subalgebras in the obvious way. The number of conservative minimal Taylor clones on domains of sizes $2, 3, 4$ is listed below.

Domain size	# up to term equiv.	# up to term equiv. and iso.
2	4	3
3	73	19
4	9829	520

4.4.1 Classification of three-element minimal Taylor algebras

As it turns out, up to term equivalence and isomorphism there are just 24 minimal Taylor algebras on a set of size 3. Of these, 19 are conservative, and the remaining 5 are easy to describe. One of the most obvious non-conservative minimal Taylor algebras of size 3 is the affine algebra $\mathbb{Z}/3^{\text{aff}}$. Three more are subdirect products of two-element minimal Taylor algebras: specifically, the free

semilattice on two generators (which is a subdirect product of two two-element semilattices), the subdirect product of a two-element semilattice and a two-element majority algebra, and the subdirect product of a two-element semilattice and $\mathbb{Z}/2^{\text{aff}}$ (all three of these are quotients of the free algebra from Proposition 4.4.10). There is no three-element subdirect product of a two-element majority algebra and $\mathbb{Z}/2^{\text{aff}}$, but the final example is nearly this: it is the algebra from Example 2.2.1, which has a 3-edge term, a $\mathbb{Z}/2^{\text{aff}}$ quotient, and a two-element centrally absorbing algebra. In this subsection we will prove that this is the complete list of minimal Taylor algebras on a three-element set.

By Theorem 4.4.12, we only have to classify the minimal Taylor algebras of size 3 which have a ternary cyclic term g, and since we have already classified the conservative ones, we just need to classify those which are generated by two elements. By Bulatov's Theorem 4.3.8, each of these algebras has a connected colored graph. Our first step will be to show that for minimal Taylor algebras of size 3, every edge of Bulatov's colored graph is a two-element subalgebra.

Proposition 4.4.17. *If $\mathbb{A}$ is a minimal Taylor algebra of size 3 other than $\mathbb{Z}/3^{\text{aff}}$, then the graph on $\mathbb{A}$ with edges given by the two-element subalgebras of $\mathbb{A}$ is connected - in other words, $\mathbb{A}$ has at least two subalgebras of size two.*

Proof. Suppose for contradiction that the graph of two-element subalgebras of $\mathbb{A}$ is disconnected, and suppose the underlying set of $\mathbb{A}$ is $\{a, b, c\}$. By Theorem 4.3.8, Bulatov's colored graph of $\mathbb{A}$ must be connected, so if $\mathbb{A}$ is not $\mathbb{Z}/3^{\text{aff}}$ then there must be some nontrivial congruence $\theta \in \text{Con}(\mathbb{A})$ such that $\mathbb{A}/\theta$ is either $\mathbb{Z}/2^{\text{aff}}$ or the two-element majority algebra (it can't be the two-element semilattice by Proposition 4.3.2). Suppose without loss of generality that $\{b, c\}$ is the congruence class of θ which has size 2, so that $\{b, c\}$ is a two-element subalgebra of $\mathbb{A}$. By our assumption that the graph of two-element subalgebras is disconnected, neither $\{a, b\}$ nor $\{a, c\}$ can be a subalgebra of $\mathbb{A}$. Let g be a ternary cyclic operation on $\mathbb{A}$, which exists by Theorem 4.4.12 (or directly from Theorem 4.1.8).

Suppose first that $\mathbb{A}/\theta$ is $\mathbb{Z}/2^{\text{aff}}$. Then since any ternary cyclic term of $\mathbb{Z}/2^{\text{aff}}$ acts as the minority operation, we must have

$$g(a, b, b) = g(a, c, c) = a.$$

Since $\{a, b\}$ and $\{a, c\}$ are not closed under g, we must then have

$$g(a, a, b) = c, \quad g(a, a, c) = b.$$

Define a ternary term t by

$$t(x, y, z) = g(g(x, x, y), g(y, y, z), g(z, z, x)).$$

Then t is also cyclic, and we have

$$t(a, a, b) = g(a, c, a) = b,$$

so $\{a, b\}$ is closed under t, and similarly $\{a, c\}$ is also closed under t. Thus t generates a strictly smaller Taylor clone, contradicting the assumption that $\mathbb{A}$ is minimal Taylor.

Now suppose that $\mathbb{A}/\theta$ is the two-element majority algebra. Then since any ternary cyclic term acts on a majority algebra as a majority operation, we must have

$$g(a, a, b) = g(a, a, c) = a,$$

248

and
$$g(a,b,c), g(a,c,b) \in \{b,c\}.$$

Since $\{a,b\}$ and $\{a,c\}$ are not closed under g, we must have

$$g(a,b,b) = c, \quad g(a,c,c) = b.$$

Assume without loss of generality that $g(a,b,c) = b$, otherwise swap b,c and reorder the last two arguments to g. Define a ternary term t by

$$t(x,y,z) = g(g(x,x,y), g(y,y,z), g(z,z,x)).$$

Then t is also cyclic, and we have

$$t(a,b,b) = g(a,b,c) = b,$$

so $\{a,b\}$ is closed under t. As before, this contradicts the assumption that $\mathbb{A}$ is minimal Taylor. $\quad\square$

From here on, we just need to classify the minimal Taylor algebras on the set $\{a,b,c\}$ such that a and b generate the algebra, while $\{a,c\}$ and $\{b,c\}$ are two-element subalgebras. First we handle the case where one of these subalgebras is a two-element semilattice.

Proposition 4.4.18. *If $\mathbb{A}$ is a minimal Taylor algebra with underlying set $\{a,b,c\}$ such that $\mathbb{A}$ is generated by a and b, then $\mathbb{A}$ does not have a semilattice subalgebra of the form $c \to a$.*

Proof. Suppose for contradiction that $\mathbb{A}$ is generated by a and b, but that $c \to a$. Let g be a ternary cyclic term of $\mathbb{A}$, which exists by Theorem 4.4.12.

First suppose that we do not also have $c \to b$. We will initially attempt to prove that $(a,a) \in \mathrm{Sg}_{\mathbb{A}^2}\{(a,b),(b,a)\}$, so that by Proposition 4.2.5 $\{a,b\}$ will be a two-element semilattice subalgebra with $b \to a$, which will contradict the assumption that a and b generate $\mathbb{A}$. Let $s \in \mathrm{Clo}(g)$ be a partial semilattice term of $\mathbb{A}$ with $s(a,c) = s(c,a) = a$. Since by assumption we do not have $a \to b$, we have $s(a,b) = a$ as well, and since we assumed that $\{b,c\}$ is a subalgebra and that we do not have $c \to b$, we have $s(c,b) = c$. Define $\mathbb{S}$ by

$$\mathbb{S} = \mathrm{Sg}_{\mathbb{A}^2} \left\{ \begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} b \\ a \end{bmatrix} \right\}.$$

If we have $(a,c) \in \mathbb{S}$, then by symmetry we also have $(c,a) \in \mathbb{S}$, so

$$s\left(\begin{bmatrix} a \\ c \end{bmatrix}, \begin{bmatrix} c \\ a \end{bmatrix} \right) = \begin{bmatrix} a \\ a \end{bmatrix}$$

is in $\mathbb{S}$ as well. If we have $(c,c) \in \mathbb{S}$, then

$$s\left(\begin{bmatrix} c \\ c \end{bmatrix}, \begin{bmatrix} a \\ b \end{bmatrix} \right) = \begin{bmatrix} a \\ c \end{bmatrix},$$

so $(a,c) \in \mathbb{S}$, and then we have $(a,a) \in \mathbb{S}$ as before. Since $c \in \mathrm{Sg}_{\mathbb{A}}\{a,b\} = \pi_2(\mathbb{S})$, if neither (a,c) nor (c,c) are in $\mathbb{S}$, then we must have $(b,c) \in \mathbb{S}$. So the only way to avoid a contradiction in this case is for $\mathbb{S}$ to be given by

$$\mathbb{S} = \left\{ \begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} b \\ a \end{bmatrix}, \begin{bmatrix} b \\ c \end{bmatrix}, \begin{bmatrix} c \\ b \end{bmatrix} \right\}.$$

249

Then the linking congruence θ of $\mathbb{S}$ corresponds to the partition $\{a, c\}, \{b\}$, and $\mathbb{A}/\theta$ is isomorphic to the two-element subalgebra $\{b, c\}$. Since $(b, b), (c, c) \notin \mathbb{S}$, $\mathbb{A}/\theta \cong \{b, c\}$ is either a majority algebra or is $\mathbb{Z}/2^{\mathrm{aff}}$. Define a ternary term t by

$$t(x, y, z) = g(s(g(x, y, z), x), s(g(x, y, z), y), s(g(x, y, z), z)).$$

Then t is also cyclic, and we will show that $\{a, b\}$ is closed under t, contradicting the assumption that $\mathbb{A}$ is minimal Taylor. If $\mathbb{A}/\theta$ is $\mathbb{Z}/2^{\mathrm{aff}}$, then we have

$$t(a, a, b) = g(a, a, b) = b,$$

and since $\{a, b\}$ is not closed under g we have

$$g(a, b, b) = c,$$

so

$$t(a, b, b) = g(s(c, a), s(c, b), s(c, b)) = g(a, c, c) = a.$$

If $\mathbb{A}/\theta$ is a majority algebra, then we have

$$t(a, b, b) = g(a, b, b) = b,$$

and since $\{a, b\}$ is not closed under g we have

$$g(a, a, b) = c,$$

so

$$t(a, a, b) = g(s(c, a), s(c, a), s(c, b)) = g(a, a, c) = a.$$

Either way, $\{a, b\}$ is closed under t, which gives us a contradiction.

Now suppose that we have both $c \to a$ and $c \to b$. Consider the binary relation $\mathbb{S} = \mathrm{Sg}_{\mathbb{A}^2}\{(a, b), (b, a)\}$ as we did before - if either (a, c) or (b, c) are in $\mathbb{S}$, then we easily see that one of $(a, a), (b, b)$ is in $\mathbb{S}$, which together with Proposition 4.2.5 would contradict the assumption that $\mathbb{A}$ is generated by a and b. Thus the only way to immediately avoid a contradiction is for $\mathbb{S}$ to be given by

$$\mathbb{S} = \left\{ \begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} b \\ a \end{bmatrix}, \begin{bmatrix} c \\ c \end{bmatrix} \right\}.$$

In particular, there must be some binary term $f \in \mathrm{Clo}(g)$ such that $f(a, b) = f(b, a) = c$, and since $\mathbb{A}$ is prepared by Proposition 4.2.5, this f must be the commutative binary operation described in the following table.

f	a	b	c
a	a	c	a
b	c	b	b
c	a	b	c

Then $(\{a, b, c\}, f)$ is isomorphic to the algebra $(\{-, 0, +\}, s_2)$ from Example 1.6.8, with the isomorphism given by $a \mapsto +, b \mapsto -, c \mapsto 0$. This algebra is not minimal Taylor: $\mathrm{Clo}(s_2)$ properly

contains the clone of the conservative bounded width algebra $(\{-, 0, +\}, g)$ described in Example 3.6.1. Explicitly, consider the ternary term t on $\mathbb{A}$ given by

$$t(x, y, z) = f(f(f(x, y), f(y, z)), f(x, z)).$$

It is easy to check that this t is symmetric, and that $\{a, b\}$ is closed under t, contradicting the assumption that $\mathbb{A}$ is minimal Taylor. In fact, even this operation t does not generate a minimal Taylor clone (this claim is left as an exercise). □

Proposition 4.4.19. *If $\mathbb{A}$ is a minimal Taylor algebra with underlying set $\{a, b, c\}$ such that $\mathbb{A}$ is generated by a and b, and if $\mathbb{A}$ has a semilattice subalgebra of the form $a \to c$, then $\mathbb{A}$ is isomorphic to a subdirect product of its two-element subalgebras $\{a, c\}$ and $\{b, c\}$.*

Proof. By assumption, we do not have $b \to a$ or $c \to a$, so $s(\{b, c\}, \mathbb{A}) \subseteq \{b, c\}$ for every partial semilattice term $s \in \mathrm{Clo}(\mathbb{A})$. Since $a \to c$ and $c \in \mathrm{Sg}_{\mathbb{A}}\{a, b\}$, we can apply Proposition 3.10.19 to see that $\{b, c\}$ is a binary absorbing subalgebra of $\mathbb{A}$. Then by Proposition 4.2.10, we see that there is a congruence θ on $\mathbb{A}$ corresponding to the partition $\{a\}, \{b, c\}$, such that $\mathbb{A}/\theta \cong \{a, c\}$ is a two-element semilattice. To finish, we just need to show that the equivalence relation ψ on $\mathbb{A}$ corresponding to the partition $\{a, c\}, \{b\}$ is also a congruence of $\mathbb{A}$.

If $b \to c$, then the same argument as in the last paragraph shows that $\{a, c\}$ is a binary absorbing subalgebra of $\mathbb{A}$ and that ψ is a congruence of $\mathbb{A}$ - in this case, $\mathbb{A}$ is the free semilattice on two generators. Additionally, the previous proposition shows that we can't have $c \to b$. Thus the only remaining cases are the case where $\{b, c\}$ is a majority algebra and the case where $\{b, c\}$ is $\mathbb{Z}/2^{\mathrm{aff}}$, and we assume from here on that we are in one of these two cases.

Consider the binary relation $\mathbb{S} = \mathrm{Sg}_{\mathbb{A}^2}\{(a, b), (b, a)\}$. Letting s be a partial semilattice operation of $\mathbb{A}$ with $s(a, c) = c$, we must have $s(a, b) \equiv s(a, c) = c \pmod{\theta}$, so since we do not have $a \to b$ we must have $s(a, b) = c$. Then by our assumption that we do not have $b \to c$, we have

$$s\left(\begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} b \\ a \end{bmatrix}\right) = \begin{bmatrix} c \\ b \end{bmatrix},$$

so

$$\mathbb{S} \supseteq \left\{ \begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} b \\ a \end{bmatrix}, \begin{bmatrix} b \\ c \end{bmatrix}, \begin{bmatrix} c \\ b \end{bmatrix} \right\}.$$

If this containment is an equality, then the linking congruence of $\mathbb{S}$ is ψ, which would prove that ψ is a congruence of $\mathbb{A}$ and finish the proof. Otherwise, since $(a, a), (b, b)$ are not contained in $\mathbb{S}$ by Proposition 4.2.5, at least one of $(a, c), (c, a)$ or (c, c) must be an element of $\mathbb{S}$. Since $a \to c$ and $\mathbb{S}$ is symmetric, in each case we see that $(c, c) \in \mathbb{S}$. From here on we assume that $(c, c) \in \mathbb{S}$.

From $(b, c), (c, c), (c, b) \in \mathbb{S}$ but $(b, b) \notin \mathbb{S}$, we see that $\{b, c\}$ can't be $\mathbb{Z}/2^{\mathrm{aff}}$ (since the parallelogram property fails for a binary relation on $\{b, c\}$). We are left with the case where $\{b, c\}$ is a majority algebra. Let g be a cyclic ternary term on $\mathbb{A}$, which exists by Theorem 4.4.12, and let f be a binary term with $f(a, b) = f(b, a) = c$, which must exist if $(c, c) \in \mathbb{S}$. Define a ternary term t by

$$t(x, y, z) = g(f(x, y), f(y, z), f(z, x)).$$

Then t is also cyclic, and we have

$$t(a, a, b) = g(a, c, c) = c, \quad t(a, b, b) = g(c, b, c) = c, \quad t(a, b, c) \in g(c, \{b, c\}, c) = \{c\}.$$

251

This completely determines t, and shows that $\{a, c\}$ is a ternary absorbing subalgebra of $\mathbb{A}$ with absorbing operation t. Therefore $\{a, c\}$ is a centrally absorbing subalgebra of $\mathbb{A}$ by Theorem 4.2.20. If we then define the cyclic ternary term u by

$$u(x, y, z) = t(s(x, y), s(y, z), s(z, x)),$$

then it is easy to check that $\{b\}$ is a ternary absorbing subalgebra of $\mathbb{A}$ with absorbing operation u, so by Theorem 4.2.20 and Corollary 4.2.21 the equivalence relation ψ is a congruence of $\mathbb{A}$ (note that this actually contradicts the assumption $(c, c) \in \mathbb{S}$). $\qquad \square$

So far we have classified every minimal Taylor algebra of size 3 which contains at least one semilattice subalgebra. The remaining cases are the cases where our algebra $\mathbb{A}$ has two two-element subalgebras, each of which is either a majority algebra or a copy of $\mathbb{Z}/2^{\mathrm{aff}}$.

Proposition 4.4.20. *If $\mathbb{A}$ is a minimal Taylor algebra with underlying set $\{a, b, c\}$ which is generated by a and b, then at least one of $\{a, c\}, \{b, c\}$ is not a $\mathbb{Z}/2^{\mathrm{aff}}$-subalgebra of $\mathbb{A}$.*

Proof. Suppose for contradiction that $\{a, c\}, \{b, c\}$ are both $\mathbb{Z}/2^{\mathrm{aff}}$-subalgebras of $\mathbb{A}$. By Corollary 4.2.7, $\mathbb{A}$ either has a proper absorbing subalgebra or an affine quotient. If $\mathbb{A}$ has a proper absorbing subalgebra, then one of $\{a, c\}, \{b, c\}$ has a proper absorbing subalgebra, which is impossible if they are both copies of $\mathbb{Z}/2^{\mathrm{aff}}$. Therefore there is a congruence θ of $\mathbb{A}$ such that $\mathbb{A}/\theta$ is affine, and we can assume without loss of generality that θ corresponds to the partition $\{a\}, \{b, c\}$ of $\mathbb{A}$.

Let g be a cyclic ternary term on $\mathbb{A}$, which exists by Theorem 4.4.12. Then since $\mathbb{A}/\theta \cong \{a, c\}$ is a copy of $\mathbb{Z}/2^{\mathrm{aff}}$ we must have

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

and since $\{a, b\}$ is not closed under g we must have

$$g(a, a, b) = c$$

as well. This, together with the fact that $\{b, c\}$ is also a copy of $\mathbb{Z}/2^{\mathrm{aff}}$, completely determines g. Define a ternary term t by

$$t(x, y, z) = g(x, g(x, y, y), g(x, y, z)),$$

and note that

$$t\left(\begin{bmatrix} a \\ a \\ b \end{bmatrix}, \begin{bmatrix} a \\ b \\ a \end{bmatrix}, \begin{bmatrix} b \\ a \\ a \end{bmatrix} \right) = g\left(\begin{bmatrix} a \\ a \\ b \end{bmatrix}, \begin{bmatrix} a \\ a \\ c \end{bmatrix}, \begin{bmatrix} c \\ c \\ c \end{bmatrix} \right) = \begin{bmatrix} c \\ c \\ b \end{bmatrix}.$$

Thus the ternary term u defined by

$$u(x, y, z) = g(t(x, y, z), t(y, z, x), t(z, x, y))$$

is a cyclic term with

$$u(a, a, b) = g(c, c, b) = b,$$

so $\{a, b\}$ is closed under u, which contradicts the assumption that $\mathbb{A}$ is a minimal Taylor algebra. $\qquad \square$

Proposition 4.4.21. *If $\mathbb{A}$ is a minimal Taylor algebra with underlying set $\{a, b, c\}$ which is generated by a and b, then at least one of $\{a, c\}, \{b, c\}$ is not a majority subalgebra of $\mathbb{A}$.*

Proof. Suppose for contradiction that $\{a, c\}, \{b, c\}$ are both majority subalgebras of $\mathbb{A}$. Then no subquotient of $\mathbb{A}$ can be affine, so by Theorem 3.13.8 CSP($\mathbb{A}$) has bounded width, and so by Theorem 3.14.8 $\mathbb{A}$ has a binary term f and a ternary term g satisfying the identities

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

We may assume without loss of generality that this g is also cyclic, by Theorem 4.4.12 and the cyclic composition trick. Since $\{a, b\}$ is not closed under g, we may assume without loss of generality that $f(b, a) = c$.

First we show that $f(a, b) \neq b$. If $f(a, b) = b$, then we have $f(a, f(a, b)) = f(a, b) = b$ and

$$f(b, f(b, a)) = f(f(a, b), f(b, a)) = f(a, b) = b,$$

so $(b, b) \in \mathrm{Sg}_{\mathbb{A}^2}\{(a, b), (b, a)\}$, and then by Proposition 4.2.5 we have $a \to b$, which contradicts the assumption that $\mathbb{A}$ is generated by a and b.

Now suppose that $f(a, b) = f(b, a) = c$. Let t be the ternary term defined by

$$t(x, y, z) = g(f(x, f(y, z)), f(y, f(z, x)), f(z, f(x, y))).$$

Then t is cyclic, and we have

$$t(a, a, b) = g(f(a, c), f(a, c), f(b, a)) = g(a, a, c) = a$$

and similarly $t(a, b, b) = b$, so $\{a, b\}$ is closed under t, contradicting the assumption that $\mathbb{A}$ is a minimal Taylor algebra.

Finally, suppose that $f(a, b) = a, f(b, a) = c$. Let h be the ternary term defined by

$$h(x, y, z) = g(f(x, y), f(y, x), g(x, y, z)),$$

let i be the ternary term defined by

$$i(x, y, z) = g(x, h(x, y, z), h(x, z, y)),$$

and let t be the cyclic ternary term defined by

$$t(x, y, z) = g(i(x, y, z), i(y, z, x), i(z, x, y)).$$

Then we have $h(a, a, b) = h(a, b, a) = h(b, a, a) = a$, so

$$t(a, a, b) = g(g(a, a, a), g(a, a, a), g(b, a, a)) = g(a, a, a) = a,$$

and $h(a, b, b) = h(b, a, b) = c, h(b, b, a) = b$, so

$$t(a, b, b) = g(g(a, c, c), g(b, b, c), g(b, c, b)) = g(c, b, b) = b.$$

Thus $\{a, b\}$ is closed under t, contradicting the assumption that $\mathbb{A}$ is a minimal Taylor algebra. $\square$

Proposition 4.4.22. *Suppose that $\mathbb{A}$ is a minimal Taylor algebra with underlying set $\{a, b, c\}$ which is generated by a and b, that $\{a, c\}$ is a $\mathbb{Z}/2^{\mathrm{aff}}$-subalgebra, and that $\{b, c\}$ is a majority subalgebra of $\mathbb{A}$. Then $\mathbb{A}$ is the same as the algebra described in Example 2.2.1, up to isomorphism and term equivalence. In particular, $\{a, c\}$ is a centrally absorbing subalgebra of $\mathbb{A}$, and $\mathbb{A}$ has a congruence θ corresponding to the partition $\{a\}, \{b, c\}$ such that $\mathbb{A}/\theta \cong \{a, c\}$ is $\mathbb{Z}/2^{\mathrm{aff}}$.*

Proof. Let $\mathbb{S} = \mathrm{Sg}_{\mathbb{A}^2}\{(a,b),(b,a)\}$, and let g be a ternary cyclic term of $\mathbb{A}$, which exists by Theorem 4.4.12. By Proposition 4.2.5, we have $(a,a),(b,b) \notin \mathbb{S}$. If $(c,c) \in \mathbb{S}$, then we have

$$\begin{bmatrix} a \\ c \end{bmatrix} = g\left(\begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} c \\ c \end{bmatrix}\right) \in \mathbb{S},$$

and then that

$$\begin{bmatrix} a \\ a \end{bmatrix} = g\left(\begin{bmatrix} a \\ c \end{bmatrix}, \begin{bmatrix} c \\ a \end{bmatrix}, \begin{bmatrix} c \\ c \end{bmatrix}\right) \in \mathbb{S},$$

which is a contradiction. Thus we have $(c,c) \notin \mathbb{S}$, so $\mathbb{S} \cap \Delta_{\mathbb{A}} = \emptyset$. If both $(a,c),(b,c) \in \mathbb{S}$, then we have

$$g\left(\begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} b \\ c \end{bmatrix}, \begin{bmatrix} c \\ a \end{bmatrix}\right) \in \mathbb{S} \cap \Delta_{\mathbb{A}},$$

which we just showed is impossible. Since $\pi_2(\mathbb{S}) = \mathrm{Sg}_{\mathbb{A}}\{a,b\} = \mathbb{A}$, exactly one of $(a,c),(b,c)$ is in $\mathbb{S}$.

Suppose first that $(b,c) \in \mathbb{S}$. Then the linking congruence θ of $\mathbb{S}$ corresponds to the partition $\{a,c\},\{b\}$ of $\mathbb{A}$, and $\mathbb{A}/\theta \cong \{b,c\}$ is a majority algebra. This implies that

$$g(a,b,b) = g(b,b,c) = b, g(b,c,c) = c, g(a,c,c) = a.$$

Since $\{a,b\}$ isn't closed under g, we must have

$$g(a,a,b) = c.$$

Let $f(x,y) = g(x,x,y)$, and define a ternary term t by

$$t(x,y,z) = g(f(g(x,y,z),f(x,y)), f(g(x,y,z),f(y,z)), f(g(x,y,z),f(z,x))).$$

Then t is cyclic,

$$t(a,a,b) = g(f(c,a),f(c,c),f(c,b)) = g(a,c,c) = a,$$

and

$$t(a,b,b) = g(f(b,c),f(b,b),f(b,b)) = g(b,b,b) = b.$$

Thus $\{a,b\}$ is closed under t, contradicting the assumption that $\mathbb{A}$ is a minimal Taylor algebra.

Now suppose that $(a,c) \in \mathbb{S}$. Then the linking congruence θ of $\mathbb{S}$ corresponds to the partition $\{a\},\{b,c\}$ of $\mathbb{A}$, and $\mathbb{A}/\theta \cong \{a,c\}$ is $\mathbb{Z}/2^{\mathrm{aff}}$. This implies that

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

Since $\{a,b\}$ is not closed under g, we must have

$$g(a,a,b) = c.$$

This completely determines g, and we see that $\{a,c\}$ is a ternary absorbing subalgebra of $\mathbb{A}$ with absorbing operation g, so by Theorem 4.2.20 $\{a,c\}$ is a centrally absorbing subalgebra of $\mathbb{A}$. After swapping a and c, g is exactly the same operation as the one described in Example 2.2.1. To see that $\mathbb{A}$ is really minimal Taylor in this case, note that any ternary cyclic $g' \in \mathrm{Clo}(g)$ must also satisfy all of the above identities, including $g'(a,a,b) = c$ since $\{a,c\}$ is centrally absorbing. $\qquad \square$

Putting all of the pieces together, we have completed the classification of minimal Taylor algebras on a three-element domain.

Theorem 4.4.23. *If $\mathbb{A}$ is a minimal Taylor algebra on a set of size 3, then up to term equivalence and isomorphism $\mathbb{A}$ is one of the following 24 algebras:*

- *one of the 19 conservative minimal Taylor algebras classified in the previous section,*

- *the affine algebra $\mathbb{Z}/3^{\mathrm{aff}} = (\mathbb{Z}/3, x - y + z)$,*

- *the free semilattice on two generators,*

- *the three-element subdirect product of $(\{0,1\}, x \vee y \vee z)$ with $(\{0,1\}, \mathrm{maj}(x, y, z))$,*

- *the three-element subdirect product of $(\{0,1\}, x \vee y \vee x)$ with $\mathbb{Z}/2^{\mathrm{aff}} = (\mathbb{Z}/2, x + y + z)$,*

- *the three-element algebra from Example 2.2.1, which has a 3-edge term, a two-element centrally absorbing subalgebra, and a $\mathbb{Z}/2^{\mathrm{aff}}$ quotient.*

Problem 4.4.1. For each one of the 24 minimal Taylor algebras on a set of size 3, find a generating set of relations for the corresponding relational clone. Are they all finitely related?

4.5 The strands of an unlinked CSP instance, and a new recursive strategy

Generally speaking, in order to guarantee a polynomial running time for solving CSPs we attempt to avoid recursion. There is a form of recursion which can be safely applied, however: we can recursively solve polynomially many subproblems as long as the size of *every* variable's domain is strictly reduced in each subproblem. The resulting algorithms will then have the property that the exponent in the running time will depend on the size of the largest domain of any variable. This approach seems to have been introduced with the solutions to the CSP dichotomy conjecture for conservative algebras by Bulatov [41], [42] and Barto [8], as well as Miklós Maróti's "Tree on top of Malcev" algorithm [107] (which used this sort of recursive strategy in a *very* different way from what we will consider in this section).

The challenge now is to find situations where we can usefully reduce to subproblems in which every single variable domain is reduced. The prototypical example of how this may occur is when a CSP instance is *unlinked*.

Definition 4.5.1. Let $\mathbf{X}$ be an instance of a CSP, with variable domains $\mathbb{A}_x$ for each variable x. We say that $\mathbf{X}$ is *unlinked* at a variable x of $\mathbf{X}$ if there are some $a, b \in \mathbb{A}_x$ such that there are *no* cycles p of $\mathbf{X}$ from x to x with $b \in \{a\} + p$. We say that $\mathbf{X}$ is *unlinked* if it is unlinked at every variable.

If an instance $\mathbf{X}$ is unlinked at a variable x, then we define the *linking relation* $\theta_{\mathbf{X}}$ of $\mathbf{X}$ at x as the equivalence relation on $\mathbb{A}_x$ defined by $(a, b) \in \theta_{\mathbf{X}}$ iff there is some cycle p from x to x such that $b \in \{a\} + p$.

Proposition 4.5.2. *If $\mathbf{X}$ is cycle-consistent and unlinked at x, then the linking relation $\theta_{\mathbf{X}}$ is a congruence of $\mathbb{A}_x$.*

Proof. We just need to check that $\theta_{\mathbf{X}}$ is closed under unary polynomial operations of $\mathbb{A}_x$. So suppose that $(a, b) \in \theta_{\mathbf{X}}$, $c_1, ..., c_n \in \mathbb{A}_x$, and f is some $n + 1$-ary polymorphism of $\mathbb{A}_x$. Since $(a, b) \in \theta_{\mathbf{X}}$, there is some cycle p from x to x such that $b \in \{a\} + p$. By cycle-consistency, we also have $c_i \in \{c_i\} + p$ for each i, so if $\mathbb{P}_p \leq \mathbb{A}_x^2$ is the binary relation corresponding to the cycle p, then we have

$$f\left(\begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} c_1 \\ c_1 \end{bmatrix}, \cdots, \begin{bmatrix} c_n \\ c_n \end{bmatrix}\right) \in \mathbb{P}_p.$$

Thus we have $(f(a, c_1, ..., c_n), f(b, c_1, ..., c_n)) \in \theta_{\mathbf{X}}$, which completes the proof. $\qquad\square$

As a consequence, one way to discover unlinked subinstances of a cycle-consistent instance $\mathbf{X}$ is to go through each maximal congruence θ on some variable domain $\mathbb{A}_x$, and to greedily build up a subinstance $\mathbf{X}'$ which is as large as possible subject to the condition $\theta_{\mathbf{X}'} \leq \theta$. The resulting subinstance $\mathbf{X}'$ only depends on θ and not on the choices we make during the greedy construction of $\mathbf{X}'$, so long as $\mathbf{X}$ is cycle-consistent: if there are two paths p, q from x to y such that $p - p$ and $q - q$ each have linking congruences contained in θ, then cycle-consistency applied to $-q + p$ shows that the linking congruence of $p - q$ is also contained in θ. This is the approach Zhuk used in his proof of the general CSP dichotomy conjecture [142].

Proposition 4.5.3. *Suppose that we have a multisorted CSP template $\mathrm{CSP}(\mathbb{A}_1, ..., \mathbb{A}_n)$ such that the CSP template $\mathrm{CSP}(\mathbb{B}_1, ..., \mathbb{B}_m)$ can be solved in polynomial time, where $\mathbb{B}_1, ..., \mathbb{B}_m$ is the collection of all algebras $\mathbb{B}$ which are isomorphic to a proper subalgebra of some $\mathbb{A}_i$ (note that $\max |\mathbb{B}_i| < \max |\mathbb{A}_j|$). Then we can solve any unlinked instance $\mathbf{X}$ of $\mathrm{CSP}(\mathbb{A}_1, ..., \mathbb{A}_n)$ in polynomial time.*

Proof. We assume that every pair of variables of $\mathbf{X}$ can be connected by some path without loss of generality, and we shrink $\mathbf{X}$ by enforcing cycle-consistency (if the shrunk instance is no longer unlinked, then every single variable domain has been reduced, and we can solve the instance). Pick any variable x of $\mathbf{X}$, and let $\theta_{\mathbf{X}}$ be the linking congruence on $\mathbb{A}_x$. Then for each congruence class $a/\theta_{\mathbf{X}} \in \mathbb{A}_x/\theta_{\mathbf{X}}$, if we restrict the possible values of x to a/θ and enforce arc-consistency, then every other variable y of $\mathbf{X}$ has its domain restricted to some subset of a congruence class of $\mathbb{A}_y/\theta_{\mathbf{X}}$ (where here we interpret $\theta_{\mathbf{X}}$ as the linking congruence of $\mathbf{X}$ on $\mathbb{A}_y$). Since the value of x must be in *some* congruence class of $\mathbb{A}_x/\theta_{\mathbf{X}}$, and since there are only a constant number of such congruence classes to check, we can solve $\mathbf{X}$ by solving a constant number of instances of CSPs with templates of the form $\mathrm{CSP}(\mathbb{B}_1, ..., \mathbb{B}_m)$, where for each variable y the algebra $\mathbb{B}_i$ is contained in an congruence class of $\mathbb{A}_y/\theta_{\mathbf{X}}$. $\qquad\square$

Of course, an entire instance being unlinked is fairly rare. Additionally, the assumption of cycle-consistency is probably unnecessarily strong. The approach used in the algorithms for the conservative CSP dichotomy is a variant of the strategy of looking for unlinked subinstances, starting from the idea of looking for *any* useful way of properly restricting the domains of some subset of the variables. This will naturally lead to different consistency principles (which are likely to also be unecessarily strong).

The most general thing we could do along these lines is the following. For each variable x, and for each element $a \in \mathbb{A}_x$, restrict the domain of x to the singleton $\{a\}$, and run arc-consistency (or cycle-consistency, etc.). Restrict our attention to the set Y of variables y whose domain has shrunk as a result of this restriction (replacing every relation which involves variables outside of Y to its projection onto the variables contained in Y), and solve the resulting instance recursively to see if we can rule out the element $a \in \mathbb{A}_x$ as a possible value for x.

Although this scheme is readily implemented, it is hard to algebraically control what happens as a result. Instead, we will consider the digraph of implications between restrictions on the individual domains, and ask under which conditions this has a nice structure.

Definition 4.5.4. If $\mathbf{X}$ is an instance of a multisorted CSP with variable domains $\mathbb{A}_x$, then we define the *implication digraph* to be the directed graph on $\mathbf{X}$ where the vertices are pairs $(x, \mathbb{B})$ such that x is a variable and $\mathbb{B} \leq \mathbb{A}_x$, and where we have a directed edge from $(x, \mathbb{B})$ to $(y, \mathbb{C})$ if there is a path p of length 1 from x to y in $\mathbf{X}$ such that $\mathbb{B} + p = \mathbb{C}$.

We write $(x, \mathbb{B}) \preceq (y, \mathbb{C})$ if there is a path p of any length from x to y such that $\mathbb{B} + p = \mathbb{C}$. The resulting quasiorder is the *implication qoset*.

Let $\mathcal{E}$ be a subqoset of the implication qoset. A *strand* of the subqoset $\mathcal{E}$ is just an equivalence class of $\mathcal{E}$. Often we might take $\mathcal{E}$ to be the subqoset of pairs $(x, \mathbb{B})$ such that $\mathbb{B}$ is a proper subalgebra of $\mathbb{A}_x$ (or a proper absorbing subalgebra, etc.). We say that $\mathcal{S}$ is a *maximal* strand of $\mathcal{E}$ if it is a maximal equivalence class of the qoset $\mathcal{E}$. If $\mathcal{E}$ is not specified, then a strand can be any subset of any equivalence class of the implication qoset.

A strand is called *absorbing* if for all $(x, \mathbb{B}) \in \mathcal{S}$ we have $\mathbb{B} \triangleleft \mathbb{A}_x$. Note that as long as $\mathbf{X}$ is arc-consistent, if this occurs for *some* $(x, \mathbb{B}) \in \mathcal{S}$ then it occurs for *all* $(x, \mathbb{B}) \in \mathcal{S}$.

We define the *partial restriction* of $\mathbf{X}$ to the strand $\mathcal{S}$ by reducing the domain of each variable x to $\bigcap_{(x, \mathbb{B}) \in \mathcal{S}} \mathbb{B}$ if some $(x, \mathbb{B}) \in \mathcal{S}$, and leaving the domain of x unchanged otherwise.

In order to restrict an instance $\mathbf{X}$ to a strand $\mathcal{S}$ in a way that guarantees that the restricted instance is arc-consistent, it seems like we should require that the strand $\mathcal{S}$ includes at most one pair $(x, \mathbb{B})$ for each variable $x \in \mathbf{X}$. This will be guaranteed as long as the instance $\mathbf{X}$ is sufficiently consistent.

Proposition 4.5.5. *If $\mathbf{X}$ is an arc-consistent instance, then each strand $\mathcal{S}$ of $\mathbf{X}$ includes at most one pair $(x, \mathbb{B})$ for each variable x as long as $\mathbf{X}$ satisfies property (P3) from Definiion 3.13.12:*

$$\mathbb{B} + p + q = \mathbb{B} \implies \mathbb{B} + p = \mathbb{B}$$

for all pairs of cycles p, q from x to x and all $\mathbb{B} \leq \mathbb{A}_x$. In particular, this always occurs if $\mathbf{X}$ is pq-consistent.

Definition 4.5.6. Suppose that $\mathbf{X}$ is an instance, and let $\mathcal{S}$ be any strand of $\mathbf{X}$ such that for each variable x, there is at most one $\mathbb{B}_x$ such that $(x, \mathbb{B}_x) \in \mathcal{S}$. We define the *full restriction* of $\mathbf{X}$ to $\mathcal{S}$ to be the instance $\mathbf{X}'$ whose variable set is the set of variables x such that some $(x, \mathbb{B}_x) \in \mathcal{S}$, with each variable domain reduced to $\mathbb{A}_x$, where for each relation of $\mathbf{X}$ we project it onto the set of variables in $\mathbf{X}'$ and restrict each of its coordinates to $\mathbb{B}_x$.

In order to mimic the situation of an unlinked instance as well as possible, we might also like to have the property that for each variable x such that some $(x, \mathbb{B})$ is in the strand $\mathcal{S}$, we have parallel strands $\mathcal{S}'$ which partition the domain $\mathbb{A}_x$ into disjoint subalgebras $\mathbb{B}'$ with each $(x, \mathbb{B}') \in \mathcal{S}'$.

Proposition 4.5.7. *If $\mathbf{X}$ is an arc-consistent instance, then for each strand $\mathcal{S}$, each $(x, \mathbb{B})$, and each $a \notin \mathbb{B}$, there is a strand $\mathcal{S}'$, involving the same set of variables as $\mathcal{S}$, such that $(x, \mathbb{B}') \in \mathcal{S}'$ for some $\mathbb{B}' \leq \mathbb{A}_x$ with $a \in \mathbb{B}'$ and $\mathbb{B} \cap \mathbb{B}' = \emptyset$, as long as $\mathbf{X}$ also satisfies property (P2) from Definition 3.13.12:*

$$\mathbb{B} + p = \mathbb{B} \implies \mathbb{B} - p = \mathbb{B}$$

for all cycles p from x to x and all $\mathbb{B} \leq \mathbb{A}_x$.

In this case, a more precise statement is true: for each $(x, \mathbb{B}) \in \mathcal{S}$ there is a congruence $\theta_{\mathcal{S}}$ on $\mathbb{A}_x$ such that $\mathbb{B}$ is a union of congruence classes of $\theta_{\mathcal{S}}$, and each congruence class $\mathbb{B}'$ of $\theta_{\mathcal{S}}$ is contained in a parallel strand $\mathcal{S}'$.

Proof. Consider the set $\mathcal{C}$ of all cycles p from x to x such that $\mathbb{B} + p = \mathbb{B}$. Property (P2) is the guarantee that for each $p \in \mathcal{C}$, $\mathbb{B}$ will be a union of congruence classes of the linking congruence of $\mathbb{P}_p$. Define $\theta_{\mathcal{S}}$ to be the join of the collection of all linking congruences of $\mathbb{P}_p$ for $p \in \mathcal{C}$. Then $\mathbb{B}$ is still a union of congruence classes of $\theta_{\mathcal{S}}$.

For any congruence class $\mathbb{B}'$ of $\theta_{\mathcal{S}}$, we define the parallel strand $\mathcal{S}'$ as follows: for every way of splitting some $p \in \mathcal{C}$ as $p = q + r$, where q is a path from x to a variable y, we include $(y, \mathbb{B}' + q)$ in $\mathcal{S}'$. Since $\mathbb{B}'$ is a congruence class of $\theta_{\mathcal{S}}$, $\mathbb{B}'$ will be a union of congruence classes of the linking congruence of $\mathbb{P}_p$, so $\mathbb{B}' + p - p = \mathbb{B}'$, so $(\mathbb{B}' + q) + r - p = \mathbb{B}'$, and we see that $(x, \mathbb{B}')$ and $(y, \mathbb{B}' + q)$ are indeed part of the same equivalence class of the implication qoset of $\mathbf{X}$. $\square$

To guarantee that the restriction of our instance $\mathbf{X}$ to any strand is arc-consistent, we need a still stronger consistency principle. This consistency principle was used to give one of the original proofs of the fact that $(2, 3)$-consistency was strong enough to ensure satisfiability of any CSP with bounded width [12]. I will define it a bit differently here, but will show that the definition given here is equivalent to the original definition by following an argument from [8].

Definition 4.5.8. An arc-consistent instance $\mathbf{X}$ is called a *Prague instance* if for every $(x, \mathbb{B}), (y, \mathbb{C})$ in the same equivalence class of the implication qoset, and for every path p of length 1 from x to y, we have $\mathbb{B} + p = \mathbb{C}$.

Proposition 4.5.9. *If $\mathbf{X}$ is a Prague instance, then $\mathbf{X}$ satisfies conditions (P2) and (P3) of Definition 3.13.12, so $\mathbf{X}$ is a weak Prague instance.*

Proposition 4.5.10 (Barto and Kozik [10], [8])**.** *If $\mathbf{X}$ is arc-consistent, then the following are equivalent:*

(a) *for every $(x, \mathbb{B}), (y, \mathbb{C})$ in the same equivalence class of the implication qoset, and for every path p of length 1 from x to y, we have $\mathbb{B} + p \subseteq \mathbb{C}$.*

(b) $\mathbf{X}$ *is a Prague instance,*

(c) *for every variable x and every pair of paths p, q from x to x such that the variables involved in p are a subset of the variables involved in q, if $b \in \{a\} + p$ then there is some $j \geq 0$ such that $b \in \{a\} + jq$,*

(d) *for every variable x and every pair of paths p, q from x to x such that the variables involved in p are a subset of the variables involved in q, then for every sufficiently large j we have $\{a\} + p \subseteq \{a\} + jq$.*

Proof. For the implication from (a) to (b), note that from $\mathbb{B} + p \subseteq \mathbb{C}$ and $\mathbb{C} - p \subseteq \mathbb{B}$ we can immediately conclude that $\mathbb{B} + p = \mathbb{C}$.

For (b) $\implies$ (c), let p, q be as in (c). Find a $j \geq 1$ such that $\{a\} + jq = \{a\} + 2jq$, write $\mathbb{B} = \{a\} + jq$, and let $\mathcal{S}$ be the strand of the implication digraph containing $(x, \mathbb{B})$. Since every Prague instance satisfies property (P3), for each variable y which shows up in the cycle q there is

a unique $\mathbb{C}$ such that $(y, \mathbb{C}) \in \mathcal{S}$. Since every step of the cycle p goes between variables involved in $\mathcal{S}$ and $\mathbf{X}$ is a Prague instance, we have $\mathbb{B} + p = \mathbb{B}$, and for the same reason we have $\mathbb{B} - jq = \mathbb{B}$. From

$$a \in \{a\} + jq - jq = \mathbb{B} - jq = \mathbb{B}$$

and $b \in \{a\} + p$, we get

$$b \in \{a\} + p \subseteq \mathbb{B} + p = \mathbb{B} = \{a\} + jq.$$

For (c) $\implies$ (d), it's enough to prove that (c) implies $a \in \{a\} + jq$ for every sufficiently large j. Since $\mathbf{X}$ is arc-consistent, there is some b such that $b \in \{a\} + q$, and since $a \in \{b\} - q$ (c) implies there is some $j \geq 0$ such that $a \in \{b\} + jq$. Since $b \in \{a\} + q$, we have $a \in \{a\} + (j+1)q$. Now set $r = (j+1)q$, and by applying (c) again we see that there is some $k \geq 0$ such that

$$a \in \{b\} + kr \subseteq \{a\} + q + k(j+1)q = \{a\} + (k(j+1) + 1)q.$$

Since $j + 1$ and $k(j+1) + 1$ are relatively prime positive integers, every sufficiently large number can be written as a positive combination of $j + 1$ and $k(j+1) + 1$, which proves (d).

For (d) $\implies$ (a), let q be a path from y to x such that $\mathbb{C} + q = \mathbb{B}$, and let r be a path from x to y such that $\mathbb{B} + r = \mathbb{C}$. Then for every j we have $\mathbb{C} + j(q + r) = \mathbb{C}$. Since the cycle $q + p$ from y to y only involves a subset of the variables involved in $q + r$, by part (d) we have $\mathbb{C} + (q + p) \subseteq \mathbb{C} + j(q + r)$ for every sufficiently large j, so

$$\mathbb{B} + p = \mathbb{C} + q + p \subseteq \mathbb{C} + j(q + r) = \mathbb{C}. \qquad \square$$

Proposition 4.5.11. *If $\mathbf{X}$ is a Prague instance, and if $\mathcal{S}$ is a strand of $\mathbf{X}$, then the full restriction of $\mathbf{X}$ to $\mathcal{S}$ (which restricts both the set of variables and the domains) will be arc-consistent - in fact, it will also be a Prague instance.*

If $\mathcal{S}$ is also a maximal *strand in the qoset of pairs $(x, \mathbb{B})$ such that $\mathbb{B} < \mathbb{A}_x$, then the partial restriction of $\mathbf{X}$ to $\mathcal{S}$ (which restricts the domains but not the set of variables) is also arc-consistent. If $\mathcal{S}$ is additionally absorbing, then the partial restriction of $\mathbf{X}$ to $\mathcal{S}$ will even be a Prague instance.*

Proof. The statement about the full restriction follows from the fact that for any tuple r of any constraint relation $\mathbb{R} \leq_{sd} \mathbb{A}_{x_1} \times \cdots \times \mathbb{A}_{x_n}$, if $\pi_{x_i}(r) \in \mathbb{B}_i$ and $(x_i, \mathbb{B}_i) \in \mathcal{S}$, then we necessarily have $\pi_{x_j}(r) \in \mathbb{B}_j$ for any $(x_j, \mathbb{B}_j) \in \mathcal{S}$, by the definition of a Prague instance. For the statement about the partial restriction, suppose that some variable x_i does not not occur in the strand $\mathcal{S}$, and that the relation $\mathbb{R}$ involves a variable x_j with $(x_j, \mathbb{B}_j) \in \mathcal{S}$. Then by the maximality of $\mathcal{S}$, we must have $\mathbb{B}_j + \pi_{x_j x_i}(\mathbb{R}) = \mathbb{A}_{x_i}$, so for each $a \in \mathbb{A}_{x_i}$, there is a tuple $r \in \mathbb{R}$ with $\pi_{x_j}(r) \in \mathbb{B}_j$, and then by the first part we have $\pi_{x_k}(r) \in \mathbb{B}_k$ for each $(x_k, \mathbb{B}_k) \in \mathcal{S}$.

Now suppose that $\mathcal{S}$ is absorbing. By the equivalence between parts (b) and (c) of Proposition 4.5.10, it's enough to show that for every cycle p from x to x in $\mathbf{X}$, if we let p' be the corresponding path in the partially restricted instance, then for any $a, b \in \mathbb{B}_x$ with $(x, \mathbb{B}_x) \in \mathcal{S}$ and $b \in \{a\} + p$, there is some $j \geq 1$ such that $b \in \{a\} + jp'$. For this, we let $\mathbb{R} = \mathbb{P}_p \leq \mathbb{A}_x \times \mathbb{A}_x$ and $\mathbb{S} = \mathbb{P}_{p'} \leq \mathbb{B}_x \times \mathbb{B}_x$, and we aim to apply Corollary 3.7.10.

Since $\mathcal{S}$ is absorbing, we have $\mathbb{S} \lhd \mathbb{R}$. Since the restriction of $\mathbf{X}$ to $\mathcal{S}$ is arc-consistent, $\mathbb{S}$ is subdirect in $\mathbb{B}_x \times \mathbb{B}_x$. Thus, there is some directed cycle $\mathcal{C}_a$ of $\mathbb{S}$ which can be reached from a in $\mathbb{S}$, and there is some directed cycle $\mathcal{C}_b$ of $\mathbb{S}$ with a directed path from $\mathcal{C}_b$ to b in $\mathbb{S}$. Since $\mathbf{X}$ is a Prague instance and $\mathcal{C}_a, \mathcal{C}_b$ are contained in the same weakly connected component of $\mathbb{R}$, there is some directed path from $\mathcal{C}_a$ to $\mathcal{C}_b$ in $\mathbb{R}$ by Proposition 3.13.13. Now we can apply Corollary 3.7.10 to see that there is a directed path from $\mathcal{C}_a$ to $\mathcal{C}_b$ in $\mathbb{S}$, so there is a directed path from a to b in $\mathbb{S}$, and we are done. $\qquad \square$

In practice, the local consistency algorithm used to reduce general instances to Prague instances or to cycle-consistent instances actually produces an instance with an even greater level of consistency.

**Definition 4.5.12.** An instance $\mathbf{X}$ is (l, k)-*minimal*, for $k \geq l$, if

- every set of at most k variables of $\mathbf{X}$ is in the scope of some constraint, and

- for any set S of at most l variables and any pair of constraints C_1, C_2 of $\mathbf{X}$ whose scopes contain S, the existential projections of C_1 and C_2 to the variables in S are the same.

Proposition 4.5.13. *If $\mathbf{X}$ is $(2, 3)$-minimal, then $\mathbf{X}$ is a Prague instance, and $\mathbf{X}$ is cycle-consistent.*

Proof. We leave cycle-consistency to the reader, and will prove that $\mathbf{X}$ is a Prague instance. By the equivalence of (a) and (b) in Proposition 4.5.10, it's enough to show that for $(x, \mathbb{B}) \preceq (y, \mathbb{C})$ and p a 1 step path from x to y, we always have $\mathbb{B} + p \subseteq \mathbb{C}$. Let q be the path from x to y with $\mathbb{B} + q = \mathbb{C}$, and suppose that the variables which occur along the path q are $x = x_0, x_1, ..., x_n = y$, and let $q = q_1 + \cdots + q_n$ be the decomposition of q into single step paths. For each i, let $\mathbb{B}_i = \mathbb{B} + q_1 + \cdots + q_i$, and let p_i be a single step path from x_i to y (which exists by $(2, 3)$-minimality of $\mathbf{X}$).

We prove by induction on i that $\mathbb{B} + p \subseteq \mathbb{B}_i + p_i$. For the inductive step, we need to show that $\mathbb{B}_i + p_i \subseteq \mathbb{B}_{i+1} + p_{i+1}$, that is, $\mathbb{B}_i + p_i \subseteq \mathbb{B}_i + q_{i+1} + p_{i+1}$. This follows from the fact that there is some constraint whose scope contains the three variables x_i, x_{i+1}, and y, together with the fact that the two-variable projections of this constraint onto pairs of variables from $\{x_i, x_{i+1}, y\}$ agree with the binary relations $\mathbb{P}_{p_i}, \mathbb{P}_{q_{i+1}}, \mathbb{P}_{p_{i+1}}$, by $(2, 3)$-minimality. $\square$

The relationships between the various types of consistency introduced so far are summarized in the Hasse diagram below.
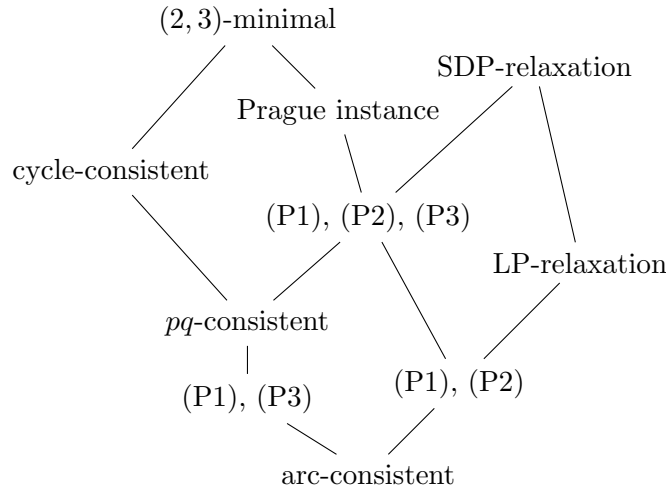
When our instance consists of just a single relation $\mathbb{R}$ (with no repeated variables), all of these consistency conditions become equivalent to subdirectness of the relation $\mathbb{R}$. Studying this very special case is what leads to the main algebraic ingredient we will need for the conservative CSP dichotomy.

Definition 4.5.14. If $\mathbb{R} \leq_{sd} \mathbb{A}_1 \times \cdots \times \mathbb{A}_n$, and if $\mathbb{B}_i \leq \mathbb{A}_i$ for all i, then an $(\mathbb{R}, \mathbb{B})$-*strand* is an equivalence class of the quasiorder $\preceq$ on $[n]$ which is the transitive closure of

$$\mathbb{B}_i + \pi_{ij}(\mathbb{R}) = \mathbb{B}_j \quad \implies \quad i \preceq j.$$

The equivalence classes of this quasiorder are the same as the equivalence classes of the more permissive quasiorder

$$i \preceq' j \quad \Longleftrightarrow \quad \mathbb{B}_i + \pi_{ij}(\mathbb{R}) \subseteq \mathbb{B}_j,$$

and these quasiorders are the same in the special case where $\mathbb{B}_i$ is a minimal absorbing subalgebra of $\mathbb{A}_i$ for each i.

Using either the fact that the instance consisting of just $\mathbb{R}$ is cycle-consistent, or the fact that it satisfies (P1) and (P2), we have the following result.

Proposition 4.5.15. *If $\mathbb{R} \leq_{sd} \mathbb{A}_1 \times \cdots \times \mathbb{A}_n$, and if $\mathbb{B}_i \leq \mathbb{A}_i$ for all i, then for each $(\mathbb{R}, \mathbb{B})$-strand $S \subseteq [n]$ there is a congruence $\theta_i \in \mathrm{Con}(\mathbb{A}_i)$ for each $i \in S$ such that each $\mathbb{B}_i$ is a union of congruence classes of θ_i, and for any $i, j \in S$, the binary relation*

$$\pi_{ij}(\mathbb{R})/(\theta_i \times \theta_j) \leq_{sd} \mathbb{A}_i/\theta_i \times \mathbb{A}_j/\theta_j$$

is the graph of an isomorphism.

The algebraic input needed for conservative CSPs is to show that if we take the $\mathbb{B}_i$s to be minimal absorbing subalgebras of the $\mathbb{A}_i$s such that $\mathbb{R} \cap (\mathbb{B}_1 \times \cdots \times \mathbb{B}_n) \neq \emptyset$, then the $(\mathbb{R}, \mathbb{B})$-strands do not interact with each other. This algebraic miracle is a special property of conservative Taylor algebras - it doesn't seem to hold in general.

4.6 The Rectangularity Theorem for conservative Taylor algebras

There are two versions of the Rectangularity Theorem for conservative algebras: one uses the theory of absorbing subalgebras and is proved in [8], while the other uses Bulatov's theory of affine-semilattice components (shortened to *as-components*) of the colored graph and is proved in [42]. Generally speaking, the affine-semilattice components of conservative Taylor algebras (and of subdirect products of conservative Taylor algebras) tend to behave like minimal absorbing subalgebras. We will mainly focus on the absorbing subalgebra approach.

First we will state both versions of the Rectangularity Theorem, before diving into the proofs.

Theorem 4.6.1 (Rectangularity Theorem for conservative Taylor algebras, absorbing version [8])**.** *If $\mathbb{R} \leq_{sd} \mathbb{A}_1 \times \cdots \times \mathbb{A}_n$ is a subdirect product of conservative Taylor algebras $\mathbb{A}_i$, and if a system of minimal absorbing subalgebras $\mathbb{B}_i \lhd \mathbb{A}_i$ for each i satisfies*

$$\mathbb{R} \cap (\mathbb{B}_1 \times \cdots \times \mathbb{B}_n) \neq \emptyset,$$

then

$$\mathbb{R} \cap (\mathbb{B}_1 \times \cdots \times \mathbb{B}_n) = \prod_{S \ an \ (\mathbb{R},\mathbb{B})\text{-}strand} \left(\pi_S(\mathbb{R}) \cap \prod_{i \in S} \mathbb{B}_i \right).$$

The choice of absorption concept used in the Rectangularity Theorem is fairly arbitrary - we could use Jónsson absorption or central absorption (as long as the algebras in question are minimal Taylor) instead, and it would still be true. We will use $\lhd_X$ in this section to refer to any choice of an absorption concept as in Section 3.9 for which the Absorption Theorem 3.11.1 applies (so if we take $\lhd_X$ to be central absorption, then we need to assume that we are in a context where binary absorption implies central absorption, such as the context of minimal Taylor algebras).

Definition 4.6.2. If $\mathbb{A}$ is a subdirect product of conservative minimal Taylor algebras, then we define a quasiorder on the elements of $\mathbb{A}$ by $a \preceq_{as} b$ if there is a sequence $a = a_0, a_1, ..., a_n = b$ such that for each i either $\{a_i, a_{i+1}\}$ is a $\mathbb{Z}/2^{\mathrm{aff}}$-subalgebra or $a_i \to a_{i+1}$.

We say that a subset $B \subseteq \mathbb{A}$ is an *as-component* of $\mathbb{A}$ if B is a maximal equivalence class of the quasiorder $\preceq_{as}$. We say that $\mathbb{A}$ is *strongly as-connected* if $\mathbb{A}$ consists of a single equivalence class of the quasiorder $\preceq_{as}$.

Theorem 4.6.3 (Rectangularity Theorem for conservative Taylor algebras, as-component version [42]). *If $\mathbb{R} \leq_{sd} \mathbb{A}_1 \times \cdots \times \mathbb{A}_n$ is a subdirect product of conservative minimal Taylor algebras $\mathbb{A}_i$, and if a system of as-components $\mathbb{B}_i \subseteq \mathbb{A}_i$ for each i satisfies*

$$\mathbb{R} \cap (\mathbb{B}_1 \times \cdots \times \mathbb{B}_n) \neq \emptyset,$$

then

$$\mathbb{R} \cap (\mathbb{B}_1 \times \cdots \times \mathbb{B}_n) = \prod_{S \ an \ (\mathbb{R},\mathbb{B})\text{-}strand} \left(\pi_S(\mathbb{R}) \cap \prod_{i \in S} \mathbb{B}_i \right).$$

We will frequently use the following consequence of the Absorption Theorem 3.11.1 throughout the proof. Note that every strongly as-connected algebra is certainly absorption-free, so it applies in that context as well (it's a good exercise to give a direct proof of the analogue for strongly as-connected algebras, along the lines of Theorem 3.3.1 - if you can't solve it, see [42]).

Proposition 4.6.4. *Let $\mathbb{R} \leq_{sd} \mathbb{A}_1 \times \mathbb{A}_2$ be a subdirect product of absorption-free idempotent algebras, and let θ_1 be a maximal congruence on $\mathbb{A}_1$. Then either θ_1 contains the linking congruence of $\mathbb{R}$ on $\mathbb{A}_1$, in which case there is a maximal congruence θ_2 on $\mathbb{A}_2$ such that*

$$\mathbb{R}/(\theta_1 \times \theta_2) \leq \mathbb{A}_1/\theta_1 \times \mathbb{A}_2/\theta_2$$

is the graph of an isomorphism from $\mathbb{A}_1/\theta_1 \xrightarrow{\sim} \mathbb{A}_2/\theta_2$, or else we have

$$(a/\theta_1) + \mathbb{R} = \mathbb{A}_2$$

for each congruence class a/θ_1 of θ_1.

Proof. This is a restatement of Corollary 3.11.4. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

The next lemma is one of the key places where conservativity is really used in the argument. Its analogue for as-components of conservative minimal Taylor algeras (replacing "absorption-free" with "strongly as-connected") is an easy exercise.

Lemma 4.6.5 (Barto [8]). *If $\mathbb{A}$ is an absorption-free conservative algebra, θ is a proper congruence on $\mathbb{A}$, and $\mathbb{B} \leq \mathbb{A}$ is any subalgebra such that $\mathbb{B}$ has at least one element from each congruence class of θ, then $\mathbb{B}$ is also absorption-free.*

Proof. Suppose for contradiction that $\mathbb{C} \lhd_X \mathbb{B}$, with $\mathbb{C} \neq \mathbb{B}$. Pick some $b \in \mathbb{B} \setminus \mathbb{C}$ such that $\mathbb{C} \not\subseteq b/\theta$ (this is possible as long as θ has at least two congruence classes), and let $\mathbb{B}'$ be $(\mathbb{B} \setminus (b/\theta)) \cup \{b\}$, that is, $\mathbb{B}'$ is the subalgebra of $\mathbb{B}$ formed by removing every element which is congruent to b other than b itself (that $\mathbb{B}'$ is a subalgebra follows from the fact that $\mathbb{B}$ is conservative). Then since $\lhd_X$ is compatible with pp-formulas, if we set $\mathbb{C}' = \mathbb{C} \cap \mathbb{B}'$, we have $\mathbb{C}' \lhd_X \mathbb{B}'$. Applying compatibility with pp-formulas again, we have

$$\mathbb{C}'/\theta \lhd_X \mathbb{B}'/\theta = \mathbb{A}/\theta,$$

and by the construction of $\mathbb{C}'$ we see that $\mathbb{C}'/\theta \neq \mathbb{A}/\theta$, since $b/\theta \notin \mathbb{C}'/\theta$. Therefore $\mathbb{A}$ has a proper absorbing subalgebra (i.e. the preimage of $\mathbb{C}'/\theta$ under the quotient homomorphism $\mathbb{A} \twoheadrightarrow \mathbb{A}/\theta$), which is a contradiction. $\square$

The next result is also specific to conservative algebras (see Example 3.3.1 for a counterexample in the 2-semilattice case). An analogue for strongly as-connected algebras can be proved using the same argument (see [42]).

Theorem 4.6.6 (Barto [8]). *If* $\mathbb{R} \leq_{sd} \mathbb{A}_1 \times \cdots \times \mathbb{A}_n$ *is a subdirect product of absorption-free conservative algebras, then* $\mathbb{R}$ *is also absorption-free.*

Proof. We induct on n and on the sizes of the $\mathbb{A}_i$s. Suppose that $\mathbb{S} \lhd_X \mathbb{R}$ - we aim to prove that $\mathbb{S} = \mathbb{R}$. By compatibility with pp-formulas we have $\pi_i(\mathbb{S}) \lhd_X \pi_i(\mathbb{R}) = \mathbb{A}_i$ for each i, so since the $\mathbb{A}_i$s are absorption-free $\mathbb{S}$ is also a subdirect product of the $\mathbb{A}_i$s.

Let θ_1 be a maximal congruence on $\mathbb{A}_1$. By Proposition 4.6.4, for each i either

- there is a maximal congruence $\theta_i \in \mathrm{Con}(\mathbb{A}_i)$ such that $\pi_{1i}(\mathbb{R})/\theta_1 \times \theta_i$ is the graph of an isomorphism between $\mathbb{A}_1/\theta_1$ and $\mathbb{A}_i/\theta_i$, or

- we have $a/\theta_1 + \pi_{1i}(\mathbb{R}) = \mathbb{A}_i$ for all congruence classes a/θ_1 of θ_1.

Rearrange the coordinates so that $\mathbb{A}_1, ..., \mathbb{A}_k$ are in the first case, with corresponding maximal congruences θ_i, while $\mathbb{A}_{k+1}, ..., \mathbb{A}_n$ are in the second case. Define a congruence θ on the product by

$$\theta = \theta_1 \times \cdots \times \theta_k \times 0_{\mathbb{A}_{k+1}} \times \cdots \times 0_{\mathbb{A}_n} \in \mathrm{Con}(\mathbb{A}_1 \times \cdots \times \mathbb{A}_n).$$

First suppose that θ is the trivial congruence (i.e. each $\theta_i = 0_{\mathbb{A}_i}$). In this case each $\mathbb{A}_i$ with $1 < i \leq k$ is a redundant coordinate (since $\pi_{1i}(\mathbb{R})$ is the graph of an isomorphism between $\mathbb{A}_1$ and $\mathbb{A}_i$), so we can assume without loss of generality that $k = 1$. Let a be any element of $\mathbb{A}$, and consider the relations $\mathbb{S}_a, \mathbb{R}_a$ given by

$$\mathbb{R}_a = \{(x_2, ..., x_n) \mid (a, x_2, ..., x_n) \in \mathbb{R}\} \leq \mathbb{A}_2 \times \cdots \times \mathbb{A}_n,$$

with $\mathbb{S}_a$ defined similarly. By compatibility with pp-formulas, we have $\mathbb{S}_a \lhd_X \mathbb{R}_a$, and since $a + \pi_{1i}(\mathbb{R}) = \mathbb{A}_i$ for each $i \geq 2$, $\mathbb{R}_a$ is a subdirect product of $\mathbb{A}_2, ..., \mathbb{A}_n$. By the induction hypothesis, we then have $\mathbb{S}_a = \mathbb{R}_a$, and since this is true for every $a \in \mathbb{A}$, we have $\mathbb{S} = \mathbb{R}$.

Now suppose that θ is nontrivial - suppose without loss of generality that θ_1 is a nontrivial congruence of $\mathbb{A}_1$, with $(a, a') \in \theta_1$ for some $a \neq a'$. Let $\mathbb{B} = \mathbb{A} \setminus \{a\}$, and let $\mathbb{B}' = \mathbb{A} \setminus \{a'\}$. By Lemma 4.6.5, each of $\mathbb{B}, \mathbb{B}'$ is absorption-free. Define $\mathbb{R}_\mathbb{B}$ by

$$\mathbb{R}_\mathbb{B} = \{(x_2, ..., x_n) \mid \exists b \in \mathbb{B} \ (b, x_2, ..., x_n) \in \mathbb{R}\} \leq \mathbb{A}_2 \times \cdots \times \mathbb{A}_n,$$

and similarly define $\mathbb{S}_\mathbb{B}$, $\mathbb{R}_{\mathbb{B}'}, \mathbb{S}_{\mathbb{B}'}$. We have $\mathbb{S}_\mathbb{B} \lhd_X \mathbb{R}_\mathbb{B}$ and $\mathbb{S}_{\mathbb{B}'} \lhd_X \mathbb{R}_{\mathbb{B}'}$ by compatibility with pp-formulas. For each $i \leq k$, we have

$$\pi_i(\mathbb{R}_\mathbb{B})/\theta_i = \pi_i(\mathbb{R}_\mathbb{B}/\theta) = \mathbb{B}/\theta_1 + \pi_{1i}(\mathbb{R}/\theta) = \mathbb{A}_1/\theta_1 + \pi_{1i}(\mathbb{R}/\theta) = \mathbb{A}_i/\theta_i,$$

so by Lemma 4.6.5, $\pi_i(\mathbb{R}_\mathbb{B})$ is absorption-free for each $i \leq k$. For $i > k$, we have

$$\pi_i(\mathbb{R}_\mathbb{B}) = \mathbb{B} + \pi_{1i}(\mathbb{R}) = \mathbb{A}_i,$$

since $\mathbb{B}$ contains at least one full congruence class of θ_1. Thus for each $i \in [n]$, $\pi_i(\mathbb{R}_\mathbb{B})$ is absorption-free, so since $|\mathbb{B}| < |\mathbb{A}_1|$ we can apply the induction hypothesis to see that $\mathbb{S}_\mathbb{B} = \mathbb{R}_\mathbb{B}$. A similar argument shows that $\mathbb{S}_{\mathbb{B}'} = \mathbb{R}_{\mathbb{B}'}$, and since $\mathbb{B} \cup \mathbb{B}' = \mathbb{A}_1$ we see that $\mathbb{S} = \mathbb{R}$. □

In order to get a foothold into the Rectangularity Theorem, we start with the case of binary relations. We will only assume that one of the algebras involved is conservative - this way we can later apply the result with the other algebra equal to a larger subdirect product of conservative algebras. Once again, an analogous argument works for as-components (we need a version of Theorem 3.3.1 for as-components to push the argument through, see [42]).

Lemma 4.6.7 (Barto [8]). *Suppose $\mathbb{R} \leq_{sd} \mathbb{A}_1 \times \mathbb{A}_2$, where $\mathbb{A}_1$ is conservative, $\mathbb{A}_2$ is idempotent, and both are finite Taylor algebras. Suppose further that $\mathbb{B}_i \lll_X \mathbb{A}_i$ for $i = 1, 2$, that $\mathbb{R} \cap (\mathbb{B}_1 \times \mathbb{B}_2) \neq \emptyset$, and that there is some $(a, b) \in \mathbb{R}$ with $a \in \mathbb{A}_1 \setminus \mathbb{B}_1$ and $b \in \mathbb{B}_2$. Then $\mathbb{B}_1 \times \mathbb{B}_2 \subseteq \mathbb{R}$.*

Proof. Since $(\mathbb{B}_1 + \mathbb{R}) \cap \mathbb{B}_2$ is a nonempty absorbing subalgebra of $\mathbb{B}_2$ and $\mathbb{B}_2$ is absorption-free, we have $\mathbb{B}_1 + \mathbb{R} \supseteq \mathbb{B}_2$, and similarly $\mathbb{B}_1 \subseteq \mathbb{B}_2 - \mathbb{R}$. Thus $\mathbb{R} \cap (\mathbb{B}_1 \times \mathbb{B}_2)$ is subdirect in $\mathbb{B}_1 \times \mathbb{B}_2$. Additionally, $\mathbb{B}_1 \cup \{a\}$ is a subalgebra of $\mathbb{A}_1$ since $\mathbb{A}_1$ is conservative, so by the assumption $a \in \mathbb{B}_2 - \mathbb{R}$ we can assume without loss of generality that $\mathbb{A}_1 = \mathbb{B}_1 \cup \{a\}$ and $\mathbb{A}_2 = \mathbb{B}_2$.

If $\mathbb{R}$ is linked, then by Theorem 3.7.12 so is $\mathbb{R} \cap (\mathbb{B}_1 \times \mathbb{B}_2)$, and then by the Absorption Theorem 3.11.1 we have $\mathbb{B}_1 \times \mathbb{B}_2 \subseteq \mathbb{R}$. Otherwise, the linking congruence of $\mathbb{R}$ is a proper congruence θ_1 on $\mathbb{A}_1$. If $c \in (\{b\} - \mathbb{R}) \cap \mathbb{B}_1$, then we have $(a, c) \in \theta_1$, so $\mathbb{A}_1/\theta_1 = \mathbb{B}_1/\theta_1$. Let

$$\mathbb{A}_1' = (\mathbb{A}_1 \setminus (a/\theta_1)) \cup \{a\},$$

then by compatibility with pp-formulas we have

$$\mathbb{B}_1 \setminus (a/\theta_1) = \mathbb{B}_1 \cap \mathbb{A}_1' \lhd_X \mathbb{A}_1',$$

so

$$(\mathbb{B}_1/\theta_1) \setminus (a/\theta_1) \lhd_X \mathbb{A}_1'/\theta_1 = \mathbb{B}_1/\theta_1,$$

which implies that $\mathbb{B}_1 \setminus (a/\theta_1) \lhd_X \mathbb{B}_1$, which is a contradiction. □

Now we bootstrap our way up.

Lemma 4.6.8 (Barto [8]). *Suppose $\mathbb{R} \leq_{sd} \mathbb{A}_1 \times \cdots \times \mathbb{A}_n \times \mathbb{A}_{n+1}$, where $\mathbb{A}_1, ..., \mathbb{A}_n$ are conservative, $\mathbb{A}_{n+1}$ is idempotent, and each $\mathbb{A}_i$ is a finite Taylor algebra. Suppose that we have $\mathbb{B}_i \lll_X \mathbb{A}_i$ for all $i \in [n+1]$, that*

$$\mathbb{R} \cap (\mathbb{B}_1 \times \cdots \times \mathbb{B}_n \times \mathbb{B}_{n+1}) \neq \emptyset,$$

that $[n]$ is an $(\mathbb{R}, \mathbb{B})$-strand, and that there is some $(a_1, ..., a_n, b_{n+1}) \in \mathbb{R}$ such that $a_i \in \mathbb{A}_i \setminus \mathbb{B}_i$ for $i \in [n]$ while $b_{n+1} \in \mathbb{B}_{n+1}$. Then we have

$$\mathbb{R} \cap (\mathbb{B}_1 \times \cdots \times \mathbb{B}_n \times \mathbb{B}_{n+1}) = (\pi_{[n]}(\mathbb{R}) \cap (\mathbb{B}_1 \times \cdots \times \mathbb{B}_n)) \times \mathbb{B}_{n+1}.$$

Proof. We induct on n and on the sizes of the $\mathbb{A}_i$s, and note that we have already proved the case $n = 1$ in the previous lemma. By Proposition 4.5.15 we can find maximal congruences $\theta_i \in \mathrm{Con}(\mathbb{A}_i)$ for $i \in [n]$ such that

$$\pi_{ij}(\mathbb{R})/(\theta_i \times \theta_j) \leq \mathbb{A}_i/\theta_i \times \mathbb{A}_j/\theta_j$$

is the graph of an isomorphism for all $i, j \in [n]$. If any θ_i is trivial (i.e. if $\mathbb{A}_i$ is simple for some $i \in [n]$) then the ith coordinate of $\mathbb{R}$ is redundant, so we can apply the induction hypothesis. Otherwise, we have $\theta_1 \neq 0_{\mathbb{A}_1}$.

Let b_1 be any element of $\mathbb{B}_1$. Let $\mathbb{A}'_1$ be any proper subalgebra of $\mathbb{A}_1$ such that $b_1, a_1 \in \mathbb{A}'_1$ and such that $\mathbb{A}'_1/\theta_1 = \mathbb{A}_1/\theta_1$, and let $\mathbb{B}'_1 = \mathbb{B}_1 \cap \mathbb{A}'_1$. For each $i \in [n+1]$ define $\mathbb{A}'_i, \mathbb{B}'_i$ by

$$\mathbb{A}'_i = \mathbb{A}'_1 + \pi_{1i}(\mathbb{R}), \quad \mathbb{B}'_i = \mathbb{B}_i \cap \mathbb{A}'_i.$$

Then since $\pi_{1i}(\mathbb{R})/(\theta_1 \times \theta_i)$ is the graph of an isomorphism for each $i \in [n]$, we have $\mathbb{A}'_i/\theta_i = \mathbb{A}_i/\theta_i$ and $\mathbb{B}'_i = \mathbb{B}'_1 + \pi_{1i}(\mathbb{R})$ for all $i \in [n]$. Then since each $\mathbb{B}'_i/\theta_i = \mathbb{B}_i/\theta_i$, we can apply Lemma 4.6.5 to see that $\mathbb{B}'_i$ is absorption-free for each $i \in [n]$. Additionally, by the previous lemma (i.e., the $n = 1$ case) we have

$$\mathbb{B}_1 \times \mathbb{B}_{n+1} \subseteq \pi_{1,n+1}(\mathbb{R}),$$

so $\mathbb{B}'_{n+1} = \mathbb{B}_{n+1}$. Thus if we set

$$\mathbb{R}' = \mathbb{R} \cap (\mathbb{A}'_1 \times \cdots \times \mathbb{A}'_n \times \mathbb{A}'_{n+1}),$$

then $(a_1, ..., a_n, b_{n+1}) \in \mathbb{R}'$ and we can apply the induction hypothesis to $\mathbb{R}'$ to see that

$$\mathbb{R}' \cap (\mathbb{B}'_1 \times \cdots \times \mathbb{B}'_n \times \mathbb{B}_{n+1}) = (\pi_{[n]}(\mathbb{R}') \cap (\mathbb{B}'_1 \times \cdots \times \mathbb{B}'_n)) \times \mathbb{B}_{n+1}.$$

In particular, any tuple in $(\pi_{[n]}(\mathbb{R}) \cap \prod_{i \in [n]} \mathbb{B}_i) \times \mathbb{B}_{n+1}$ such that the first coordinate is b_1 is contained in $\mathbb{R}'$, and therefore is also contained in $\mathbb{R}$. Since b_1 was an arbitrary element of $\mathbb{B}_1$, we are done. $\square$

Proof of the Rectangularity Theorem 4.6.1, following [8]. We induct on n, the number of algebras occuring in the product. Suppose for the sake of contradiction that $\mathbb{R}$ is a counterexample, i.e. that there is some

$$b = (b_1, ..., b_n) \in \prod_{S \text{ an } (\mathbb{R}, \mathbb{B})\text{-strand}} \left(\pi_S(\mathbb{R}) \cap \prod_{i \in S} \mathbb{B}_i \right)$$

such that $b \notin \mathbb{R}$. By the induction hypothesis, we have $\pi_{[n] \setminus \{i\}}(b) \in \pi_{[n] \setminus \{i\}}(\mathbb{R})$ for each $i \in [n]$.

Consider the quasiorder $\preceq$ on $[n]$ from Definition 4.5.14 defined by

$$i \preceq j \iff \mathbb{B}_i + \pi_{ij}(\mathbb{R}) = \mathbb{B}_j,$$

and suppose without loss of generality that $[k]$ is a $\preceq$-minimal $(\mathbb{R}, \mathbb{B})$-strand for some $k \leq n$. If there is any tuple

$$(a_1, ..., a_k, b'_{k+1}, ..., b'_n) \in \mathbb{R}$$

such that $a_i \in \mathbb{A}_i \setminus \mathbb{B}_i$ for $i \in [k]$ and $b'_j \in \mathbb{B}_j$ for $j > k$, then we can apply the previous lemma to the situation

$$\mathbb{R} \leq_{sd} \mathbb{A}_1 \times \cdots \times \mathbb{A}_k \times \pi_{[k+1,n]}(\mathbb{R})$$

with

$$\pi_{[k+1,n]}(\mathbb{R}) \cap \prod_{j > k} \mathbb{B}_j \lhd_X \pi_{[k+1,n]}(\mathbb{R}),$$

265

by compatibility with pp-formulas and Theorem 4.6.6, to finish the proof. To arrange for this situation, we consider the relation

$$\mathbb{R}' = \mathbb{R} \cap \left(\left(\{(b_1, ..., b_k)\} \times \prod_{i \in [k]} (\mathbb{A}_i \setminus \mathbb{B}_i) \right) \times \prod_{j > k} \mathbb{A}_j \right),$$

set $\mathbb{A}'_i = \pi_i(\mathbb{R}')$, and set $\mathbb{B}'_i = \mathbb{A}'_i \cap \mathbb{B}_i$ for each i. By the induction hypothesis applied to $\pi_{[k] \cup \{j\}}(\mathbb{R})$, we see that $\mathbb{B}'_j = \mathbb{B}_j$ for each $j > k$, so we have $\mathbb{B}'_i \lhd_X \mathbb{A}'_i$ for all $i \leq n$. We will apply the induction hypothesis to $\pi_{[k+1,n]}(\mathbb{R}')$, but first we need to check that $\pi_{[k+1,n]}(\mathbb{R}')$ has more than one strand.

Let $S \subset [n]$ be some $(\mathbb{R}, \mathbb{B})$-strand which is disjoint from $[k]$. By the assumption that $[k]$ was a $\preceq$-minimal $(\mathbb{R}, \mathbb{B})$-strand, there must be some $(c_1, ..., c_n) \in \mathbb{R}$ such that $c_i \notin \mathbb{B}_i$ for $i \in [k]$ and $c_j \in \mathbb{B}_j$ for $j \in S$. Suppose without loss of generality that there is some $m \geq k$ such that $c_j \in \mathbb{B}_j$ iff $j \in [m+1, n]$. If $m = k$, then we take $(a_1, ..., a_k, b'_{k+1}, ..., b'_n) = c$ to finish. Otherwise, since c is an element of $\mathbb{R}'$, we see that every strand of $\pi_{[k+1,n]}(\mathbb{R}')$ is either contained in $[k+1, m]$ or contained in $[m+1, n]$. Thus, by the induction hypothesis we have

$$\pi_{[k+1,n]}(\mathbb{R}') \cap \prod_{j > k} \mathbb{B}_j = \left(\pi_{[k+1,m]}(\mathbb{R}') \cap \prod_{j \in [k+1,m]} \mathbb{B}_j \right) \times \left(\pi_{[m+1,n]}(\mathbb{R}') \cap \prod_{j \in [m+1,n]} \mathbb{B}_j \right).$$

Since we have $\pi_{[k+1,m]}(b) \in \pi_{[k+1,m]}(\mathbb{R}')$ and $\pi_{[m+1,n]}(b) \in \pi_{[m+1,n]}(\mathbb{R}')$ by the induction hypothesis applied to $\pi_{[1,m]}(\mathbb{R})$ and $\pi_{[k] \cup [m+1,n]}(\mathbb{R})$, we see that $\pi_{[k+1,n]}(b) \in \mathbb{R}'$. Thus either $b \in \mathbb{R}$, or there are some $a_i \in \mathbb{A}_i \setminus \mathbb{B}_i$ for $i \in [k]$ such that

$$(a_1, ..., a_k, b_{k+1}, ..., b_n) \in \mathbb{R},$$

which allows us to apply the previous lemma to finish the proof. $\qquad\square$

4.7 The algorithm for conservative CSPs

In this section we present Barto's simple algorithm from [8]. Bulatov's algorithm from [42] is similar in spirit, but it relies on ideas from Maróti's "Tree on top of Malcev" algorithm [107] which we haven't covered yet.

The main idea of Barto's algorithm for conservative CSPs is to try to reduce to the case where all edges of the colored graphs occuring in each of the variable domains $\mathbb{A}_x$ are affine. In this case, any daisy chain term will be a Mal'cev term for each variable domain, and we can solve the problem by using the algorithm for CSPs with a Mal'cev polymorphism. In Bulatov's algorithm from [42], the main idea is to reduce to the case where there are no semilattice edges instead, in which case any daisy chain term will be a ternary generalized majority-minority operation (that a ternary generalized majority-minority operation exists in this case also follows from Theorem 2.1.5).

In order to accomplish this, we aim to show that if any semilattice or majority edge occurs in any variable domain of an instance $\mathbf{X}$, then we can reduce some variable domain by solving an instance where every variable domain has been strictly decreased. We assume that our instance is $(2, 3)$-minimal (or perhaps just that it is a Prague instance), and we consider the subqoset $\mathcal{E}$ of the implication qoset consisting of pairs $(x, \mathbb{B})$ such that that $\mathbb{B} \leq \mathbb{A}_x$ is an algebra with at least one proper absorbing subalgebra. The qoset $\mathcal{E}$ will be nonempty as long as any algebra $\mathbb{A}_x$ has any non-affine edge. We pick any maximal strand $\mathcal{S}$ of the qoset $\mathcal{E}$, and we note that the full restriction of our instance $\mathbf{X}$ to the strand $\mathcal{S}$ is then a Prague instance in which every single domain has

a proper absorbing subalgebra by Proposition 4.5.11. We can then repeatedly apply Proposition 4.5.11 (or, alternatively, we could apply Kozik's [95] result from Section 3.9, using the fact that every Prague instance is *pq*-consistent), to find an arc-consistent absorbing reduction $\mathbf{X}'$ of the full restriction of $\mathbf{X}$ to the strand $\mathcal{S}$, such that every variable domain in $\mathbf{X}'$ is absorption free - and as a consequence, such that each variable domain in $\mathbf{X}'$ is a proper subagebra of the corresponding variable domain in the original instance $\mathbf{X}$. We can then apply the following consequence of the Rectangularity Theorem 4.6.1.

Theorem 4.7.1 (Barto [8]). *Suppose that $\mathbf{X}$ is a Prague instance such that each variable domain $\mathbb{A}_x$ is a conservative Taylor algebra. Let $\mathcal{E}$ be the subqoset of the implication qoset which consists of pairs $(x, \mathbb{B})$ such that $\mathbb{B} \le \mathbb{A}_x$ and $\mathbb{B}$ has a proper absorbing subalgebra, and let $\mathcal{S}$ be any maximal strand of $\mathcal{E}$.*

Suppose that for each $(x, \mathbb{B}_x) \in \mathcal{S}$ we choose a minimal absorbing subalgebra $\mathbb{C}_x \lhd \mathbb{B}_x$, such that the system of variable domains $\mathbb{C}_x$ defines an arc-consistent reduction of the full restriction of $\mathbf{X}$ to the strand $\mathcal{S}$. Then one of the following is true:

- *the instance $\mathbf{X}$ has no solutions with any variable x assigned to any value in $\mathbb{B}_x$, for any $(x, \mathbb{B}_x) \in \mathcal{S}$,*

- *the instance $\mathbf{X}'$ which we get by restricting to the variables in $\mathcal{S}$ and by restricting each variable domain to the corresponding $\mathbb{C}_x$ has a solution, and every solution of $\mathbf{X}'$ extends to a solution of $\mathbf{X}$, or*

- *for some strand $\mathcal{T}$ of the subqoset $\mathcal{C}$ of $\mathcal{E}$ consisting of the pairs $(x, \mathbb{C}_x)$ for x occuring in $\mathcal{S}$, the full restriction of $\mathbf{X}$ to $\mathcal{T}$ has no solutions.*

Proof. Suppose that $a \in \prod_x \mathbb{A}_x$ is a solution to the instance $\mathbf{X}$ such that $a_x \in \mathbb{B}_x$ for some $(x, \mathbb{B}_x) \in \mathcal{S}$, and suppose that for each strand $\mathcal{T}$ of the subqoset $\mathcal{C}$ there is a solution $c^{\mathcal{T}} \in \prod_{(x, \mathbb{C}_x) \in \mathcal{T}} \mathbb{C}_x$ of the full restriction of $\mathbf{X}$ to the strand $\mathcal{T}$. We construct a tuple $b \in \prod_x \mathbb{A}_x$ by stitching these solutions together:

$$b_x = \begin{cases} a_x & x \text{ does not occur in the strand } \mathcal{S}, \\ c_x^{\mathcal{T}} & x \text{ occurs in the strand } \mathcal{T} \text{ of the subqoset } \mathcal{C}. \end{cases}$$

We claim that the tuple b is also a solution to the instance $\mathbf{X}$. For this, we can focus our attention on any particular constraint relation

$$\mathbb{R} \le_{sd} \mathbb{A}_{x_1} \times \cdots \times \mathbb{A}_{x_n}$$

of the instance $\mathbf{X}$. Suppose without loss of generality that the variables $x_1, ..., x_k$ occur in the strand $\mathcal{S}$ and that the variables $x_{k+1}, ..., x_n$ do not, and suppose that $k \ge 1$. For each $i \in [n]$, define $\mathbb{B}_i$ by

$$\mathbb{B}_i = \mathbb{B}_{x_1} + \pi_{1i}(\mathbb{R}).$$

Since $\mathbf{X}$ is a Prague instance, we have $\mathbb{B}_i = \mathbb{B}_{x_i}$ for all $i \le k$, and since $\mathcal{S}$ is a maximal strand of $\mathcal{E}$ each $\mathbb{B}_j$ with $j \ge k + 1$ is absorption-free. Define the relation $\mathbb{R}_{\mathbb{B}} \le \mathbb{R}$ by

$$\mathbb{R}_{\mathbb{B}} = \mathbb{R} \cap \left(\prod_i \mathbb{B}_i \right) \le_{sd} \mathbb{B}_1 \times \cdots \times \mathbb{B}_n,$$

where subdirectness follows from Proposition 4.5.11. Set $\mathbb{C}_i = \mathbb{C}_{x_i}$ for $i \leq k$, and set $\mathbb{C}_j = \mathbb{B}_j$ for $j \geq k+1$, so we have $\mathbb{C}_i \lll \mathbb{B}_i$ for all $i \in [n]$. Then since $\mathbf{X}$ is a Prague instance, the $(\mathbb{R}_{\mathbb{B}}, \mathbb{C})$-strands are given by $[k+1, n]$ and by the intersections of the strands of $\mathcal{C}$ to $\{x_1, ..., x_k\}$, and we have

$$\mathbb{R}_{\mathbb{B}} \cap \prod_i \mathbb{C}_i \neq \emptyset$$

since the system of variable domains $\mathbb{C}_x$ defines an arc-consistent reduction of the full restriction of $\mathbf{X}$ to the strand $\mathcal{S}$. Then the Rectangularity Theorem 4.6.1 says that

$$\mathbb{R} \supseteq \prod_{T \text{ an } (\mathbb{R}_{\mathbb{B}}, \mathbb{C})\text{-strand}} \left(\pi_T(\mathbb{R}_{\mathbb{B}}) \cap \prod_{i \in T} \mathbb{C}_i \right),$$

so the tuple b satisfies the constraint relation $\mathbb{R}$. $\qquad\square$

Using this result, we get the following algorithm for solving conservative CSPs.

Algorithm 12 Algorithm for solving an instance $\mathbf{X}$ of a CSP with conservative Taylor variable domains $\mathbb{A}_x$, from [8].

1: Run a local consistency algorithm until $\mathbf{X}$ is a Prague instance.
2: Let $\mathcal{E}$ be the subqoset of the implication qoset consisting of pairs $(x, \mathbb{B})$ such that $\mathbb{B} \leq \mathbb{A}_x$ and $\mathbb{B}$ has a proper absorbing subalgebra.
3: **if** $\mathcal{E}$ is non-empty **then**
4:     Let $\mathcal{S}$ be any maximal strand of $\mathcal{E}$.
5:     Define $\mathbb{C}_x = \mathbb{B}_x$ for each $(x, \mathbb{B}_x) \in \mathcal{S}$.
6:     **while** some $\mathbb{C}_x$ is not absorption-free **do**
7:         Let $\mathbf{X}_{\mathbb{C}}$ be the Prague instance we get by restricting to variables in $\mathcal{S}$ and restricting each variable domain to $\mathbb{C}_x$.
8:         Pick any maximal strand $\mathcal{T}$ of the subqoset of the implication qoset of $\mathbf{X}_{\mathbb{C}}$ consisting of $(x, \mathbb{C}')$ such that $\mathbb{C}' \lhd \mathbb{C}_x$ and $\mathbb{C}' \neq \mathbb{C}_x$.
9:         Set $\mathbb{C}_x \leftarrow \mathbb{C}'$ for each $(x, \mathbb{C}') \in \mathcal{T}$.
10:     Let $\mathcal{C}$ be the subqoset consisting of $(x, \mathbb{C}_x)$ for x occuring in the strand $\mathcal{S}$.
11:     **for all** strands $\mathcal{T}$ of $\mathcal{C}$ **do**
12:         Let $\mathbf{X}_{\mathcal{T}}$ be the full restriction of $\mathbf{X}$ to the strand $\mathcal{T}$.
13:         Solve the instance $\mathbf{X}_{\mathcal{T}}$ recursively. $\rhd$ $\mathbb{C}_x < \mathbb{B}_x$ for all $(x, \mathbb{C}_x) \in \mathcal{T}$.
14:         **if** $\mathbf{X}_{\mathcal{T}}$ has no solutions **then**
15:             Set $\mathbb{A}_x \leftarrow \mathbb{A}_x \setminus \mathbb{C}_x$ for each $(x, \mathbb{C}_x) \in \mathcal{T}$.
16:             **go to** Step 1.
17:         **else**
18:             Let $c^{\mathcal{T}}$ be a solution to the instance $\mathbf{X}_{\mathcal{T}}$.
19:     Set $\mathbb{A}_x \leftarrow (\mathbb{A}_x \setminus \mathbb{B}_x) \cup \{c_x^{\mathcal{T}}\}$ for all $(x, \mathbb{B}_x) \in \mathcal{S}$, where $\mathcal{T}$ is the strand of $\mathcal{C}$ which contains $(x, \mathbb{C}_x)$.
20:     **go to** Step 1.
21: Solve $\mathbf{X}$ by using the algorithm for CSPs with a Mal'cev polymorphism.

Theorem 4.7.2. *Algorithm 4.7 correctly solves every instance* **X** *of any multisorted CSP where each variable domain is a conservative Taylor algebra. If each variable domain has size at most k, then Algorithm 4.7 runs in time $\|\mathbf{X}\|^{O(k)}$, where $\|\mathbf{X}\|$ is the number of bits needed to describe the instance* **X**.

It seems plausible that a much more careful analysis of the algorithm for conservative CSPs might show that it runs in time $\|\mathbf{X}\|^{O(1)}$, regardless of the sizes of the variable domains.

Corollary 4.7.3. *The CSP dichotomy conjecture is true for all CSP templates on a domain of size at most 3.*

Proof. By the classification of minimal Taylor algebras of size 3 from Subsection 4.4.1, every minimal Taylor algebra of size at most 3 is either a subdirect product of conservative Taylor algebras, or has a 3-edge term. □

4.8 The meta-problem for conservative CSP templates

In this section we will go over Carbonnel's solution to the meta-problem for conservative CSPs from their thesis [45]. Recall that in the meta-problem, we are given a CSP template as a relational structure $\mathbf{A} = (A, \Gamma)$ (which we usually assume to be a core), and we wish to either prove that $\mathrm{CSP}(\mathbf{A})$ is NP-complete or to find a Taylor polymorphism of $\mathbf{A}$, in time polynomial in the total size $\|\mathbf{A}\|$ of the description of $\mathbf{A}$, which we define as

$$\|\mathbf{A}\| := \sum_{R \in \Gamma} |R|.$$

In the meta-problem for conservative CSPs, we restrict our attention to CSP templates where Γ contains the unary relation $A \setminus \{a\}$ for each $a \in A$ (in particular, any such $\mathbf{A}$ is automatically a rigid core). Note that by our classification of conservative minimal Taylor algebras, if a conservative CSP template $\mathbf{A}$ has a Taylor polymorphism, then it has a ternary weak near-unanimity polymorphism, i.e. an idempotent ternary polymorphism w satisfying the identities

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

Furthermore, we can consider the case of 3-conservative CSP templates without any additional difficulty, since any 3-conservative Taylor algebra has a conservative Taylor reduct (by a 3-conservative template, we mean a CSP template such that Γ contains every unary relation of size at most 3). These facts were not known at the time that Carbonnel wrote their thesis, and by using them we can simplify Carbonnel's algorithm.

In Carbonnel's thesis [45], the strategy for solving the meta-problem was described as being similar to a treasure hunt (or perhaps a puzzle hunt): we have a sequence of locked boxes, and a single key which opens the first box, such that each box contains the key to opening the next box. Here, the key is a metaphor for the Taylor polymorphism - once we know a Taylor polymorphism, we can use it to solve instances of our CSP. More specifically, the key is a metaphor for a *partial description* of a Taylor polymorphism. In order to see how a partial description of a Taylor polymorphism can make sense, we first describe a useful rephrasing of the meta-problem for a 3-conservative template in terms of a meta-problem for a multisorted CSP template.

Definition 4.8.1. Suppose that $\mathbf{A} = (A, \Gamma)$ is a 3-conservative CSP template. We define the *associated multisorted template* $\mathbf{A}_3$ to have a sort for each subset of A of size at most 3, with two types of relations:

- for each relation $R \in \Gamma$ of arity m, and for every triple of elements $u, v, w \in \mathbb{R}$ (not necessarily distinct), we have a multisorted relation

$$R \cap (\{u_1, v_1, w_1\} \times \cdots \times \{u_n, v_n, w_n\}) \subseteq \{u_1, v_1, w_1\} \times \cdots \times \{u_n, v_n, w_n\},$$

- for every $a, b, c \in A$ (not necessarily distinct), we have the binary inclusion relation

$$\{(a, a), (b, b)\} \subseteq \{a, b\} \times \{a, b, c\}.$$

Note that the size of the associated multisorted template $\mathbf{A}_3$ is bounded by

$$\|\mathbf{A}_3\| \leq \sum_{R \in \Gamma} \sum_{u, v, w \in R} |R| + \sum_{a, b, c \in A} |\{a, b\}| \leq \|\mathbf{A}\|^4 + 2|A|^3.$$

A polymorphism of a multisorted relational structure is defined to be an operation with a different interpretation on each sort of the structure, such that applying the operation componentwise preserves each multisorted relation.

Proposition 4.8.2. *There is a bijection between conservative ternary polymorphisms of $\mathbf{A}$ and ternary polymorphisms of the associated multisorted structure $\mathbf{A}_3$ which preserves height 1 identities.*

Proof. There is an obvious way to convert any conservative ternary polymorphism of $\mathbf{A}$ into a ternary polymorphism of $\mathbf{A}_3$. Conversely, any ternary polymorphism f of $\mathbf{A}_3$ can be stitched together into a (necessarily conservative) ternary polymorphism $\tilde{f}$ of $\mathbf{A}$: the fact that f preserves the binary inclusion relations guarantees that the values of $\tilde{f}$ are well-defined on two-element sets, and for every $u, v, w \in R \in \Gamma$ the fact that f preserves the multisorted relation

$$R \cap (\{u_1, v_1, w_1\} \times \cdots \times \{u_n, v_n, w_n\})$$

guarantees that $\tilde{f}(u, v, w) \in R$. $\qquad\qquad\square$

So we have reduced the problem of determining whether a given relational structure $\mathbf{A}$ has a conservative Taylor polymorphism to the problem of determining whether the associated multisorted relational structure $\mathbf{A}_3$ has a weak near-unanimity polymorphism. Now we have a way to understand what a partial description of a weak near-unanimity polymorphism on $\mathbf{A}_3$ should be.

Definition 4.8.3. Suppose that $\mathbf{A} = (A, \Gamma)$ is a 3-conservative CSP template, and let $\mathcal{U} \subseteq \mathcal{P}(A)$ be a collection nonempty subsets of A each of size at most 3 which is closed under taking nonempty subsets. We define the multisorted template $\mathbf{A}_{\mathcal{U}}$ to be the template whose sorts are exactly the elements of $\mathcal{U}$, such that for each m-ary relation R of $\mathbf{A}_3$ with at least one coordinate of a sort in $\mathcal{U}$, if $S \subseteq [m]$ is the set of coordinates of R which have a sort in $\mathcal{U}$, then the relation $\pi_S(R)$ is a relation of $\mathbf{A}_{\mathcal{U}}$.

Later we will want to fix certain coordinates of various relations of $\mathbf{A}_3$ to have certain values, before projecting to the coordinates with sorts in $\mathcal{U}$. To reassure ourselves that this will not cause unexpected problems, we have the following result.

Proposition 4.8.4. *Suppose that* $\mathbf{A} = (A, \Gamma)$ *is a 3-conservative CSP template. Let* $\mathcal{U} \subseteq \mathcal{V} \subseteq \mathcal{P}(A)$ *be collections of nonempty subsets of* A *each of size at most* 3 *which are closed under taking nonempty subsets. Suppose that* f *is a ternary polymorphism of the multisorted structure* $\mathbf{A}_{\mathcal{U}}$, *and that* R *is an* m-*ary relation of* $\mathbf{A}_{\mathcal{V}}$. *Let* $S \subseteq [m]$ *be the set of coordinates of the relation* R *with sorts from* $\mathcal{U}$. *Then for any* $y \in \pi_{[m] \setminus S}(R)$, *the relation*

$$R_y := \{x \mid \exists r \in R \text{ s.t. } \pi_S(r) = x, \pi_{[m] \setminus S}(r) = y\} \subseteq \pi_S(R)$$

is preserved by the polymorphism f.

Proof. We just need to check that for every three tuples $u, v, w \in R_y$, we have $f(u, v, w) \in R_y$. If R is one of the binary inclusion relations, then this follows from the fact that f is conservative on the sorts in $\mathcal{U}$ (which folows from the fact that $\mathcal{U}$ is closed under taking nonempty subsets). Otherwise, R originally came from some relation $\tilde{R} \in \Gamma$. Then there are lifts $\tilde{u}, \tilde{v}, \tilde{w} \in \tilde{R}$ such that projecting to the coordinates of $\tilde{R}$ corresponding to S gives us u, v, w, and such that projecting to the coordinates of $\tilde{R}$ corresponding to $[m] \setminus S$ gives us y in each case. Then there is a relation R' of $\mathbf{A}_{\mathcal{U}}$ given by restricting the ith coordinate of $\tilde{R}$ to the set $\{\tilde{u}_i, \tilde{v}_i, \tilde{w}_i\}$ for all i and projecting to the coordinates corresponding to S, and we have

$$f(u, v, w) \in R' \subseteq R_y$$

since f preserves R'. $\qquad\square$

In order to find ternary weak near-unanimity polymorphisms of the structure $\mathbf{A}_{\mathcal{U}}$, we use the idea of solving an *indicator instance* (we used this idea once already to solve the meta-problem for bounded width templates - this idea appears to have shown up for the first time in [78]).

Proposition 4.8.5. *If* $\mathbf{B} = (\mathcal{U}, \Gamma)$ *is a multisorted relational structure with sorts* $\mathcal{U}$, *then* $\mathbf{B}$ *has a ternary weak near-unanimity polymorphism* f *iff the following instance of* $\mathrm{CSP}(\mathbf{B})$ *has a solution:*

$$\exists f = \{f_U \in U^{U^3}\}_{U \in \mathcal{U}} \text{ s.t. } \bigwedge_{U \in \mathcal{U}} \left(\bigwedge_{a \in U} f_U(a, a, a) \in \{a\} \wedge \bigwedge_{a, b \in U} f_U(a, a, b) = f_U(a, b, a) = f_U(b, a, a) \right)$$

$$\wedge \bigwedge_{R \in \Gamma} \bigwedge_{u, v, w \in R} f(u, v, w) \in R.$$

Now we can describe the treasure hunt algorithm: we iteratively build ternary weak near-unanimity polymorphisms $f_{\mathcal{U}}$ of the multisorted structures $\mathbf{A}_{\mathcal{U}}$ for progressively larger collections $\mathcal{U}$ of subsets of $\mathbb{A}$ of size at most 3. In each step, we add a single new subset V of A to $\mathcal{U}$ to produce a larger collection $\mathcal{V}$.

In order to solve the indicator instance for $\mathbf{A}_{\mathcal{V}}$, we brute force over all possible ternary weak near-unanimity operations on V, and for each one, we check if it extends to a solution to the indicator instance, using the fact that all of the remaining variables have sorts in $\mathcal{U}$. In fact, we don't need to consider *all* weak near-unanimity operations on V - we only need to check the 73 specific operations from the classification of conservative minimal Taylor algebras of size 3 (if V has size 2, we only need to consider 4 possible operations).

Theorem 4.8.6. *If* $\mathbf{A}$ *is a 3-conservative relational structure, then Algorithm 4.8 runs in time polynomial in* $\|\mathbf{A}\|$, *and either correctly determines that* $\mathrm{CSP}(\mathbf{A})$ *is NP-complete or produces a ternary weak near-unanimity polymorphism of* $\mathbf{A}$.

Algorithm 13 Treasure hunt algorithm for solving the meta-problem for a 3-conservative relational structure $\mathbf{A} = (A, \Gamma)$, from [45].

1: Set $n \leftarrow \binom{|A|}{3} + \binom{|A|}{2}$.

2: Pick a sequence $\mathcal{U}_0 \subseteq \mathcal{U}_1 \subseteq \mathcal{U}_2 \subseteq \cdots \subseteq \mathcal{U}_n = \{U \subseteq A \mid |U| \in [3]\}$ such that $\mathcal{U}_0$ is the set of singleton subsets of A, each $\mathcal{U}_{i+1}$ contains exactly one more subset of A then $\mathcal{U}_i$, and each $\mathcal{U}_i$ is closed under taking nonempty subsets.

3: Let $f_{\mathcal{U}_0}$ be the unique ternary polymorphism of $\mathbf{A}_{\mathcal{U}_0}$, given by $f_{\{a\}}(a, a, a) = a$ for all $a \in A$.

4: **for** $i \in [n]$ **do**

5:     Let $\mathbf{X}$ be the indicator instance from Proposition 4.8.5 for the multisorted structure $\mathbf{A}_{\mathcal{U}_i}$.

6:     Let V be the new set in $\mathcal{U}_i \setminus \mathcal{U}_{i-1}$.

7:     **for all** ternary weak near-unanimity operations g_V on V **do**

8:         Let $\mathbf{X}'$ be the instance we get from $\mathbf{X}$ by replacing each variable $f_V(a, b, c)$ of $\mathbf{X}$ with sort V by the constant $g_V(a, b, c)$.

9:         Solve the instance $\mathbf{X}'$ by using Algorithm 4.7 with the Taylor operation $f_{\mathcal{U}_{i-1}}$.

10:         **if** $\mathbf{X}'$ has a solution **then**

11:             Let $f_{\mathcal{U}_i}$ be any solution to $\mathbf{X}'$.

12:     **if** $f_{\mathcal{U}_i}$ hasn't been defined **then**

13:         **return** "NP-complete".

14: Stitch $f_{\mathcal{U}_n}$ into a ternary polymorphism f of $\mathbf{A}$ using Proposition 4.8.2.

15: **return** f.

Proof. Given what we have already proved, the only thing left to check is that the algorithm runs in polynomial time. The only step which looks dangerous is the step where we apply Algorithm 4.7, since the degree of the polynomial in the running time of that algorithm depends on the size of the largest sort which shows up as a variable domain. However, we only ever apply Algorithm 4.7 to multisorted structures where every sort has size at most 3. □

Bibliography

[1] Erhard Aichinger, Peter Mayr, and Ralph McKenzie. On the number of finite algebraic structures. *Journal of the European Mathematical Society*, 016(8):1673–1686, 2014.

[2] D. Angluin and M. Kharitonov. When won't membership queries help? *Journal of Computer and System Sciences*, 50(2):336 – 355, 1995.

[3] Dana Angluin. Queries and concept learning. *Machine Learning*, 2(4):319–342, Apr 1988.

[4] Michael Aschbacher. Near subgroups of finite groups. *J. Group Theory*, 1(2):113–129, 1998.

[5] Albert Atserias and Víctor Dalmau. A combinatorial characterization of resolution width. *Journal of Computer and System Sciences*, 74(3):323 – 334, 2008. Computational Complexity 2003.

[6] Per Austrin, Venkatesan Guruswami, and Johan Håstad. $(2+\varepsilon)$-Sat is NP-hard. *SIAM Journal on Computing*, 46(5):1554–1573, 2017.

[7] Kirby A. Baker and Alden F. Pixley. Polynomial interpolation and the Chinese remainder theorem for algebraic systems. *Math. Z.*, 143(2):165–174, 1975.

[8] L. Barto. The dichotomy for conservative constraint satisfaction problems revisited. In *2011 IEEE 26th Annual Symposium on Logic in Computer Science*, pages 301–310, 2011.

[9] L. Barto, J. Bulín, A. Krokhin, and J. Opršal. Algebraic approach to promise constraint satisfaction. *arXiv e-prints*, November 2018.

[10] L. Barto and M. Kozik. Constraint satisfaction problems of bounded width. In *2009 50th Annual IEEE Symposium on Foundations of Computer Science*, pages 595–603, 2009.

[11] Libor Barto. Finitely related algebras in congruence distributive varieties have near unanimity terms. *Canadian Journal of Mathematics*, 65(1):3–21, 2013.

[12] Libor Barto. The collapse of the bounded width hierarchy. *Journal of Logic and Computation*, 2014.

[13] Libor Barto and Jakub Bulín. Deciding absorption in relational structures. *Algebra universalis*, 78(1):3–18, Sep 2017.

[14] Libor Barto and Ondřej Draganov. The minimal arity of near unanimity polymorphisms. *Mathematica Slovaca*, 69(2):297–310, 2019.

[15] Libor Barto and Alexandr Kazda. Deciding absorption. *International Journal of Algebra and Computation*, 26(05):1033–1060, 2016.

[16] Libor Barto and Marcin Kozik. Cyclic terms for SD∨ varieties revisited. *Algebra universalis*, 64(1):137–142, 2010.

[17] Libor Barto and Marcin Kozik. Absorbing subalgebras, cyclic terms, and the constraint satisfaction problem. *Log. Methods Comput. Sci.*, 8(1):1:07, 27, 2012.

[18] Libor Barto and Marcin Kozik. Robust satisfiability of constraint satisfaction problems. In *Proceedings of the Forty-fourth Annual ACM Symposium on Theory of Computing*, STOC '12, pages 931–940, New York, NY, USA, 2012. ACM.

[19] Libor Barto and Marcin Kozik. Constraint satisfaction problems solvable by local consistency methods. *J. ACM*, 61(1):Art. 3, 19, 2014.

[20] Libor Barto, Marcin Kozik, Miklós Maróti, Ralph McKenzie, and Todd Niven. Congruence modularity implies cyclic terms for finite algebras. *Algebra universalis*, 61(3):365–380, 2009.

[21] Libor Barto, Marcin Kozik, and David Stanovský. Mal'tsev conditions, lack of absorption, and solvability. *Algebra universalis*, 74(1):185–206, Sep 2015.

[22] Libor Barto, Marcin Kozik, and Ross Willard. Near unanimity constraints have bounded pathwidth duality. In *2012 27th Annual IEEE Symposium on Logic in Computer Science*, pages 125–134. IEEE, 2012.

[23] Libor Barto, Jakub Opršal, and Michael Pinsker. The wonderland of reflections. *Israel Journal of Mathematics*, 223(1):363–398, 2018.

[24] Libor Barto and Michael Pinsker. Topology is irrelevant (in the infinite domain dichotomy conjecture for constraint satisfaction problems). *Preprint*, 2018.

[25] Joel Berman, Paweł Idziak, Petar Marković, Ralph McKenzie, Matthew Valeriote, and Ross Willard. Varieties with few subalgebras of powers. *Transactions of the American Mathematical Society*, 362(3):1445–1473, 2010.

[26] Garrett Birkhoff. On the structure of abstract algebras. In *Mathematical proceedings of the Cambridge philosophical society*, volume 31, pages 433–454. Cambridge University Press, 1935.

[27] Garrett Birkhoff. *Lattice theory*, volume 25. American Mathematical Soc., 1940.

[28] Garrett Birkhoff. Subdirect unions in universal algebra. *Bull. Amer. Math. Soc.*, 50:764–768, 1944.

[29] Garrett Birkhoff and Stephen A Kiss. A ternary operation in distributive lattices. *Bulletin of the American Mathematical Society*, 53(8):749–752, 1947.

[30] Anselm Blumer, A. Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis Dimension. *J. ACM*, 36(4):929–965, October 1989.

[31] Manuel Bodirsky. Complexity classification in infinite-domain constraint satisfaction. *arXiv preprint arXiv:1201.0856*, 2012.

[32] Manuel Bodirsky and Bertalan Bodor. Structures with small orbit growth. *arXiv preprint arXiv:1810.05657*, 2018.

[33] Joshua Brakensiek and Venkatesan Guruswami. Promise constraint satisfaction: Algebraic structure and a symmetric boolean dichotomy. *arXiv preprint arXiv:1704.01937*, 2017.

[34] A. A. Bulatov. A graph of a relational structure and constraint satisfaction problems. In *Proceedings of the 19th Annual IEEE Symposium on Logic in Computer Science, 2004.*, pages 448–457, July 2004.

[35] Andrei Bulatov, Hubie Chen, and Víctor Dalmau. Learnability of relatively quantified generalized formulas. In Shoham Ben-David, John Case, and Akira Maruoka, editors, *Algorithmic Learning Theory*, pages 365–379, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.

[36] Andrei Bulatov and Víctor Dalmau. A simple algorithm for mal'tsev constraints. *SIAM Journal on Computing*, 36(1):16–27, 2006.

[37] Andrei Bulatov and Peter Jeavons. Algebraic structures in combinatorial problems. 2001.

[38] Andrei Bulatov, Peter Mayr, and Ágnes Szendrei. The subpower membership problem for finite algebras with cube terms. *arXiv preprint arXiv:1803.08019*, 2018.

[39] Andrei A. Bulatov. Combinatorial problems raised from 2-semilattices. *J. Algebra*, 298(2):321–339, 2006.

[40] Andrei A. Bulatov. Bounded relational width. *manuscript.* http://www.cs.sfu.ca/~abulatov/papers/relwidth.pdf, 2009.

[41] Andrei A. Bulatov. Complexity of conservative constraint satisfaction problems. *ACM Trans. Comput. Logic*, 12(4), July 2011.

[42] Andrei A. Bulatov. Conservative constraint satisfaction re-revisited. *Journal of Computer and System Sciences*, 82(2):347–356, 2016.

[43] Andrei A. Bulatov. Graphs of finite algebras, edges, and connectivity. *CoRR*, abs/1601.07403, 2016.

[44] Andrei A Bulatov. A dichotomy theorem for nonuniform CSPs. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 319–330. IEEE, 2017.

[45] Clément Carbonnel. *Harnessing tractability in constraint satisfaction problems.* PhD thesis, Institut National Polytechnique de Toulouse, 2016.

[46] Catarina Carvalho and Andrei Krokhin. On algebras with many symmetric operations. *International Journal of Algebra and Computation*, 26(05):1019–1031, 2016.

[47] Hubie Chen. The expressive rate of constraints. *Annals of Mathematics and Artificial Intelligence*, 44(4):341–352, Aug 2005.

[48] Hubie Chen, Victor Dalmau, and Berit Grußien. Arc consistency and friends. *Journal of Logic and Computation*, 23(1):87–108, 2013.

[49] Hubie Chen and Matthew Valeriote. Learnability of solutions to conjunctive queries. *Journal of Machine Learning Research*, 20(67):1–28, 2019.

[50] V Dalmau. *Computational complexity of problems over generalized formulas, 2000.* PhD thesis, PhD thesis, Universitat Politécnica de Catalunya.

[51] Victor Dalmau. Generalized majority-minority operations are tractable. In *20th Annual IEEE Symposium on Logic in Computer Science (LICS'05)*, pages 438–447. IEEE, 2005.

[52] Víctor Dalmau. There are no pure relational width 2 constraint satisfaction problems. *Information Processing Letters*, 109(4):213 – 218, 2009.

[53] Víctor Dalmau, Andrei Krokhin, and Rajsekar Manokaran. Towards a characterization of constant-factor approximable finite-valued CSPs. *Journal of Computer and System Sciences*, 97:14 – 27, 2018.

[54] Víctor Dalmau and Justin Pearson. Closure functions and width 1 problems. In *International Conference on Principles and Practice of Constraint Programming*, pages 159–173. Springer, 1999.

[55] Rina Dechter. From local to global consistency. *Artificial intelligence*, 55(1):87–107, 1992.

[56] Richard Dedekind. Über die von drei moduln erzeugte dualgruppe. *Mathematische Annalen*, 53(3):371–403, 1900.

[57] Irit Dinur, Oded Regev, and Clifford Smyth. The hardness of 3-uniform hypergraph coloring. *Combinatorica*, 25(5):519–535, 2005.

[58] Beno Eckmann and Peter J Hilton. Group-like structures in general categories I multiplications and comultiplications. *Mathematische Annalen*, 145(3):227–255, 1962.

[59] Samuel Eilenberg and Marcel P Schützenbérger. *On pseudovarieties*. IRIA. Laboratoire de Recherche en Informatique et Automatique, 1975.

[60] Tomas Feder. Constraint satisfaction on finite groups with near subgroups. In *Electronic Colloquium on Computational Complexity (ECCC), TR05-005*, 2005.

[61] Tomás Feder and Moshe Y Vardi. The computational structure of monotone monadic SNP and constraint satisfaction: A study through Datalog and group theory. *SIAM Journal on Computing*, 28(1):57–104, 1998.

[62] Ralph Freese and Ralph McKenzie. *Commutator theory for congruence modular varieties*, volume 125. CUP Archive, 1987.

[63] Merrick Furst, John Hopcroft, and Eugene Luks. Polynomial-time algorithms for permutation groups. In *21st Annual Symposium on Foundations of Computer Science (sfcs 1980)*, pages 36–41. IEEE, 1980.

[64] David Geiger. Closed systems of functions and predicates. *Pacific journal of mathematics*, 27(1):95–100, 1968.

[65] Oded Goldreich. Valiant's polynomial-size monotone formula for majority, 2011.

[66] G Grätzer and JB Nation. Prime intervals and maximal chains in finite dimensional semi-modular lattices. 2010.

[67] Martin Grohe. The complexity of homomorphism and constraint satisfaction problems seen from the other side. *Journal of the ACM (JACM)*, 54(1):1, 2007.

[68] Heinz Peter Gumm. *Geometrical methods in congruence modular algebras*, volume 286. American Mathematical Soc., 1983.

[69] Venkatesan Guruswami and Yuan Zhou. Tight bounds on the approximability of almost-satisfiable Horn SAT and Exact Hitting Set. In *Proceedings of the twenty-second annual ACM-SIAM symposium on Discrete algorithms*, pages 1574–1589. Society for Industrial and Applied Mathematics, 2011.

[70] Pavol Hell and Jaroslav Nešetřil. On the complexity of H-coloring. *J. Combin. Theory Ser. B*, 48(1):92–110, 1990.

[71] Pavol Hell and Jaroslav Nešetřil. The core of a graph. *Discrete Mathematics*, 109(1):117 – 126, 1992.

[72] Christian Herrmann. On the word problem for the modular lattice with four free generators. *Mathematische Annalen*, 265(4):513–527, 1983.

[73] Graham Higman. Ordering by divisibility in abstract algebras. *Proceedings of the London Mathematical Society*, s3-2(1):326–336, 1952.

[74] David Hobby and Ralph McKenzie. *The structure of finite algebras*, volume 76 of *Contemporary Mathematics*. American Mathematical Society, Providence, RI, 1988.

[75] Johan Håstad. Some optimal inapproximability results. *J. ACM*, 48(4):798–859, July 2001.

[76] Paweł Idziak, Petar Marković, Ralph McKenzie, Matthew Valeriote, and Ross Willard. Tractability and learnability arising from algebras with few subpowers. *SIAM Journal on Computing*, 39(7):3023–3037, 2010.

[77] Peter Jeavons. On the algebraic structure of combinatorial problems. *Theoretical Computer Science*, 200(1-2):185–204, 1998.

[78] Peter Jeavons, David Cohen, and Marc Gyssens. Closure properties of constraints. *J. ACM*, 44(4):527–548, July 1997.

[79] Přemysl Jedlička, Agata Pilitowska, David Stanovský, and Anna Zamojska-Dzienio. Subquandles of affine quandles. *Journal of Algebra*, 510:259 – 288, 2018.

[80] Bjarni Jónsson. Algebras whose congruence lattices are distributive. *Mathematica Scandinavica*, pages 110–121, 1968.

[81] Jelena Jovanović. On terms describing omitting unary and affine types. *Filomat*, 27(1):183–199, 2013.

[82] Jelena Jovanović, Petar Marković, Ralph McKenzie, and Matthew Moore. Optimal strong mal'cev conditions for congruence meet-semidistributivity in locally finite varieties. *Algebra universalis*, pages 1–21, 2016.

[83] Alexandr Kazda, Marcin Kozik, Ralph McKenzie, and Matthew Moore. *Absorption and directed Jónsson terms*, pages 203–220. Springer International Publishing, Cham, 2018.

[84] Keith Kearnes, Petar Marković, and Ralph McKenzie. Optimal strong Mal'cev conditions for omitting type 1 in locally finite varieties. *Algebra Universalis*, 72(1):91–100, 2014.

[85] Keith A Kearnes. A quasi-affine representation. *International Journal of Algebra and Computation*, 5:673–702, 1995.

[86] Keith A Kearnes. Varieties with a difference term. *Journal of Algebra*, 177(3):926–960, 1995.

[87] Keith A. Kearnes. Idempotent simple algebras. In *Logic and algebra (Pontignano, 1994)*, volume 180 of *Lecture Notes in Pure and Appl. Math.*, pages 529–572. Dekker, New York, 1996.

[88] Keith A Kearnes and Ágnes Szendrei. The relationship between two commutators. *International Journal of Algebra and Computation*, 8(04):497–531, 1998.

[89] Keith A Kearnes and Ágnes Szendrei. Clones of algebras with parallelogram terms. *International Journal of Algebra and Computation*, 22(01):1250005, 2012.

[90] E. Kiss and M. Valeriote. Strongly abelian varieties and the hamiltonian property. *Canadian Journal of Mathematics*, 43(2):331–346, 1991.

[91] Vladimir Kolmogorov, Andrei Krokhin, and Michal Rolinek. The complexity of general-valued CSPs. *SIAM Journal on Computing*, 46(3):1087–1110, 2017.

[92] Alexander Kozachinskiy and Vladimir Podolskii. Multiparty Karchmer-Wigderson games and threshold circuits. *arXiv preprint arXiv:2002.07444*, 2020.

[93] Marcin Kozik. A finite set of functions with an EXPTIME-complete composition problem. *Theoretical Computer Science*, 407(1):330 – 341, 2008.

[94] Marcin Kozik. Weaker consistency notions for all the CSPs of bounded width. *CoRR*, abs/1605.00565, 2016.

[95] Marcin Kozik. Solving CSPs using weak local consistency. 2018.

[96] Marcin Kozik, Andrei Krokhin, Matt Valeriote, and Ross Willard. Characterizations of several Maltsev conditions. *Algebra Universalis*, 73(3-4):205–224, 2015.

[97] Marcin Kozik and Joanna Ochremiak. Algebraic properties of valued constraint satisfaction problem. In *International Colloquium on Automata, Languages, and Programming*, pages 846–858. Springer, 2015.

[98] Gabor Kun, Ryan O'Donnell, Suguru Tamaki, Yuichi Yoshida, and Yuan Zhou. Linear programming, width-1 CSPs, and robust satisfaction. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 484–495. ACM, 2012.

[99] Richard E Ladner. On the structure of polynomial time reducibility. *Journal of the ACM (JACM)*, 22(1):155–171, 1975.

[100] Benoit Larose, Matt Valeriote, and László Zádori. Omitting types, bounded width and the ability to count. *Internat. J. Algebra Comput.*, 19(5):647–668, 2009.

[101] Dietlinde Lau. *Function algebras on finite sets: Basic course on many-valued logic and clone theory*. Springer Science & Business Media, 2006.

[102] Paolo Lipparini. Difference terms and commutators.

[103] Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2(4):285–318, Apr 1988.

[104] Petar Marković, Miklós Maróti, and Ralph McKenzie. Finitely related clones and algebras with cube terms. *Order*, 29(2):345–359, 2012.

[105] Miklós Maróti. The existence of a near-unanimity term in a finite algebra is decidable. *The Journal of Symbolic Logic*, 74(3):1001–1014, 2009.

[106] Miklós Maróti. Malcev on top. *Manuscript, available at* `http://www.math.u-szeged.hu/~mmaroti/pdf/200x%20Maltsev%20on%20top.pdf`, 2011.

[107] Miklós Maróti. Tree on top of Malcev. *Manuscript, available at* `http://www.math.u-szeged.hu/~mmaroti/pdf/200x%20Tree%20on%20top%20of%20Maltsev.pdf`, 2011.

[108] Peter Mayr. The subpower membership problem for mal'cev algebras. *International Journal of Algebra and Computation*, 22(07):1250075, 2012.

[109] Ralph McKenzie. Finite forbidden lattices. In Ralph S. Freese and Octavio C. Garcia, editors, *Universal Algebra and Lattice Theory*, pages 176–205, Berlin, Heidelberg, 1983. Springer Berlin Heidelberg.

[110] Ralph McKenzie and John Snow. Congruence modular varieties: commutator theory and its uses. In *Structural theory of automata, semigroups, and universal algebra*, pages 273–329. Springer, 2005.

[111] Aaron Meyerowitz. Maximal intersecting families. *European Journal of Combinatorics*, 16(5):491 – 501, 1995.

[112] Matthew Moore. Finite degree clones are undecidable. *Theoretical Computer Science*, 796:237–271, 2019.

[113] Miroslav Olšák. The weakest nontrivial idempotent equations. *Bulletin of the London Mathematical Society*, 49(6):1028–1047, 2017.

[114] A Ju Ol'šanskiĭ. Varieties of finitely approximable groups. *Mathematics of the USSR-Izvestiya*, 3(4):867, 1969.

[115] Peter Ouwehand. Commutator theory and abelian algebras. *arXiv preprint arXiv:1309.0662*, 2013.

[116] Péter Pál Pálfy. Unary polynomials in algebras, I. *Algebra Universalis*, 18(3):262–273, 1984.

[117] Péter Pál Pálfy and Pavel Pudlák. Congruence lattices of finite algebras and intervals in subgroup lattices of finite groups. *Algebra Universalis*, 11(1):22–27, 1980.

[118] H. Peter Gumm. Algebras in permutable varieties: Geometrical properties of affine algebras. *algebra universalis*, 9(1):8–34, Dec 1979.

[119] Michael Pinsker. *Rosenberg's characterization of maximal clones*. na, 2002.

[120] Alden F Pixley. Distributivity and permutability of congruence relations in equational classes of algebras. *Proceedings of the American Mathematical Society*, 14(1):105–109, 1963.

[121] Reinhard Pöschel. A general galois theory for operations and relations and concrete characterization of related algebraic structures. 1980.

[122] Emil L Post. *The Two-Valued Iterative Systems of Mathematical Logic*. Princeton University Press, 1942.

[123] Robert W. Quackenbush. Quasi-affine algebras. *algebra universalis*, 20(3):318–327, Oct 1985.

[124] Prasad Raghavendra. Optimal algorithms and inapproximability results for every CSP? In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 245–254. ACM, 2008.

[125] Prasad Raghavendra. *Approximating NP-hard problems: efficient algorithms and their limits*. University of Washington, 2009.

[126] Jan Reiterman. The Birkhoff theorem for finite algebras. *Algebra universalis*, 14(1):1–10, 1982.

[127] Martin Roller. Poc sets, median algebras and group actions. *arXiv preprint arXiv:1607.07747*, 2016.

[128] Ivo Rosenberg. *Über die funktionale Vollständigkeit in den mehrwertigen Logiken*. Academia, 1970.

[129] Arto Salomaa. *On essential variables of functions, especially in the algebra of logic*. Suomalainen Tiedeakatemia, 1963.

[130] Thomas J. Schaefer. The complexity of satisfiability problems. In *Conference Record of the Tenth Annual ACM Symposium on Theory of Computing (San Diego, Calif., 1978)*, pages 216–226. ACM, New York, 1978.

[131] Jeff Shriner. Hardness results for the subpower membership problem. *International Journal of Algebra and Computation*, 28(05):719–732, 2018.

[132] Mark H Siggers. A strong mal'cev condition for locally finite varieties omitting the unary type. *Algebra universalis*, 64(1-2):15–20, 2010.

[133] Michał Stronkowski and David Stanovský. Embedding general algebras into modules. *Proceedings of the American Mathematical Society*, 138(8):2687–2699, 2010.

[134] S Świerczkowski. Algebras which are independently generated by every n elements. *Fundamenta Mathematicae*, 49:93–104, 1960.

[135] Walter Taylor. Varieties obeying homotopy laws. *Canadian Journal of Mathematics*, 29(3):498–527, 1977.

[136] L. G. Valiant. A theory of the learnable. In *Proceedings of the Sixteenth Annual ACM Symposium on Theory of Computing*, STOC '84, pages 436–445, New York, NY, USA, 1984. ACM.

[137] L.G Valiant. Short monotone formulae for the majority function. *Journal of Algorithms*, 5(3):363 – 366, 1984.

[138] Douglas Wiedemann. Solving sparse linear equations over finite fields. *IEEE transactions on information theory*, 32(1):54–62, 1986.

[139] Ernst Witt. Über die kommutativität endlicher schiefkörper. In *Abhandlungen aus dem Mathematischen Seminar der Universität Hamburg*, volume 8, pages 413–413. Springer, 1931.

[140] Yu I Yanov and AA Muchnik. On the existence of k-valued closed classes that do not have a basis. In *Soviet Acad. Sci. Dokl*, volume 127, pages 144–146, 1959.

[141] Dmitriy Zhuk. The lattice of all clones of self-dual functions in three-valued logic. *Journal of Multiple-Valued Logic & Soft Computing*, 24, 2015.

[142] Dmitriy Zhuk. A proof of CSP dichotomy conjecture. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 331–342. IEEE, 2017.

[143] Dmitriy Zhuk. Strong subalgebras and the constraint satisfaction problem, 2020.

[144] Dmitriy N Zhuk. The existence of a near-unanimity function is decidable. *Algebra universalis*, 71(1):31–54, 2014.

[145] Dmitriy N Zhuk. Key (critical) relations preserved by a weak near-unanimity function. *Algebra universalis*, 77(2):191–235, 2017.

[146] I. Ágoston, J. Demetrovics, and L. Hannák. On the number of clones containing all constants (a problem of R. McKenzie). In L. Szabó and Á. Szendrei, editors, *Lectures in Universal Algebra*, Colloquia Mathematica Societatis Janos Bolyai, pages 21–25. North-Holland, Amsterdam, 1986.

Appendix 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.0.1. 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.0.2. 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 $a = b \wedge c$, and we may start by considering sublattices of the interval $[\![a, b \vee c]\!]$. The claim is that in this scenario, we need $b \vee c$ to cover both b and c, i.e. the sublattice generated by b and c must have length two. If $b \vee c$ does *not* cover c, say $c < d < b \vee c$ for some d, then we have a problem: the sublattice generated by b, c, d is a copy of the pentagon lattice $\mathcal{N}_5$, which is not ranked. The only hard part of verifying this is checking that $b \wedge d = a$, but this follows from $a \leq b \wedge d \leq b$ and $b \not\leq d$.

Definition A.0.3. 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.0.4. *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.0.5 (Birkhoff [27]). *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.0.6 (Birkhoff [27]). *A ranked lattice of finite length is upper semimodular if and only if its height function satisfies the inequality*

$$h(x) + h(y) \geq h(x \vee y) + h(x \wedge y).$$

Proof. The inequality clearly implies upper semimodularity. Now suppose our lattice is upper semimodular, and pick maximal chains

$$x \wedge y = x_0 < x_1 < \cdots < x_m = x,$$
$$x \wedge y = y_0 < y_1 < \cdots < y_n = y.$$

We claim that for each i, j, $x_i \vee y_j$ is either covered by or equal to $x_{i+1} \vee y_j$ and $x_i \vee y_{j+1}$. We can prove this by induction on i, j: if it's true for i, j, then by upper semimodularity $x_{i+1} \vee y_{j+1}$ will either cover or be equal to both of $x_{i+1} \vee y_j$ and $x_i \vee y_{j+1}$.

Thus, the sequence

$$x = x \vee y_0 \leq x \vee y_1 \leq \cdots \leq x \vee y_n = x \vee y$$

has every adjacent pair either equal or a cover, so

$$h(x \vee y) - h(x) \leq h(y) - h(x \wedge y). \qquad \square$$

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.0.7 (Birkhoff [27]). *A lattice of finite length is modular iff it satisfies the Jordan-Dedekind chain condition and its height function satisfies*

$$h(x) + h(y) = h(x \vee y) + h(x \wedge y).$$

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 $\mathcal{N}_5$ (recall from the discussion around Definition 1.7.7 that a lattice is modular iff it doesn't have $\mathcal{N}_5$ as a sublattice).

Suppose this sublattice is generated by a, b, c, with $b < c$ and $a \wedge b = a \wedge c$, $a \vee b = a \vee c$. Then we have

$$h(a) + h(b) = h(a \vee b) + h(a \wedge b) = h(a \vee c) + h(a \wedge c) = h(a) + h(c),$$

so $h(b) = h(c)$, contradicting $b < c$. $\square$

The next result can be viewed as a strengthening of the fact that a modular lattice is both upper and lower semimodular.

Theorem A.0.8 (Diamond Isomorphism Theorem). *If a, b are elements of a modular lattice, then the maps $\phi : [\![a, a \vee b]\!] \to [\![a \wedge b, b]\!]$ and $\varphi : [\![a \wedge b, b]\!] \to [\![a, a \vee b]\!]$ given by*

$$\phi : x \mapsto x \wedge b \text{ and } \varphi : y \mapsto y \vee a$$

are lattice isomorphisms.

Proof. First we check that ϕ, φ are inverse to each other. By the modular law, for $x \in [\![a, a \vee b]\!]$ we have

$$\varphi(\phi(x)) = (x \wedge b) \vee a = x \wedge (b \vee a) = x,$$

and for $y \in [\![a \wedge b, b]\!]$ we have

$$\phi(\varphi(y)) = (y \vee a) \wedge b = y \vee (a \wedge b) = y.$$

It is clear that ϕ respects meets and that φ respects joins, so from the fact that they are inverse to each other we see that they are both lattice isomorphisms. $\square$

Definition A.0.9. If a, b are elements of a lattice, then we say that the intervals $[\![a, a \vee b]\!]$ and $[\![a \wedge b, b]\!]$ are *perspective* to each other, and we abbreviate this with either the notation

$$[\![a, a \vee b]\!] \searrow [\![a \wedge b, b]\!]$$

or the notation

$$[\![a \wedge b, b]\!] \nearrow [\![a, a \vee b]\!].$$

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.0.10 (Jordan-Hölder for semimodular lattices [66]). *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 $\nearrow$, $\searrow$) 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 $c_i = a_1 \vee b_i$. 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 $\nearrow$, $\searrow$ to $[\![c_{\sigma'(i)-1}, c_{\sigma'(i)}]\!]$. Since $[\![c_{\sigma'(i)-1}, c_{\sigma'(i)}]\!] \searrow [\![b_{\sigma'(i)-1}, b_{\sigma'(i)}]\!]$, and since $[\![a_0, a_1]\!] \nearrow [\![b_k, b_{k+1}]\!]$, we can take σ to be the extension of σ' given by setting $\sigma(1) = k+1$. $\square$

To relate this to the usual Jordan-Hölder Theorem, we have to consider the lattice of *subnormal subgroups* of a group. A subgroup $\mathbb{M} \leq \mathbb{G}$ is called subnormal if there is a finite chain of subgroups connecting it to $\mathbb{G}$, such that each is a normal subgroup of the next.

Proposition A.0.11. *A subgroup $\mathbb{M} \leq \mathbb{G}$ is subnormal iff the sequence of groups $\mathbb{G} = \mathbb{G}_0 \rhd \mathbb{G}_1 \rhd \cdots$ defined by taking $\mathbb{G}_{i+1}$ to be the normal closure of $\mathbb{M}$ inside $\mathbb{G}_i$ eventually reaches $\mathbb{M}$. As a consequence, the intersection of two subnormal subgroups is also subnormal.*

Proposition A.0.12. *If $\mathbb{G}$ is a group of finite composition length, then the collection of subnormal subgroups of $\mathbb{G}$ forms a lower semimodular lattice. If $[\![\mathbb{N}_1, \mathbb{M}_1]\!]$, $[\![\mathbb{N}_2, \mathbb{M}_2]\!]$ are $\searrow, \nearrow$ projective covers in this lattice, then $\mathbb{M}_1/\mathbb{N}_1 \cong \mathbb{M}_2/\mathbb{N}_2$.*

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 $a \wedge b$ with a:

$$(a \wedge b) \vee (c \wedge b) \approx ((a \wedge b) \vee c) \wedge b.$$

Thus modular lattices form a variety of lattices. We finish our review of modular lattices by mentioning a famous result of Dedekind.

Proposition A.0.13 (Dedekind [56])**.** *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 $\mathcal{M}_3$.*
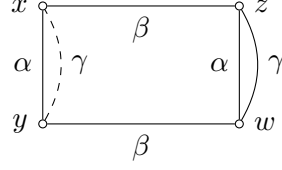
In particular, one can test whether a given 3-variable lattice identity is a consequence of modularity in finite time, by testing whether it holds on $\mathcal{M}_3$.

A corresponding result for 4 generators does not exist: 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 [72].

A.1 The Shifting Lemma and the Day terms

We will follow Freese and McKenzie [62], with some arguments taken from Gumm [68] and some from [115]. 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.1.1 (Shifting Lemma). *If $\mathbb{A}$ is congruence modular, $x, y, z, w \in \mathbb{A}$ and $\alpha, \beta, \gamma \in \mathrm{Con}(\mathbb{A})$ with $\alpha \wedge \beta \leq \gamma$ 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 $(x, y) \in \alpha \wedge (\beta \circ (\alpha \wedge \gamma) \circ \beta) \subseteq \alpha \wedge (\beta \vee (\alpha \wedge \gamma))$. Since $\alpha \wedge \gamma \leq \alpha$, we can apply the modular law to get $\alpha \wedge (\beta \vee (\alpha \wedge \gamma)) = (\alpha \wedge \beta) \vee (\alpha \wedge \gamma)$, and this is contained in γ by the assumption $\alpha \wedge \beta \leq \gamma$, so $(x, y) \in \gamma$. $\square$

Corollary A.1.2 (Day terms). *In any congruence modular variety $\mathcal{V}$, if $\mathcal{F}_\mathcal{V}(x, y, z, w)$ 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 $m_0, ..., m_n \in \mathcal{F}_\mathcal{V}(x, y, z, w)$ such that*

$$
\begin{aligned}
&m_0 = x, \\
&m_i \; (\theta_{x,y} \vee \theta_{z,w}) \wedge (\theta_{x,z} \vee \theta_{y,w}) \; m_{i+1} \text{ for } i \text{ even}, \\
&m_i \; \theta_{z,w} \; m_{i+1} \text{ for } i \text{ odd}, \\
&m_n = y.
\end{aligned}
$$

In other words, the m_i satisfy the following system of identities:

$$
\begin{aligned}
&m_0(x, y, z, w) \approx x, \\
&m_i(x, x, z, z) \approx x \text{ for all } i, \\
&m_i(x, y, x, y) \approx m_{i+1}(x, y, x, y) \text{ for } i \text{ even}, \\
&m_i(x, y, z, z) \approx m_{i+1}(x, y, z, z) \text{ for } i \text{ odd}, \\
&m_n(x, y, z, w) \approx y.
\end{aligned}
$$

Proof. Apply the Shifting Lemma with $\alpha = \theta_{x,y} \vee \theta_{z,w}, \beta = \theta_{x,z} \vee \theta_{y,w}$, and $\gamma = (\alpha \wedge \beta) \vee \theta_{z,w}$ to see that $(x, y) \in (\alpha \wedge \beta) \vee \theta_{z,w} = \bigcup_n ((\alpha \wedge \beta) \circ \theta_{z,w})^{\circ n}$. $\square$

Lemma A.1.3. *Let $\mathbb{A}$ be an algebra with Day terms $m_0, ..., m_n$, $\theta \in \mathrm{Con}(\mathbb{A})$, 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 $m_i(a, b, a, b) \equiv_\theta m_i(a, b, c, d)$.*

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 $m_i(a, b, c, d) \equiv_\theta m_i(a, a, c, c) = a$. 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 $c \equiv_\theta d$. $\square$

The existence of Day terms implies a result slightly stronger than the Shifting Lemma, called the Shifting Principle.

Lemma A.1.4 (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 $\alpha, \gamma \in \mathrm{Con}(\mathbb{A})$ 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 $z \equiv_\gamma w \implies x \equiv_\gamma y$.*
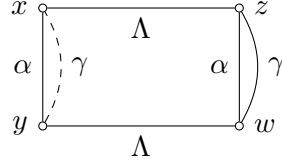
Proof. By Lemma A.1.3, it's enough to show that $m_i(x, y, x, y) \equiv_\gamma m_i(x, y, z, w)$ for each i. Since Λ is preserved by the m_i and is reflexive, we have

$$\begin{bmatrix} m_i(x, y, x, y) \\ m_i(x, y, z, w) \end{bmatrix} = m_i\left(\begin{bmatrix} x \\ x \end{bmatrix}, \begin{bmatrix} y \\ y \end{bmatrix}, \begin{bmatrix} x \\ z \end{bmatrix}, \begin{bmatrix} y \\ w \end{bmatrix}\right) \in \Lambda,$$

while $m_i(x, y, x, y) \equiv_\alpha m_i(x, y, z, w)$ by Lemma A.1.3, so $(m_i(x, y, x, y), m_i(x, y, z, w)) \in \alpha \cap \Lambda \subseteq \gamma$. $\qquad\square$

Lemma A.1.5. *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 $\alpha, \beta, \gamma \in \mathrm{Con}(\mathbb{A})$ with $\alpha \geq \gamma \geq \alpha \wedge \beta$, then to verify congruence modularity we just need to check that $\alpha \wedge (\beta \vee \gamma) \leq \gamma$, as this rules out the existence of a sublattice of $\mathrm{Con}(\mathbb{A})$ isomorphic to the pentagon $\mathcal{N}_5$. Defining reflexive, symmetric relations Λ_i by $\Lambda_i = \beta \circ (\gamma \circ \beta)^{\circ i}$, 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 α, Λ_i, and γ see that if $\alpha \cap \Lambda_i \subseteq \gamma$, then

$$\alpha \cap \Lambda_{2i+1} = \alpha \cap (\Lambda_i \circ \gamma \circ \Lambda_i) = \alpha \cap (\Lambda_i \circ (\alpha \wedge \gamma) \circ \Lambda_i) \subseteq \gamma. \qquad\square$$

Corollary A.1.6. *A variety is congruence modular iff it has Day terms.*

Example A.1.1. If $p(x, y, z)$ is a Mal'cev term, then we can take

$$m_0(x, y, z, w) = x,$$
$$m_1(x, y, z, w) = p(z, w, y),$$
$$m_2(x, y, z, w) = y$$

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 $x \ (\alpha \wedge \beta) \ p(z, w, y) \ \gamma \ y$, so from $\alpha \wedge \beta \leq \gamma$ we get $x \ \gamma \ y$.

Example A.1.2. If $g(x, y, z)$ is a majority term, then we can take

$$m_0(x, y, z, w) = x,$$
$$m_1(x, y, z, w) = g(x, y, z),$$
$$m_2(x, y, z, w) = g(x, y, w),$$
$$m_3(x, y, z, w) = y$$

as a sequence of Day terms. Again, in the Shifting Lemma setup, we have $x \ (\alpha \wedge \beta) \ g(x, y, z) \ \gamma \ g(x, y, w) \ (\alpha \wedge \beta) \ y$, so $x \ \gamma \ y$.

The next corollary gives us a large class of examples of congruence modular varieties, generalizing groups and rings.

Definition A.1.7. An algebra is *congruence regular* if every congruence on $\mathbb{A}$ is uniquely determined by any of its congruence classes.

Corollary A.1.8 (Gumm [68]). *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 Λ. 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 α, Λ, γ.

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. $\square$

To finish this section, 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.1.9 (The Cube Lemma [68]). *Suppose every subalgebra of $\mathbb{A}^2$ satisfies the Shifting Lemma. If $\alpha, \beta, \gamma \in \mathrm{Con}(\mathbb{A})$ with $\gamma \geq \alpha \wedge \beta$, 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 $(d, d') \in \gamma$.*
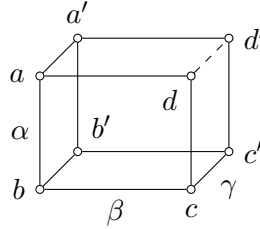
Proof. We apply the Shifting Lemma to the algebra $\beta \leq \mathbb{A}^2$, and the congruences $\gamma \times 1_\mathbb{A}|_\beta, \alpha \times \alpha|_\beta$, and $\gamma \times \gamma|_\beta$. By the Shifting Lemma applied to α, β, γ, we have $(\alpha \times \alpha|_\beta) \wedge (\gamma \times 1_\mathbb{A}|_\beta) \leq \gamma \times \gamma|_\beta$, so the Shifting Lemma applies to $\gamma \times 1_\mathbb{A}|_\beta, \alpha \times \alpha|_\beta, \gamma \times \gamma|_\beta$.

Thus, from $((b, c), (b', c')) \in \gamma \times \gamma|_\beta, ((a, d), (a', d')) \in \gamma \times 1_\mathbb{A}|_\beta$, 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 $(d, d') \in \gamma$. $\square$

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 $\mathbb{M}(\alpha, \beta)$ used to visualize the term condition (Definition 1.9.28) and Lemma A.1.3.

Definition A.2.1. Suppose $\mathbb{A}$ has Day terms $m_0, ..., m_n$. For $\alpha, \beta \in \mathrm{Con}(\mathbb{A})$, we define $X(\alpha, \beta)$ to be the set of pairs $(m_i(a, b, a, b), m_i(a, b, c, d))$ for $\begin{bmatrix} a & c \\ b & d \end{bmatrix} \in \mathbb{M}(\alpha, \beta)$ and $i \leq n$.

Example A.2.1. 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 $\begin{bmatrix} a & c \\ b & d \end{bmatrix} \in \mathbb{M}(\alpha, \beta)$.

Example A.2.2. 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 $\begin{bmatrix} a & c \\ b & d \end{bmatrix} \in \mathbb{M}(\alpha, \beta)$. For $(a, b) \in \alpha \wedge \beta$, we have

$$g\left(\begin{bmatrix} a & a \\ b & b \end{bmatrix}, \begin{bmatrix} a & b \\ a & b \end{bmatrix}, \begin{bmatrix} b & b \\ b & b \end{bmatrix}\right) = \begin{bmatrix} a & b \\ b & b \end{bmatrix} \in \mathbb{M}(\alpha, \beta),$$

so $(a, g(a, b, b)) = (a, b) \in X(\alpha, \beta)$. Thus $X(\alpha, \beta) = \alpha \wedge \beta$ for majority algebras.

Theorem A.2.2 (Commutator via Day terms). *If $\mathbb{A}$ has Day terms $m_0, ..., m_n$ and $\alpha, \beta, \delta \in \mathrm{Con}(\mathbb{A})$, 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 $\begin{bmatrix} a & c \\ b & d \end{bmatrix} \in \mathbb{M}(\alpha, \beta)$, then

$$m_i\left(\begin{bmatrix} a & a \\ a & a \end{bmatrix}, \begin{bmatrix} a & a \\ b & b \end{bmatrix}, \begin{bmatrix} a & c \\ a & c \end{bmatrix}, \begin{bmatrix} a & c \\ b & d \end{bmatrix}\right) = \begin{bmatrix} a & a \\ m_i(a, b, a, b) & m_i(a, b, c, d) \end{bmatrix} \in \mathbb{M}(\alpha, \beta),$$

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.1.3 to see that if $\begin{bmatrix} a & c \\ b & d \end{bmatrix} \in \mathbb{M}(\alpha, \beta)$, $(c, d) \in \delta$, and $X(\alpha, \beta) \subseteq \delta$, then we must have $(a, b) \in \delta$ as well, so $C(\beta, \alpha; \delta)$ holds. $\quad\square$

Now we can finally prove some useful properties of commutators.

Proposition A.2.3. *If $\mathbb{A}$ is contained in a congruence modular variety, then for congruences on $\mathbb{A}$ we have*

(a) *$[\alpha, \beta] = [\beta, \alpha]$,*

(b) *$[\alpha \wedge \gamma, \beta] \leq [\alpha, \beta] \wedge \gamma$,*

(c) $[\bigvee_i \alpha_i, \beta] = \bigvee_i [\alpha_i, \beta],$

(d) if $f : \mathbb{A} \twoheadrightarrow \mathbb{B}$ is surjective, then $f([\alpha, \beta] \vee \ker f) = [f(\alpha \vee \ker f), f(\beta \vee \ker f)],$

(e) if $\mathbb{B} \leq \mathbb{A}$, then $[\alpha|_{\mathbb{B}}, \beta|_{\mathbb{B}}] \leq [\alpha, \beta]|_{\mathbb{B}},$

(f) if $\mathbb{A} = \prod_{i \in I} \mathbb{A}_i$, then $[\bigoplus_i \alpha_i, \bigoplus_i \beta_i] = \bigoplus_i [\alpha_i, \beta_i],$ where $\bigoplus_i \alpha_i$ is the set of pairs (a, b) in $\prod_i \alpha_i$ such that for all but finitely many i we have $a_i = b_i,$

(g) if $\mathbb{A} = \prod_{i \in I} \mathbb{A}_i$, then $[\prod_i \alpha_i, \prod_i \beta_i] \leq \prod_i [\alpha_i, \beta_i].$

Proof. Part (a) follows from Theorem A.2.2, part (b) follows from Proposition 1.9.30(d), and part (e) is Proposition 1.9.30(g). For part (c), Theorem A.2.2 shows that $C(\alpha_j, \beta; \bigvee_i [\alpha_i, \beta])$ holds for each j, so we can use Proposition 1.9.30(e) to see that $[\bigvee_i \alpha_i, \beta] \leq \bigvee_i [\alpha_i, \beta]$, while the other inequality follows from monotonicity of the commutator.

For part (d), note that part (c) implies $[\alpha, \beta] \vee \ker f = [\alpha \vee \ker f, \beta \vee \ker f] \vee \ker f$, so we may assume that $\alpha, \beta \geq \ker f$ without loss of generality. By Theorem A.2.2, $[\alpha, \beta] \vee \ker f$ is the congruence generated by $X(\alpha, \beta) \cup \ker f$, and $[f(\alpha), f(\beta)]$ is the congruence generated by $X(f(\alpha), f(\beta)) = f(X(\alpha, \beta))$, so $f([\alpha, \beta] \vee \ker f) = [f(\alpha), f(\beta)]$.

Parts (f) and (g) follow directly from Theorem A.2.2, but they can also be proved using only parts (a) - (d) (left as an exercise to the reader). $\square$

Proposition A.2.3(d) tells us that we can compute commutators on quotients of $\mathbb{A}$ directly in $\mathbb{A}$. Since $\mathrm{Con}(\mathbb{A}/\pi)$ is naturally isomorphic to the interval $[\![\pi, 1_{\mathbb{A}}]\!]$ in $\mathrm{Con}(\mathbb{A})$, 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* $[\alpha, \beta]_\pi$ is defined to be the least $\delta \geq \pi$ which satisfies the term condition $C(\alpha, \beta; \delta)$.

Corollary A.2.4. *If $\alpha, \beta \geq \pi$ are congruences in a congruence modular variety, then their relative commutator is given by the formula $[\alpha, \beta]_\pi = [\alpha, \beta] \vee \pi$.*

Theorem A.2.5 (Diamond Isomorphism Theorem for relative commutators)**.** *If $\mathbb{A}$ is in a congruence modular variety and $\alpha, \beta \in \mathrm{Con}(\mathbb{A})$, then the maps $\phi : [\![\alpha, \alpha \vee \beta]\!] \to [\![\alpha \wedge \beta, \beta]\!]$ and $\varphi : [\![\alpha \wedge \beta, \beta]\!] \to [\![\alpha, \alpha \vee \beta]\!]$ given by*

$$\phi : x \mapsto x \wedge \beta \quad and \quad \varphi : y \mapsto y \vee \alpha$$

are lattice isomorphisms which respect the relative commutators $[\cdot, \cdot]_\alpha, [\cdot, \cdot]_{\alpha \wedge \beta}$.

Furthermore, in this case we have the equality of relative centralizers $(\alpha : \alpha \vee \beta) = (\alpha \wedge \beta : \beta)$.

Proof. By Theorem A.0.8, ϕ and φ are lattice isomorphisms. If $\gamma, \delta \geq \alpha \wedge \beta$, then from $[\gamma \vee \alpha, \delta \vee \alpha] \leq [\gamma, \delta] \vee \alpha$ we have

$$\varphi([\gamma, \delta]_{\alpha \wedge \beta}) = [\gamma, \delta]_{\alpha \wedge \beta} \vee \alpha = [\gamma, \delta] \vee \alpha = [\gamma \vee \alpha, \delta \vee \alpha] \vee \alpha = [\varphi(\gamma), \varphi(\delta)]_\alpha.$$

If $\gamma, \delta \in [\![\alpha, \alpha \vee \beta]\!]$, then from $\gamma = \phi(\gamma) \vee \alpha, \delta = \phi(\delta) \vee \alpha$ and $[\phi(\gamma), \phi(\delta)] \leq \beta$ we have

$$\phi([\gamma, \delta]_\alpha) = [\gamma, \delta]_\alpha \wedge \beta = [\phi(\gamma) \vee \alpha, \phi(\delta) \vee \alpha]_\alpha \wedge \beta$$
$$= ([\phi(\gamma), \phi(\delta)] \vee \alpha) \wedge \beta = [\phi(\gamma), \phi(\delta)] \vee (\alpha \wedge \beta) = [\phi(\gamma), \phi(\delta)]_{\alpha \wedge \beta}.$$

For the last statement, note that

$$[\delta, \alpha \vee \beta] \leq \alpha \iff [\delta, \alpha] \vee [\delta, \beta] \leq \alpha \iff [\delta, \beta] \leq \alpha \iff [\delta, \beta] \leq \alpha \wedge \beta,$$

so $\delta \leq (\alpha : \alpha \vee \beta) \iff \delta \leq (\alpha \wedge \beta : \beta)$. $\qquad\qquad\square$

For the second approach to the commutator, we will follow Gumm [68] and take the transitive closure of $\mathbb{M}(\alpha, \beta)$ to produce a congruence on α, considered as a subalgebra of $\mathbb{A}^2$.

Definition A.2.6. For $\alpha, \beta \in \mathrm{Con}(\mathbb{A})$, if we consider $\alpha \leq \mathbb{A}^2$ as an algebra of column vectors then we can treat $\mathbb{M}(\alpha, \beta)$ (from Definition 1.9.28) as a binary relation on α, so $\left(\begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} c \\ d \end{bmatrix} \right) \in \mathbb{M}(\alpha, \beta)$ means that $\begin{bmatrix} a & c \\ b & d \end{bmatrix} \in \mathbb{M}(\alpha, \beta)$. We define Δ_α^β to be the transitive closure of this binary relation on α.

Note that Δ_α^β is the least congruence on α which contains the binary relation $\beta \times \beta|_{\Delta_\mathbb{A}}$. When $\mathbb{A}$ has a Mal'cev term, Δ_α^β simplifies to $\mathbb{M}(\alpha, \beta)$.

Proposition A.2.7. *If $\mathbb{A}$ has a Mal'cev polynomial p and $\alpha, \beta \in \mathrm{Con}(\mathbb{A})$, then $\Delta_\alpha^\beta = \mathbb{M}(\alpha, \beta)$.*

Proof. We just need to check that $\mathbb{M}(\alpha, \beta)$ is transitively closed, so supposed that $\begin{bmatrix} a & c \\ b & d \end{bmatrix}, \begin{bmatrix} c & e \\ d & f \end{bmatrix} \in \mathbb{M}(\alpha, \beta)$. Then we have

$$\begin{bmatrix} a & e \\ b & f \end{bmatrix} = p\left(\begin{bmatrix} a & c \\ b & d \end{bmatrix}, \begin{bmatrix} c & c \\ d & d \end{bmatrix}, \begin{bmatrix} c & e \\ d & f \end{bmatrix} \right) \in \mathbb{M}(\alpha, \beta). \qquad\qquad\square$$

Theorem A.2.8. *Suppose that the Shifting Lemma holds for every subalgebra of $\mathbb{A}^2$. Then for $x, y \in \mathbb{A}$ and $\alpha, \beta \in \mathrm{Con}(\mathbb{A})$, the following are equivalent:*

(a) $(x, y) \in [\beta, \alpha]$,

(b) $\begin{bmatrix} x & y \\ y & y \end{bmatrix} \in \Delta_\alpha^\beta$,

(c) there exists $a \in \mathbb{A}$ such that $\begin{bmatrix} x & a \\ y & a \end{bmatrix} \in \Delta_\alpha^\beta$,

(d) there exists $b \in \mathbb{A}$ such that $\begin{bmatrix} x & y \\ b & b \end{bmatrix} \in \Delta_\alpha^\beta$.

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 Δ_α^β. That (c) implies (b) follows from the fact that $(a, y) \in \beta \implies \begin{bmatrix} a & y \\ a & y \end{bmatrix} \in \mathbb{M}(\alpha, \beta)$, and since Δ_α^β is the transitive closure of $\mathbb{M}(\alpha, \beta)$ we have $\begin{bmatrix} x & y \\ y & y \end{bmatrix} \in \Delta_\alpha^\beta \circ \mathbb{M}(\alpha, \beta) = \Delta_\alpha^\beta$.

For (d) $\implies$ (b) we apply the Shifting Lemma to the algebra $\alpha \leq_{sd} \mathbb{A} \times \mathbb{A}$, the congruences $\ker \pi_1, \ker \pi_2, \Delta_\alpha^\beta \in \mathrm{Con}(\alpha)$, 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 $x \equiv_\alpha b \equiv_\alpha y$).

291
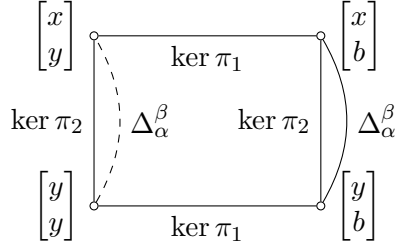
For (a) $\implies$ (b), we will show that the relation Θ defined by $(x, y) \in \Theta \iff \begin{bmatrix} x & y \\ y & y \end{bmatrix} \in \Delta_\alpha^\beta$ is a congruence which satisfies $C(\beta, \alpha; \Theta)$, which will show that $[\beta, \alpha] \leq \Theta$. That Θ is reflexive is obvious, 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 Δ_α^β 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 Θ satisfies $C(\beta, \alpha; \Theta)$, that is, if $\begin{bmatrix} a & c \\ b & d \end{bmatrix} \in \mathbb{M}(\alpha, \beta)$ 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 Δ_α^β is the transitive closure of $\mathbb{M}(\alpha, \beta)$, 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 $(a, b) \in \Theta$. $\quad\square$

Corollary A.2.9. *If every subalgebra of $\mathbb{A}^2$ satisfies the Shifting Lemma, then for $\alpha, \beta_i \in \mathrm{Con}(\mathbb{A})$ we have $[\bigvee_i \beta_i, \alpha] = \bigvee_i [\beta_i, \alpha]$.*

Proof. We have $\Delta_\alpha^{\bigvee_i \beta_i} = \bigvee_i \Delta_\alpha^{\beta_i}$, 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 $\begin{bmatrix} x_i & y_i \\ y_i & y_i \end{bmatrix} \in \bigvee_j (\Delta_\alpha^{\beta_j} \wedge \ker \pi_2)$, this will show that $(x_i, y_i) \in \bigvee_j [\beta_j, \alpha]$. For the inductive step, we apply the Shifting Lemma to $\alpha \leq \mathbb{A}^2$ with the congruences $\ker \pi_2, \Delta_\alpha^{\beta_{j_i}}, \bigvee_j (\Delta_\alpha^{\beta_j} \wedge \ker \pi_2)$.
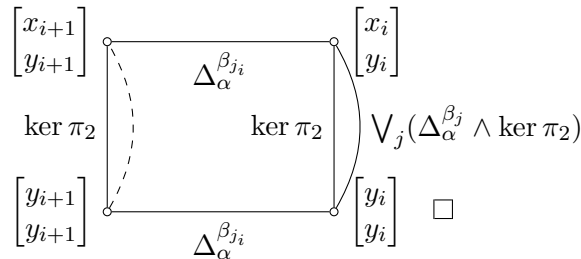
Corollary A.2.10. *If every subalgebra of $\mathbb{A}^2$ satisfies the Shifting Lemma, then for $f : \mathbb{A} \twoheadrightarrow \mathbb{B}$ surjective and $\alpha, \beta \geq \ker f$, we have $f([\beta, \alpha] \vee \ker f) = [f(\beta), f(\alpha)]$.*

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 $\begin{bmatrix} x & y \\ y & y \end{bmatrix} \in \Delta_\alpha^\beta \vee (\ker f \times \ker f|_\alpha)$. Using a similar argument to the previous corollary, we can show that this implies $\begin{bmatrix} x & y \\ y & y \end{bmatrix} \in (\Delta_\alpha^\beta \wedge \ker \pi_2) \vee (\ker f \times \ker f|_\alpha)$ by repeatedly applying the Shifting Lemma on α. Thus we have $(x, y) \in [\beta, \alpha] \vee \ker f$ by Theorem A.2.8. $\square$

To prove the symmetry of the commutator, we will actually prove a stronger statement: Δ_α^β is in fact the *transpose* of Δ_β^α. In particular, if we view Δ_α^β as a binary relation on *row* vectors in β, then Δ_α^β will be transitively closed (which is far from obvious from the definition!).

Theorem A.2.11. *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 Δ_α^β, then $\overline{\Delta}_\alpha^\beta$ is transitively closed as a binary relation on β 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 β, as we will then have $\Delta_\beta^\alpha = \bigcup_n \mathbb{M}(\beta, \alpha)^{\circ n} \subseteq \overline{\Delta}_\alpha^\beta$, and a symmetric argument with α, β swapped will show that $\Delta_\alpha^\beta \subseteq \overline{\Delta}_\beta^\alpha$, so $\Delta_\beta^\alpha \subseteq \overline{\Delta}_\alpha^\beta \subseteq \Delta_\beta^\alpha$.

Suppose that $\begin{bmatrix} a & b \\ c & d \end{bmatrix}, \begin{bmatrix} c & d \\ e & f \end{bmatrix} \in \Delta_\alpha^\beta$. To finish, we just need to show that $\begin{bmatrix} a & b \\ e & f \end{bmatrix} \in \Delta_\alpha^\beta$. This follows from the following application of the Cube Lemma (Lemma A.1.9) applied to the congruences $\ker \pi_1, \ker \pi_2, \Delta_\alpha^\beta$ on α.
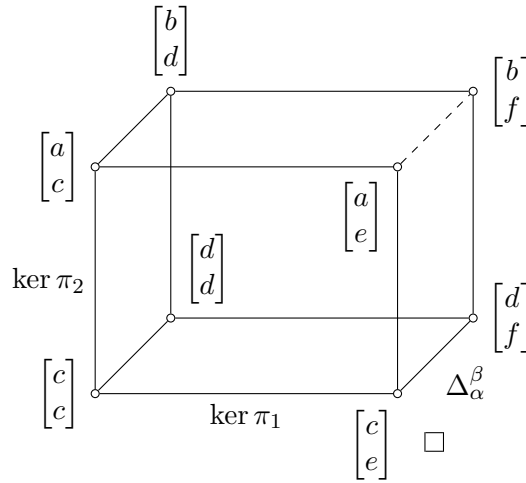
Theorem A.2.12. *In a congruence modular variety, any alternative commutator $[\cdot, \cdot]'$ which satisfies $[\alpha, \beta]' \leq \alpha \wedge \beta$ and $f([\alpha, \beta]' \vee \ker f) = [f(\alpha), f(\beta)]'$ for f surjective and $\alpha, \beta \geq \ker f$ has $[\alpha, \beta]' \leq [\alpha, \beta]$ for all α, β.*

Proof. Consider congruences on $\alpha \leq_{sd} \mathbb{A} \times \mathbb{A}$. We have $[\Delta_\alpha^\beta, \ker \pi_2]' \leq \Delta_\alpha^\beta \wedge \ker \pi_2 \leq \pi_1^{-1}[\beta, \alpha]$ by Theorem A.2.8. Also, $\alpha = \pi_1(\ker \pi_2 \vee \ker \pi_1), \beta = \pi_1(\Delta_\alpha^\beta \vee \ker \pi_1)$, so

$$[\beta, \alpha]' = [\pi_1(\Delta_\alpha^\beta \vee \ker \pi_1), \pi_1(\ker \pi_2 \vee \ker \pi_1)]' = \pi_1([\Delta_\alpha^\beta, \ker \pi_2]' \vee \ker \pi_1) \leq [\beta, \alpha]. \qquad \square$$

A.3 The Gumm difference term

In this section 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.3.1 (Gumm difference term). *For any variety with Day terms $m_0, ..., m_n$, 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 $(x, y) \in \theta$, θ 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 α, β, γ with $\alpha \wedge \beta \leq \gamma$, the implication in the following picture holds.
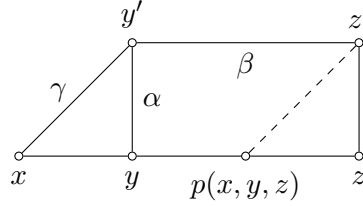
Finally, if a variety has a term p which satisfies (iii), then it is congruence modular.

Proof. Recall the identities satisfied by Day terms:

$$m_0(x, y, z, w) \approx x,$$
$$m_i(x, x, z, z) \approx x \text{ for all } i,$$
$$m_i(x, y, x, y) \approx m_{i+1}(x, y, x, y) \text{ for } i \text{ even,}$$
$$m_i(x, y, z, z) \approx m_{i+1}(x, y, z, z) \text{ for } i \text{ odd,}$$
$$m_n(x, y, z, w) \approx y.$$

We inductively define a sequence of ternary terms $q_i(x, y, z)$ by $q_0(x, y, z) = z$, and

$$q_{i+1}(x, y, z) = \begin{cases} m_{i+1}(q_i(x, y, z), q_i(x, y, z), y, x) & i \text{ odd,} \\ m_{i+1}(q_i(x, y, z), q_i(x, y, z), x, y) & i \text{ even,} \end{cases}$$

and we set $p(x, y, z) = q_n(x, y, z)$.

To see that (i) holds, we just check inductively that $q_i(y, y, x) \approx x$:

$$q_{i+1}(y, y, x) = m_{i+1}(q_i(y, y, x), q_i(y, y, x), y, y) \approx m_{i+1}(x, x, y, y) \approx x.$$

For (ii), we will inductively check that

$$q_i(y, x, x) \; [\theta, \theta] \; \begin{cases} m_i(x, y, x, y) & i \text{ even,} \\ m_i(x, y, x, x) & i \text{ odd.} \end{cases}$$

294

Taking $i = n$, this will give us $p(y, x, x)\ [\theta, \theta]\ m_n(x, y, x, ?) = y$.

The base case is easy: $q_0(y, x, x) = x = m_0(x, y, x, y)$. For the inductive step, we divide into cases based on whether i is even or odd.

If i is even, then the induction hypothesis gives

$$q_{i+1}(y, x, x) = m_{i+1}(q_i(y, x, x), q_i(y, x, x), y, x)\ [\theta, \theta]\ m_{i+1}(m_i(x, y, x, y), m_i(x, y, x, y), y, x).$$

Using the term condition $C(\theta, \theta; [\theta, \theta])$, from

$$m_{i+1}(m_i(x, y, x, y), m_i(x, y, x, y), y, \boxed{y}) = m_i(x, y, x, y) = m_{i+1}(x, y, x, y)$$
$$= m_{i+1}(m_i(x, x, x, x), m_i(y, y, y, y), x, \boxed{y}),$$

we conclude

$$m_{i+1}(m_i(x, y, x, y), m_i(x, y, x, y), y, \boxed{x})\ [\theta, \theta]\ m_{i+1}(m_i(x, x, x, x), m_i(y, y, y, y), x, \boxed{x})$$
$$= m_{i+1}(x, y, x, x),$$

so $q_{i+1}(y, x, x)\ [\theta, \theta]\ m_{i+1}(x, y, x, x)$.

When i is odd, the proof is very similar. Inductively, we have

$$q_{i+1}(y, x, x) = m_{i+1}(q_i(y, x, x), q_i(y, x, x), x, y)\ [\theta, \theta]\ m_{i+1}(m_i(x, y, x, x), m_i(x, y, x, x), x, y).$$

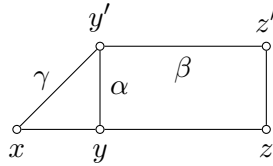Using the term condition $C(\theta, \theta; [\theta, \theta])$, from

$$m_{i+1}(m_i(x, y, x, x), m_i(x, y, x, x), x, \boxed{x}) = m_i(x, y, x, x) = m_{i+1}(x, y, x, x)$$
$$= m_{i+1}(m_i(x, x, x, x), m_i(y, y, y, y), x, \boxed{x}),$$

we conclude

$$m_{i+1}(m_i(x, y, x, x), m_i(x, y, x, x), x, \boxed{y})\ [\theta, \theta]\ m_{i+1}(m_i(x, x, x, x), m_i(y, y, y, y), x, \boxed{y})$$
$$= m_{i+1}(x, y, x, y),$$

so $q_{i+1}(y, x, x)\ [\theta, \theta]\ m_{i+1}(x, y, x, y)$. This conclude the proof of (ii).

Now we show that (i) and (ii) imply (iii). Suppose we have the configuration

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 $(z, z') \in \alpha \wedge (\beta \vee \gamma)$, we have

$$p(z', z, z)\ [\alpha \wedge (\beta \vee \gamma), \alpha \wedge (\beta \vee \gamma)]\ z'.$$

The commutator above is bounded by

$$[\alpha \wedge (\beta \vee \gamma), \alpha \wedge (\beta \vee \gamma)] \leq [\alpha, \beta \vee \gamma] = [\alpha, \beta] \vee [\alpha, \gamma] \leq (\alpha \wedge \beta) \vee (\alpha \wedge \gamma) = \alpha \wedge \gamma.$$

Thus, we have $(p(y', y, z), z') \in \alpha \wedge (\beta \vee (\alpha \wedge \gamma))$, and by the modular law this is $(\alpha \wedge \beta) \vee (\alpha \wedge \gamma) = \alpha \wedge \gamma \leq \gamma$.

Finally, assume that d is a term which satisfies (iii). Taking $x = y = y', z = z', \alpha = \gamma = 0_\mathbb{A}, \beta = 1_\mathbb{A}$, 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 θ. Applying (iii) in the picture
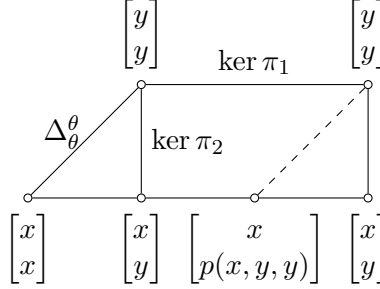
we see that $\begin{bmatrix} x & y \\ p(x,y,y) & y \end{bmatrix} \in \Delta_\theta^\theta$, so by Theorem A.2.8 we have $p(x, y, y) \ [\theta, \theta] \ x$. $\qquad \square$

Corollary A.3.2 (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 $\gamma \in \mathrm{Con}(\mathbb{A})$.*

Proof. A pair of congruences $\alpha, \beta \in \mathrm{Con}(\mathbb{A})$ correspond to a pair of factor congruences iff they satisfy $\alpha \wedge \beta = 0_\mathbb{A}$ 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 $\gamma \geq 0_\mathbb{A} = \alpha \wedge \beta$ we can use property (iii) of a difference term to see that $x \ \alpha \ d(x, y', z') \ \gamma \ z'$, so $(x, z') \in \gamma \circ \alpha \implies (x, z') \in \alpha \circ \gamma$. $\qquad \square$

Corollary A.3.3. *Any abelian algebra which is contained in a congruence modular variety is affine.*

Corollary A.3.4. *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 $\mathrm{Con}(\mathbb{B})$.*

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 $\mathrm{Con}(\mathbb{B})$, since then $\mathbb{B}$ is affine and $\mathbb{A}$ is a quotient of $\mathbb{B}$, so $\mathbb{A}$ is also affine.

Let $\alpha, \beta, \gamma \in \mathrm{Con}(\mathbb{B})$ generate a copy of $\mathcal{M}_3$ which is a $0, 1$-sublattice. Then

$$[1, 1] = [\alpha \vee \beta, \alpha \vee \gamma] = [\alpha, \alpha] \vee [\alpha, \gamma] \vee [\beta, \alpha] \vee [\beta, \gamma] \leq \alpha \vee (\beta \wedge \gamma) = \alpha.$$

Similarly we have $[1, 1] \leq \beta$, so $[1, 1] \leq \alpha \wedge \beta = 0$. $\qquad \square$

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.3.5. For any congruence α, define $[\alpha]^n$ inductively by $[\alpha]^0 = \alpha, [\alpha]^{n+1} = [[\alpha]^n, [\alpha]^n]$.

Proposition A.3.6. *If p is a Gumm difference term, and if we define terms p_n inductively by $p_0 = p$ and*

$$p_{n+1}(x, y, z) = p_n(x, p_n(x, y, y), p_n(x, y, z)),$$

then each p_n is also a Gumm difference term, and for any $(x, y) \in \theta$ we have $p_n(x, y, y) \ [\theta]^{2^n} \ x$.

Proof. Inductively, we have

$$p_{n+1}(y, y, x) = p_n(y, p_n(y, y, y), p_n(y, y, x)) = p_n(y, y, x) = x,$$

and from $(x, p_n(x, y, y)) \in [\theta]^{2^n}$, we have

$$p_{n+1}(x, y, y) = p_n(x, p_n(x, y, y), p_n(x, y, y)) \; [[\theta]^{2^n}]^{2^n} \; x. \qquad \square$$

Corollary A.3.7. *Any solvable algebra in a congruence modular variety is Mal'cev.*

The last result of this section is useful for understanding the center of an algebra in terms of the difference term.

Theorem A.3.8. *Suppose p is a Gumm difference term for a congruence modular variety and $\alpha \geq \beta$. Then Δ_β^α is given by*
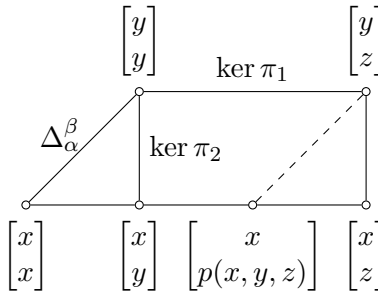
$$\begin{bmatrix} x & w \\ y & z \end{bmatrix} \in \Delta_\beta^\alpha \iff (p(x, y, z) \; [\alpha, \beta] \; w) \wedge (x \; \beta \; y \; \alpha \; z).$$

Proof. If $\begin{bmatrix} x & w \\ y & z \end{bmatrix} \in \Delta_\beta^\alpha$ then clearly $(x \; \beta \; y \; \alpha \; z)$, and from

$$p\left(\begin{bmatrix} x & x \\ x & x \end{bmatrix}, \begin{bmatrix} x & x \\ y & y \end{bmatrix}, \begin{bmatrix} x & w \\ y & z \end{bmatrix} \right) = \begin{bmatrix} x & w \\ p(x, y, y) & p(x, y, z) \end{bmatrix} \in \Delta_\beta^\alpha,$$

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 α to see the implication in the following picture.

Taking transposes, we have $\begin{bmatrix} x & p(x, y, z) \\ y & z \end{bmatrix} \in \Delta_\beta^\alpha$ by Theorem A.2.11. By Theorems A.2.8 and A.2.11, if $p(x, y, z) \; [\alpha, \beta] \; w$ then $\begin{bmatrix} p(x, y, z) & w \\ z & z \end{bmatrix} \in \Delta_\beta^\alpha$, so $\begin{bmatrix} x & w \\ y & z \end{bmatrix} \in \Delta_\beta^\alpha$ by the fact that Δ_β^α is transitively closed. $\qquad \square$

Corollary A.3.9. *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 $\mathbb{A}$.*

Using this, it's possible to show that if $\mathbb{A}$ has center ζ, then we can write $\mathbb{A}$ as an extension of the quotient $\mathbb{A}/\zeta$ by the abelian algebra ζ/Δ_ζ^1 after making a choice of a section $s : \mathbb{A}/\zeta \to \mathbb{A}$, with each n-ary basic operation f inducing a map $t : (\mathbb{A}/\zeta)^n \to \zeta/\Delta_\zeta^1$ so that the action of f on $\mathbb{A}$ can be decomposed as $(x, y) \mapsto (f^{\zeta/\Delta_\zeta^1}(x) + t(y), f^{\mathbb{A}/\zeta}(y))$. If $\mathbb{A}$ is idempotent, then we can simplify this description slightly by noting that in this case, ζ/Δ_ζ^1 is isomorphic to any congruence class of ζ.

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'cev 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 [62] for details.

A.4 (Directed) Jónsson and Gumm terms

First we give Jónsson's [80] characterization of congruence distributive varieties.

Definition A.4.1. A variety $\mathcal{V}$ is congruence distributive if for every $\mathbb{A} \in \mathcal{V}$, $\mathrm{Con}(\mathbb{A})$ is a distributive lattice, that is, if the inequality

$$\alpha \wedge (\beta \vee \gamma) \leq (\alpha \wedge \beta) \vee (\alpha \wedge \gamma)$$

holds for all $\alpha, \beta, \gamma \in \mathrm{Con}(\mathbb{A})$.

The prototypical modular lattice which is *not* distributive is the lattice $\mathcal{M}_3$, as the next proposition shows.

Proposition A.4.2 (Birkhoff [27])**.** *In any modular lattice, if a, b, c do not satisfy the distributive law, and if we define elements d, e, f by*

$$d = (b \wedge c) \vee (a \wedge (b \vee c)) = ((b \wedge c) \vee a) \wedge (b \vee c),$$

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

$$d \wedge e = e \wedge f = f \wedge d = (a \wedge b) \vee (b \wedge c) \vee (c \wedge a)$$

and

$$d \vee e = e \vee f = f \vee d = (a \vee b) \wedge (b \vee c) \wedge (c \vee a).$$

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

$$a \wedge ((a \vee b) \wedge (b \vee c) \wedge (c \vee a)) = a \wedge (b \vee c)$$

and (using the modular law again)

$$a \wedge ((a \wedge b) \vee (b \wedge c) \vee (c \wedge a)) = (a \wedge b) \vee (a \wedge c). \qquad \square$$

Proposition A.4.3. *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 $[\alpha, \beta] = \alpha \wedge \beta$ and $p(x, y, z) = z$ is a Gumm difference term in any congruence distributive variety.*

Proof. If α is an abelian congruence, then $\ker \pi_1, \ker \pi_2, \Delta_\alpha^\alpha \in \mathrm{Con}(\alpha)$ generate a sublattice isomorphic to $\mathcal{M}_3$, with top element $\alpha \times \alpha|_\alpha$. The other direction follows from Proposition 1.9.32, since $\mathcal{M}_3$ does not satisfy the meet-semidistributive law $\mathrm{SD}(\wedge)$. $\qquad \square$

Example A.4.1. 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.4.4 (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 $\mathcal{F}(x, y, z)$ in a congruence distributive variety. From $(x, z) \in \theta_{x,z} \wedge (\theta_{x,y} \vee \theta_{y,z})$ and distributivity, we have

$$x \ (\theta_{x,z} \wedge \theta_{x,y}) \vee (\theta_{x,z} \wedge \theta_{y,z}) \ z.$$

Thus there exist $q_0, ..., q_n \in \mathcal{F}(x, y, z)$ with $q_0 = x$, $q_i \ (\theta_{x,z} \wedge \theta_{x,y}) \ q_{i+1}$ for i even, $q_i \ (\theta_{x,z} \wedge \theta_{y,z}) \ q_{i+1}$ 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 α, β, γ are congruences on any algebra and that $(a, c) \in \alpha \wedge (\beta \vee \gamma)$. We need to show that $(a, c) \in (\alpha \wedge \beta) \vee (\alpha \wedge \gamma)$.

From $(a, c) \in \beta \vee \gamma$, there is a sequence $b_0, ..., b_m$ with $a = b_0$, $b_j \ \beta \cup \gamma \ b_{j+1}$ 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

$$q_i(a, b_j, c) \ (\alpha \wedge \beta) \cup (\alpha \wedge \gamma) \ q_i(a, b_{j+1}, c)$$

for each i, j, so $q_i(a, a, c) \ (\alpha \wedge \beta) \vee (\alpha \wedge \gamma) \ q_i(a, c, c)$ for all i. Stringing these together with the identities relating q_i to q_{i+1}, we see that $a = q_0(a, c, c) \ (\alpha \wedge \beta) \vee (\alpha \wedge \gamma) \ q_n(a, a, c) = c$. $\qquad \square$

Example A.4.2. 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 $(x \wedge y) \vee (y \wedge z) \vee (z \wedge x)$. 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.4.5. *If α, β are any two congruences in a congruence modular variety, then*

$$\alpha \circ \beta \subseteq [\alpha, \alpha] \circ \beta \circ \alpha.$$

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

$$a \; [\alpha, \alpha] \; p(a,b,b) \; \beta \; p(a,b,c) \; \alpha \; p(b,b,c) = c. \qquad \square$$

Corollary A.4.6. *If α, β, γ are congruences in a congruence modular variety, then*

$$(\alpha \circ \beta) \cap \gamma \subseteq ((\alpha \wedge \beta) \vee (\alpha \wedge \gamma)) \circ \beta \circ \alpha.$$

Proof. We have $(\alpha \circ \beta) \cap \gamma = ((\alpha \wedge (\beta \vee \gamma)) \circ \beta) \cap \gamma$, and

$$[\alpha \wedge (\beta \vee \gamma), \alpha \wedge (\beta \vee \gamma)] \leq [\alpha, \beta \vee \gamma] = [\alpha, \beta] \vee [\alpha, \gamma] \leq (\alpha \wedge \beta) \vee (\alpha \wedge \gamma).$$

Thus, by the previous theorem we have

$$(\alpha \circ \beta) \cap \gamma \subseteq (\alpha \wedge (\beta \vee \gamma)) \circ \beta \subseteq ((\alpha \wedge \beta) \vee (\alpha \wedge \gamma)) \circ \beta \circ \alpha. \qquad \square$$

A very similar argument shows that

$$(\alpha \circ \beta) \cap \gamma \subseteq ((\alpha \wedge \gamma) \vee (\beta \wedge \gamma)) \circ \beta \circ \alpha,$$

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.

Theorem A.4.7 (Gumm terms). *A variety is congruence modular iff it has ternary terms $q_0, ..., q_n, p$ satisfying the system of identities*

$$
\begin{aligned}
q_0(x,y,z) &\approx x, \\
q_i(x,y,x) &\approx x \text{ for all } i, \\
q_i(x,y,y) &\approx q_{i+1}(x,y,y) \text{ for } i \text{ odd}, \\
q_i(x,x,y) &\approx q_{i+1}(x,x,y) \text{ for } i \text{ even}, \\
q_n(x,y,y) &\approx p(x,y,y), \\
p(x,x,y) &\approx y.
\end{aligned}
$$

Furthermore, a ternary term p is a Gumm difference term iff there exist terms $q_0, ..., q_n$ satisfying the above system of identities.

Proof. Consider the congruences $\theta_{x,y}, \theta_{y,z}, \theta_{x,z}$ corresponding to identifying pairs of variables on the free algebra $\mathcal{F}(x,y,z)$ in a congruence distributive variety. From $(x,z) \in \theta_{x,z} \wedge (\theta_{x,y} \vee \theta_{y,z})$ and

$$[\theta_{x,z} \wedge (\theta_{x,y} \vee \theta_{y,z}), \theta_{x,z} \wedge (\theta_{x,y} \vee \theta_{y,z})] \leq (\theta_{x,z} \wedge \theta_{x,y}) \vee (\theta_{x,z} \wedge \theta_{y,z}),$$

which is proved as in the previous corollary, we see that for any Gumm difference term p we have

$$x \; (\theta_{x,z} \wedge \theta_{x,y}) \vee (\theta_{x,z} \wedge \theta_{y,z}) \; p(x,z,z).$$

Thus there exist $q_0, ..., q_n \in \mathcal{F}(x,y,z)$ with $q_0 = x$, $q_i \; (\theta_{x,z} \wedge \theta_{x,y}) \; q_{i+1}$ for i even, $q_i \; (\theta_{x,z} \wedge \theta_{y,z}) \; q_{i+1}$ for i odd, and $q_n(x,y,z) = p(x,z,z)$. Therefore $q_0, ..., q_n, p$ 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

$$m_0(x, y, z, w) = x,$$
$$m_{2i-1}(x, y, z, w) = q_i(x, w, y) \text{ for } i \text{ even,}$$
$$m_{2i}(x, y, z, w) = q_i(x, z, y) \text{ for } i \text{ even,}$$
$$m_{2i-1}(x, y, z, w) = q_i(x, z, y) \text{ for } i \text{ odd,}$$
$$m_{2i}(x, y, z, w) = q_i(x, w, y) \text{ for } i \text{ odd,}$$
$$m_{2n+1}(x, y, z, w) = p(z, w, y),$$
$$m_{2n+2}(x, y, z, w) = y,$$

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 $m_0, ..., m_{2n+2}$ 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 $p(x, y, y)$ $[\theta, \theta]$ x. We will show by induction that $q_i(x, y, y)$ $[\theta, \theta]$ x for all i. For the inductive step, we just need to show that for all i, we have $q_i(x, y, y)$ $[\theta, \theta]$ $q_i(x, x, y)$. This follows from the term condition for $[\theta, \theta]$:

$$q_i(x, y, \boxed{x}) = q_i(x, x, \boxed{x}) \implies q_i(x, y, \boxed{y}) \ [\theta, \theta] \ q_i(x, x, \boxed{y}). \qquad \square$$

The need to constantly divide into cases for even vs. odd i can be eliminated by the main result of [83], 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

$$f(x, y, y) \approx g(x, x, y),$$

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,

$$f(x, y, y) \approx h(x, y, y),$$
$$h(x, x, y) \approx g(x, x, y).$$

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 ϕ 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 ϕ is idempotent they will satisfy $f_1(x, x, y) \approx x$ and $f_n(x, y, y) \approx y$. Finally, ϕ will be a near-unanimity term iff each f_i satisfies $f_i(x, y, x) \approx x$.

Theorem A.4.8 (Directed Gumm terms [83]). *A variety is congruence modular iff it has ternary terms $f_1, ..., f_m, p$ with*

$$f_1(x, x, y) \approx x,$$
$$f_i(x, y, x) \approx x \text{ for all } i,$$
$$f_i(x, y, y) \approx f_{i+1}(x, x, y) \text{ for all } i,$$
$$f_m(x, y, y) \approx p(x, y, y),$$
$$p(x, x, y) \approx y,$$

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

$$q_1(x, x, y) \approx x,$$
$$q_i(x, y, x) \approx x \text{ for all } i,$$
$$q_{2i-1}(x, y, y) \approx q_{2i}(x, y, y) \text{ for all } i,$$
$$q_{2i}(x, x, y) \approx q_{2i+1}(x, x, y) \text{ for all } i,$$
$$q_{2k+1}(x, y, y) \approx p_1(x, y, y),$$
$$p_1(x, x, y) \approx y.$$

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 $t_n(x, y, y) = b(x, y)$.

Additionally, let $\rightarrow$ be the relation on $\mathcal{F}$ with $a \rightarrow b$ iff there is a sequence of ternary terms t_i with $t_1(x, x, y) = a(x, y)$, $t_i(x, y, y) = t_{i+1}(x, x, y)$, $t_n(x, y, y) = b(x, y)$, 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

$$q(\rightarrow, \rightsquigarrow, \rightarrow) \subseteq \rightarrow .$$

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

$$\exists b \in \mathcal{F} \quad x \rightarrow b_k^{2^k}(b(x, y), b_k^{2^k - 1}(x, y)).$$

It will then be easy to construct a ternary term p with $p(x, y, y) = b_k^{2^k}(b, b_k^{2^k - 1})$ and $p(x, x, y) = y$, 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 $a \rightsquigarrow b$ and $c(x, y) \rightarrow d(x, y)$, then $c(a, b) \rightarrow d(a, b)$.

Proof of Claim 1: We just have to check this in the case where $c \rightarrow d$ in one step. So suppose that $t(x, x, y) = c(x, y), t(x, y, y) = d(x, y), t(x, y, x) = x$. Then

$$\begin{bmatrix} c(a, b) \\ d(a, b) \end{bmatrix} = t\left(\begin{bmatrix} a \\ a \end{bmatrix}, \begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} b \\ b \end{bmatrix} \right) \in t(\rightarrow, \rightsquigarrow, \rightarrow) \subseteq \rightarrow .$$

Claim 1.5: If $a \rightsquigarrow b$ and $c(x, y) \leftarrow d(x, y)$, then $c(b, a) \rightarrow d(b, a)$.

Proof of Claim 1.5: This follows from Claim 1 and the fact that $c(x, y) \leftarrow d(x, y) \iff c(y, x) \rightarrow d(y, x)$.

Claim 2: If $a \rightarrow b$, then $a^n \rightarrow b^n$ for every n.

Proof of Claim 2: Induct on n. For the inductive step, we have

$$a^{n+1}(x, y) = a(x, a^n(x, y)) \rightarrow b(x, a^n(x, y)) \rightarrow b(x, b^n(x, y)) = b^{n+1}(x, y),$$

where the first $\rightarrow$ follows from $x = a^n(x, x) \rightsquigarrow a^n(x, y)$ and Claim 1, while the second $\rightarrow$ follows from the fact that $\rightarrow$ is preserved by b and the inductive hypothesis.

The sequence of Gumm terms $q_1, ..., q_{2k+1}$ gives us a *k-fence*:

$$x = a_0 \rightarrow b_0 \leftarrow a_1 \rightarrow b_1 \leftarrow a_2 \rightarrow \cdots \leftarrow a_k \rightarrow b_k,$$

where $a_i(x, y) = q_{2i+1}(x, x, y) = q_{2i}(x, x, y)$, $b_i(x, y) = q_{2i+1}(x, y, y) = q_{2i+2}(x, y, y)$. Our strategy will be to use Claims 1 and 1.5 to iteratively reduce the length of the fence.

Claim 3: If $x \rightarrow b \leftarrow a \rightarrow c$ is a 1-fence, then $x \rightarrow b_k(b, c(b, c))$.

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

- $d_i \rightsquigarrow a$, $d_i \rightarrow d_{i+1}$, $d_i \rightsquigarrow e_i$, $d_i \rightarrow b$,

- $e_i \rightarrow d_{i+1}$, $e_i \rightarrow e_{i+1}$, $e_i \rightarrow c(b, c)$.

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

$$x \to b_{0,i} \leftarrow a_{1,i} \to b_{1,i} \leftarrow a_{2,i} \to \cdots \leftarrow a_{k-i,i} \to b_{k-i,i} = b_k^{2^{i+1}-1}.$$

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

$$x \to b_k(b_{0,i}, b_{1,i}(b_{0,i}, b_{1,i})),$$

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 $b_{1,i}^2 \leftarrow a_{2,i}^2 \to b_{2,i}^2 \leftarrow \cdots$, 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

$$x \to b_{0,i+1} \leftarrow a_{1,i+1} \to b_{1,i+1} \leftarrow a_{2,i+1} \to \cdots \leftarrow a_{k-i-1,i+1} \to b_{k-i-1,i+1},$$

and

$$b_{k-i-1,i+1} = b_k(x, b_{k-i,i}^2) = b_k(x, (b_k^{2^{i+1}-1})^2) = b_k^{2^{i+2}-1}.$$

This completes the proof of Claim 4.

By Claim 4 applied with $i = k - 1$, we get a 1-fence

$$x \to b_{0,k-1} \leftarrow a_{1,k-1} \to b_{1,k-1} = b_k^{2^k-1}.$$

Applying Claim 3, we get

$$x \to b_k(b_{0,k-1}, b_k^{2^k-1}(b_{0,k-1}, b_k^{2^k-1})) = b_k^{2^k}(b_{0,k-1}, b_k^{2^k-1}).$$

Letting $b = b_{0,k-1}$, we see that we have succeeded in showing that $x \to b_k^{2^k}(b, b_k^{2^k-1})$. Thus there exist ternary terms f_i with

$$f_1(x, x, y) \approx x,$$
$$f_i(x, y, x) \approx x \text{ for all } i,$$
$$f_i(x, y, y) \approx f_{i+1}(x, x, y) \text{ for all } i,$$
$$f_m(x, y, y) \approx b_k^{2^k}(b(x, y), b_k^{2^k-1}(x, y))).$$

Note that if $b_k(x, y) = y$, then we also have $b_k^{2^k}(b(x, y), b_k^{2^k-1}(x, y))) = y$, so the above becomes a sequence of directed Jónsson terms.

To finish, we just need to construct p with $p(x, y, y) = b_k^{2^k}(b(x, y), b_k^{2^k-1}(x, y)))$ and $p(x, x, y) = y$. Recall that p_1 satisfied $p_1(x, y, y) = b_k(x, y)$ and $p_1(x, x, y) = y$. We construct terms p_i inductively. For $2 \le i + 1 < 2^k$, we set

$$p_{i+1}(x, y, z) = p_1(x, p_i(x, y, y), p_i(x, y, z)),$$

and for $2^k \le i + 1$, 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 $p(x, y, z) = p_{2^{k+1}-1}(x, y, z)$. $\qquad \square$

A.5 Subdirectly irreducible algebras, ultraproducts, and residually small varieties

In this section, we go over the proof of an extension of Jónsson's Lemma [80], 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 $\mathbb{A}$ is isomorphic to $\mathbb{A}$.

Before we discuss ultraproducts, we first review some basic results about subdirect representations of algebras due to Birkhoff [28]. The following result is elementary.

Proposition A.5.1. *If $\mathbb{A} \leq_{sd} \prod_{i \in I} \mathbb{A}_i$ is a subdirect product, then $\bigwedge_{i \in I} \ker \pi_i = 0_{\mathbb{A}}$. In particular, if no π_i is an isomorphism then the congruence $0_{\mathbb{A}}$ can be written as a meet of some family of nontrivial congruences.*

Conversely, if $0_{\mathbb{A}}$ can be written as a meet of congruences $\alpha_i \in \mathrm{Con}(\mathbb{A})$ for $i \in I$, then $\mathbb{A} \leq_{sd} \prod_{i \in I} \mathbb{A}/\alpha_i$.

Definition A.5.2. 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 preceeding proposition can now be rephrased as saying that $\mathbb{A}$ is subdirectly irreducible iff $0_{\mathbb{A}}$ is *meet-irreducible*.

Definition A.5.3. An element α 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 α_i is equal to α. In this case, we define the *cover* of α, written α^*, 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.5.4 (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 $\mathbb{A} \leq_{sd} \prod_{a \neq b} \mathbb{A}/\theta'_{a,b}$. $\square$

Birkhoff's subdirect representation theorem has a purely lattice-theoretic generalization to *algebraic* lattices.

Definition A.5.5. An element α of a complete lattice is called *compact* if for any family α_i such that $\alpha \leq \bigvee_{i \in I} \alpha_i$, there is some finite subset $\{i_1, ..., i_k\} \subseteq I$ 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 $\mathrm{Con}(\mathbb{A})$ 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.5.6. *Let $\mathcal{L}$ be an algebraic lattice. Then every element α of $\mathcal{L}$ can be written as a meet of some family of meet-irreducible elements of $\mathcal{L}$.*

Proof. Let θ be any compact element of $\mathcal{L}$ with $\alpha \not\geq \theta$. By Zorn's Lemma and the compactness of θ, there is some $\theta' \geq \alpha$ which is maximal such that $\theta' \not\geq \theta$, and this θ' is necessarily meet-irreducible with cover $\theta' \vee \theta$. Then $\bigwedge_{\theta \not\leq \alpha} \theta'$ is $\geq \alpha$, and is not $\geq$ any compact element θ with $\alpha \not\geq \theta$, so it must be equal to α. $\qquad\square$

Corollary A.5.7. *If $\alpha < \beta$ in an algebraic lattice, then there is a meet-irreducible γ 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 section.

Definition A.5.8. 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 $\mathcal{U}$.

Proposition A.5.9. *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 $U, I \setminus U \notin \mathcal{U}$, 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 $\mathcal{U}$. $\qquad\square$

Definition A.5.10. If $\mathbb{A}_i$ is a collection of structures which share a common signature σ and are indexed by $i \in I$, and if $\mathcal{U}$ is an ultrafilter on I, then we define the *ultraproduct* $\prod_i \mathbb{A}_i / \mathcal{U}$ to be the quotient of $\prod_i \mathbb{A}_i$ by the congruence defined by

$$a \equiv_{\mathcal{U}} b \iff \{i \mid a_i = b_i\} \in \mathcal{U}.$$

That $\equiv_{\mathcal{U}}$ 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 $\prod_i \mathbb{A}_i / \mathcal{U}$ by

$$R(a^1/\mathcal{U}, ..., a^m/\mathcal{U}) \iff \{i \mid R(a_i^1, ..., a_i^m)\} \in \mathcal{U}.$$

If all the $\mathbb{A}_i$ are isomorphic to $\mathbb{A}$, then we call $\mathbb{A}^I/\mathcal{U}$ an *ultrapower* of $\mathbb{A}$.

Note that in terms of the congruence lattice $\mathrm{Con}(\prod_i \mathbb{A}_i)$, the congruence $\equiv_{\mathcal{U}}$ is equal to the join

$$\bigvee_{U \in \mathcal{U}} \ker \pi_U,$$

where $\pi_U : \prod_{i \in I} \mathbb{A}_i \to \prod_{i \in U} \mathbb{A}_i$ 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.5.11. *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.5.12. *If $|\mathbb{A}_i| \leq n$ for all $i \in I$, then $|\prod_i \mathbb{A}_i / \mathcal{U}| \leq n$ as well. If each $\mathbb{A}_i$ is finite and only finitely many isomorphism classes occur among the $\mathbb{A}_i$, then $\prod_i \mathbb{A}_i / \mathcal{U}$ is isomorphic to some $\mathbb{A}_i$.*

In fact, much more is true about ultraproducts, and the corollary above also follows from the following result from model theory.

Theorem A.5.13 (Łoś's Theorem)**.** *Let $\varphi(x_1, ..., x_n)$ be any first order formula in the signature σ with parameters $x_1, ..., x_n$, then for any $a^1, ..., a^n \in \prod_{i \in I} \mathbb{A}_i$ and any ultrafilter $\mathcal{U}$ on I, we have*

$$\prod_i \mathbb{A}_i/\mathcal{U} \models \varphi(a^1/\mathcal{U}, ..., a^n/\mathcal{U}) \iff \{i \mid \mathbb{A}_i \models \varphi(a_i^1, ..., a_i^n)\} \in \mathcal{U}.$$

Proof. If φ is atomic, then this follows directly from the definitions. Otherwise, φ can be built up from atomic formulas via $\neg, \wedge, \exists$, and we can induct on the structure of φ: for $\neg$, we use the ultrafilter property that exactly one of $U, I \setminus U$ is in $\mathcal{U}$ for each U, for $\wedge$ 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 $\mathcal{U}$. □

Now for the main result. We extend Birkhoff's H, S, P notation by the operation P_u, where $P_u(\{\mathbb{A}_i\})$ is the collection of ultraproducts of the $\mathbb{A}_i$s. Recall that for β a congruence, the centralizer $(0 : \beta)$ of β is defined as the largest α such that $[\alpha, \beta] = 0$, and more generally $(\delta : \beta)$ is defined as the largest α such that $[\alpha, \beta] \leq \delta$.

Theorem A.5.14. *Let $\{\mathbb{A}_i\}$ be a family of algebras, and let $\mathcal{V} = \mathcal{V}(\{\mathbb{A}_i\})$ be the variety they generate. If $\mathcal{V}$ is congruence modular, $\mathbb{B} \in \mathcal{V}$ is subdirectly irreducible, and $\alpha = (0_{\mathbb{B}} : 0_{\mathbb{B}}^*)$ is the centralizer of the monolith $0_{\mathbb{B}}^*$ of $\mathbb{B}$, then $\mathbb{B}/\alpha$ is a homomorphic image of a subalgebra of an ultraproduct of the $\mathbb{A}_i$s, that is, $\mathbb{B}/\alpha \in HSP_u(\{\mathbb{A}_i\})$.*

Proof. (From [62], where a stronger statement is proved.) By Birkhoff's HSP Theorem, we can write $\mathbb{B} = \mathbb{C}/\theta$ for $\mathbb{C} \leq \prod_i \mathbb{A}_i$. Then $\mathbb{B}$ will be subdirectly irreducible iff θ is meet-irreducible in $\mathrm{Con}(\mathbb{C})$, so θ will have a cover θ^*. The preimage φ of α under $\mathbb{C} \to \mathbb{B}$ is the largest congruence on $\mathbb{C}$ such that $[\varphi, \theta^*] \leq \theta$ (i.e. $\varphi = (\theta : \theta^*)$), and we have $\mathbb{B}/\alpha = \mathbb{C}/\varphi$.

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 $\mathcal{F}$ be a maximal filter such that $U \in \mathcal{F}$ implies $\ker \pi_U \leq \theta$, and let $\mathcal{U}$ be any ultrafilter which extends $\mathcal{F}$. Then for any $U \in \mathcal{U}$, we were unable to adjoin its complement to $\mathcal{F}$, so there is some $V \in \mathcal{F}$ such that $\ker \pi_{V \setminus U} \not\leq \theta$. 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 $\bigvee_{U \in \mathcal{U}} \ker \pi_U$ corresponding to $\mathcal{U}$ is also $\leq \varphi$, and we see that $\mathbb{B}/\alpha = \mathbb{C}/\varphi$ is a quotient of $\mathbb{C}/\mathcal{U} \leq \prod_i \mathbb{A}_i/\mathcal{U}$. □

Corollary A.5.15 (Jónsson's Lemma [80])**.** *Let $\{\mathbb{A}_i\}$ be a family of algebras, and let $\mathcal{V} = \mathcal{V}(\{\mathbb{A}_i\})$ be the variety they generate. If $\mathcal{V}$ is congruence distributive and $\mathbb{B} \in \mathcal{V}$ is subdirectly irreducible, then $\mathbb{B} \in HSP_u(\{\mathbb{A}_i\})$. In particular, if $\{\mathbb{A}_i\}$ is a finite set of finite algebras, then $\mathbb{B} \in HS(\{\mathbb{A}_i\})$.*

Corollary A.5.16. *For any two finite subdirectly irreducible algebras $\mathbb{A}, \mathbb{B}$ with the same signature which generate congruence distributive varieties, we have $\mathbb{A} \cong \mathbb{B}$ iff the set of identities that hold in $\mathbb{A}$ is the same as the set of identities that hold in $\mathbb{B}$.*

Example A.5.1. 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.

Proposition A.5.17. *Let* $\mathbb{G}, \mathbb{M}$ *be abelian groups and let* $\mathbb{R}$ *be a finite subgroup of* $\mathrm{Hom}(\mathbb{G}, \mathbb{M})$, *such that there is a nonzero element* $a \in \mathbb{M}$ *so that for all* $x \in \mathbb{G}$ *there is an* $r \in \mathbb{R}$ *with* $rx = a$. *Then* $|\mathbb{G}|$ *is a prime power dividing* $|\mathbb{R}|$.

Proof. First we show that $\mathbb{G}$ is finite, following [62]. Let $r_1, ..., r_k$ be the nontrivial elements of $\mathbb{R}$.

We will show by induction on k that $|\mathbb{G}| \leq (k+1)!$. For the base case, if $k = 0$ then $\mathbb{G}$ can have no nonzero elements, so $|\mathbb{G}| = 1 = (k+1)!$. For the inductive step, note that by the pigeonhole principle there is some r_i such that at least $\frac{|\mathbb{G}|-1}{k}$ elements are mapped to a by r_i, so $|\ker r_i| \geq \frac{|\mathbb{G}|-1}{k}$ (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 $|\mathbb{G}| \leq 1 + k \cdot k! \leq (k+1)!$.

Now that we know that $\mathbb{G}$ is finite, we know that every element of $\mathbb{G}$ has finite order, so some element x has order p for some prime p. Then there is some $r \in \mathbb{R}$ with $rx = a$, so a must also have order p. From this argument, we see that every element of $\mathbb{G}$ must have order a power of p, so $|\mathbb{G}|$ is also a power of p.

We may assume without loss of generality that $\mathbb{M}$ is generated by the image of $\mathbb{G}$ under all elements of $\mathbb{R}$, so in particular that $\mathbb{M}$ is finite. Then there exists an element $\pi \in \hat{\mathbb{M}} = \mathrm{Hom}(\mathbb{M}, \mathbb{Q}/\mathbb{Z})$ such that $\pi(a) \neq 0$.

Define a linear map $\phi : \mathbb{R} \to \hat{\mathbb{G}} = \mathrm{Hom}(\mathbb{G}, \mathbb{Q}/\mathbb{Z})$ by $\phi : r \mapsto \phi_r$, where ϕ_r is the linear map $\phi_r : x \mapsto \pi(rx)$. Then ϕ must be surjective, or else the image will be a proper subgroup of $\hat{\mathbb{G}}$ and so there will be some nonzero $x \in \mathbb{G}$ with $\phi_r(x) = 0$ for all $r \in \mathbb{R}$, which implies $rx \neq a$ for all r. Thus $|\mathbb{G}| = |\hat{\mathbb{G}}|$ divides $|\mathbb{R}|$. $\square$

Corollary A.5.18. *Let* $\mathbb{R}$ *be a finite ring, and let* $\mathbb{M}$ *be a subdirectly irreducible module over* $\mathbb{R}$. *Then* $|\mathbb{M}|$ *is a prime power dividing* $|\mathbb{R}|$.

Proof. If $\mathbb{M}$ is subdirectly irreducible, then it has a least nontrivial submodule $\mathbb{N}$, which is generated by some nonzero element $a \in \mathbb{N}$. Then for each nonzero $x \in \mathbb{M}$ we have $\mathbb{N} \leq \mathbb{R}x$, so there is some $r \in \mathbb{R}$ with $rx = a$. Thus we can apply the previous proposition with $\mathbb{G} = \mathbb{M}$. $\square$

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.5.19. *Suppose that* $\mathbb{B} \in \mathcal{V}$ *is subdirectly irreducible, and* $\mathcal{V}$ *is locally finite and congruence modular. If* $\alpha \in \mathrm{Con}(\mathbb{B})$ *is abelian and* $|\mathbb{B}/\alpha| = k$, *then every congruence class of* α *has size a prime power bounded by* $|\mathcal{F}_\mathcal{V}(k+1)|$.

Proof. (Adapted from [110].) Assume α is nontrivial, so $0^*_\mathbb{B} \leq \alpha$. Let p be a Gumm difference term. By Corollary A.3.9, the restriction of the graph of p to the blocks of α is preserved by every operation of $\mathbb{B}$. Choose elements $0 \neq a$ with $(0, a) \in 0^*_\mathbb{B}$, and note that $0, a$ are in the same congruence block of α.

Pick constants $c_0, ..., c_{k-1}$ with $c_0 = 0$ such that each congruence class of α contains some c_i. We will treat each congruence class c_i/α of α as an abelian group with zero element c_i, addition given by $x +_i y = p(x, c_i, y)$, and subtraction given by $x -_i y = p(x, y, c_i)$.

Suppose that $x \neq y$ with $(x, y) \in \alpha$. Then since $0^*_\mathbb{B}$ 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/α. One explicit way to construct such an f is given by

$$f(z) = p(p(\cdots p(p(f_0(z), f_1(x), f_1(z)), f_2(x), f_2(z)), \cdots), f_m(x), f_m(z)).$$

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 α, if $x, y \in c_i/\alpha$ then we have $f(x -_i y) -_0 f(c_i) = a$, and the unary polynomial $z \mapsto f(z) -_0 f(c_i)$ 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 $\mathbb{R}_{i,0}$ of linear maps in $\text{Hom}(c_i/\alpha, c_0/\alpha)$ which can be defined by unary polynomials f. Suppose that $f(z) = t(z, b_1, ..., b_m)$ for some term t and constants $b_1, ..., b_m \in \mathbb{B}$, 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

$$f'(z) = t(z, c_{j_1}, ..., c_{j_m}) -_0 t(c_i, c_{j_1}, ..., c_{j_m}).$$

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 α, we have $f'(z) = f(z)$ for $z \in c_i/\alpha$ (alternatively, we could prove this by the term condition for $[\alpha, \alpha] = 0_\mathbb{B}$). 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 $f'(z) = t'(z, c_0, ..., c_{k-1})$ for some $k + 1$-ary term t', so

$$|\mathbb{R}_{i,0}| \leq |\mathcal{F}_\mathcal{V}(k + 1)|.$$

Applying the previous proposition, we see that $|c_i/\alpha|$ is a prime power dividing $|\mathbb{R}_{i,0}|$. $\square$

Corollary A.5.20. *If $|\mathbb{A}| = m$ is finite and $\mathcal{V}(\mathbb{A})$ is congruence modular, and if $\mathbb{B} \in \mathcal{V}(\mathbb{A})$ is subdirectly irreducible with $(0_\mathbb{B} : 0^*_\mathbb{B})$ abelian, then $|\mathbb{B}| \leq m \cdot m^{m^{m+1}}$.*

Definition A.5.21. A variety $\mathcal{V}$ is called *residually small* if there is a cardinal κ such that every subdirectly irreducible algebra $\mathbb{B} \in \mathcal{V}$ has $|\mathbb{B}| < \kappa$, and *residually finite* if every subdirectly irreducible algebra in $\mathcal{V}$ is finite. An algebra $\mathbb{A}$ is called residually small if the variety $\mathcal{V}(\mathbb{A})$ generated by $\mathbb{A}$ 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.5.22. *If $\mathbb{B}$ is subdirectly irreducible, then $\mathbb{B}$ is a subalgebra of an ultraproduct of a family of finitely generated subdirectly irreducible algebras in $HS(\mathbb{B})$.*

Proof. (From [62].) Let the monolith $0_{\mathbb{B}}^*$ of $\mathbb{B}$ be generated (as a congruence) by the pair (a, b). Let I be the family of finitely generated subalgebras $\mathbb{S} \leq \mathbb{B}$ with $a, b \in \mathbb{S}$, and for each $\mathbb{S} \in I$, pick a congruence $\alpha_{\mathbb{S}}$ on $\mathbb{S}$ which is maximal among all congruences which do not contain (a, b). Then each $\mathbb{S}/\alpha_{\mathbb{S}}$ is subdirectly irreducible, since every congruence which properly contains $\alpha_{\mathbb{S}}$ contains (a, b).

Let $\mathcal{U}$ be an ultrafilter on I such that the set $U_S = \{\mathbb{S} \mid S \subseteq \mathbb{S}\}$ is in $\mathcal{U}$ for every finite $S \subseteq \mathbb{B}$. Such an ultrafilter exists since for any S_1, S_2 we have $U_{S_1} \cap U_{S_2} = U_{S_1 \cup S_2}$, and for S finite U_S is nonempty since it contains $\mathrm{Sg}_{\mathbb{B}}(S \cup \{a, b\})$.

Define a map $\varphi : \mathbb{B} \to (\prod_{\mathbb{S} \in I} \mathbb{S}/\alpha_{\mathbb{S}})/\mathcal{U}$ as the ultraproduct of the family of maps $\varphi_{\mathbb{S}}$ given by $\varphi_{\mathbb{S}}(x) = x/\alpha_{\mathbb{S}}$ for $x \in \mathbb{S}$ and $\varphi_{\mathbb{S}}(x) = a/\alpha_{\mathbb{S}}$ for $x \notin \mathbb{S}$. Then for $x_1, ..., x_k \in \mathbb{B}$ and t a k-ary term of $\mathbb{B}$, we have

$$\{\mathbb{S} \mid \varphi_{\mathbb{S}}(t(x_1, ..., x_k)) = t(\varphi_{\mathbb{S}}(x_1), ..., \varphi_{\mathbb{S}}(x_k))\} \in \mathcal{U},$$

since it contains $U_{\{x_1, ..., x_k\}}$. Thus φ is a homomorphism. To see that it is injective, just note that $\varphi_{\mathbb{S}}(a) \neq \varphi_{\mathbb{S}}(b)$ for all $\mathbb{S} \in I$. $\square$

Corollary A.5.23. *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 $\mathbb{A}$ satisfies a commutator identity *hereditarily* if every congruence lattice of every subalgebra of $\mathbb{A}$ satisfies the identity.

Proposition A.5.24. *The commutator identity $[\alpha \wedge \beta, \beta] = \alpha \wedge [\beta, \beta]$ 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 $\alpha \wedge [\beta, \beta] \leq [\beta, \beta]$ to see that

$$\alpha \wedge [\beta, \beta] = [\alpha \wedge [\beta, \beta], \beta] \leq [\alpha \wedge \beta, \beta] \leq \alpha \wedge [\beta, \beta]. \qquad \square$$

Proposition A.5.25. *If $\mathbb{A}$ is in a congruence modular variety and satisfies the commutator identity $[\alpha \wedge \beta, \beta] = \alpha \wedge [\beta, \beta]$ hereditarily, then so does every quotient $\mathbb{B}$ of $\mathbb{A}$.*

Proof. Suppose $\mathbb{B} = \mathbb{A}/\gamma$ and $\alpha, \beta \in \mathrm{Con}(\mathbb{A})$ 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 $\alpha \leq [\beta, \beta] \vee \gamma$ then

$$\alpha = \alpha \wedge ([\beta, \beta] \vee \gamma) = (\alpha \wedge [\beta, \beta]) \vee \gamma = [\alpha, \beta] \vee \gamma = [\alpha, \beta]_\gamma. \qquad \square$$

Proposition A.5.26. *If $\mathbb{A}_1, \mathbb{A}_2$ are in a congruence modular variety and satisfy the commutator identity $[\alpha \wedge \beta, \beta] = \alpha \wedge [\beta, \beta]$ 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 $\mathbb{A}_i$ with $\pi_i(\mathbb{B})$. Suppose $\alpha, \beta \in \mathrm{Con}(\mathbb{B})$ with $\alpha \leq [\beta, \beta]$, we will show that $[\alpha, \beta] = \alpha$. We have

$$\alpha \vee \ker \pi_1 \leq [\beta \vee \ker \pi_1, \beta \vee \ker \pi_1]_{\ker \pi_1},$$

so from the assumption on $\mathbb{A}_1$ we get

$$\alpha \vee \ker \pi_1 = [\alpha \vee \ker \pi_1, \beta \vee \ker \pi_1]_{\ker \pi_1} = [\alpha, \beta] \vee \ker \pi_1.$$

Thus by the modular law and $[\alpha, \beta] \leq \alpha$, we have

$$\alpha = \alpha \wedge (\ker \pi_1 \vee [\alpha, \beta]) = (\alpha \wedge \ker \pi_1) \vee [\alpha, \beta].$$

Similarly, we have $\alpha = (\alpha \wedge \ker \pi_2) \vee [\alpha, \beta]$. Since $\alpha \wedge \ker \pi_2 \leq \alpha \leq [\beta, \beta]$, we may apply the same reasoning to $\alpha \wedge \ker \pi_2$ to see that

$$\alpha \wedge \ker \pi_2 = (\alpha \wedge \ker \pi_2 \wedge \ker \pi_1) \vee [\alpha \wedge \ker \pi_2, \beta],$$

so $\alpha \wedge \ker \pi_2 \leq [\alpha, \beta]$, so

$$\alpha = (\alpha \wedge \ker \pi_2) \vee [\alpha, \beta] = [\alpha, \beta]. \qquad \square$$

Theorem A.5.27. *If $|\mathbb{A}| = m$ is finite and $\mathcal{V}(\mathbb{A})$ is congruence modular, and if $\mathbb{A}$ satisfies the commutator identity $[\alpha \wedge \beta, \beta] = \alpha \wedge [\beta, \beta]$ hereditarily, then every subdirectly irreducible algebra $\mathbb{B} \in \mathcal{V}(\mathbb{A})$ has $|\mathbb{B}| \leq m \cdot m^{m^{m+1}}$.*

Proof. By Corollary A.5.23, we just need to check the bound in the case where $\mathbb{B}$ is finite. In this case, we have $\mathbb{B} \in HSP_{fin}(\mathbb{A})$, so $\mathbb{B}$ satisfies the commutator identity $[\alpha \wedge \beta, \beta] = \alpha \wedge [\beta, \beta]$ by the previous propositions. Let $0_{\mathbb{B}}^*$ be the monolith of $\mathbb{B}$, and let $\alpha = (0_{\mathbb{B}} : 0_{\mathbb{B}}^*)$ be its centralizer.

We claim that α is abelian. To see this, note that from $[\alpha, 0_{\mathbb{B}}^*] = 0_{\mathbb{B}}$ we have

$$0_{\mathbb{B}} = [0_{\mathbb{B}}^* \wedge \alpha, \alpha] = 0_{\mathbb{B}}^* \wedge [\alpha, \alpha],$$

so $[\alpha, \alpha] = 0_{\mathbb{B}}$. Now we can apply Corollary A.5.20 to see that $|\mathbb{B}| \leq m \cdot m^{m^{m+1}}$. $\qquad \square$

Example A.5.2. 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 $\mathbb{G} \in \mathcal{V}(S_3)$ has an abelian normal subgroup $\mathbb{N}$ with $\mathbb{G}/\mathbb{N} \in HS(S_3)$, with $|\mathbb{N}|$ a prime power bounded by $|\mathcal{F}_{\mathcal{V}(S_3)}(|\mathbb{G}/\mathbb{N}| + 1)| \leq 6^{6^7}$. Since $\mathbb{N} \in \mathcal{V}(S_3)$ and every element of S_3 has order dividing 6, $\mathbb{N}$ has exponent 2 or 3. From here it is not too hard to check that the only 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.5.28. *If $\mathbb{A}$ 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 $\mathbb{B} \in HS(\mathbb{A})$ such that the centralizer of the monolith of $\mathbb{B}$ is not abelian.*

Proof. Suppose that $\mathbb{A}$ fails to satisfy the commutator identity. In this case there must be $\alpha, \beta \in \mathrm{Con}(\mathbb{A})$ with $\alpha \leq [\beta, \beta]$ and $[\alpha, \beta] < \alpha$. Let θ be a meet-irreducible congruence such that $\theta \geq [\alpha, \beta]$ but $\theta \not\geq \alpha$, and let θ^* be its cover. Then

$$\theta^* \leq \alpha \vee \theta \leq [\beta, \beta] \vee \theta \leq [\beta \vee \theta, \beta \vee \theta]_\theta$$

and

$$[\theta^*, \beta \vee \theta]_\theta \leq [\alpha \vee \theta, \beta \vee \theta]_\theta = [\alpha, \beta] \vee \theta = \theta,$$

so if we take $\mathbb{B}$ to be $\mathbb{A}/\theta$, then the monolith of $\mathbb{B}$ is θ^*/θ, and $\beta \vee \theta/\theta$ is contained in the centralizer of the monolith of $\mathbb{B}$ but is not abelian. $\qquad \square$

Theorem A.5.29. *If $\mathbb{A}$ is contained in a congruence modular variety but does not satisfy the commutator identity $[\alpha \wedge \beta, \beta] = \alpha \wedge [\beta, \beta]$, then $\mathcal{V}(\mathbb{A})$ is not residually small. In fact, for every cardinal κ, $\mathcal{V}(\mathbb{A})$ contains a subdirectly irreducible algebra whose congruence lattice has size at least κ.*

Proof. (From [62].) By the proposition, we can reduce to the case where $\mathbb{A}$ is subdirectly irreducible and the centralizer β of the monolith 0^* is not abelian.

Consider β as a subalgebra of $\mathbb{A}^2$, and $\Delta_\beta^{0^*}$ as a congruence on β. From $[\beta, 0_{\mathbb{A}}^*] = 0_{\mathbb{A}}$ 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 $\Delta_\beta^{0^*} \vee \ker \pi_i = \pi_i^{-1}(0^*)$. Set $0_i^* = \pi_i^{-1}(0^*)$ and $\theta = 0_1^* \wedge 0_2^*, \theta_i = 0_i^* \wedge \ker \pi_{\{1,2\}\setminus\{i\}}$, then (after several applications of the modular law - don't worry about the details just yet) we have the following sublattice in $\mathrm{Con}(\beta)$.
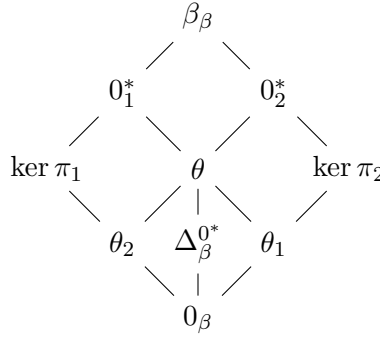
In the picture, we see that $\Delta_\beta^{0^*}$ appears to be meet-irreducible in $\mathrm{Con}(\beta)$, and the interval $[\![\Delta_\beta^{0^*}, \beta_\beta]\!]$ contains the incomparable elements $0_1^*, 0_2^*$. If $\Delta_\beta^{0^*}$ isn't meet-irreducible, we can still try to find a meet-irreducible congruence λ on β which is above $\Delta_\beta^{0^*}$ but not above θ, and then β/λ 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 β_β will come from the assumption that β 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 κ be any cardinal, considered as the set of all ordinals below κ. Define $\mathbb{B} \leq \mathbb{A}^\kappa$ by

$$\mathbb{B} = \{a \in \mathbb{A}^\kappa \mid a_i \equiv_\beta a_j \ \forall i, j \in \kappa\}.$$

Then $\mathbb{B}$ has a natural map to $\mathbb{A}/\beta$, and we call the kernel of this map $\beta_{\mathbb{B}}$. Inside $\mathrm{Con}(\mathbb{B})$, 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 λ 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 $\ker \pi_i \vee \lambda/\lambda \in \mathrm{Con}(\mathbb{B}/\lambda)$ are pairwise distinct. The congruence λ will be constructed by first constructing congruences Δ, θ 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 β. We define Δ_i by

$$(a, b) \in \Delta_i \iff \begin{bmatrix} a_0 & b_0 \\ a_i & b_i \end{bmatrix} \in \Delta_\beta^{0^*} \wedge (a_j = b_j \ \forall j \neq 0, i),$$

and define Δ by

$$\Delta = \bigvee_{0 < i < \kappa} \Delta_i.$$

312

We also define congruences θ_i by

$$(a, b) \in \theta_i \iff (a_i, b_i) \in 0_{\mathbb{A}}^* \wedge (a_j = b_j \; \forall j \neq i),$$

and define θ 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 $\mathrm{Con}(\beta)$. 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:

$$\begin{bmatrix} a_0 \\ a_i \end{bmatrix} \; \ker \pi_{i'} \; \begin{bmatrix} a_0 \\ a_0 \end{bmatrix} \; \Delta_i \; \begin{bmatrix} b_0 \\ b_0 \end{bmatrix} \; \ker \pi_{i'} \; \begin{bmatrix} b_0 \\ b_i \end{bmatrix}.$$

Then by the modular law, if we let $0_i^* = \pi_i^{-1}(0_{\mathbb{A}}^*)$ and note that $\Delta_i \leq 0_i^*$, we get

$$\theta_0 = \theta_0 \wedge 0_i^* \leq (\Delta_i \vee \ker \pi_{i'}) \wedge 0_i^* = \Delta_i \vee (\ker \pi_{i'} \wedge 0_i^*) = \Delta_i \vee \theta_i.$$

Similarly, we get $\theta_i \leq \theta_0 \vee \Delta_i$ for all i.

Next, for each i we have $\Delta \vee \theta_i = \theta$: for each $j \in \kappa$, we have

$$\Delta \vee \theta_i = \Delta \vee \Delta_i \vee \theta_i \geq \Delta \vee \theta_0 \geq \Delta_j \vee \theta_0 \geq \theta_j,$$

so $\Delta \vee \theta_i \geq \bigvee_{j \in \kappa} \theta_j = \theta$, while the other containment follows from $\Delta_i \leq \theta_0 \vee \theta_i$ for all i.

We now check that $\Delta \neq \theta$. It's enough to check that $\theta_0 \not\leq \Delta$, since $\Delta \vee \theta_0 = \theta$. Note first that θ_0 is compact, since $0_{\mathbb{A}}^*$ is compact. Thus we just need to check that $\theta_0 \not\leq \bigvee_{j \leq n} \Delta_{i_j}$ for all $i_1, ..., i_n$. In fact, we can assume that $i_1, ..., i_n$ are $1, ..., n$ by a symmetry argument.

We will show by induction on n that $\theta_0 \wedge (\Delta_1 \vee \cdots \vee \Delta_n) = 0_{\mathbb{B}}$ for all n. The base case follows from the fact that $[\beta, 0_{\mathbb{A}}^*] = 0_{\mathbb{A}} \implies \ker \pi_2 \wedge \Delta_{\beta}^{0^*} = 0_{\beta}$ in $\mathrm{Con}(\beta)$, which in turn implies $\theta_0 \wedge \Delta_1 = 0_{\mathbb{B}}$. For the inductive step, we argue as follows:

$$\begin{aligned}
\theta_0 \wedge (\Delta_1 \vee \cdots \vee \Delta_n) &= \theta_0 \wedge (\theta_0 \vee \theta_n) \wedge (\Delta_1 \vee \cdots \vee \Delta_n) \\
&= \theta_0 \wedge (((\theta_0 \vee \theta_n) \wedge (\Delta_1 \vee \cdots \vee \Delta_{n-1})) \vee \Delta_n) \\
&= \theta_0 \wedge ((\theta_0 \wedge (\Delta_1 \vee \cdots \vee \Delta_{n-1})) \vee \Delta_n) \\
&= \theta_0 \wedge \Delta_n = 0_{\mathbb{B}},
\end{aligned}$$

where the second equality used the modular law and the fact that $\Delta_n \leq \theta_0 \vee \theta_n$, the third equality used the fact that θ_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 $\Delta < \theta$. We can now apply Corollary A.5.7 to see that there is some meet-irreducible congruence λ with $\lambda \geq \Delta$ but $\lambda \not\geq \theta$. To finish, we just need to check that $\lambda \vee \ker \pi_i \not\geq \beta_{\mathbb{B}}$. To see this, note that $\lambda \not\geq \theta_i$, since otherwise we would have $\lambda \geq \Delta \vee \theta_i = \theta$, a contradiction. Since θ_i is the minimal nonzero element of the interval $[\![0_{\mathbb{B}}, \ker \pi_{i'}]\!]$, this means that $\lambda \wedge \ker \pi_{i'} = 0_{\mathbb{B}}$. Thus if (for contradiction) $\lambda \vee \ker \pi_i \geq \beta_{\mathbb{B}}$, then we would have

$$[\beta_{\mathbb{B}}, \beta_{\mathbb{B}}] \leq [\ker \pi_{i'} \vee \ker \pi_i, \lambda \vee \ker \pi_i] \leq (\lambda \wedge \ker \pi_{i'}) \vee \ker \pi_i = \ker \pi_i,$$

and applying π_i we would get $[\beta, \beta] = 0_{\mathbb{A}}$, a contradiction to the assumption that β was not abelian.

Putting it all together, we have a meet-irreducible congruence λ such that $\lambda \vee \ker \pi_i \not\geq \beta_{\mathbb{B}}$ for each i, but $\ker \pi_i \vee \ker \pi_j \geq \beta_{\mathbb{B}}$ for all $i \neq j$. Thus $\mathbb{B}/\lambda$ is subdirectly irreducible, and the congruences $\ker \pi_i \vee \lambda / \lambda$ are mutually distinct elements of $\mathrm{Con}(\mathbb{B}/\lambda)$. $\qquad\square$

Corollary A.5.30. *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 $[\alpha \wedge \beta, \beta] = \alpha \wedge [\beta, \beta]$,*

- *$\mathcal{V}$ is generated by a finite algebra $\mathbb{A}$ such that for every subdirectly irreducible $\mathbb{B} \in HS(\mathbb{A})$, the centralizer of the monolith of $\mathbb{B}$ 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.5.31. *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 $\mathrm{Con}(\mathbb{A})$ 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.3.8 that

$$\begin{bmatrix} a & p(a,b,c) \\ b & c \end{bmatrix} \in \Delta_{0_{\mathbb{A}}^*}^{1_{\mathbb{A}}}.$$

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.2.8 we have $(a, b) \in [1_{\mathbb{A}}, 0_{\mathbb{A}}^*]$, which contradicts $(0_{\mathbb{A}} : 0_{\mathbb{A}}^*) = 1_{\mathbb{A}}$. $\qquad\square$

Proposition A.5.32. *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

$$[[\alpha, \alpha], \alpha] = [[\alpha, \alpha] \wedge \alpha, \alpha] = [\alpha, \alpha] \wedge [\alpha, \alpha] = [\alpha, \alpha]. \qquad\square$$

Proposition A.5.33 (Ol'šanskiĭ [114])**.** *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 $\prod_i s_i/[\mathbb{S}, \mathbb{S}]$, where for each i, $s_i \in \mathbb{S}$ is given by $gx_i = x_j s_i$ 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 $gx_i = x_i g$ 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 $g \notin [\mathbb{G}, \mathbb{G}]$. 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. $\square$

Corollary A.5.34 (Ol'šanskiĭ [114])**.** *A finite group is residually small iff all of its Sylow subgroups are abelian.*

Proof. By Proposition A.5.32, 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.5.33, we have $Z(\mathbb{C}) \wedge [\mathbb{C}, \mathbb{C}] = 0$, so $0^* \wedge [\mathbb{C}, \mathbb{C}] = 0$, which implies that $[\mathbb{C}, \mathbb{C}] = 0$. $\square$

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 [62]. We will use a different definition of similarity than their definition, but which they prove to be equivalent.

Definition A.5.35. We say that subdirectly irreducible algebras $\mathbb{A}, \mathbb{B}$ in a congruence modular variety $\mathcal{V}$ are *similar* if there exists an algebra $\mathbb{C} \in \mathcal{V}$ with congruences $\alpha, \beta, \gamma, \delta \in \mathbb{C}$ such that $\mathbb{C}/\alpha \cong \mathbb{A}$, $\mathbb{C}/\beta \cong \mathbb{B}$, and

$$[\![\alpha, \alpha^*]\!] \searrow [\![\gamma, \delta]\!] \nearrow [\![\beta, \beta^*]\!].$$

If furthermore $\mathbb{C} \leq_{sd} \mathbb{A} \times \mathbb{B}$ and α, β are the kernels of the projections to $\mathbb{A}, \mathbb{B}$, then we say that $\mathbb{C}$ is the *graph of a similarity* from $\mathbb{A}$ to $\mathbb{B}$.

Proposition A.5.36. *If $\mathbb{A}, \mathbb{B}$ are similar, then there is a witnessing algebra $\mathbb{C} \leq_{sd} \mathbb{A} \times \mathbb{B}$ which is the graph of a similarity from $\mathbb{A}$ to $\mathbb{B}$. If α, β are the kernels of the projections to $\mathbb{A}, \mathbb{B}$, then $(\alpha : \alpha^*) = (\beta : \beta^*)$ and $\mathbb{C}/(\alpha : \alpha^*)$ is the graph of an isomorphism*

$$\mathbb{A}/(0_{\mathbb{A}} : 0_{\mathbb{A}}^*) \xrightarrow{\sim} \mathbb{B}/(0_{\mathbb{B}} : 0_{\mathbb{B}}^*).$$

If $\mathbb{A}, \mathbb{B}$ are similar but not isomorphic, then they must both have abelian monoliths.

Proof. For the first statement, let $\mathbb{C} \in \mathcal{V}$ and $\alpha, \beta, \gamma, \delta \in \mathrm{Con}(\mathbb{C})$ be as in the definition of similarity. It's enough to show that we have

$$[\![\alpha, \alpha^*]\!] \searrow [\![\alpha \wedge \beta, (\alpha \wedge \beta) \vee \delta]\!] \nearrow [\![\beta, \beta^*]\!],$$

315

since then we can replace $\mathbb{C}$ by $\mathbb{C}/(\alpha \wedge \beta)$, which is a subdirect product of $\mathbb{C}/\alpha \cong \mathbb{A}$ and $\mathbb{C}/\beta \cong \mathbb{B}$. We have

$$\alpha \vee ((\alpha \wedge \beta) \vee \delta) = \alpha \vee \delta = \alpha^*,$$

and by the modular law and the fact that $\gamma \leq \alpha \wedge \beta$, we have

$$\alpha \wedge ((\alpha \wedge \beta) \vee \delta) = (\alpha \wedge \delta) \vee (\alpha \wedge \beta) = \gamma \vee (\alpha \wedge \beta) = (\alpha \wedge \beta),$$

so $[\![\alpha, \alpha^*]\!] \searrow [\![\alpha \wedge \beta, (\alpha \wedge \beta) \vee \delta]\!]$, and the other perspectivity follows by a symmetric argument.

The remaining statements follow from the Diamond Isomorphism Theorem A.2.5: if $[\![\alpha, \alpha^*]\!] \searrow [\![\gamma, \delta]\!] \nearrow [\![\beta, \beta^*]\!]$, then $(\alpha : \alpha^*) = (\gamma : \delta) = (\beta : \beta^*)$, so

$$\mathbb{A}/(0_\mathbb{A} : 0_\mathbb{A}^*) \cong \mathbb{C}/(\alpha : \alpha^*) = \mathbb{C}/(\beta : \beta^*) \cong \mathbb{B}/(0_\mathbb{B} : 0_\mathbb{B}^*),$$

and

$$[\alpha^*, \alpha^*]_\alpha = \alpha \iff [\delta, \delta]_\gamma = \gamma \iff [\beta^*, \beta^*]_\beta = \beta,$$

so $0_\mathbb{A}^*$ is abelian iff $0_\mathbb{B}^*$ is abelian, and if neither is abelian then $\alpha = (\alpha : \alpha^*) = (\beta : \beta^*) = \beta$ and $\mathbb{A} \cong \mathbb{C}/\alpha = \mathbb{C}/\beta \cong \mathbb{B}$. $\qquad\square$

Proposition A.5.37. *If $\mathbb{A}, \mathbb{B}$ are similar such that σ is the corresponding isomorphism*

$$\sigma : \mathbb{A}/(0_\mathbb{A} : 0_\mathbb{A}^*) \xrightarrow{\sim} \mathbb{B}/(0_\mathbb{B} : 0_\mathbb{B}^*),$$

then they are similar via the algebra $\mathbb{R} = \{(x, y) \in \mathbb{A} \times \mathbb{B} \mid \sigma(x/(0_\mathbb{A} : 0_\mathbb{A}^*)) = y/(0_\mathbb{B} : 0_\mathbb{B}^*)\}$.

Proof. Suppose $\mathbb{C} \leq \mathbb{R}$ is the graph of a similarity from $\mathbb{A}$ to $\mathbb{B}$, with

$$[\![\ker \pi_1, (\ker \pi_1)^*]\!] \searrow [\![0_\mathbb{C}, \delta]\!] \nearrow [\![\ker \pi_2, (\ker \pi_2)^*]\!]$$

in $\mathrm{Con}(\mathbb{C})$. We may assume that $\mathbb{A}, \mathbb{B}$ have abelian monoliths, so $[\delta, \delta] = 0_\mathbb{C}$ by the Diamond Isomorphism Theorem A.2.5. Then by Theorem A.4.5, δ permutes with all congruences in $\mathrm{Con}(\mathbb{C})$, so in particular $(\ker \pi_1)^* = \delta \circ \ker \pi_1$. In other words, for any $(a, b) \in \mathbb{C}$ and any $a' \in a/0_\mathbb{A}^*$, 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 $\delta \wedge \ker \pi_1 = 0_\mathbb{C}$. Additionally, we must have $b' \in b/0_\mathbb{B}^*$, since $\delta \leq (\ker \pi_2)^*$.

Now we can extend δ to a congruence $\delta_\mathbb{R} \in \mathrm{Con}(\mathbb{R})$ as follows. For $(a, b), (a', b') \in \mathbb{R}$ with $a \ 0_\mathbb{A}^* \ a'$ and $b \ 0_\mathbb{B}^* \ b'$, we pick any $(u, v) \in \mathbb{C}$ with $u \ (0_\mathbb{A} : 0_\mathbb{A}^*) \ a$ and write

$$\begin{bmatrix} a & a' \\ b & b' \end{bmatrix} \in \delta_\mathbb{R} \iff \begin{bmatrix} p(a, a', u) & u \\ p(b, b', v) & v \end{bmatrix} \in \delta,$$

where p is a Gumm difference term. Note that by Corollary A.3.9, this choice of $\delta_\mathbb{R}$ is preserved by the operations of $\mathbb{A}$ 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 $u' \ (0_\mathbb{A} : 0_\mathbb{A}^*) \ a$, and apply Corollary A.3.9 again to see that

$$p\left(\begin{bmatrix} p(a, a', u) & u \\ p(b, b', v) & v \end{bmatrix}, \begin{bmatrix} p(a, a, u) & u \\ p(b, b, v) & v \end{bmatrix}, \begin{bmatrix} p(a, a, u') & u' \\ p(b, b, v') & v' \end{bmatrix}\right) = \begin{bmatrix} p(a, a', u') & u' \\ p(b, b', v') & v' \end{bmatrix},$$

316

where we have used $0^*_{\mathbb{A}}, 0^*_{\mathbb{B}}$ abelian to see that $p(a', a, a) = a'$ and $p(b', b, b) = b'$.

We need to check that $\delta_{\mathbb{R}}$ is a congruence on $\mathbb{R}$. It clearly contains the equality relation on $\mathbb{R}$. For symmetry and transitivity, note that

$$p\left(\begin{bmatrix} p(a, a', u) \\ p(b, b', v) \end{bmatrix}, \begin{bmatrix} p(a'', a', u) \\ p(b'', b', v) \end{bmatrix}, \begin{bmatrix} u \\ v \end{bmatrix}\right) = p\left(\begin{bmatrix} p(a, a', u) \\ p(b, b', v) \end{bmatrix}, \begin{bmatrix} p(a'', a', u) \\ p(b'', b', v) \end{bmatrix}, \begin{bmatrix} p(a'', a'', u) \\ p(b'', b'', v) \end{bmatrix}\right) = \begin{bmatrix} p(a, a'', u) \\ p(b, b'', v) \end{bmatrix}.$$

Finally, we need to check that $\delta_{\mathbb{R}} \wedge \ker \pi_1 = 0_{\mathbb{R}}$ and $\delta_{\mathbb{R}} \vee \ker \pi_1 = (\ker \pi_1)^*$. That $\delta_{\mathbb{R}} \wedge \ker \pi_1 = 0_{\mathbb{R}}$ follows from the fact that if we pick u such that $(u, b') \in \mathbb{C}$, then

$$\begin{bmatrix} a & a \\ b & b' \end{bmatrix} \in \delta_{\mathbb{R}} \iff \begin{bmatrix} p(a, a, u) & u \\ p(b, b', b') & b' \end{bmatrix} = \begin{bmatrix} u & u \\ b & b' \end{bmatrix} \in \delta,$$

and so this can only occur when $b = b'$ since $\delta \wedge \ker \pi_1 = 0_{\mathbb{C}}$ (by assumption). That $\delta_{\mathbb{R}} \vee \ker \pi_1 = (\ker \pi_1)^*$ follows from $\delta \subseteq \delta_{\mathbb{R}} \subseteq (\ker \pi_1)^*$ and $\delta \not\subseteq \ker \pi_1$. $\qquad\square$

Corollary A.5.38. *A similarity from $\mathbb{A}$ to $\mathbb{B}$ can be described by the following data: an isomorphism*

$$\sigma : \mathbb{A}/(0_{\mathbb{A}} : 0^*_{\mathbb{A}}) \xrightarrow{\sim} \mathbb{B}/(0_{\mathbb{B}} : 0^*_{\mathbb{B}})$$

together with a congruence $\delta \in \operatorname{Con}(\mathbb{R})$, where $\mathbb{R} = \{(x, y) \in \mathbb{A} \times \mathbb{B} \mid \sigma(x/(0_{\mathbb{A}} : 0^*_{\mathbb{A}})) = y/(0_{\mathbb{B}} : 0^*_{\mathbb{B}})\}$, such that for every $(a, b) \in \mathbb{R}$ and every $a' \in a/0^*_{\mathbb{A}}$, there exists a unique $b' \in b/0^*_{\mathbb{B}}$ such that

$$\begin{bmatrix} a & a' \\ b & b' \end{bmatrix} \in \delta.$$

In particular, if $\mathbb{A}, \mathbb{B}$ are idempotent, then for any $(a, b) \in \mathbb{R}$ the congruence classes $a/0^*_{\mathbb{A}}$ and $b/0^*_{\mathbb{B}}$ are isomorphic to each other.

Corollary A.5.39. *Similarity is an equivalence relation on subdirectly irreducible algebras.*

Proof. Suppose we have similarities from $\mathbb{A}$ to $\mathbb{B}$ and from $\mathbb{B}$ to $\mathbb{C}$, described by isomorphisms

$$\mathbb{A}/(0_{\mathbb{A}} : 0^*_{\mathbb{A}}) \xrightarrow{\sigma} \mathbb{B}/(0_{\mathbb{B}} : 0^*_{\mathbb{B}}) \xrightarrow{\sigma'} \mathbb{C}/(0_{\mathbb{C}} : 0^*_{\mathbb{C}})$$

and congruences δ, δ'. We define a congruence $\delta \circ \delta'$ by

$$\begin{bmatrix} a & a' \\ c & c' \end{bmatrix} \in \delta \circ \delta' \iff \exists (b, b') \in 0^*_{\mathbb{B}} \left(\begin{bmatrix} a & a' \\ b & b' \end{bmatrix} \in \delta\right) \wedge \left(\begin{bmatrix} b & b' \\ c & c' \end{bmatrix} \in \delta'\right).$$

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 δ, and then there is a unique c' which satisfies δ'. 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 δ, 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 δ' similarly. $\qquad\square$

We will show that every subdirectly irreducible algebra $\mathbb{A}$ with abelian monolith is similar to a subdirectly irreducible algebra $D(\mathbb{A})$ such that the monolith of $D(\mathbb{A})$ is equal to its own centralizer. The size of the algebra $D(\mathbb{A})$ can then be bounded using Theorem A.5.14 and the following proposition.

Proposition A.5.40. *If $\mathbb{B} \in \mathcal{V}(\mathbb{A})$ is subdirectly irreducible, $\mathbb{A}$ is finite, and $\mathcal{V}(\mathbb{A})$ is congruence modular, then every congruence class of $0_{\mathbb{B}}^*$ has size at most $|\mathbb{A}|$.*

Proof. By Theorem A.5.22 and Corollary A.5.12, we may assume without loss of generality that $\mathbb{B}$ is finite. By Theorem A.5.14, we may also assume that $0_{\mathbb{B}}^*$ is abelian. Take m minimal such that there exists $\mathbb{C} \leq \mathbb{A}^m$ and $\theta \in \mathrm{Con}(\mathbb{C})$ with $\mathbb{B} \cong \mathbb{C}/\theta$, 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 $\ker \pi_{1'} \not\leq \theta$. Thus we have
$$\llbracket \theta, \theta^* \rrbracket \searrow \llbracket \theta \wedge \ker \pi_{1'}, \theta^* \wedge \ker \pi_{1'} \rrbracket.$$
By Theorem A.4.5, the congruences θ and $\theta^* \wedge \ker \pi_{1'}$ permute. Thus for every congruence class C^* of θ^* containing some $c \in \mathbb{C}$, the size of C^*/θ is equal to the size of $C'/(\theta \wedge \ker \pi_{1'})$, where C' is the congruence class of $\theta^* \wedge \ker \pi_{1'}$ containing c. But $|C'/(\theta \wedge \ker \pi_{1'})| \leq |\mathbb{C}/\ker \pi_{1'}| = |\mathbb{A}|$, so every congruence class of $0_{\mathbb{B}}^*$ has size bounded by $|\mathbb{A}|$. $\qquad\square$

Definition A.5.41. Suppose $\mathbb{A}$ is a subdirectly irreducible algebra in a congruence modular variety. If $0_{\mathbb{A}}^*$ is nonabelian, define $D(\mathbb{A})$ to be $\mathbb{A}$. Otherwise, consider $0_{\mathbb{A}}^*$ as a subalgebra of $\mathbb{A}^2$ and $\Delta_{0_{\mathbb{A}}^*}^{(0:0^*)}$ as a congruence on $0_{\mathbb{A}}^*$, and define $D(\mathbb{A}) = 0_{\mathbb{A}}^*/\Delta_{0_{\mathbb{A}}^*}^{(0:0^*)}$.

Recall that by Theorem A.3.8, if $0_{\mathbb{A}}^*$ is abelian and p is a Gumm difference term, then $(0_{\mathbb{A}} : 0_{\mathbb{A}}^*) \geq 0_{\mathbb{A}}^*$ and $[(0_{\mathbb{A}} : 0_{\mathbb{A}}^*), 0_{\mathbb{A}}^*] = 0_{\mathbb{A}}$, so we have
$$\begin{bmatrix} x & w \\ y & z \end{bmatrix} \in \Delta_{0_{\mathbb{A}}^*}^{(0:0^*)} \iff (p(x,y,z) = w) \wedge (x \equiv_{0_{\mathbb{A}}^*} y \equiv_{(0:0^*)} z).$$

In this case, the subalgebra $\{(x,x)/\Delta_{0_{\mathbb{A}}^*}^{(0:0^*)}\} \leq D(\mathbb{A})$ meets every congruence class of $(0_{\mathbb{A}} : 0_{\mathbb{A}}^*)_{D(\mathbb{A})}$ (that is, the congruence $(0_{\mathbb{A}} : 0_{\mathbb{A}}^*)$ considered as a congruence on $D(\mathbb{A})$) exactly once, and is isomorphic to $\mathbb{A}/(0_{\mathbb{A}} : 0_{\mathbb{A}}^*)$.

Proposition A.5.42. *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 $(0_{\mathbb{A}} : 0_{\mathbb{A}}^*)_{D(\mathbb{A})}$, and $\mathbb{A}, D(\mathbb{A})$ are similar via the algebra $0_{\mathbb{A}}^*$ and the congruences $\ker \pi_1, \Delta_{0_{\mathbb{A}}^*}^{(0:0^*)} \in \mathrm{Con}(0_{\mathbb{A}}^*)$. Furthermore, the monolith $(0_{\mathbb{A}} : 0_{\mathbb{A}}^*)_{D(\mathbb{A})}$ 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 $\pi_1(\ker \pi_1 \vee \ker \pi_2) = 0_{\mathbb{A}}^*$. First we check that in $\mathrm{Con}(0_{\mathbb{A}}^*)$ we have the perspectivities
$$\llbracket \ker \pi_1, \ker \pi_1 \vee \ker \pi_2 \rrbracket \searrow \llbracket 0_{0_{\mathbb{A}}^*}, \ker \pi_2 \rrbracket \nearrow \llbracket \Delta_{0_{\mathbb{A}}^*}^{(0:0^*)}, (0_{\mathbb{A}} : 0_{\mathbb{A}}^*)_{0_{\mathbb{A}}^*} \rrbracket.$$

The hardest step here is checking that $\ker \pi_2 \vee \Delta_{0_{\mathbb{A}}^*}^{(0:0^*)} = (0_{\mathbb{A}} : 0_{\mathbb{A}}^*)_{0_{\mathbb{A}}^*}$: if $(x,y), (w,z) \in 0_{\mathbb{A}}^*$ with $(y,z) \in (0_{\mathbb{A}} : 0_{\mathbb{A}}^*)$, then we have
$$\begin{bmatrix} x \\ y \end{bmatrix} \Delta_{0_{\mathbb{A}}^*}^{(0:0^*)} \begin{bmatrix} p(x,y,z) \\ z \end{bmatrix} \ker \pi_2 \begin{bmatrix} w \\ z \end{bmatrix}.$$

318

To see that $\ker \pi_2 \wedge \Delta_{0_{\mathbb{A}}^*}^{(0:0^*)} = 0_{0_{\mathbb{A}}^*}$, note that by Theorem A.2.8 the inequality $\ker \pi_2 \wedge \Delta_{0_{\mathbb{A}}^*}^{(0:0^*)} \leq \ker \pi_1$ is equivalent to $[(0_{\mathbb{A}} : 0_{\mathbb{A}}^*), 0_{\mathbb{A}}^*] = 0_{\mathbb{A}}$.

Next we show that $(0_{\mathbb{A}} : 0_{\mathbb{A}}^*)_{0_{\mathbb{A}}^*}$ is the unique cover of $\Delta_{0_{\mathbb{A}}^*}^{(0:0^*)}$ in $\mathrm{Con}(0_{\mathbb{A}}^*)$. Note first that $(0_{\mathbb{A}} : 0_{\mathbb{A}}^*)_{0_{\mathbb{A}}^*}$ is *a* cover of $\Delta_{0_{\mathbb{A}}^*}^{(0:0^*)}$, since the interval $[\![\Delta_{0_{\mathbb{A}}^*}^{(0:0^*)}, (0_{\mathbb{A}} : 0_{\mathbb{A}}^*)_{0_{\mathbb{A}}^*}]\!]$ is isomorphic to $[\![\ker \pi_1, \ker \pi_1 \vee \ker \pi_2]\!] \cong [\![0_{\mathbb{A}}, 0_{\mathbb{A}}^*]\!]$ by the Diamond Isomorphism Theorem A.2.5.

Suppose that ψ is any congruence in $\mathrm{Con}(0_{\mathbb{A}}^*)$ with $\psi > \Delta_{0_{\mathbb{A}}^*}^{(0:0^*)}$. If $\psi \geq \ker \pi_2$, then $\psi \geq \Delta_{0_{\mathbb{A}}^*}^{(0:0^*)} \vee \ker \pi_2 = (0_{\mathbb{A}} : 0_{\mathbb{A}}^*)_{0_{\mathbb{A}}^*}$, and we are done. Otherwise, since $\ker \pi_2$ is a cover of $0_{0_{\mathbb{A}}^*}$, we must have $\psi \wedge \ker \pi_2 = 0_{0_{\mathbb{A}}^*}$. Then we have

$$[\psi \vee \ker \pi_1, \ker \pi_2 \vee \ker \pi_1]_{\ker \pi_1} \leq [\psi, \ker \pi_2] \vee \ker \pi_1 \leq (\psi \wedge \ker \pi_2) \vee \ker \pi_1 = \ker \pi_1.$$

Applying π_1 to both sides, we see that $\pi_1(\psi \vee \ker \pi_1) \leq (0_{\mathbb{A}} : 0_{\mathbb{A}}^*)$, so $\psi \vee \ker \pi_1 \leq (0_{\mathbb{A}} : 0_{\mathbb{A}}^*)_{0_{\mathbb{A}}^*}$. Thus $\psi \in [\![\Delta_{0_{\mathbb{A}}^*}^{(0:0^*)}, (0_{\mathbb{A}} : 0_{\mathbb{A}}^*)_{0_{\mathbb{A}}^*}]\!]$, so again we must have $\psi = (0_{\mathbb{A}} : 0_{\mathbb{A}}^*)_{0_{\mathbb{A}}^*}$. We have finished showing that $D(\mathbb{A})$ is subdirectly irreducible.

To see that the monolith $(0_{\mathbb{A}} : 0_{\mathbb{A}}^*)_{D(\mathbb{A})}$ of $D(\mathbb{A})$ is its own centralizer, note that by the Diamond Isomorphism Theorem A.2.5 we have

$$(\Delta_{0_{\mathbb{A}}^*}^{(0:0^*)} : (0_{\mathbb{A}} : 0_{\mathbb{A}}^*)_{0_{\mathbb{A}}^*}) = (\ker \pi_1 : \ker \pi_1 \vee \ker \pi_2) = \pi_1^{-1}((0_{\mathbb{A}} : 0_{\mathbb{A}}^*)) = (0_{\mathbb{A}} : 0_{\mathbb{A}}^*)_{0_{\mathbb{A}}^*}. \qquad \square$$

Proposition A.5.43. *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 $X_{\mathbb{A}}, X_{\mathbb{B}}$ which intersect their monoliths transversely, then they are similar iff they are isomorphic.

Let $\sigma : \mathbb{A}/0_{\mathbb{A}}^* \to \mathbb{B}/0_{\mathbb{B}}^*$ be the isomorphism and $\delta \in \mathrm{Con}(\mathbb{R})$, where $\mathbb{R} = \{(x, y) \in \mathbb{A} \times \mathbb{B} \mid \sigma(x/(0_{\mathbb{A}} : 0_{\mathbb{A}}^*)) = y/(0_{\mathbb{B}} : 0_{\mathbb{B}}^*)\}$, be the data describing a similarity from $\mathbb{A}$ to $\mathbb{B}$. Then σ induces an isomorphism $\sigma_X : X_{\mathbb{A}} \to X_{\mathbb{B}}$, and the graph of σ_X is a subalgebra of $\mathbb{R}$. Let $\mathbb{S}$ be the subalgebra of $(a, b) \in \mathbb{R}$ such that (a, b) is congruent to some element of σ_X modulo δ. Then $\mathbb{S}$ must be the graph of an isomorphism from $\mathbb{A}$ to $\mathbb{B}$. $\qquad \square$

Theorem A.5.44. *If $\mathbb{B} \in \mathcal{V}(\mathbb{A})$ is subdirectly irreducible, $\mathbb{A}$ is finite, and $\mathcal{V}(\mathbb{A})$ is congruence modular, then $\mathbb{B}$ is similar to a subdirectly irreducible algebra in $HS(\mathbb{A})$.*

Proof. We may as well replace $\mathbb{B}$ by $D(\mathbb{B})$, so assume without loss of generality that the monolith of $\mathbb{B}$ is either nonabelian or equal to its own centralizer. If the monolith of $\mathbb{B}$ is nonabelian, then $\mathbb{B} \in HS(\mathbb{A})$ by Theorem A.5.14, so we just need to handle the case where $0_{\mathbb{B}}^* = (0_{\mathbb{B}} : 0_{\mathbb{B}}^*)$. In this case, Theorem A.5.14 implies that $\mathbb{B}/0_{\mathbb{B}}^* \in HS(\mathbb{A})$, so by Proposition A.5.40 we have $|\mathbb{B}| \leq |\mathbb{A}|^2 < \infty$.

Since $\mathbb{B}$ is finite, we can write $\mathbb{B} = \mathbb{R}/\theta$ for some $\mathbb{R} \leq \mathbb{A}^n$ and $\theta \in \mathrm{Con}(\mathbb{R})$. Then we can write $\mathbb{R}$ as a subdirect product $\mathbb{R} \leq_{sd} \mathbb{A}_1 \times \cdots \times \mathbb{A}_m$ of finitely many subdirectly irreducible algebras $\mathbb{A}_i \in HS(\mathbb{A})$. We assume that the $\mathbb{A}_i$ are chosen such that none of them can be replaced by a subdirect product of some number of proper quotients of $\mathbb{A}_i$ while still keeping the isomorphism $\mathbb{R}/\theta \cong \mathbb{B}$.

Then for any i, we must have $\theta \wedge \ker \pi_{[m]\setminus\{i\}} = 0_{\mathbb{R}}$: if not, we could replace $\mathbb{A}_i$ with a subdirect representation of $\mathbb{R}/(\ker \pi_i \vee (\theta \wedge \ker \pi_{[m]\setminus\{i\}}))$, since by the modular law we have

$$\ker \pi_{[m]\setminus\{i\}} \wedge (\ker \pi_i \vee (\theta \wedge \ker \pi_{[m]\setminus\{i\}})) = (\ker \pi_{[m]\setminus\{i\}} \wedge \ker \pi_i) \vee (\theta \wedge \ker \pi_{[m]\setminus\{i\}}) \leq \theta.$$

Since $\ker \pi_{[m]\setminus\{i\}} \neq 0_{\mathbb{R}}$, we have $\theta \vee \ker \pi_{[m]\setminus\{i\}} \geq \theta^*$, so $\theta^* \wedge \ker \pi_{[m]\setminus\{i\}}$ is a cover of $0_{\mathbb{R}}$, and we have

$$[\![\theta, \theta^*]\!] \searrow [\![0_{\mathbb{R}}, \theta^* \wedge \ker \pi_{[m]\setminus\{i\}}]\!] \nearrow [\![\ker \pi_i, (\ker \pi_i)^*]\!],$$

so $\mathbb{B} = \mathbb{R}/\theta$ is similar to $\mathbb{A}_i = \mathbb{R}/\ker \pi_i$. $\qquad\square$

Example A.5.3. Let's work out what $D(\mathbb{G})$ is when $\mathbb{G}$ is a subdirectly irreducible group. Let $\mathbb{M} \triangleleft \mathbb{G}$ be the normal subgroup corresponding to the monolith $0_{\mathbb{G}}^*$, and let $\mathbb{N} = C_{\mathbb{G}}(\mathbb{M}) \triangleleft \mathbb{G}$ be the normal subgroup corresponding to the centralizer $(0_{\mathbb{G}} : 0_{\mathbb{G}}^*)$. First off, what is the group structure on the congruence $0_{\mathbb{G}}^*$?

By definition, we have

$$0_{\mathbb{G}}^* = \{(x, y) \in \mathbb{G}^2 \mid x^{-1}y \in \mathbb{M}\}.$$

We have a natural exact sequence of groups

$$0 \to \mathbb{M} \hookrightarrow 0_{\mathbb{G}}^* \twoheadrightarrow \mathbb{G} \to 0,$$

where the inclusion is the map $m \mapsto (1, m)$ and the quotient map is the first projection π_1. The quotient $0_{\mathbb{G}}^* \twoheadrightarrow \mathbb{G}$ has a section $\Delta : \mathbb{G} \hookrightarrow 0_{\mathbb{G}}^*$ given by $g \mapsto (g, g)$. Thus we can write $0_{\mathbb{G}}^*$ as a semidirect product

$$0_{\mathbb{G}}^* \cong \mathbb{M} \rtimes \mathbb{G},$$

where the action of $\mathbb{G}$ on $\mathbb{M}$ is the standard conjugation action.

How about the congruence $\Delta_{0_{\mathbb{G}}^*}^{(0:0^*)} \in \mathrm{Con}(0_{\mathbb{G}}^*)$? By Theorem A.3.8, we have

$$\begin{bmatrix} x & w \\ y & z \end{bmatrix} \in \Delta_{0_{\mathbb{G}}^*}^{(0:0^*)} \iff (xy^{-1}z = w) \wedge (x \equiv_{\mathbb{M}} y \equiv_{\mathbb{N}} z).$$

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 $\Delta_{0_{\mathbb{G}}^*}^{(0:0^*)}$ corresponds to the normal subgroup

$$\{(n, n) \mid n \in \mathbb{N}\},$$

so under the isomorphism $0_{\mathbb{G}}^* \cong \mathbb{M} \rtimes \mathbb{G}$ it corresponds to $\mathbb{N}$, considered as a subgroup of $\mathbb{G}$. Thus we have

$$D(\mathbb{G}) = 0_{\mathbb{G}}^* / \Delta_{0_{\mathbb{G}}^*}^{(0:0^*)} \cong (\mathbb{M} \rtimes \mathbb{G}) / \mathbb{N} \cong \mathbb{M} \rtimes (\mathbb{G}/\mathbb{N}).$$

That any of this makes sense follows from $\mathbb{N} = C_{\mathbb{G}}(\mathbb{M})$. We see that $\mathbb{M}$ is the normal subgroup corresponding to the monolith of $D(\mathbb{G})$, that $\mathbb{M}$ is equal to its own centralizer in $D(\mathbb{G})$, and that the natural map $\mathbb{G}/\mathbb{N} \hookrightarrow D(\mathbb{G})$ has image transverse to the monolith, and induces an isomorphism

$$\sigma : \mathbb{G}/\mathbb{N} \xrightarrow{\sim} D(\mathbb{G})/\mathbb{M}.$$

To complete the description of the similarity from $\mathbb{G}$ to $D(\mathbb{G})$, we let $\mathbb{R}$ be the fiber product of $\mathbb{G}$ and $D(\mathbb{G})$ over $\mathbb{G}/\mathbb{N}$, and define the congruence $\delta \in \mathrm{Con}(\mathbb{R})$ as the 4-ary relation

$$\begin{bmatrix} a & a' \\ b & b' \end{bmatrix} \in \delta \iff \begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} a' \\ b' \end{bmatrix} \in \mathbb{R} \ \wedge \ a^{-1}a' = b^{-1}b' \in \mathbb{M}.$$

That δ is closed under multiplication must be checked - it follows from the fact that $\mathbb{N}$ centralizes $\mathbb{M}$, 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 $\mathbb{G}/\mathbb{N}$.

What are the possible values for $D(\mathbb{G})$, assuming the monolith is abelian? Note that if we consider $\mathbb{M}$ as a module via the $\mathbb{G}/\mathbb{N}$ action, then it must be a *simple* module, since if it has any nontrivial submodule $\mathbb{M}'$, then $\mathbb{M}'$ will be a smaller normal subgroup of $\mathbb{G}$. Thus the general situation is that $\mathbb{M}$ is some simple module over the ring $\mathbb{Z}[\mathbb{G}/\mathbb{N}]$ (where $\mathbb{G}/\mathbb{N}$ acts faithfully on $\mathbb{M}$), and $D(\mathbb{G}) \cong \mathbb{M} \rtimes (\mathbb{G}/\mathbb{N})$.

Example A.5.4. If we take $\mathbb{G} = S_3$ in the above, we find that $D(S_3) \cong \mathbb{Z}/3 \rtimes \mathbb{Z}/2 \cong S_3$. The 4-ary relation $\delta \leq S_3^{2 \times 2}$ corresponding to the trivial similarity from S_3 to itself is given by

$$\begin{bmatrix} a & a' \\ b & b' \end{bmatrix} \in \delta \iff s(a) = s(b) = s(a') = s(b') \ \wedge \ a^{-1}a' = b^{-1}b',$$

where $s : S_3 \to \{\pm 1\}$ is the sign homomorphism.

We can think of the relation δ as having two "strands" corresponding to the two possible signs of permutations, and if we restrict to either strand then δ becomes an affine relation over $\mathbb{Z}/3$. The fact that we can multiply elements of δ which come from different strands and still get an element of δ is worth thinking about.

Now suppose that $\mathbb{G}$ is some other subdirectly irreducible group such that $D(\mathbb{G}) \cong S_3$, with monolith corresponding to $\mathbb{M} \lhd \mathbb{G}$ and $\mathbb{N} = C_{\mathbb{G}}(\mathbb{M})$. Then since $\mathbb{G}$ is similar to S_3, we must have $\mathbb{M} \cong \mathbb{Z}/3$ and $\mathbb{G}/\mathbb{N} \cong \mathbb{Z}/2$ by Corollary A.5.38, with $\mathbb{G}/\mathbb{N}$ acting on $\mathbb{M}$ by negation since $D(\mathbb{G}) \cong \mathbb{M} \rtimes (\mathbb{G}/\mathbb{N}) \cong S_3$. If the action of $\mathbb{G}/\mathbb{N}$ on $\mathbb{N}$ is given by an involution τ, then for any $n \in \mathbb{N} \setminus \{1\}$ we must have $\mathbb{M}$ contained in the normal subgroup of $\mathbb{N}$ generated by n, n^τ.

In particular, if $\mathbb{N}$ is abelian then we see that $n + n^\tau, n - n^\tau \in \mathbb{M}$ for all $n \in \mathbb{N}$, and additionally in this case $\mathbb{N}$ must have prime power order by Theorem A.5.19. Thus if $\mathbb{N}$ is abelian then we must actually have $\mathbb{N} = \mathbb{M}$, and $\mathbb{G} \cong S_3$.

Example A.5.5. If we take $\mathbb{G} = Q_8 = \{\pm 1, \pm i, \pm j, \pm k\}$ the quaternion group with $i^2 = j^2 = k^2 = ijk = -1$, 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

$$\begin{bmatrix} a & a' \\ b & b' \end{bmatrix} \in \delta \iff a' = (-1)^{b+b'}a.$$

This relation closely resembles an affine relation over $\mathbb{Z}/2$.

Appendix B

Tame Congruence Theory

Tame congruence theory was introduced by Hobby and McKenzie [74] in order to answer questions about congruence lattices of finite algebras. Since every congruence contains the diagonal, every congruence is automatically invariant under the *polynomial* clone of our algebra, and in fact the congruence lattice of an algebra is completely determined by the collection of *unary* polynomials of the algebra.

Proposition B.0.1. *An equivalence relation θ on an algebra $\mathbb{A}$ is a congruence of $\mathbb{A}$ iff θ is preserved by every unary polynomial operation of $\mathbb{A}$.*

So tame congruence theory is really about the how the identities satisfied in a (usually locally finite) variety affect the behavior of unary polynomials, and how this in turn affects the behavior of congruences. Because of the important role of polynomial operations in tame congruence theory, we will use $\mathrm{Pol}_n(\mathbb{A})$ to represent the set of n-ary polynomial operations of $\mathbb{A}$ throughout this appendix (hopefully this doesn't cause any confusion with the notation for the polymorphism clone of a relational structure).

The material in this appendix is mostly taken from Hobby and McKenzie's wonderful book [74] (some of it is my solutions to various exercises from their book).

B.1 Shrinking algebras with unary polynomials, minimal sets, and traces

As soon as you have a unary operation φ on a finite set, the natural thing to do with it is to iterate it until we get the compositionally idempotent operation

$$\varphi^{\infty} := \lim_{n \to \infty} \varphi^{\circ n!}.$$

This gives us a large collection of (compositionally) idempotent unary polynomial operations on any finite algebra. For unary operations, I will drop the qualifier "compositionally" on "idempotent", since idempotent unary operations in the usual sense are not very interesting.

Definition B.1.1. For any algebra $\mathbb{A}$, we define $E(\mathbb{A})$ to be the set of unary polynomials $e \in \mathrm{Pol}_1(\mathbb{A})$ such that $e \circ e = e$. Elements of $E(\mathbb{A})$ might be called the *idempotents* or *projections* of $\mathbb{A}$.

Recall from Section 3.2 that for any idempotent $e \in E(\mathbb{A})$, the clone of restrictions to $e(\mathbb{A})$ of polynomial operations of $\mathbb{A}$ which preserve $e(\mathbb{A})$ is essentially the same as the clone of operations of the form

$$(x_1, ..., x_n) \mapsto e(f(e(x_1), ..., e(x_n))),$$

for $f \in \mathrm{Pol}_n(\mathbb{A})$ (strictly speaking, the es on the inside are not really necessary, they are only there to stop us from caring about how f behaves outside the set $e(\mathbb{A})$). This gives us a rich enough source of polynomial operations which preserve $e(\mathbb{A})$ to make it worth studying the restricted clone and introducing notation for it.

Definition B.1.2. If $\mathbb{A}$ is an algebra and $U \subseteq \mathbb{A}$, then we define the restriction $\mathrm{Pol}(\mathbb{A})|_U$ to be the set of restrictions $f|_U$ of polynomial operations $f \in \mathrm{Pol}(\mathbb{A})$ which preserve the subset U, and we define the *induced algebra* $\mathbb{A}|_U$ to be $(U, \mathrm{Pol}(\mathbb{A})|_U)$ (up to polynomial equivalence).

Restrictions are related to the congruence lattice using the following result.

Lemma B.1.3 (Pálfy and Pudlák [117], [74]). *For any idempotent $e \in E(\mathbb{A})$, if we set $U = e(\mathbb{A})$, then the map taking the congruence $\theta \in \mathrm{Con}(\mathbb{A})$ to $e(\theta) = \theta|_U$ defines a surjective lattice homomorphism:*

$$\theta \mapsto \theta|_U : \mathrm{Con}(\mathbb{A}) \twoheadrightarrow \mathrm{Con}(\mathbb{A}|_U).$$

More generally, if $N \subseteq U$, then:

- $\mathbb{A}|_N = (\mathbb{A}|_U)|_N$,

- *if N is a union of $\theta|_U$ congruence classes, then the map $\alpha \mapsto \alpha|_N$ defines a lattice homomorphism from the interval $[\![0_{\mathbb{A}}, \theta]\!]$ of $\mathrm{Con}(\mathbb{A})$ to the interval $[\![0_N, \theta|_N]\!]$ of $\mathrm{Con}(\mathbb{A}|_N)$, and*

- *if N is equal to a congruence class of $\theta|_U$, then the map $\alpha \mapsto \alpha|_N$ is a surjective lattice homomorphism $[\![0_{\mathbb{A}}, \theta]\!] \twoheadrightarrow \mathrm{Con}(\mathbb{A}|_N)$.*

Proof. (Following [74]) First, we will prove the statements about the map $\theta \mapsto \theta|_U$. For any $\alpha \in \mathrm{Con}(\mathbb{A}|_U)$, we define the equivalence relation $\hat{\alpha}$ on $\mathbb{A}$ by

$$\hat{\alpha} := \{(x, y) \mid \forall f \in \mathrm{Pol}_1(\mathbb{A}), (e(f(x)), e(f(y))) \in \alpha\}.$$

Then for any $\alpha \in \mathrm{Con}(\mathbb{A}|_U)$ and $\theta \in \mathrm{Con}(\mathbb{A})$ we have $\hat{\alpha} \in \mathrm{Con}(\mathbb{A})$, $\hat{\alpha}|_U = \alpha$, and

$$\theta|_U \leq \alpha \iff \theta \leq \hat{\alpha}.$$

Since restriction obviously preserves meets of congruences, we just need to check that it preserves joins. For this, let $\alpha = \theta_1|_U \vee \theta_2|_U$, and note that

$$\theta_i|_U \leq \alpha \implies \theta_i \leq \hat{\alpha} \implies \theta_1 \vee \theta_2 \leq \hat{\alpha} \implies (\theta_1 \vee \theta_2)|_U \leq \alpha,$$

while the inequality $\alpha \leq (\theta_1 \vee \theta_2)|_U$ is obvious.

To see that $\mathbb{A}|_N = (\mathbb{A}|_U)|_N$, note that if $f \in \mathrm{Pol}(\mathbb{A})$ preserves N, then $e \circ f$ preserves U and $(e \circ f)|_N = f|_N$. Since we have $\alpha|_N = (\alpha|_U)|_N$ for $\alpha \in \mathrm{Con}(\mathbb{A})$, to prove the remaining claims we just need to think about the restriction map $[\![0_U, \theta|_U]\!] \to [\![0_N, \theta|_N]\!]$.

If N is a union of congruence classes of $\theta|_U$, then for any $\theta_1, \theta_2 \leq \theta|_U$ neither θ_i connects any element inside N to any element outside N, so $(\theta_1 \vee \theta_2)|_N = \theta_1|_N \vee \theta_2|_N$. Thus the map $[\![0_U, \theta|_U]\!] \to [\![0_N, \theta|_N]\!]$ is a lattice homomorphism.

323

To finish, we need to show that if N is a congruence class of $\theta|_U$ then this map is surjective. For this, we extend a congruence $\alpha \in \mathrm{Con}(\mathbb{A}|_N)$ to a congruence on $\mathbb{A}|_U$ by

$$\check{\alpha} := \{(x,y) \mid \forall f \in \mathrm{Pol}_1(\mathbb{A}|_U), f(x) \in N \text{ or } f(y) \in N \implies (f(x), f(y)) \in \alpha\}.$$

Since every $f \in \mathrm{Pol}_1(\mathbb{A}|_U)$ preserves $\theta|_U$, if $f(x) \in N$ for any $x \in N$ then f must preserve N, so we have $\check{\alpha}|_N = \alpha$. $\qquad\square$

In order to get any use out of this to study an interval $[\![\alpha, \beta]\!]$ of the congruence lattice, we need to find idempotents e such that $e(\alpha) \neq e(\beta)$, or equivalently such that $e(\beta) \not\subseteq \alpha$.

Definition B.1.4. If $\alpha < \beta \in \mathrm{Con}(\mathbb{A})$, then we define $U_\mathbb{A}(\alpha, \beta)$ to be the collection of sets of the form $f(\mathbb{A})$ for $f \in \mathrm{Pol}_1(\mathbb{A})$ such that $f(\beta) \not\subseteq \alpha$. We define $M_\mathbb{A}(\alpha, \beta)$ to be the collection of minimal sets in $U_\mathbb{A}(\alpha, \beta)$, and we call the sets in $M_\mathbb{A}(\alpha, \beta)$ the (α, β)-*minimal sets* of $\mathbb{A}$.

Proposition B.1.5. *If β is a cover of α in $\mathrm{Con}(\mathbb{A})$ and $\mathbb{A}$ is finite, then for each (α, β)-minimal set $U \in M_\mathbb{A}(\alpha, \beta)$, there is some $e \in E(\mathbb{A})$ such that $U = e(\mathbb{A})$.*

Proof. Pick any $g \in \mathrm{Pol}_1(\mathbb{A})$ such that $g(\mathbb{A}) = U$ and $g(\beta) \not\subseteq \alpha$. Then since β covers α and $g(\beta) \subseteq \beta$, the congruence generated by $g(\beta) \cup \alpha$ must be β. Thus for any $(x,y) \in \beta$, there must be some $h_i \in \mathrm{Pol}_1(\mathbb{A})$ and $(u_i, v_i) \in g(\beta) \cup \alpha$ such that $x = h_1(u_1)$, $h_i(v_i) = h_{i+1}(u_{i+1})$, and $h_n(v_n) = y$ for some n. Then if we choose $(x,y) \in \beta$ such that $(g(x), g(y)) \notin \alpha$, we see that there must be some i such that $(g(h_i(u_i)), g(h_i(v_i))) \notin \alpha$, and for this i we must have $(u_i, v_i) \in g(\beta)$. Setting $f = g \circ h_i$, we see that $f(g(\beta)) \notin \alpha$ and $f(g(\mathbb{A})) \subseteq U$. Since U is (α, β)-minimal, we must have $f(g(\mathbb{A})) = U$, so $f(U) = U$. Iterating f gives us $e = f^\infty \in E(\mathbb{A})$ with $e(\mathbb{A}) = U$. $\qquad\square$

We will also want to prove similar results for certain other pairs of congruences $\alpha < \beta$. Precisely stating what we are going for requires a bit more work.

Definition B.1.6. A $0,1$-*lattice* is defined to be a lattice with constants 0 and 1 which satisfy $0 \leq x \leq 1$ for all x. Note that every interval $[\![\alpha, \beta]\!]$ in a lattice can be regarded as a $0,1$-lattice, with 0 interpreted as α and 1 interpreted as β.

A $0,1$-*separating homomorphism* is a lattice homomorphism such that $f^{-1}(f(0)) = \{0\}$ and $f^{-1}(f(1)) = 1$.

Definition B.1.7. A *congruence quotient* is defined to be an ordered pair of congruences (α, β) such that $\alpha < \beta$. A congruence quotient (α, β) is called *prime* if β covers α.

A congruence quotient (α, β) is called *tame* if there is some $U \in M_\mathbb{A}(\alpha, \beta)$ and some $e \in E(\mathbb{A})$ such that $e(\mathbb{A}) = U$, and such that the restriction homomorphism

$$[\![\alpha, \beta]\!] \twoheadrightarrow [\![\alpha|_U, \beta|_U]\!]$$

is a $0,1$-separating homomorphism, that is, for $\alpha \leq \gamma \leq \beta$ we have $\gamma|_U = \alpha|_U \implies \gamma = \alpha$ and $\gamma|_U = \beta|_U \implies \gamma = \beta$. An algebra $\mathbb{A}$ is called *tame* if $(0_\mathbb{A}, 1_\mathbb{A})$ is tame.

So far we have shown that every prime quotient on a finite algebra is tame. There is a more general lattice theoretic condition that implies tameness, but first we should try to see what being tame is good for. The first important result is that all of the minimal sets for a tame quotient look the same as each other.

324

Definition B.1.8. If $U, V \subseteq \mathbb{A}$, then we say that U, V are *polynomially isomorphic* in $\mathbb{A}$ if there are unary polynomials $f, g \in \text{Pol}_1(\mathbb{A})$ such that

$$f(U) = V, \ g(V) = U, \ g \circ f|_U = \text{id}_U, \ f \circ g|_V = \text{id}_V .$$

In this case we write $f : U \simeq V$.

Proposition B.1.9. *If $f : U \simeq V$ in $\mathbb{A}$, then $f|_U$ defines an isomorphism from $\mathbb{A}|_U$ to $\mathbb{A}|_V$ (up to polynomial equivalence). Furthermore, for any $\theta \in \text{Con}(\mathbb{A})$ we have $f|_U(\theta|_U) = \theta|_V$.*

Proposition B.1.10. *If $U, V \subseteq \mathbb{A}$ and $\mathbb{A}$ is finite, then $U \simeq V$ iff there are $f, g \in \text{Pol}_1(\mathbb{A})$ such that $f(U) = V$ and $g(V) = U$. If additionally we have $f(\mathbb{A}) = V$, then there is some idempotent $e \in E(\mathbb{A})$ such that $e(\mathbb{A}) = U$.*

Proof. Take $g' = (g \circ f)^{\infty - 1} \circ g$ and $e = g' \circ f = (g \circ f)^\infty$, so $e \in E(\mathbb{A})$. Then $g'(V) = U$, $g' \circ f|_U = \text{id}_U$, $f \circ g'|_V = \text{id}_V$, and if $f(\mathbb{A}) = V$ then $e(\mathbb{A}) = g'(f(\mathbb{A})) = g'(V) = U$. $\qquad\square$

Theorem B.1.11 (Minimal sets for tame quotients [74]). *If (α, β) is a tame congruence quotient on a finite algebra $\mathbb{A}$, then all of the following are true.*

(a) *For all $U, V \in M_{\mathbb{A}}(\alpha, \beta)$, we have $U \simeq V$ in $\mathbb{A}$.*

(b) *For all $U \in M_{\mathbb{A}}(\alpha, \beta)$, there is some $e \in E(\mathbb{A})$ such that $e(\mathbb{A}) = U$, and the restriction homomorphism $[\![\alpha, \beta]\!] \twoheadrightarrow [\![\alpha|_U, \beta|_U]\!]$ is a $0, 1$-separating homomorphism.*

(c) *For all $U \in M_{\mathbb{A}}(\alpha, \beta)$ and $(x, y) \in \beta \setminus \alpha$, there is some $f \in \text{Pol}_1(\mathbb{A})$ such that $f(\mathbb{A}) = U$ and $(f(x), f(y)) \notin \alpha$.*

(d) *For all $U \in M_{\mathbb{A}}(\alpha, \beta)$, β is the transitive closure of $\alpha \cup \bigcup_{g \in \text{Pol}_1(\mathbb{A})} g(\beta|_U)$.*

(e) *For all $U \in M_{\mathbb{A}}(\alpha, \beta)$ and $f \in \text{Pol}_1(\mathbb{A})$ such that $f(\beta|_U) \not\subseteq \alpha$, we have $f(U) \in M_{\mathbb{A}}(\alpha, \beta)$ and $f : U \simeq f(U)$.*

(f) *For any $f \in \text{Pol}_1(\mathbb{A})$ such that $f(\beta) \not\subseteq \alpha$, there is some $U \in M_{\mathbb{A}}(\alpha, \beta)$ such that $f : U \simeq f(U)$.*

(g) *For any $(x, y) \in \beta \setminus \alpha$, there is some $U \in M_{\mathbb{A}}(\alpha, \beta)$ and $e \in E(\mathbb{A})$ such that $e(\mathbb{A}) = U$ and $(e(x), e(y)) \notin \alpha$.*

Proof. (Following [74]) By the definition of tameness, there is *some* $U \in M_{\mathbb{A}}(\alpha, \beta)$ that satisfies (b). We will first show that (c) and (d) hold for this U, and then use this to prove (a), which will imply that (b) is true in general. Then we will use these to prove (e), (f), and (g).

Suppose that (b) holds for U. To prove (c), let $\gamma \in [\![\alpha, \beta]\!]$ be the congruence generated by α and (x, y). Then since $\gamma \neq \alpha$ we must have $\gamma|_U \neq \alpha|_U$ (since restriction is $0, 1$-separating). Since $\gamma|_U = e(\gamma)$, this means that $\gamma \not\subseteq e^{-1}(\alpha)$, so there must be some $g \in \text{Pol}_1(\mathbb{A})$ such that $(g(x), g(y)) \notin e^{-1}(\alpha)$. Taking $f = e \circ g$ proves (c).

To see that (b) implies (d), let $\gamma \in [\![\alpha, \beta]\!]$ be the transitive closure of $\alpha \cup \bigcup_{g \in \text{Pol}_1(\mathbb{A})} g(\beta|_U)$. Then $\gamma|_U = \beta|_U$, so we must have $\gamma = \beta$ (since restriction is $0, 1$-separating).

Now suppose that $U, V \in M_{\mathbb{A}}(\alpha, \beta)$ and that V satisfies (b), (c), (d). Since $U \in M_{\mathbb{A}}(\alpha, \beta)$, there must be some $h \in \text{Pol}_1(\mathbb{A})$ and $(x, y) \in h(\beta) \setminus \alpha$ with $h(\mathbb{A}) = U$, so by (c) applied to V there is some $f \in \text{Pol}_1(\mathbb{A})$ such that $f(\mathbb{A}) = V$ and $(f(x), f(y)) \notin \alpha$. Then from $f(h(\mathbb{A})) \subseteq V$ and

325

$f(h(\beta)) \not\subseteq \alpha$ we have $f(h(\mathbb{A})) = V$ by (α, β)-minimality of V, so $f(U) = V$. Next, by (d) applied to V we see that there must be some $g \in \text{Pol}_1(\mathbb{A})$ such that $g(\beta|_V) \not\subseteq h^{-1}(\alpha)$. Then if $e \in E(\mathbb{A})$ has $e(\mathbb{A}) = V$, we see that $h(g(e(\beta))) \not\subseteq \alpha$ and $h(g(e(\mathbb{A}))) \subseteq U$, so since U is (α, β)-minimal we see that $h(g(V)) = U$. Now we can apply the previous proposition to see that $f : U \simeq V$ and that U satisfies (b) as well.

To prove (e), we use (b) to see that there is some $e \in E(\mathbb{A})$ with $e(\mathbb{A}) = U$, and note that $f(e(\beta)) = f(\beta|_U) \not\subseteq \alpha$, so $f(U) = f(e(\mathbb{A}))$ must contain some minimal $V \in M_{\mathbb{A}}(\alpha, \beta)$. By (a) we see that $|V| = |U|$, so we must in fact have $f(U) = V$ and $f : U \simeq V$.

To prove (f), we apply (d) to any $V \in M_{\mathbb{A}}(\alpha, \beta)$ to see that there is some $g \in \text{Pol}_1(\mathbb{A})$ such that $g(\beta|_V) \not\subseteq f^{-1}(\alpha)$. Then by applying (e) twice we see that we can take $U = g(V)$.

To prove (g), we apply (c) to any $V \in M_{\mathbb{A}}(\alpha, \beta)$ to see that there is some $f \in \text{Pol}_1(\mathbb{A})$ such that $f(\mathbb{A}) = V$ and $(f(x), f(y)) \notin \alpha$. By (f), there is some $U \in M_{\mathbb{A}}(\alpha, \beta)$ such that $f : U \simeq f(U)$, and since $f(U) \subseteq f(\mathbb{A}) = V$, we must have $f(U) = V$ by (α, β)-minimality. Thus there is some $g \in \text{Pol}_1(\mathbb{A})$ such that $g \circ f|_U = \text{id}_U$, and we can take $e = g \circ f$. $\qquad \square$

Corollary B.1.12. *If (α, β) is a tame congruence quotient on a finite algebra $\mathbb{A}$ and $U \in M_{\mathbb{A}}(\alpha, \beta)$, then every unary polynomial f of $\mathbb{A}|_U$ is either a permutation of U or has $f(\beta|_U) \subseteq \alpha|_U$.*

If we restrict to a congruence class of $\beta|_U$, we get an even stronger result.

Definition B.1.13. *If (α, β) is a tame congruence quotient on $\mathbb{A}$ and $U \in M_{\mathbb{A}}(\alpha, \beta)$, then a set $N \subseteq U$ is called an (α, β)-trace in U if N is a congruence class of $\beta|_U$ which is not also a congruence class of $\alpha|_U$.*

We define the *body* of the (α, β)-minimal set U to be the union of the (α, β)-traces, and we define the *tail* of U to be the set of congruence classes of $\beta|_U$ which are also congruence classes of $\alpha|_U$.

Since $\beta|_U \not\subseteq \alpha|_U$ by the definition of an (α, β)-minimal set, we see that every (α, β)-minimal set U has a nonempty body, i.e. there is at least one (α, β)-trace N in U.

Corollary B.1.14. *If (α, β) is a tame congruence quotient on a finite algebra $\mathbb{A}$ and N is an (α, β)-trace, then every unary polynomial f of $\mathbb{A}|_N$ is either a permutation, or has $f(N)$ contained in some congruence class of $\alpha|_N$. In particular, every unary polynomial of $\mathbb{A}|_N/\alpha|_N$ is either a permutation or is constant.*

Definition B.1.15. We say that an algebra is *permutational* or *minimal* if every unary polynomial is either a permutation or is constant.

More generally, if (α, β) is a congruence quotient on $\mathbb{A}$, we say that $\mathbb{A}$ is (α, β)-*minimal* if every unary polynomial f of $\mathbb{A}$ is either a permutation or has $f(\beta) \subseteq \alpha$. Note that in this case, (α, β) is necessarily tame.

The general strategy will be to understand an algebra by first understanding the structure of the traces, and then reconstructing the algebra from the traces. When we apply unary polynomials of $\mathbb{A}$ to (α, β)-traces, the result will often also be an (α, β)-trace.

Corollary B.1.16. *If (α, β) is a tame congruence quotient on a finite algebra $\mathbb{A}$ and N is an (α, β)-trace, then for every unary polynomial f of $\mathbb{A}$ either $f(N)$ is contained in some congruence class of α, or $f(N)$ is another (α, β)-trace and $f : N \simeq f(N)$.*

Proof. This follows directly from Theorem B.1.11(e). $\qquad\square$

For the purpose of stitching together the traces to reconstruct the algebra, we have another consequence of Theorem B.1.11.

Corollary B.1.17. *If (α, β) is a tame congruence quotient on a finite algebra $\mathbb{A}$, then β is the transitive closure of*

$$\alpha \cup \{N^2 \mid N \text{ is an } (\alpha, \beta)\text{-trace}\}.$$

Proof. This follows from Theorem B.1.11(d) and the previous corollary, once we note that for each (α, β)-minimal set U, every congruence class of $\beta|_U$ is either a congruence class of $\alpha|_U$ or is an (α, β)-trace. $\qquad\square$

Proposition B.1.18. *If (α, β) is a prime congruence quotient of a finite algebra $\mathbb{A}$ (i.e., if β is a cover of α), then all of the (α, β)-traces are polynomially isomorphic in $\mathbb{A}$.*

Proof. Let U be any (α, β)-minimal set. By Theorem B.1.11(a) it's enough to show that any pair of (α, β)-traces N, K contained in U are polynomially isomorphic in $\mathbb{A}|_U$. By Lemma B.1.3, the restriction homomorphism $[\![\alpha, \beta]\!] \to [\![\alpha|_U, \beta|_U]\!]$ is surjective, so $\beta|_U$ covers $\alpha|_U$ in $\mathbb{A}|_U$. Then since N is not contained in a congruence class of α, the congruence of $\mathbb{A}|_U$ generated by $\alpha|_U \cup N^2$ must be $\beta|_U$. In particular, there must be some unary polynomial $f \in \mathrm{Pol}_1(\mathbb{A}|_U)$ such that for $f(N) \subseteq K$ and $f(N)$ is not contained in any congruence class of α. Then f must be a permutation of U, and so we have $f : N \simeq K$. $\qquad\square$

Tameness can also be derived from some of its consequences.

Proposition B.1.19. *If $\alpha < \beta \in \mathrm{Con}(\mathbb{A})$ and there is some finite (α, β)-minimal set U which satisfies Theorem B.1.11(c) and B.1.11(d), then (α, β) is tame.*

Proof. If we make a digraph on $(x, y) \in \beta|_U \setminus \alpha|_U$ with an edge from (x, y) to (u, v) whenever there is some $f \in \mathrm{Pol}_1(\mathbb{A})$ with $f(\mathbb{A}) = U$ and $f(x) = u, f(y) = v$, then Theorem B.1.11(c) for U implies that every vertex in this digraph has outdegree at least one, so there must be a directed cycle. By composing the unary polymorphisms corresponding to the edges of this directed cycle, we find an $f \in \mathrm{Pol}_1(\mathbb{A})$ such that $f(\mathbb{A}) \subseteq U$ and $f(x) = x, f(y) = y$ for some $(x, y) \in \beta|_U \setminus \alpha|_U$. Taking $e = f^\infty$, we see that $e \in E(\mathbb{A})$ with $e(\mathbb{A}) \subseteq U$ and $e(\beta) \not\subseteq \alpha$. Finally, Theorem B.1.11(c) and B.1.11(d) for U directly imply that the restriction homomorphism $[\![\alpha, \beta]\!] \twoheadrightarrow [\![\alpha|_U, \beta|_U]\!]$ is $0, 1$-separating. $\qquad\square$

Many of the previous results simplify if $\alpha = 0_\mathbb{A}$. To relate the general case to that situation, we need to know how tameness behaves when we pass to a quotient.

Proposition B.1.20. *If $\delta \leq \alpha < \beta$ are congruences on a finite algebra $\mathbb{A}$, then (α, β) is tame on $\mathbb{A}$ iff $(\alpha/\delta, \beta/\delta)$ is tame on $\mathbb{A}/\delta$. If (α, β) is tame, then we have*

$$M_{\mathbb{A}/\delta}(\alpha/\delta, \beta/\delta) = \{U/\delta \mid U \in M_\mathbb{A}(\alpha, \beta)\}.$$

In particular, the $(\alpha/\delta, \beta/\delta)$-traces are exactly the quotients of the (α, β)-traces by δ.

Proof. (Following [74]) For any unary polynomial $f \in \mathrm{Pol}_1(\mathbb{A})$, we have $f(\beta) \subseteq \alpha$ iff $f(\beta/\delta) \subseteq \alpha/\delta$ in $\mathbb{A}/\delta$, and for any $\gamma, \gamma' \in [\![\alpha, \beta]\!]$ we have $\gamma = \gamma' \iff \gamma/\delta = \gamma'/\delta$ and similarly for restrictions. The challenge is showing that the minimal sets correspond.

First we show that if $U \in M_\mathbb{A}(\alpha, \beta)$ and there is an $e \in E(\mathbb{A})$ with $e(\mathbb{A}) = U$, then $U/\delta \in M_{\mathbb{A}/\delta}(\alpha/\delta, \beta/\delta)$. To see this, suppose that some $f \in \mathrm{Pol}_1(\mathbb{A})$ has $f(\beta) \not\subseteq \alpha$ and $f(\mathbb{A}/\delta) \subseteq U/\delta$. Then $e(f(\mathbb{A})) \subseteq U$ and $e(f(\beta)) \not\subseteq \alpha$, so $e(f(\mathbb{A})) = U$, so $f(\mathbb{A}/\delta) = U/\delta$. This shows that if (α, β) is tame then $(\alpha/\delta, \beta/\delta)$ is tame.

Now suppose $(\alpha/\delta, \beta/\delta)$ is tame, and let U be any (α, β)-minimal set. Pick $f \in \mathrm{Pol}_1(\mathbb{A})$ with $f(\mathbb{A}) = U$ and $f(\beta) \subseteq \alpha$. By Theorem B.1.11(f) we see that there is some $V \in M_{\mathbb{A}/\delta}(\alpha/\delta, \beta/\delta)$ such that $f : V \simeq f(V)$. Pick $g \in \mathrm{Pol}_1(\mathbb{A})$ such that $g : f(V) \to V$ inverts $f : V \to f(V)$, then by iterating $f \circ g$ we get an idempotent $e \in E(\mathbb{A})$ with $e(f(V)) = f(V)$. Thus $e(\beta) \not\subseteq \alpha$, and from $e(\mathbb{A}) \subseteq f(\mathbb{A}) = U$ we get $e(\mathbb{A}) = U$. Then by the previous paragraph we see that U/δ is $(\alpha/\delta, \beta/\delta)$-minimal, which allows us to conclude that the restriction homomorphism $[\![\alpha, \beta]\!] \twoheadrightarrow [\![\alpha|_U, \beta|_U]\!]$ is $0, 1$-separating.

To finish, we need to show that any $V \in M_{\mathbb{A}/\delta}(\alpha/\delta, \beta/\delta)$ is a quotient of an (α, β)-minimal set when (α, β) is tame. Pick any $U \in M_\mathbb{A}(\alpha, \beta)$, then since U/δ is $(\alpha/\delta, \beta/\delta)$-minimal we can apply Theorem B.1.11(a) to see that there is an $f \in \mathrm{Pol}_1(\mathbb{A})$ with $f : U/\delta \simeq V$. Then $f(\beta) \not\subseteq \alpha$, so by Theorem B.1.11(e) we have $f(U) \in M_\mathbb{A}(\alpha, \beta)$, and $V = f(U)/\delta$. $\qquad\square$

Proposition B.1.21. *If $\alpha \le \gamma < \beta$ and $(\alpha, \beta), (\gamma, \beta)$ are both tame quotients on a finite algebra $\mathbb{A}$, then $M_\mathbb{A}(\alpha, \beta) = M_\mathbb{A}(\gamma, \beta)$.*

Proof. If $U \in M_\mathbb{A}(\alpha, \beta)$, then $\beta|_U \not\subseteq \gamma|_U$ since the restriction map is $0, 1$-separating (by Theorem B.1.11(b)). If $f(\mathbb{A}) \subseteq U$ and $f(\beta) \not\subseteq \gamma$, then we have $f(\beta) \not\subseteq \alpha$, so $f(\mathbb{A}) = U$ by (α, β)-minimality, so $U \in M_\mathbb{A}(\gamma, \beta)$. Conversely, if $V \in M_\mathbb{A}(\gamma, \beta)$, then by Theorem B.1.11(a) we have $V \simeq U$ for some $U \in M_\mathbb{A}(\alpha, \beta)$, so $V \in M_\mathbb{A}(\alpha, \beta)$ by Theorem B.1.11(e). $\qquad\square$

Recall that two intervals are perspective, written $[\![\alpha, \beta]\!] \searrow [\![\gamma, \delta]\!]$, if $\alpha \wedge \delta = \gamma$ and $\alpha \vee \delta = \beta$.

Proposition B.1.22. *If $[\![\alpha, \beta]\!] \searrow [\![\gamma, \delta]\!]$ in $\mathrm{Con}(\mathbb{A})$, then $M_\mathbb{A}(\alpha, \beta) = M_\mathbb{A}(\gamma, \delta)$ and $M_\mathbb{A}(\gamma, \beta) \subseteq M_\mathbb{A}(\gamma, \alpha) \cup M_\mathbb{A}(\gamma, \delta)$.*

Proof. For any $f \in \mathrm{Pol}_1(\mathbb{A})$, we have $f(\beta) = f(\alpha \vee \delta) \subseteq \alpha$ iff $f(\alpha) \cup f(\delta) \subseteq \alpha$ iff $f(\delta) \subseteq \alpha$ iff $f(\delta) \subseteq \alpha \cap \delta = \gamma$, so $M_\mathbb{A}(\alpha, \beta) = M_\mathbb{A}(\gamma, \delta)$. Similarly, we have $f(\beta) \subseteq \gamma$ iff $f(\beta) \subseteq \alpha$ and $f(\beta) \subseteq \delta$, so $M_\mathbb{A}(\gamma, \beta) \subseteq M_\mathbb{A}(\alpha, \beta) \cup M_\mathbb{A}(\delta, \beta) = M_\mathbb{A}(\gamma, \alpha) \cup M_\mathbb{A}(\gamma, \delta)$. $\qquad\square$

Corollary B.1.23. *If α, β are congruences on a finite algebra and the interval $[\![\alpha, \beta]\!]$ is isomorphic to the diamond lattice $\mathcal{M}_n$ for some $n \ge 3$, then (α, β) is tame and every congruence quotient contained in $[\![\alpha, \beta]\!]$ has the same collection of minimal sets.*

Proposition B.1.24. *If (α, β) is a tame congruence quotient on a finite algebra $\mathbb{A}$ and U is an (α, β)-minimal set, then for an $\gamma' < \delta' \in [\![\alpha|_U, \beta|_U]\!]$, there are lifts $\gamma < \delta \in [\![\alpha, \beta]\!]$ such that $\gamma|_U = \gamma', \delta|_U = \delta'$, with (γ, δ) a tame quotient. For any such γ, δ we have $M_\mathbb{A}(\gamma, \delta) = M_\mathbb{A}(\alpha, \beta)$.*

Proof. Since any unary polynomial which collapses β into α necessarily collapses δ into γ for any $\gamma, \delta \in [\![\alpha, \beta]\!]$, we see that U is (γ, δ)-minimal for any $\gamma < \delta$ which restrict to γ', δ', so we just need to ensure that the restriction homomorphism from $[\![\gamma, \delta]\!]$ is $0, 1$-separating. Since restriction to U is a lattice homomorphism, we can take γ to be maximal among congruences which restrict to γ' and

are $\leq \beta$, and take δ to be minimal among congruences which restrict to δ' and are $\geq \gamma$. By Theorem B.1.11(a) and B.1.11(e) every (γ, δ)-minimal set V has $U \simeq V$ and so is also (α, β)-minimal, and similarly every (α, β)-minimal set is (γ, δ)-minimal. □

Some basic examples to keep in mind follow.

Example B.1.1. If $\mathbb{A}$ is a finite lattice and $\alpha < \beta \in \mathrm{Con}(\mathbb{A})$, then the (α, β)-minimal sets all have the form $\{a, b\}$ with $a < b$ and $(a, b) \in \beta \setminus \alpha$, and any pair $\{a, b\}$ which satisfies those conditions and additionally has b covering a is an (α, β)-minimal set. Each such set with b covering a is the image of the idempotent unary polynomial $x \mapsto (x \wedge b) \vee a$, however, in order for the restriction homomorphism to be $0, 1$-separating, β must be a cover of α. Thus the tame congruence quotients of $\mathbb{A}$ are exactly the same as the prime quotients. Additionally, every (α, β)-minimal set consists of just a single (α, β)-trace (i.e., the minimal sets have no tails), and every trace is the image of some pair $\{a, b\}$ with b covering a under some unary polynomial of $\mathbb{A}$, and is polynomially equivalent to a two element lattice.

Example B.1.2. If $\mathbb{A}$ is a finite module over a ring $\mathbb{R}$ (which we may assume acts faithfully on $\mathbb{A}$ without loss of generality) and $\alpha < \beta \in \mathrm{Con}(\mathbb{A})$, then we can represent the congruences α, β by the submodules $\mathbb{M}_\alpha = 0/\alpha$, $\mathbb{M}_\beta = 0/\beta$ of $\mathbb{A}$. Every unary polynomial f of $\mathbb{A}$ has the form $f : x \mapsto rx + c$, and we have $f(\beta) \subseteq \alpha$ iff $r\mathbb{M}_\beta \subseteq \mathbb{M}_\alpha$. The set of $r \in \mathbb{R}$ such that $r\mathbb{M}_\beta \subseteq \mathbb{M}_\alpha$ is called the *annihilator* of $\mathbb{M}_\beta/\mathbb{M}_\alpha$, and forms a (two-sided) ideal $\mathbb{I}$ of $\mathbb{R}$. Then $\mathbb{M}_\beta/\mathbb{M}_\alpha$ is a module over $\mathbb{R}/\mathbb{I}$, and $\mathbb{R}/\mathbb{I}$ acts faithfully on $\mathbb{M}_\beta/\mathbb{M}_\alpha$.

We will show that (α, β) is a tame congruence quotient if and only if $\mathbb{I}$ is maximal among two-sided ideals of $\mathbb{R}$, that is, iff $\mathbb{R}/\mathbb{I}$ is a simple ring. By the classification of finite simple rings, this is equivalent to proving that $\mathbb{R}/\mathbb{I}$ is isomorphic to a matrix ring $M_n(\mathbb{F}_{p^k})$ over a finite field $\mathbb{F}_{p^k}$ for some n and some prime power p^k. Additionally, we will show that $\mathbb{M}_\beta/\mathbb{M}_\alpha$ is one of the modules $\mathbb{F}_{p^k}^{n \times m}$ for some m, with the action of $\mathbb{R}/\mathbb{I}$ on $\mathbb{M}_\beta/\mathbb{M}_\alpha$ given by matrix multiplication. To prove this, it is simpler to think only about the module $\mathbb{M}_\beta/\mathbb{M}_\alpha$ - note that $\mathbb{M}_\beta/\mathbb{M}_\alpha$ is a tame algebra, with $(e\mathbb{A} \cap \mathbb{M}_\beta)/\mathbb{M}_\alpha = e\mathbb{M}_\beta/\mathbb{M}_\alpha$ as a $(0, 1)$-minimal set, for any $e \in \mathbb{R}$ such that $e^2 = e$ and $e\mathbb{A}$ is an (α, β)-minimal set.

Proposition B.1.25. *. Suppose that $\mathbb{M}$ is a finite module over a finite ring $\mathbb{R}$ which acts faithfully on $\mathbb{M}$, and suppose that $\mathbb{M}$ is a tame algebra. Then $\mathbb{R}$ is isomorphic to a matrix ring $M_n(\mathbb{F}_{p^k})$ over a finite field $\mathbb{F}_{p^k}$ for some n and some prime power p^k, and $\mathbb{M}$ is $\mathbb{F}_{p^k}^{n \times m}$ for some m.*

Proof. By Corollary B.1.14, every $(0_{\mathbb{M}}, 1_{\mathbb{M}})$-trace N has $\mathbb{M}|_N$ a permutational algebra. If $N = e\mathbb{M}$ for some $e \in E(\mathbb{A})$, the restriction $\mathrm{Pol}(\mathbb{M})|_N$ consists of the linear functions with coefficients in the ring $e\mathbb{R}e$, so we see that every nonzero element of the ring $e\mathbb{R}e$ is invertible, that is, $e\mathbb{R}e$ is a division ring. Since every finite division ring is a field by Wedderburn's little theorem, we have $e\mathbb{R}e \cong \mathbb{F}_{p^k}$ for some finite field $\mathbb{F}_{p^k}$ (note, the rest of the argument still works over a division ring rather than a field).

We claim that we can pick $e_1, ..., e_n \in \mathbb{R}$ idempotent such that $\sum_i e_i = 1$, $e_i e_j = e_j e_i = 0$ for $i \neq j$, and each $e_i\mathbb{M}$ is $(0_{\mathbb{M}}, 1_{\mathbb{M}})$-minimal. Suppose that $e_1, ..., e_n$ is a maximal collection of idempotents such that $e_i e_j = e_j e_i = 0$ for $i \neq j$ and each $e_i\mathbb{M}$ is $(0_{\mathbb{M}}, 1_{\mathbb{M}})$-minimal, and let $f = 1 - \sum_i e_i$ (that such a maximal set exists and is finite follows from the fact that $e_i\mathbb{M} \cap e_j\mathbb{M} = \emptyset$ for $i \neq j$, since $e_i e_j = 0$). Then we have $f^2 = f$, and if $f \neq 0$ then there must be some $(0_{\mathbb{M}}, 1_{\mathbb{M}})$-minimal set $U \subseteq f\mathbb{M}$. If $e' \in E(\mathbb{M})$ has $e'\mathbb{M} = U$, then we set $e_{n+1} = e'f$, and note that we have $e_{n+1}\mathbb{M} = e'\mathbb{M} = U$, and $e_{n+1}e_i = e_i e_{n+1} = 0$ for each $i \leq n$, contradicting the maximalilty of the

collection $e_1, ..., e_n$. Therefore we must have $f = 0$, that is, $\sum_i e_i = 1$. Note that every element $x \in \mathbb{M}$ has a unique decomposition

$$x = \sum_i e_i x_i,$$

so as a group, $\mathbb{M}$ is the direct sum of the $e_i \mathbb{M}$s.

By Theorem B.1.11(a), for each pair $i, j \leq n$ we have $e_i \mathbb{M} \simeq e_j \mathbb{M}$. Pick $f_i : e_1 \mathbb{M} \simeq e_i \mathbb{M}$ for each i, and pick inverses $g_i : e_i \mathbb{M} \simeq e_1 \mathbb{M}$ to each f_i (with $f_1 = g_1 = e_1$ for $i = 1$). Then for any i, j, define the matrix element e_{ij} by

$$e_{ij} = f_i g_j e_j,$$

and note that each e_{ij} is an isomorphism $e_{ij} : e_j \mathbb{M} \simeq e_i \mathbb{M}$, with $e_i e_{ij} = e_{ij}$ and $e_{ij} e_j = e_{ij}$, with $e_{ij} e_{ji} = e_{ii} = e_i$, with $e_{ij} e_{jk} = e_{ik}$, and $e_{ij} e_{kl} = 0$ for $j \neq k$. For each $r_1 \in e_1 \mathbb{R} e_1$ we can additionally define the corresponding scalar $r \in \mathbb{R}$ by

$$r_1 \in e_1 \mathbb{R} e_1 \mapsto r = \sum_i e_{i1} r_1 e_{1i},$$

and we identify the set of such scalars r with $\mathbb{F}_{p^k}$, noting that $r e_{ij} = e_{ij} r$ for all $r \in \mathbb{F}_{p^k}$ and $i, j \leq n$, and that the multiplication in $\mathbb{F}_{p^k}$ is the same as the multiplication in $e_1 \mathbb{R} e_1$. We claim that every element $m \in \mathbb{R}$ can be written uniquely in the form

$$m = \sum_{i,j} r_{i,j} e_{ij}$$

for some $r_{i,j} \in \mathbb{F}_{p^k}$. To prove this, note that since $\sum_i e_i = 1$, we have

$$m = \sum_{i,j} e_i m e_j,$$

and each $e_i m e_j$ defines a map $e_j \mathbb{M} \to e_i \mathbb{M}$. If we define the element $r_{i,j} \in \mathbb{F}_{p^k}$ by

$$r_{i,j} := \sum_k e_{ki} m e_{jk},$$

then we have

$$r_{i,j} e_{ij} = \sum_k e_{ki} m e_{jk} e_{ij} = e_{ii} m e_{ji} e_{ij} = e_i m e_j,$$

so $m = \sum_{i,j} r_{i,j} e_{ij}$. For the uniqueness, note that $\sum_{i,j} r_{i,j} e_{ij} = 0$ implies that each $r_{i,j} e_{ij} = 0$, and if $r_{i,j} \neq 0$ then $r_{i,j}$ is invertible, since $\mathbb{F}_{p^k}$ is a field. Thus we have an explicit isomorphism $M_n(\mathbb{F}_{p^k}) \cong \mathbb{R}$. Finally, $e_1 \mathbb{M}$ is a vector space of some dimension m over $\mathbb{F}_{p^k}$, and since $\mathbb{M}$ is the direct sum of n copies of $e_1 \mathbb{M}$ as a vector space over $\mathbb{F}_{p^k}$ we have $\mathbb{M} \cong \mathbb{F}_{p^k}^{n \times m}$, with the action of $\mathbb{R}$ on $\mathbb{M}$ corresponding to matrix multiplication. $\qquad \square$

For the sake of completeness, we include Witt's proof of Wedderburn's little theorem here.

Theorem B.1.26 (Wedderburn's little theorem)**.** *If $\mathbb{R}$ is a finite division ring, then $\mathbb{R}$ is a field.*

Proof. (Following Witt [139]) Let $\mathbb{R}^\times$ be the group of nonzero elements of $\mathbb{R}$, and let $Z(\mathbb{R}^\times)$ be the center of the group $\mathbb{R}^\times$. Then $\mathbb{F} = Z(\mathbb{R}) = Z(\mathbb{R}^\times) \cup \{0\}$ is a finite field, of some prime power order $q = p^k$. Since $\mathbb{R}$ is an $\mathbb{F}$-algebra, $\mathbb{R}$ is in particular a vector space over $\mathbb{F}$ of some dimension n, so $|\mathbb{R}| = q^n$ and $|\mathbb{R}^\times| = q^n - 1$.

We consider the conjugation action of $\mathbb{R}^\times$ on itself: if $x \in \mathbb{R} \setminus \mathbb{F}$, then the centralizer $C_\mathbb{R}(x) = \{r \in \mathbb{R} \mid rx = xr\}$ is a proper $\mathbb{F}$-subalgebra of $\mathbb{R}$, so $|C_\mathbb{R}(x)| = q^k$ for some $k < n$, and since $\mathbb{R}$ can be thought of as a module over the division ring $C_\mathbb{R}(x)$, we have $k \mid n$. Then the conjugacy class of x in $\mathbb{R}^\times$ has size $\frac{q^n - 1}{q^k - 1}$, so we have

$$q^n - 1 = |\mathbb{R}^\times| = |Z(\mathbb{R}^\times)| + \sum_{\text{conj. classes of } \mathbb{R} \setminus \mathbb{F}} \frac{q^n - 1}{q^{k_i} - 1} = q - 1 + \sum_{\text{conj. classes of } \mathbb{R} \setminus \mathbb{F}} \frac{q^n - 1}{q^{k_i} - 1}.$$

If we let $\Phi_n(x)$ be the nth cyclotomic polynomial, then we have $\Phi_n(q) \mid \frac{q^n - 1}{q^{k_i} - 1}$ for each conjugacy class, so $q - 1$ must be a multiple of $\Phi_n(q)$. However, $|\Phi_n(q)|$ is the product of $|q - \zeta|$ over various nth roots of unity ζ, so $|\Phi_n(q)| > q - 1$ for $n > 1$, a contradiction. $\qquad\square$

B.2   Pálfy's classification of finite permutational algebras: the five types

In the last section we proved that if (α, β) is a tame congruence quotient of a finite algebra $\mathbb{A}$, then for every (α, β)-trace N the restriction $\mathbb{A}|_N$ is permutational, i.e. every unary polynomial of $\mathbb{A}|_N$ is either a constant or a permutation. In [116], Pálfy gave a complete classification of the finite permutational algebras (up to polynomial equivalence), which was one of the key ingredients needed for tame congruence theory.

The classification splits into two very different cases: algebras of size 2, and algebras of size ≥ 3. Since every unary operation on a set of size 2 is either constant or is a permutation, the classification of permutational algebras on a set of size 2 is the same as the classification of *all* algebras on a set of size 2, up to polynomial equivalence. There turn out to be exactly 7 of these. On the other hand, the number of polynomial clones on any set of size ≥ 3 is uncountable [146] - but as we will see, the permutational algebras on a set of size ≥ 3 are all either unary or affine algebras, so they end up being much simpler than general algebras.

We start by giving some definitions in order to rule out the least interesting case - the case of unary operations only.

Definition B.2.1. An operation f of arity n *depends on* its ith input if there is some tuple $a_1, ..., a_n$ and some b_i such that
$$f(a_1, ..., a_n) \neq f(a_1, ..., a_{i-1}, b_i, a_{i+1}, ..., a_n).$$

An operation f is *essentially unary* if it only depends on one of its inputs - equivalently, f is essentially unary if it can be written as the composition of a projection π_i and a unary operation.

If f does not depend on its ith input, then we can express f in terms of the function we get by replacing its ith input by some other input, such as its first input. So there is no need to ever think too deeply about functions which do not depend on all their inputs. In order to gain a foothold, it is helpful to start by considering the case of a binary operation which depends on all of its inputs

- for this, we will replace one of the inputs of a higher arity polynomial with some constant to get a lower arity polynomial which also depends on all its inputs.

The next result is much stronger than what we will need: all we really need is the fact that if f depends on at least two of its inputs, then there is a way to plug in constants for some subset of the inputs to f to get a polynomial in two variables that depends on both of its inputs.

Proposition B.2.2 (Salomaa [129])**.** *If a polynomial f of arity n depends on all of its inputs, then it is possible to substitute a constant for one of its inputs to get a polynomial of arity $n-1$ which also depends on all of its inputs.*

In fact, if $n \geq 2$, then it is possible to find at least two different inputs to f where constants can be substituted to get polynomials depending on all $n-1$ of their inputs.

Proof. Following [74], we write $f[a,i]$ for the polynomial we get by substituting a for the ith input of f. Suppose that for some a and i,j,k, $f[a,i]$ does not depend on the jth input but does depend on the kth input. Then for *every* b we see that $f[b,j]$ depends on the kth input, by considering the case where we plug in a in the ith input and b in the jth input. Additionally, since f depends on all its inputs there must be some a' such that $f[a',i]$ depends on its jth input, so there must be some b such that $f[b,j]$ depends on the ith coordinate (consider plugging in a tuple with an a' in the ith input such that varying the jth input changes the value, and note that if we change a' to a in the ith position then varying the jth input no longer changes the value of f). Thus if $f[a,i]$ does not depend on its jth input, then there is some b such that $f[b,j]$ depends on a strictly larger subset of its inputs than $f[a,i]$ does, which proves the first claim.

For the second claim, note that for each $i \leq n$ and each $j \neq i$, there is some a such that $f[a,j]$ depends on the ith input, as long as f depends on its ith input. Then if we choose a pair a,j such that $j \neq i$, $f[a,j]$ depends on the ith input, and $f[a,j]$ depends on as many inputs as possible subject to the previous constraints, then the argument of the previous paragraph shows that $f[a,j]$ must depend on all of its inputs. $\square$

Lemma B.2.3. *Suppose that $f \in \mathrm{Pol}_2(\mathbb{A})$ is a binary polynomial of a finite permutational algebra $\mathbb{A}$ which depends on both of its inputs, and suppose that $|\mathbb{A}| \geq 3$. Then f is a quasigroup operation, that is, every unary polynomial of the form $f(a, \cdot)$ or $f(\cdot, b)$ is a permutation.*

Proof. Suppose for the sake of contradiction that f depends on both of its inputs, but that there is some a such that $f(a, \cdot)$ is constant, with $f(a,y) = e$ for all $y \in \mathbb{A}$. Since f depends on its second coordinate, there must be some $a' \neq a$ such that $f(a', \cdot)$ is a permutation, which implies that there is some $b \in \mathbb{A}$ such that $f(a', b) = e$. Then since $f(a,b) = e = f(a',b)$, we must have $f(x,b) = e$ for all $x \in \mathbb{A}$ as well.

For any $a' \neq a$, if $f(a', \cdot)$ is constant, then since $f(a',b) = e$, we must have $f(a',y) = e$ for all $x \in \mathbb{A}$, and then for each y from $f(a',y) = e = f(a,y)$ we conclude that $f(x,y) = e$ for all x, so f is constant. This contradicts the assumption that f depends on its inputs, so for all $a' \neq a$ the unary polynomial $f(a', \cdot)$ must be a permutation.

So far we have not used the fact that $|\mathbb{A}| \geq 3$, and we have not fully exploited the fact that $\mathbb{A}$ is finite and permutational. For this, we iterate f on its second argument: define $f^1 = f$, and for each n define $f^{n+1}(x,y)$ by

$$f^{n+1}(x,y) = f(x, f^n(x,y)),$$

and take $f^\infty(x,y) = \lim_{n \to \infty} f^{n!}(x,y)$, so

$$f^\infty(x,y) = f^\infty(x, f^\infty(x,y)).$$

Then $f^\infty(a, \cdot)$ is constant, while for $a' \neq a$ we have $f^\infty(a', y) = y$ for all $y \in \mathbb{A}$ since each $f(a', \cdot)$ is a permutation. But then for any distinct $a', a'' \neq a$, we have $f^\infty(a', y) = y = f^\infty(a'', y)$, so $f^\infty(\cdot, y)$ must be constant for all y, and in particular $f^\infty(a, y) = y$ for all y, which contradicts the fact that $f^\infty(a, \cdot)$ is constant. $\qquad\square$

Corollary B.2.4. *Suppose $\mathbb{A}$ is a finite permutational algebra with $|\mathbb{A}| \geq 3$, and suppose that some operation of $\mathbb{A}$ is not essentially unary. Then $\mathbb{A}$ has a Mal'cev polynomial $p(x, y, z)$.*

Proof. This follows from the previous lemma and Proposition 1.7.10. $\qquad\square$

Corollary B.2.5. *Suppose $\mathbb{A}$ is a finite permutational algebra with $|\mathbb{A}| \geq 3$, and suppose $f \in \mathrm{Pol}_n(\mathbb{A})$ has $f(a_1, ..., a_n) = f(a_1, ..., a_{i-1}, b_i, a_{i+1}, ..., a_n)$ for some $a_1, ..., a_n$ and some $b_i \neq a_i$. Then f does not depend on its ith coordinate.*

Proof. We will show that for any $j \neq i$, any a'_j, and any b'_i, we have

$$f(a_1, ..., a_{j-1}, a'_j, a_{j+1}, ..., a_n) = f(a_1, ..., a_{i-1}, b'_i, a_{i+1}, ..., a_{j-1}, a'_j, a_{j+1}, ..., a_n).$$

For this, we define a two-variable polynomial from f by substituting the kth input of f with a_k for all $k \neq i, j$, and apply the previous lemma to this two variable polynomial to see that it can't depend on its ith input. Applying this repeatedly, we can mutate the tuple $a_1, ..., a_n$ into any tuple $a'_1, ..., a'_n$, so f does not depend on its ith coordinate. $\qquad\square$

Lemma B.2.6. *Suppose $\mathbb{A}$ is a finite permutational algebra with $|\mathbb{A}| \geq 3$, $f, g \in \mathrm{Pol}_n(\mathbb{A})$, and suppose that for some $0 \in \mathbb{A}$ we have $f(x_1, ..., x_n) = g(x_1, ..., x_n)$ for all $x_1, ..., x_n \in \mathbb{A}$ such that all but one x_i is 0. Then $f = g$.*

Proof. If every operation of $\mathbb{A}$ is essentially unary, then this is obvious. Otherwise, let $p \in \mathrm{Pol}_3(\mathbb{A})$ be the Mal'cev operation from Corollary B.2.4. Then the polynomial

$$h(x_1, ..., x_n) = p(f(x_1, ..., x_n), g(x_1, ..., x_n), 0)$$

is 0 whenever $f(x_1, ..., x_n) = g(x_1, ..., x_n)$, and since any Mal'cev operation must depend on its second input, we can apply Corollary B.2.5 to $p(x, y, z)$ to see that $h(x_1, ..., x_n) = 0$ if and only if $f(x_1, ..., x_n) = g(x_1, ..., x_n)$. For any input i, since we have

$$h(0, ..., 0, x_i, 0, ..., 0) = 0$$

for all x_i, we can apply Corollary B.2.5 to see that h does not depend on any of its inputs, so h must be constantly 0, which implies that $f = g$. $\qquad\square$

Theorem B.2.7 (Pálfy [116]). *Suppose $\mathbb{A}$ is a finite permutational algebra with $|\mathbb{A}| \geq 3$, and suppose that some operation of $\mathbb{A}$ is not essentially unary. Then $\mathbb{A}$ is affine, and in fact $\mathbb{A}$ is polynomially equivalent to a vector space over a finite field.*

Proof. Let $p(x, y, z) \in \mathrm{Pol}_3(\mathbb{A})$ be the Mal'cev operation from Corollary B.2.4, pick an element to call 0 in $\mathbb{A}$, and define a binary polynomial $+ \in \mathrm{Pol}_2(\mathbb{A})$ by

$$x + y = p(x, 0, y).$$

Then since p is Mal'cev, we have $0+x = x = x+0$ for all x, so by Lemma B.2.6 we have $x+y = y+x$ for all x, y. Similarly, from

$$0 + (0 + x) = x = (0 + 0) + x, \quad 0 + (x + 0) = x = (0 + x) + 0, \quad x + (0 + 0) = x = (x + 0) + 0,$$

we can apply Lemma B.2.6 to conclude that $x + (y + z) = (x + y) + z$ for all x, y, z. Since $x + 0 = x = 0 + x$ for all x, $+$ depends on both of its arguments, so by Lemma B.2.3, we see that every element $x \in \mathbb{A}$ has an inverse $-x$ such that $x + (-x) = 0$. Thus $+$ defines an abelian group structure on $\mathbb{A}$ with identity element 0.

For any $f \in \mathrm{Pol}_1(\mathbb{A})$, if $f(0) = c$, then we can define $r \in \mathrm{Pol}_1(\mathbb{A})$ by $r(x) = f(x) - c$, so that $r(0) = 0$. We will show that any such r distributes over addition: since

$$r(x + 0) = r(x) = r(x) + r(0), \quad r(0 + y) = r(y) = r(0) + r(y),$$

we can apply Lemma B.2.6 to conclude that $r(x + y) = r(x) + r(y)$ for all x, y. Thus every unary polynomial $f \in \mathrm{Pol}_1(\mathbb{A})$ can be written in the form $f(x) = r(x) + c$, where r distributes over addition and takes 0 to 0. Letting $\mathbb{F}$ be the ring of $r \in \mathrm{Pol}_1(\mathbb{A})$ such that $r(0) = 0$, we see that every nonzero element of $\mathbb{F}$ is invertible, so $\mathbb{F}$ is a finite division ring, and therefore $\mathbb{F}$ is a finite field by Wedderburn's little theorem B.1.26.

Now suppose $f \in \mathrm{Pol}_n(\mathbb{A})$ is any n-ary polynomial. Then if we define unary polynomials r_i by

$$r_i(x_i) = f(0, ..., 0, x_i, 0, ..., 0) - f(0, ..., 0),$$

then each r_i has $r_i(0) = 0$, so $r_i \in \mathbb{F}$ for all i. If we define $g \in \mathrm{Pol}_n(\mathbb{A})$ by

$$g(x_1, ..., x_n) = r_1(x_1) + \cdots + r_n(x_n) + f(0, ..., 0),$$

then we can apply Lemma B.2.6 to see that $f = g$, so every operation of $\mathbb{A}$ is linear over the finite field $\mathbb{F}$. $\qquad \square$

To complete the classification, we just need to classify the polynomial clones on the two-element set $\{0, 1\}$. We use $\neg$ to denote the unary negation operation on $\{0, 1\}$, $\oplus$ to denote xor, and of course $\wedge, \vee$ to denote and and or.

Proposition B.2.8. *If $\mathbb{A}$ has underlying set $\{0, 1\}$, then $\mathbb{A}$ is polynomially complete iff $\neg, \wedge \in \mathrm{Pol}(\mathbb{A})$. Additionally, we have $\oplus \in \mathrm{Pol}(\mathbb{A}) \implies \neg \in \mathrm{Pol}(\mathbb{A})$.*

Lemma B.2.9. *If $\mathbb{A}$ has underlying set $\{0, 1\}$ and if there is some $f \in \mathrm{Pol}(\mathbb{A})$ which is not monotone, then the unary negation $\neg$ is a polynomial of $\mathbb{A}$.*

Proof. If f is not monotone, then there is some tuple $a_1, ..., a_n \in \{0, 1\}$ and some i such that

$$f(a_1, ..., a_{i-1}, 0, a_{i+1}, ..., a_n) > f(a_1, ..., a_{i-1}, 1, a_{i+1}, ..., a_n).$$

Then the left hand side of the displayed inequality must be 1 and the right hand side must be 0, so we have

$$\neg(x) = f(a_1, ..., a_{i-1}, x, a_{i+1}, ..., a_n). \qquad \square$$

Lemma B.2.10. *If $\mathbb{A}$ has underlying set $\{0, 1\}$ and $\oplus \in \mathrm{Pol}(\mathbb{A})$, and if there is any $f \in \mathrm{Pol}(\mathbb{A})$ which is not affine-linear over $\mathbb{Z}/2$, then $\wedge \in \mathrm{Pol}(\mathbb{A})$, so $\mathbb{A}$ is polynomially complete.*

Proof. By xoring with an affine-linear function over $\mathbb{Z}/2$, we may assume without loss of generality that $f(x_1, ..., x_n) = 0$ whenever at most one x_i is nonzero. Since f is not identically 0, there must be some $a_1, ..., a_n \in \{0, 1\}$ with $f(x_1, ..., x_n) = 1$, and we may suppose that $\sum_i a_i$ is minimal. By our assumption on f, the number of nonzero a_i must be at least 2, so there is some pair of coordinates $i \neq j$ such that $a_i = a_j = 1$. If we decrease either a_i or a_j, then by the minimality assumption f becomes 0, so we have

$$x \wedge y = f(a_1, ..., a_{i-1}, x, a_{i+1}, ..., a_{j-1}, y, a_{j+1}, ..., a_n). \qquad \square$$

Proposition B.2.11. *The polynomial clone of $(\{0, 1\}, \wedge, \vee)$ is exactly the clone of all monotone functions.*

Proof. We prove that every monotone function f of arity n is in the clone generated by $\wedge, \vee$ by induction on n:
$$f(x_1, ..., x_n) = f(x_1, ..., x_{n-1}, 0) \vee (f(x_1, ..., x_{n-1}, 1) \wedge x_n). \qquad \square$$

Lemma B.2.12. *If $\mathbb{A}$ has underlying set $\{0, 1\}$ and $\vee \in \mathrm{Pol}(\mathbb{A})$, and if $f \in \mathrm{Pol}(\mathbb{A})$ is monotone but is not contained in the clone generated by $\vee$, then $\wedge \in \mathrm{Pol}(\mathbb{A})$, so $\mathrm{Pol}(\mathbb{A})$ contains the clone of all monotone functions.*

Proof. We may suppose without loss of generaility that f depends on all of its inputs. If f is not contained in the clone generated by $\vee$, then in particular $f(x_1, ..., x_n) \neq x_1 \vee \cdots \vee x_n$, so since f is monotone there must be some input i such that $f(0, ..., 0, 1, 0, ..., 0) = 0$. Since f is monotone and depends on its ith input, there must be some $a_1, ..., a_n \in \{0, 1\}$ such that

$$f(a_1, ..., a_{i-1}, 0, a_{i+1}, ..., a_n) = 0, \quad f(a_1, ..., a_{i-1}, 1, a_{i+1}, ..., a_n) = 1.$$

Choose the $a_1, ..., a_n$ such that $\sum_j a_j$ is minimized subject to the displayed equations above. Then there is at least one j such that $a_j = 1$, by the choice of i, and for this j we have

$$x \wedge y = f(a_1, ..., a_{i-1}, x, a_{i+1}, ..., a_{j-1}, y, a_{j+1}, ..., a_n). \qquad \square$$

Putting everything together, we have the following classification of finite permutational algebras.

Theorem B.2.13. *If $\mathbb{A}$ is a finite permutational algebra, then up to isomorphism and polynomial equivalence, $\mathbb{A}$ is one of the following:*

(1) *a unary algebra, with the set of unary operations equal to a finite permutation group,*

(2) *a vector space over a finite field,*

(3) *the boolean algebra $(\{0, 1\}, \vee, \wedge, \neg)$,*

(4) *the lattice $(\{0, 1\}, \vee, \wedge)$, or*

(5) *the semilattice $(\{0, 1\}, \vee)$.*

Proof. If every polynomial of $\mathbb{A}$ is essentially unary, then we are in case (1). Otherwise, by Proposition B.2.2 there is some binary polynomial $f \in \mathrm{Pol}_2(\mathbb{A})$ which depends on both of its inputs. If $|\mathbb{A}| \geq 3$, then Theorem B.2.7 shows that we are in case (2). Otherwise, we assume that the underlying set of $\mathbb{A}$ is $\{0, 1\}$.

If $\oplus \in \mathrm{Pol}_2(\mathbb{A})$, then Lemma B.2.10 shows that we are either in case (2) or case (3). If $\oplus \notin \mathrm{Pol}_2(\mathbb{A})$, then by Lemma B.2.9 every polynomial operation of $\mathbb{A}$ is monotone, and f is either $\vee$ or $\wedge$. By possibly swapping 0 and 1, we may assume without loss of generality that $f = \vee$. Then Lemma B.2.12 shows that we are either in case (4) or (5). $\qquad\square$

Corollary B.2.14. *If $\mathbb{A}$ is a finite permutational algebra, then $\mathrm{Pol}(\mathbb{A})$ is generated by the binary polynomials of $\mathbb{A}$.*

The previous corollary is a general feature of tame congruence theory: most concrete computations in tame congruence theory depend only on the set of binary polynomials.

Definition B.2.15. If (α, β) is a tame congruence quotient of a finite algebra $\mathbb{A}$, and if N is an (α, β)-trace, then we say that N has

- *unary type*, or type **1**, if $\mathbb{A}|_N/\alpha|_N$ is polynomially equivalent to a unary algebra,

- *affine type*, or type **2**, if $\mathbb{A}|_N/\alpha|_N$ is polynomially equivalent to a vector space over a finite field,

- *boolean type*, or type **3**, if $\mathbb{A}|_N/\alpha|_N$ is polynomially equivalent to a boolean algebra,

- *lattice type*, or type **4**, if $\mathbb{A}|_N/\alpha|_N$ is polynomially equivalent to a lattice,

- *semilattice type*, or type **5**, if $\mathbb{A}|_N/\alpha|_N$ is polynomially equivalent to a semilattice.

We say that the tame congruence quotient (α, β) has type **i** if there is any (α, β)-trace with type **i**. As we will see in the next section, all of the traces of a tame congruence quotient (α, β) have the same type as each other.

The numbering of the five types can be remembered with the following visual mnemonic lattice, where the ordering corresponds to the richness of the operations in the polynomial clone.
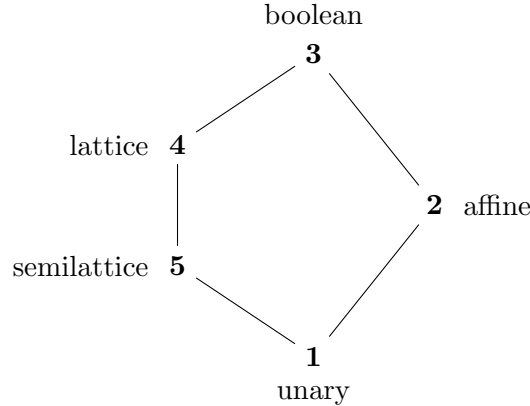
B.3 The structure of minimal sets

So far we have classified the traces of tame congruence quotients, by classifying the permutaitonal algebras. In order to classify the minimal sets of an algebra, we note that for any (α, β)-minimal set U, the restriction $\mathbb{A}|_U$ has the property that for each unary polynomial $f \in \mathrm{Pol}_1(\mathbb{A}|_U)$, either f is a permutation or $f(\beta|_U) \subseteq \alpha|_U$. Thus the restricted algebra $\mathbb{A}|_U$ is $(\alpha|_U, \beta|_U)$-minimal:

Definition B.3.1. A finite algebra $\mathbb{A}$ is called (α, β)-*minimal*, for $\alpha < \beta \in \mathrm{Con}(\mathbb{A})$, if for every unary polynomial $f \in \mathrm{Pol}_1(\mathbb{A})$, either f is a permutation or $f(\beta) \subseteq \alpha$.

By Proposition B.1.20, an algebra $\mathbb{A}$ is (α, β)-minimal iff $\mathbb{A}/\alpha$ is $(0_{\mathbb{A}/\alpha}, \beta/\alpha)$-minimal, and for each (α, β)-trace N there is a corresponding $(0_{\mathbb{A}/\alpha}, \beta/\alpha)$-trace N/α of the same type, so we may can oftend reduce to the case $\alpha = 0$ without loss of generality. We can simplify some of the arguments of [74] about types **3**, **4**, and **5** by using the concept of a partial semilattice operation from Section 3.2.

Definition B.3.2. 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.

We write $a \to_s b$ if s is a partial semilattice and $s(a, b) = b$.

Proposition B.3.3. *If $\mathbb{A}$ is (α, β)-minimal and has a trace N of type **3**, **4**, or **5** (that is, of either boolean, lattice, or semilattice type), then $\mathbb{A}$ has a partial semilattice polynomial $s \in \mathrm{Pol}_2(\mathbb{A})$ such that N is closed under s, and such that $(N/\alpha, s)$ is a two-element semilattice. Furthermore, if N has type **3** or **4** (i.e. boolean or lattice type), then there is another partial semilattice $s' \in \mathrm{Pol}_2(\mathbb{A})$ such that (N, s, s') is a two-element lattice.*

If s is a partial semilattice term and $a, b \in N$ have $s(a, b) \not\equiv_\alpha a$, then $a \to_s x$ for all $x \in \mathbb{A}$ and $a/\alpha = \{a\}$.

Proof. Let $t \in \mathrm{Pol}_2(\mathbb{A})$ be such that N is closed under t and $(N/\alpha, t)$ is a two-element semilattice. Since the unary polynomial $t(x, x)$ is not constant on N/α, (α, β)-minimality implies the unary polynomial $t(x, x)$ must be invertible, so we may assume without loss of generality that t is idempotent. Then we may apply the semilattice iteration argument from Proposition 3.2.9 to produce a partial semilattice polynomial $s \in \mathrm{Clo}(t)$ such that the restriction of s and t to N/α agree.

For the last statement, if $a, b \in N$ have $s(a, b) \not\equiv_\alpha s(a, a) = a$, then by (α, β)-minimality the unary polynomial $x \mapsto s(a, x)$ must be a permutation, and since $s(a, s(a, x)) = s(a, x)$, it must be the identity, so $s(a, x) = x$ for all $x \in \mathbb{A}$. If $a' \in a/\alpha$, then the same argument shows that $s(a', x) = x$ for all x, so $a \to_s a'$ and $a' \to_s a$, which is only possible if $a' = a$. $\qquad\square$

Hobby and McKenzie [74] like to think of their semilattices as meet-semilattices, so they call the partial semilattice polynomial s from Proposition B.3.3 a *pseudo-meet* operation. If the type is **3** or **4**, then they call the second partial semilattice operation s' a *pseudo-join* operation. If the type is **3** (i.e. boolean), then you can additionally find a unary polynomial f which preserves N and swaps the elements of N, and any such f will be invertible. We can assume without loss of generality that this f has even order, and we might call such an f a *pseudo-negation* operation.

Proposition B.3.4. *If $\mathbb{A}$ is (α, β)-minimal and has at least two different (α, β)-traces, then all of the (α, β)-traces have type **1** or **2** (that is, they all have either unary or affine type).*

Proof. We assume without loss of generality that $\alpha = 0_{\mathbb{A}}$. Suppose that N is a $(0_{\mathbb{A}}, \beta)$-trace of type **3**, **4**, or **5** (that is, of either boolean, lattice, or semilattice type). Then N has two elements, call them a, b, and by Proposition B.3.3 there is some partial semilattice polynomial $s \in \mathrm{Pol}_2(\mathbb{A})$ such that $N = \{a, b\}$ is closed under s and such that s acts as a semilattice operation on $\{a, b\}$,

337

say with $s(a,b) = b$. By the second part of Proposition B.3.3, we then have $s(a,x) = s(x,a) = x$ for all $x \in \mathbb{A}$.

Now suppose, for the sake of a contradiction, that K is a different $(0_\mathbb{A}, \beta)$-trace. Since $s(a,b) = s(b,b) = b$, $(0_\mathbb{A}, \beta)$-minimality implies that the unary polynomial $f : x \mapsto s(x,b)$ has $f(\beta) \subseteq 0_\mathbb{A}$, so $s(K,b)$ is a singleton. Thus there must be some $c \in K$ such that $s(c,b) \neq c$. Since $s(c,a) = c \neq s(c,b)$, the unary polynomial $g : x \mapsto s(c,x)$ must be a permutation by $(0_\mathbb{A}, \beta)$-minimality. However, we have $g(K) = s(c,K) \subseteq s(c,c)/\beta = c/\beta = K$ and $g(N) = s(c,N) \subseteq s(c,a)/\beta = c/\beta = K$, so g can't be a permutation, which is a contradiction. $\qquad \square$

Proposition B.3.5. *If $\mathbb{A}$ is (α, β)-minimal, then β is abelian over α if and only if all of the (α, β)-traces have type **1** or **2** (i.e., unary or affine type).*

Proof. We assume without loss of generality that $\alpha = 0_\mathbb{A}$. If some (α, β)-trace N has type **3**, **4**, or **5** (i.e., boolean, lattice, or semilattice), then there is a partial semilattice polynomial $s \in \mathrm{Pol}_2(\mathbb{A})$ which acts nontrivially on N by Proposition B.3.3. Since semilattices aren't abelian, $\mathbb{A}|_N$ is not abelian, and therefore β isn't abelian either (since N is a congruence class of β).

Now suppose for contradiction that all the traces have type **1** or **2**, but that β is not abelian. The plan is to transport the nonabelianness of β into one of the $(0_\mathbb{A}, \beta)$-traces to contradict the fact that traces of type **1** or **2** must be abelian. Recall that β not being abelian means that there is some polynomial $g \in \mathrm{Pol}(\mathbb{A})$ and some $(u,v), (a_1, b_1), ..., (a_n, b_n) \in \beta$ such that

$$g(u, a_1, ..., a_n) = g(u, b_1, ..., b_n)$$

but

$$g(v, a_1, ..., a_n) \neq g(v, b_1, ..., b_n),$$

and we may assume without loss of generality that n is minimal. By the minimality of n, we have $a_i \neq b_i$ for all i (else we could just substitute a_i for the ith argument), and we clearly have $u \neq v$, so there are $(0_\mathbb{A}, \beta)$-traces $N_0, N_1, ..., N_n$ such that $u, v \in N_0$ and $a_i, b_i \in N_i$ for each i. Let K be the $(0_\mathbb{A}, \beta)$-trace which contains $g(u, a_1, ..., a_n)$, then since g is compatible with β we have

$$g(N_0, N_1, ..., N_n) \subseteq K.$$

The restriction of g to $N_0 \times N_1 \times \cdots \times N_n$ must depend on all of its inputs by the minimality of n, so for each i there are $c_j \in N_j$ such that the unary polynomial

$$f_i : x \mapsto g(c_0, ..., c_{i-1}, x, c_{i+1}, ..., c_n)$$

is not constant on N_i. Each such f_i must be a permutation by $(0_\mathbb{A}, \beta)$-minimality, so we have $f_i : N_i \simeq K$ for each i. Then the polynomial h given by

$$h(x_0, ..., x_n) = g(f_0^{-1}(x_0), ..., f_n^{-1}(x_n))$$

preserves K, and we have

$$h(f_0(u), f_1(a_1), ..., f_n(a_n)) = h(f_0(u), f_1(b_1), ..., f_n(b_n))$$

but

$$h(f_0(v), f_1(a_1), ..., f_n(a_n)) \neq h(f_0(v), f_1(b_1), ..., f_n(b_n)).$$

Thus $\mathbb{A}|_K$ is not abelian, so $\mathbb{A}|_K$ can't be polynomially equivalent to a unary or affine algebra, which is a contradiction. $\qquad \square$

The next challenge is to construct a *pseudo-Mal'cev* operation when the type is **2**, and to use it to prove that all of the (α, β)-traces are isomorphic when at least one of them has type **2**.

Lemma B.3.6. *If $\mathbb{A}$ is (α, β)-minimal and has an (α, β)-trace N of type **2**, then there is a ternary polynomial $p \in \mathrm{Pol}_3(\mathbb{A})$ such that, if B is the union of all (α, β)-traces (the "body"), we have*

(a) *N is closed under p, and the restriction of $p(x, y, z)$ to $\mathbb{A}|_N/\alpha$ is the Mal'cev operation $x - y + z$,*

(b) *p is idempotent, that is $p(a, a, a) = a$ for all $a \in \mathbb{A}$,*

(c) *for all $a \in \mathbb{A}, b \in B$ we have $p(a, b, b) = a$, and*

(d) *for all $a \in \mathbb{A}, b \in B$ we have $p(b, b, a) = a$.*

Proof. (Following [74]) We construct p in stages, in each step getting a ternary polynomial which satisfies one more of (a), (b), (c), (d). To start, since N has type **2**, there is a polynomial $f \in \mathrm{Pol}_3(\mathbb{A})$ satisfying (a). Next, since the restriction of the unary polynomial $g(x) = f(x, x, x)$ to $\mathbb{A}|_N/\alpha$ is nonconstant, (α, β)-minimality implies that $g(x)$ is a permutation, and since the restriction of g to $\mathbb{A}|_N/\alpha$ is the identity, the polynomial $h(x, y, z) = g^{-1}(f(x, y, z))$ satisfies (a) and (b).

Claim. Suppose that f is any polynomial satisfying (a) and (b). For any $b \in B$, the polynomials $x \mapsto f(x, b, b)$ and $x \mapsto f(b, b, x)$ are permutations.

Proof of claim. Suppose not - suppose for contradiction that the unary polynomial $x \mapsto f(x, b, b)$ is not a permutation for some $b \in B$, and let K be the (α, β)-trace which contains b. Iterate f on its first argument to get $f^\infty \in \mathrm{Pol}_3(\mathbb{A})$ such that

$$f^\infty(f^\infty(x, y, z), y, z) = f^\infty(x, y, z)$$

for all x, y, z, define a unary polynomial g by

$$g(x) = f^\infty(x, b, b),$$

and note that if $x \mapsto f(x, b, b)$ is not a permutation, then g is also not a permutation. By (a) and (α, β)-minimality, K can't be N: for any $a \in N$, the restriction of $x \mapsto f^\infty(x, a, a)$ to $\mathbb{A}|_N/\alpha$ is the identity, so (α, β)-minimality and the identity $f^\infty(f^\infty(x, a, a), a, a) = f^\infty(x, a, a)$ imply that

$$f^\infty(x, a, a) = x$$

for all $x \in \mathbb{A}$ and $a \in N$. Since g is not a permutation, (α, β)-minimality implies that $g(K)$ is contained in a single α-congruence class, so in particular there is some $c \in K$ such that $g(c) \not\equiv_\alpha c$. Then if we define the unary polynomial h by $h(x) = f^\infty(c, x, x)$, we have

$$h(c) = f^\infty(c, c, c) = c \not\equiv_\alpha g(c) = f^\infty(c, b, b) = h(b),$$

so by (α, β)-minimality h is a permutation. But then

$$h(c) = f^\infty(c, c, c) = c = f^\infty(c, a, a) = h(a)$$

for any $a \in N$, so h is not injective, a contradiction.

Now we use the claim to upgrade an f satisfying (a) and (b) to one which also satisfies (c). Let $t(x, y) = f(x, y, y)$, and iterate t on its first argument, to get $t^\infty \in \mathrm{Pol}_2(\mathbb{A})$ with $t^\infty(t^\infty(x, y), y) = t^\infty(x, y)$. By the claim, for any $b \in B$ we have $t^\infty(x, b) = x$. Now define $g \in \mathrm{Pol}_3(\mathbb{A})$ by

$$g(x, y, z) = t^{\infty -1}(f(x, y, z), z).$$

The restriction of t to $\mathbb{A}|_N / \alpha$ is just first projection, so g satisfies (a), since f is idempotent g will be idempotent as well, and by construction we have $g(x, b, b) = t^\infty(x, b) = x$ for any $b \in B$.

Finally, we use the claim to upgrade an f satisfying (a), (b), (c) to one which also satisfies (d). By swapping the first and third inputs to f, this is equivalent to upgrading an f which satisfies (a), (b), (d) to one which also satisfies (c). We use the exact same construction for this as in the previous step - we just have to check that the resulting g also satisfies (d): for $b \in B$, we have

$$g(b, b, x) = t^{\infty -1}(f(b, b, x), x) = t^{\infty -1}(x, x) = x,$$

where the last step follows from idempotence. $\qquad \square$

Definition B.3.7. If $\mathbb{A}$ is (α, β)-minimal and has a trace of type **2** (i.e. affine type), and if B is the union of the (α, β)-traces, then we call any idempotent ternary polynomial $p \in \mathrm{Pol}_3(\mathbb{A})$ such that $p(a, b, b) = p(b, b, a) = a$ for all $a \in \mathbb{A}$ and all $b \in B$ a *pseudo-Mal'cev* polynomial for $\mathbb{A}$.

Theorem B.3.8. *Suppose that $\mathbb{A}$ is (α, β)-minimal and has a trace of type **2** (i.e. affine type), let B be the union of the (α, β)-traces, and let p be any pseudo-Mal'cev polynomial for $\mathbb{A}$. Then*

- *for all $a, b \in B$, the unary polynomials $x \mapsto p(x, a, b), p(a, x, b), p(a, b, x)$ are all permutations,*

- *B is closed under p and the restriction of p to B is Mal'cev, and*

- *all of the (α, β)-traces are polynomially isomorphic.*

*In particular, if one of the (α, β)-traces has type **2** then they all do.*

Proof. (Following [74]) We assume without loss of generality that $\alpha = 0_\mathbb{A}$. First we show that for $a, b \in B$, the unary polynomial $x \mapsto p(x, a, b)$ is a permutation iff $x \mapsto p(a, x, b)$ is. For this, let c be any element of the $(0_\mathbb{A}, \beta)$-trace $N = a/\beta$, and note that since β is abelian (by Proposition B.3.5) and $(a, c) \in \beta$, we have

$$p(c, a, b) = p(a, a, b) = b \quad \Longleftrightarrow \quad p(a, c, b) = p(c, c, b) = b,$$

so $x \mapsto p(x, a, b)$ is not a permutation iff $p(N, a, b) = \{b\}$, which happens iff $p(a, N, b) = \{b\}$, which happens iff $x \mapsto p(a, x, b)$ is not a permutation (and in fact, these all occur iff $p(N, N, b) = \{b\}$). Similarly, $x \mapsto p(a, x, b)$ is a permutation iff $x \mapsto p(a, b, x)$ is a permutation.

Now suppose for a contradiction that $x \mapsto p(a, x, b)$ and $x \mapsto p(a, b, x)$ are not permutations, and consider the unary poynomial $f(x) = p(a, p(a, x, b), x)$. If $x \in N = a/\theta$, then we have

$$f(x) = p(a, p(a, a, b), x) = p(a, b, x) = p(a, b, a),$$

so f is not a permutation. If $x \in b/\theta$, then we have

$$f(x) = p(a, p(a, b, b), x) = p(a, a, x) = x,$$

340

so by $(0_{\mathbb{A}}, \beta)$-minimality f must be a permutation, which is a contradiction.

To see that B is closed under p, let $a, b \in B$, then since the unary polynomial $x \mapsto p(a, b, x)$ is a permutation and therefore takes $(0_{\mathbb{A}}, \beta)$-traces to $(0_{\mathbb{A}}, \beta)$-traces, we see that it takes B to B. Finally, if N, K are two $(0_{\mathbb{A}}, \beta)$-traces and $a \in N, b \in K$, then the unary polynomial $g(x) = p(a, x, b)$ takes N to K bijectively. $\qquad\square$

Putting these results together, we have proved the main result of this section.

Theorem B.3.9. *If (α, β) is a tame congruence of a finite algebra $\mathbb{A}$, then all of the (α, β)-traces have the same type. If this type is not 1, or if (α, β) is a prime quotient, then all of the (α, β)-traces are polynomially isomorphic to each other. If the type is not 1 or 2, then each (α, β)-minimal set has just one trace, and if the type is 3 or 4 then every (α, β)-trace has size two.*

In order to rule out type **1** in most cases, we introduce a stronger version of abelianness which is characteristic of unary algebras.

Definition B.3.10. If $\alpha \leq \beta \in \mathrm{Con}(\mathbb{A})$, then β is *strongly abelian* over α if for all $f \in \mathrm{Pol}(\mathbb{A})$, all $(u, v) \in \beta$ and all tuples x, y, z with $x_i \equiv y_i \equiv z_i \pmod{\beta}$, we have

$$f(u, x_1, ..., x_n) \equiv_\alpha f(v, y_1, ..., y_n) \implies f(u, z_1, ..., z_n) \equiv_\alpha f(v, z_1, ..., z_n).$$

An algebra $\mathbb{A}$ is *strongly abelian* if $1_{\mathbb{A}}$ is strongly abelian over $0_{\mathbb{A}}$.

It's easy to see that every unary algebra is strongly abelian, while any group or semilattice is not strongly abelian. We can now characterize type **1** in terms of strong abelianness.

Proposition B.3.11. *If $\mathbb{A}$ is (α, β)-minimal, then β is strongly abelian over α iff all of the (α, β)-traces have type 1 (i.e. unary type).*

Proof. The proof is almost identical to the proof of Proposition B.3.5. $\qquad\square$

Proposition B.3.12. *Suppose that $\mathbb{A}$ has a Taylor polynomial and that $\alpha < \beta \in \mathrm{Con}(\mathbb{A})$. Then β is not strongly abelian over α. As a consequence, if $\mathbb{A}$ is a finite Taylor algebra then no tame congruence quotient of $\mathbb{A}$ has type 1.*

Proof. Suppose for contradiction that β is strongly abelian over α, and pick some $(a, b) \in \beta \setminus \alpha$. Let t be a Taylor term, then for each input of t we have an equation of the form

$$t(?, ..., ?, a, ?, ..., ?) = t(?, ..., ?, b, ?, ..., ?),$$

where each ? is either an a or a b, so strong abelianness of β over α implies that

$$t(a, ..., a, a, b, ..., b) \equiv_\alpha t(a, ..., a, b, b, ..., b)$$

at each input. Stringing these equations together and using the idempotence of t, we get

$$a = t(a, ..., a) \equiv_\alpha \cdots \equiv_\alpha t(b, ..., b) = b,$$

which contradicts the assumption $(a, b) \notin \alpha$.

For the last statement, note that if t is a Taylor polynomial for $\mathbb{A}$ and $e \in E(\mathbb{A})$ has $e(\mathbb{A}) = U$ for some (α, β)-minimal set U, then $e(t(x_1, ..., x_n))$ is a Taylor polynomial for $\mathbb{A}|_U$. In fact, by the idempotence of t, the restriction of $e \circ t$ to any (α, β)-trace N would be a Taylor polynomial for the unary algebra $\mathbb{A}|_N / \alpha|_N$, which gives an even simpler contradiction. $\qquad\square$

The following reformulation of strong abelianness from [109] should give a more concrete idea of just how strong it is.

Proposition B.3.13. *An algebra $\mathbb{A}$ is strongly abelian iff, for each n-ary polynomial t of $\mathbb{A}$, there are equivalence relations R_i on $\mathbb{A}$ such that*

$$t(a_1, ..., a_n) = t(b_1, ..., b_n) \iff \forall i \leq n, \ (a_i, b_i) \in R_i.$$

In particular, if $\mathbb{A}$ is finite and strongly abelian then every polynomial of $\mathbb{A}$ depends on at most $\log_2 |\mathbb{A}|$ of its inputs.

Corollary B.3.14. *Every finite, idempotent, strongly abelian algebra can be written as a product of algebras where every operation is a projection.*

Example B.3.1 (From [90]). There is a 5-element strongly abelian algebra $\mathbb{A} = (\{a, b, c, d, e\}, \cdot_1, \cdot_2)$ which is not quasiaffine. The basic binary operations of $\mathbb{A}$ are given below.

$\cdot_1$	a	b	c	d	e
a	a	b	a	b	b
b	a	b	a	b	b
c	c	d	c	d	d
d	c	d	c	d	d
e	c	d	c	d	d

$\cdot_2$	a	b	c	d	e
a	a	b	a	b	b
b	a	b	a	b	b
c	c	e	c	e	e
d	c	e	c	e	e
e	c	e	c	e	e

This algebra fails to be quasiaffine because it fails to satisfy the two term condition:

$$a \cdot_1 a = a \cdot_2 a, \ \ a \cdot_1 b = a \cdot_2 b, \ \ c \cdot_1 a = c \cdot_2 a, \ \ \text{but} \ \ c \cdot_1 b \neq c \cdot_2 b.$$

To see that it is strongly abelian, note that for each i, j we have the identities

$$x \cdot_i (y \cdot_j z) \approx x \cdot_i z, \ \ (x \cdot_i y) \cdot_j z \approx x \cdot_j z,$$

so every term of $\mathbb{A}$ is one of $x, x \cdot_i x, x \cdot_i y$ for some $i \in \{1, 2\}$, up to permuting its inputs.