CATEGORIFICATION, TERM REWRITING AND THE KNUTH–BENDIX PROCEDURE

TIBOR BEKE

ABSTRACT. An axiomatization of a finitary, equational universal algebra by a convergent term rewrite system gives rise to a finite, coherent categorification of the algebra.

INTRODUCTION

MacLane's notion of a (symmetric) monoidal category forms a paradigm of categorification: the notion of a certain universal algebra — that of (commutative) monoid — is lifted to the category of small categories where, instead of equational axioms, a finite set of coherence diagrams specify the structure. However, those coherence diagrams are not "translations" or "categorifications" of the axioms of monoids. Where do the MacLane pentagon and hexagon come from? Notice that the associative axiom $x(yz) = (xy)z$ made *asymmetric* — occurrences of $x(yz)$ are permitted to be replaced by $(xy)z$, but not the other way around — is what is called a *self-normalizing rewrite rule*: by re-parenthesizing to the left, in any order but going as long as one can, any fully parenthesized expression becomes transformed into a normal form; and normal forms biject with equivalence classes of terms under associativity. However, this still does not explain MacLane coherence. If one thinks of re-association as a natural transformation, why is it that as a consequence of the pentagon axiom, any two re-association paths between the same source and target will compose to the same natural transformation?

The answer is that the directed associativity axiom $x(yz) \Rightarrow (xy)z$ is an example of a convergent rewrite system, and the MacLane pentagon consists of the two rewrite paths leading from the so-called critical pair of this system to their common normal form. MacLane's coherence theorem follows automatically from these two facts.

Thankfully, monoids are not the only example of universal algebras axiomatizable by convergent rewrite systems. The latter are extensively researched in computer science, due to their role in automated proof theory and algebraic decision problems. This link was discovered in the groundbreaking paper of Knuth–Bendix [KB70] that also introduced the notion of critical pair and the semi-decision algorithm now known as the Knuth–Bendix procedure. The connection to recursion theory plays no direct role in this paper, but Knuth's fundamental discoveries on the combinatorics of terms do.

We start with a rapid overview of term rewriting, aimed at the reader who knows little or nothing about the subject. (The material is standard — see for example Baader and Nipkow [BN98] —

Date: March 28, 2010.

and included mainly to make the discussion self-contained and to establish notation.) The next two sections develop the formalism of axiomatizations of algebraic theories in categories via generators (functors and natural transformations) and relations (commutative diagrams), and define coherence for an axiomatization. This is gruesomely formal material that, unfortunately, seems to be unavailable elsewhere in the literature.

Section 4 contains the main result. An interesting case study is that of groups, a detailed investigation of which occupies much of Knuth–Bendix [KB70]; note that this example is already beyond operadic techniques. Other examples include lax homomorphisms and diagrams of various shapes, such as monoids and groups with convergent presentations, and Reedy diagrams. The closing section is devoted to discussion and open questions.

Family resemblances. Since the subject matter lies at the crossroads of several well-researched areas, it may be worthwhile to point out differences and similarities.

• Term rewriting is concerned with a set equipped with a relation; the set is that of terms well-formed from variables and function symbols, and the relation is that of 'elementary rewriting', a formal string-replacement operation that imitates the application of an equational identity. Elementary rewriting generates a preorder on terms, and the main concern is understanding that preorder — *which* term can be generated from which one.

In 2-algebras such as monoidal categories, terms are interpreted by functors and rewrites by natural transformations. Each sequence of elementary rewrites corresponds, potentially, to a different natural transformation. In fact, such ambiguity may be present already for elementary rewrites: as in Malcev algebras, let f be a tertiary operation subject to the rewrite rules

$$f\mathbf{xxy} \Rightarrow \mathbf{y}$$

and

$$f\mathbf{xyy} \Rightarrow \mathbf{x}$$

Then $f\mathbf{xxx}$ can be rewritten into $\mathbf{x}$ for two reasons that, at the level of natural transformations, better be kept apart. Similarly, terms can have non-trivial rewrites *into themselves* (interpreted by natural endomorphisms of objects). Syntactically, terms and rewrites form a graph, and the main concern is understanding that graph — *how* terms can be transformed into each other.

• The flourishing area of higher-order categorical rewriting is part of the theory of n-categories, and is concerned with the combinatorics of concatenation of (higher) arrows. The geometric objects (n-graphs) thus generated have applications in homological algebra, chiefly in building resolutions of groups and monoids. The 0-dimensional version would be string rewriting (also called word rewriting), where the underlying symbols are interpreted as generators of a free monoid. String rewriting can be thought of as a special case of term rewriting — namely, term rewriting with only unary functions present — while the converse is not true. It seems that the full arsenal of term rewriting is needed to handle categorical coherence.

It is possible to embed term rewriting for universal algebras into a theory of higher-dimensional graphical rewriting; see Burroni [Bur93], Lafont [Laf95]. That formalism seems to be very different from the one employed in this paper, particularly as regards the role of convergent presentations and categorification.

• *Proof-nets*, originating in the work of Girard, can be thought of as categories with structure associated to term rewrite systems modeling proofs in propositional logic and its various extensions, such as linear logic; cf. Schneck [Sch99]. They are similar to the graph of terms introduced here, but without the explicit link to universal algebras, equational logic and categorification.

• Homotopical algebra, originating in the theories of Boardman–Vogt and May, provides well-known machines for defining 'up to coherent homotopy' replacements of operadic algebras. Those machines, by default, produce homotopical algebras with infinitely many sorts, generated by countably many operations subject to a (recursive, or at least recursively enumerable) set of identities. Categorification — in the sense of passage from a set-based to a groupoid-based universal algebra — can certainly be thought of as a case of homotopical universal algebra, whose general notions (Quillen model categories, cofibrant replacement etc.) are applicable here as well. But the emphasis in this paper is on the much more delicate and combinatorial issue of *finite* axiomatization of categorical algebras.

• The work on categorical coherence in the 70's is close in spirit, if not in notation and underlying mathematics, to the present one; research on coherence inevitably becomes research on word problems. It should probably be pointed out that the work by Knuth and Bendix on term rewriting and by MacLane and his students on categorical coherence occurred independently right around the same time, and the ensuing separation of mathematical cultures has a lot to do with the fact that it's not been recognized just how close they are.

The article [Bek99] asked which axiomatizations of universal algebras (in terms of functions and equational identities) allows a categorification. So, an answer — a sufficient but not necessary condition — is that the axiomatization be convergent; more precisely, that it can be embedded in a convergent term rewrite system.

Remark. After this paper was completed, it came to my attention that the unpublished 2008 Ph.D. thesis of Jon Cohen [Coh08] contains a very similar (though not identical) development of rewriting, 2-theories, coherence and categorification. The syntactic details and proof of the main theorem are different enough, I believe, for both works to warrant attention.

1. Term rewriting

Well-formed terms. We assume given a finite set S of *sorts*. (S should be thought of as indexing the types of ingredients out of which our universal algebra is built. For example, a vector space involves two sorts of things: scalars and vectors.) For each sort $s \in S$, there is to exist an infinite set of variables $x_i^{(s)}$ of that sort. There is a finite set of *function symbols*, each of which carries a sort (the sort of the 'output' the function produces) and an arity, which is an n-tuple of sorts, specifying the sorts of the respective inputs. Here n is any natural number; function symbols with a 0-tuple as input are called constants. The set of *well-formed terms* (or just *terms* for brevity) and their sorts is defined inductively. A variable symbol $x_i^{(s)}$ is a term of sort s. If $\mathtt{t}_1, \mathtt{t}_2, \dots, \mathtt{t}_n$ are terms respectively of sorts s_i, and f is a function symbol of sort s with arity $\langle s_1, s_2, \dots, s_n \rangle$, then the string $f\mathtt{t}_1\mathtt{t}_2 \dots \mathtt{t}_n$ is a term of sort s. Note that constants are well-formed terms on their own. Terms should be thought of as names of composite functions.

A *subterm* of a term t is a substring of t that is a well-formed term as a string itself.

Example 1.1. Thinking of '$+$' and '$\times$' as binary function symbols in a one-sorted universal algebra, x, y, z as variables,

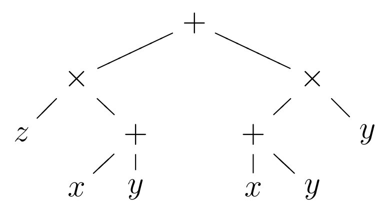
$$+ \times z + x\,y \times +x\,y\,y$$

is a well-formed term. Three of its subterms are '$\times z + x\,y$' and '$+xy$' and 'y'.

Remark. Though we defined terms using parenthesis-free prefix (also called 'Polish') notation, what is essential is only that they be specified by an unambiguous context-free grammar. Fully parenthesized infix notation

$$((z \times (x + y)) + ((x + y) \times y))$$

or representation of terms as partially ordered sets of strings (the 'derivation trees' of computer science, the 'rooted planar trees', decorated with variables, familiar from the topological literature)



would do as well. What would be the "pruning and grafting of (sub)trees" in the tree notation, becomes in prefix notation the replacement of subterms by well-formed terms.

Replacement. If u is a subterm of sort s of the term t, then replacing the substring u in t by another term of sort s results in a well-formed term. In particular, one can replace every occurrence of a variable $x_i^{(s_i)}$ in t by a term t_i of the corresponding sort s_i. We write the result of several such global replacements performed independently as $\mathsf{t}[x_1^{(s_1)} \mapsto \mathsf{t}_1, x_2^{(s_2)} \mapsto \mathsf{t}_2, \ldots, x_k^{(s_k)} \mapsto \mathsf{t}_k]$. For the sake of brevity, we often pretend that the formal variables are equipped with a conventional ordering, and use function-composition notation $\mathsf{t}(\mathsf{t}_1, \mathsf{t}_2, \ldots, \mathsf{t}_k)$. Note that replacements allow for empty occurrences.

Term rewriting can be thought of as the application of equational axioms in a *given direction* only.

Definition 1.2. A *rewrite rule* is an ordered pair $\langle \mathsf{u}, \mathsf{v} \rangle$ of terms of the same sort. (We will often typeset rewrite rules as $\mathsf{u} \Rightarrow \mathsf{v}$.) A *substitution instance* (or just *instance*) of a rewrite rule results from making an identical substitution into variables on both sides: $\langle \mathsf{u}(\mathsf{t}_1, \mathsf{t}_2, \ldots, \mathsf{t}_k), \mathsf{v}(\mathsf{t}_1, \mathsf{t}_2, \ldots, \mathsf{t}_k) \rangle$. Let $\langle \mathsf{U}, \mathsf{V} \rangle$ be an instance of a rewrite rule, and let the term t contain an occurrence of U; say, $\mathsf{t} = \mathsf{t}_1 \mathsf{U} \mathsf{t}_2$. Then t permits to be rewritten into $\mathsf{r} = \mathsf{t}_1 \mathsf{V} \mathsf{t}_2$.

In the literature, the condition is often added that any variable occurring on the right-hand side v of a rewrite rule $\langle \mathsf{u}, \mathsf{v} \rangle$ should also occur in the left-hand term u. Indeed, this condition is necessary for the rewrite rule to be noetherian (see below). All our examples will satisfy this condition.

Convergence. A rewrite system is given by a finite set of rewrite rules. Within this section, let us observe the computer science tradition by denoting by an arrow the relation between terms induced by rewriting; so $\mathsf{x} \to \mathsf{y}$ denotes the fact that some rewrite of x results in y. The transitive-reflexive closure of the relation $\to$ is denoted $\to^*$. That is, $\mathsf{x} \to^* \mathsf{y}$ means that there exist some (possibly zero) terms t_i such that $\mathsf{x} \to \mathsf{t}_1 \to \mathsf{t}_2 \to \ldots \to \mathsf{t}_k \to \mathsf{y}$, or that $\mathsf{x} = \mathsf{y}$.

A rewrite system is *noetherian* if there do not exist infinitely many terms t_i such that

$$\mathsf{t}_0 \to \mathsf{t}_1 \to \mathsf{t}_2 \to \ldots \to \mathsf{t}_i \to \ldots$$

It is *confluent* if, whenever w, t_1 and t_2 are terms such that $\mathsf{w} \to^* \mathsf{t}_1$ and $\mathsf{w} \to^* \mathsf{t}_2$, then a term z exists such that $\mathsf{t}_1 \to^* \mathsf{z}$ and $\mathsf{t}_2 \to^* \mathsf{z}$. A rewrite system that is both noetherian and confluent is called *convergent*.

In a convergent rewrite system, the equivalence class of any term t under the equivalence relation generated by $\to$ contains a canonical representative $\mathrm{nf}(\mathsf{t})$. It is the only term in its equivalence class that is in 'normal form', i.e. rewrites no further: there exists no w such that $\mathrm{nf}(\mathsf{t}) \to \mathsf{w}$. By applying rewrite rules (any that applies, in any order, as long as any applies), every term is transformed into its normal form in finitely many steps. In particular, in a convergent rewrite system the 'word problem', or equivalence problem under the equivalence relation generated by $\to$, is recursively solvable.

Convergent presentations of equational universal algebras. An *equational universal algebra* $\mathfrak{U}$ is one axiomatized by finitely many universal axioms that state the equality of two terms of the same sort, for all values of the variables:

$$\forall x_1 \forall x_2 \ldots \forall x_k \; \mathsf{u} = \mathsf{v}$$

A rewrite system, that is, set of pairs $\langle \mathsf{u}, \mathsf{v} \rangle$ of terms of like sort, gives rise to equational axioms, simply by replacing the arrow by equality and preceding it by universal quantification. It is the converse problem that motivated Knuth and Bendix in [KB70]: whether, given a finite set of equational axioms, these could be 'oriented' (i.e. endowed with a preferred rewrite direction) and possibly augmented by *finitely many* more rewrite rules that are (when thought of as formulas) consequences of the original axioms, such that a convergent rewrite system results. If a convergent axiomatization exists for the theory, then it is decidable whether a given universal-equational formula is a consequence of the axioms; and the word problem in a free algebra on a finite set is also decidable. To be sure, that is not always the case, so not all equational universal algebras have convergent presentations. The Knuth-Bendix procedure is a deterministic algorithm that takes as input finitely many equational axioms and a suitable well-ordering of terms. The algorithm may or may not terminate. If it does, it yields a convergent presentation of the corresponding universal algebra. The dependence on the initial well-ordering and sufficient and necessary conditions for termination are difficult questions that, within this article, we do not need to be concerned with.

Example 1.3. Formalize the notion of group as one-sorted structure with a constant symbol **1**, a binary function $\cdot$ and a unary function symbol $(-)^{-1}$. The following ten rules give a convergent presentation of the axioms of groups (in the usual infix notation):

$$\begin{array}{rclcrcl} \mathbf{1}\cdot x & \Rightarrow & x & \qquad & \mathbf{1}^{-1} & \Rightarrow & \mathbf{1} \\ x\cdot\mathbf{1} & \Rightarrow & x & & (x^{-1})^{-1} & \Rightarrow & x \\ x^{-1}\cdot x & \Rightarrow & \mathbf{1} & & (x\cdot y)^{-1} & \Rightarrow & y^{-1}\cdot x^{-1} \\ x\cdot x^{-1} & \Rightarrow & \mathbf{1} & & x^{-1}\cdot(x\cdot y) & \Rightarrow & y \\ (x\cdot y)\cdot z & \Rightarrow & x\cdot(y\cdot z) & & x\cdot(x^{-1}\cdot y) & \Rightarrow & y \end{array}$$

The left-hand column contains the usual axioms of groups, oriented appropriately. If one considered the unoriented (i.e. equational) versions of these, the right-hand column would be redundant (and in fact, either the third or the fourth axiom could also be omitted from the left-hand column). As a rewrite system, however, this collection is not redundant. (At the same time, it is not the only minimal convergent presentation of the group axioms.) This example is due to Knuth-Bendix [KB70] and was historically also the first example of Knuth–Bendix completion.

2. Presenting a 2-theory

2-theories are a categorification of equational universal algebras. They are axiomatized by a set of generating functors, a set of generating natural transformations between composites of these functors, and commutativity diagrams whose arrows are well-formed composites of the generating natural transformations with each other and with the generating functors. The goal is to exploit the obvious idea: composites of functors will be thought of as terms; the generating natural transformations as rewrite rules; composable sequences of natural transformations formed from the generating ones, as rewrites; and commutativity axioms become commuting pairs of rewrites. As mentioned in the introduction, slightly more complicated bookkeeping is called for than in the case of set-based term rewriting.

Given a set of sorts S, set of sorted variables, and set F of sorted function symbols, the set T of well-formed terms is defined as before. Let R be a labelled set of pairs of terms, i.e. a function $I \xrightarrow{R} T\times T$ from some index set I to pairs of terms of the same sort. *Rewrite data* with source x and target y consist of an element $i\in I$, a substitution instance $\langle \mathsf{U},\mathsf{V}\rangle$ of the rule $\langle \mathsf{u},\mathsf{v}\rangle = R(i)$, and an occurrence of U in x, say $\mathsf{x}=\mathsf{x}_1\mathsf{U}\mathsf{x}_2$ such that $\mathsf{y}=\mathsf{x}_1\mathsf{V}\mathsf{x}_2$.

Definition 2.1. The *graph of terms* corresponding to $\langle S,F,R\rangle$ is the directed graph whose vertices are the well-formed terms, and whose edges from vertex x to vertex y are the rewrite data with source x and target y.

Definition 2.2. A presentation of a 2-theory is a tuple $\langle S,F,R,C\rangle$ where S is a set of sorts, F a set of sorted function symbols, R a labelled set of pairs of terms of the same sort, and C a set of (unordered) pairs of paths in the graph of terms of $\langle S,F,R\rangle$. Each such pair is to contain two paths with the same initial and terminal vertex.

$$\begin{array}{ccccc} & & \mathsf{t}_1 \longrightarrow \mathsf{t}_2 \longrightarrow \cdots \longrightarrow \mathsf{t}_i & & \\ & \nearrow & & \searrow & \\ \mathsf{v} & & & & \mathsf{w} \\ & \searrow & & \nearrow & \\ & & \mathsf{u}_1 \longrightarrow \cdots \longrightarrow \mathsf{u}_{j-1} \longrightarrow \mathsf{u}_j & & \end{array}$$

For ease of language, we will refer to such pairs of edge paths as *cycles*.

Some clarifying remarks:

- In the graph of terms, note that edges can only occur between terms of the same sort, but there may be both multiple edges between terms and 'loops' on a vertex. For readability, we will often suppress the edge-labels.
- In the definition of rewrite data, the actual variable substitutions that take $\langle \mathsf{u}, \mathsf{v} \rangle$ to $\langle \mathsf{U}, \mathsf{V} \rangle$ are not part of the data. (It turns out that substitutions taking a given term into another, if they exist at all, are uniquely determined by those variables that actually occur in the source term.)
- The edge paths, of course, may repeat vertices or edges. For any vertex t, there is, by definition, a unique path of length 0 beginning and ending at t.

An *interpretation* of the data $\langle S, F \rangle$ is a category $\mathcal{C}_s$ of 'objects of sort s' for each $s \in S$, and a functor $f : \mathcal{C}_{s_1} \times \mathcal{C}_{s_2} \times \cdots \times \mathcal{C}_{s_n} \to \mathcal{C}_s$ of the appropriate arity for each $f \in F$. This induces an interpretation of well-formed terms as follows. Let $\prod_{x_i^{(s)}} \mathcal{C}_s$ be the product, over *all* the variables, of the categories corresponding to their sorts. A term t of sort r will be interpreted as a certain functor $\prod_{x_i^{(s)}} \mathcal{C}_s \to \mathcal{C}_r$. Now if $x_k^{(s_k)}$, $k \in K$, is the (finite) set of variables that actually occur in t, then the interpretation will factor as $\prod_{x_i^{(s)}} C_s \to \prod_{k \in K} C_{s_k} \to \mathcal{C}_r$ where the first functor is projection; thus we will only indicate the second part of the interpretation. A variable of sort s is to be interpreted as the identity $\mathcal{C}_s \xrightarrow{\mathrm{id}} \mathcal{C}_s$. A constant symbol of sort s is to be interpreted as a functor from the terminal category to $\mathcal{C}_s$, that is to say, as an object of $\mathcal{C}_s$. The interpretation of terms $f\mathsf{t}_1\mathsf{t}_2 \ldots \mathsf{t}_k$ is then defined inductively, composing the functor f with the products of the interpretations of the t_i and diagonal functors for the variables.

An interpretation of the rewrite rules is to assign to each $i \in I$, say with $R(i) = \langle \mathsf{u}, \mathsf{v} \rangle$, a natural transformation from the functor corresponding to u to the functor corresponding to v. This induces an interpretation for each substitution instance $\langle \mathsf{U}, \mathsf{V} \rangle$ of $\langle \mathsf{u}, \mathsf{v} \rangle$ as the natural transformation obtained by precomposing $\mathsf{u} \overset{i}{\Rightarrow} \mathsf{v}$ with the functors interpreting the terms occurring in the variable substitutions $x_i \mapsto \mathsf{t}_i$. Rewrite data from $\mathsf{x} = \mathsf{x}_1\mathsf{U}\mathsf{x}_2$ to $\mathsf{y} = \mathsf{x}_1\mathsf{V}\mathsf{x}_2$, that is, an edge from x to y, will be interpreted by a natural transformation too; obtained this time by postcomposing, at the location x, the natural isomorphism from U to V by the functor corresponding to $\mathsf{x}_1 x \mathsf{x}_2$. (Here x is just a place-marker variable of the same sort as U and V.)

The interpretation of the path of length 0 at t is the identity natural transformation on t.

The entire graph of terms therefore becomes interpreted by functors $\prod_{x_i^{(s)}} \mathcal{C}_s \to \mathcal{C}_r$ and natural transformations. If each cycle commutes (that is, the composites of the natural transformations assigned to the edges are equal within each of the pairs of paths contained in C) then the interpretation is said to be a *model of* (or *algebra for*) the axioms $\langle S, F, R, C \rangle$.

Remark. The reason for having $\prod_{x_i^{(s)}} \mathcal{C}_s$, the product of the category-sorts over *all* the variables, as the domain of interpretation for terms is that not necessarily the same variables are present in the source and target of a rewrite rule; consider, for example, $x \cdot x^{-1} \Rightarrow \mathbf{1}$. The desire to interpret each term as a specific functor and each rewrite as a natural transformation then pretty much forces the selection of a 'universal domain'. Another solution, much more in the spirit of categorical logic, is to consider a well-formed term to be a term $\mathtt{t}$ together with a *context*, i.e. a tuple of variables $\langle x_{i_1}, x_{i_2}, \ldots, x_{i_k} \rangle$ that contain all the variables of $\mathtt{t}$ (see e.g. Definition D.1.1.4. of Johnstone [Joh02]). One could then demand that rewrites only act between terms with the same context, but one would have to equip the graph of terms with operations of omitting and introducing variables into contexts, increasing its complexity quite a bit. Nonetheless, rewrite rules, hence edges, preserve the set of free variables quite often in practice.

Example 2.3. A *monad* structure is an example of a 2-theory. In the above terms, it is one-sorted with one unary function symbol f and two rewrite rules: $\mathbf{x} \overset{\eta}{\Rightarrow} f\mathbf{x}$ and $ff\mathbf{x} \overset{\mu}{\Rightarrow} f\mathbf{x}$. The set C of commutative cycles contains two pairs of paths,

$$\left\{ fff\mathbf{x} \xrightarrow{f\mu} ff\mathbf{x} \xrightarrow{\mu} f\mathbf{x} \quad ; \quad fff\mathbf{x} \xrightarrow{\mu f} ff\mathbf{x} \xrightarrow{\mu} f\mathbf{x} \right\}$$

and

$$\left\{ f\mathbf{x} \xrightarrow{f\eta} ff\mathbf{x} \xrightarrow{\mu} f\mathbf{x} \quad ; \quad f\mathbf{x} \xrightarrow{\eta f} ff\mathbf{x} \xrightarrow{\mu} f\mathbf{x} \right\}$$

where $f\mu$ is shorthand for rewrite data consisting of the function symbol f being applied to the rewrite rule μ etc.

Discussion. The 2-dimensional syntax of functors and natural transformations is richer than that of rewrites, so a word must be said why Def. 2.2 (which is stated with a view towards the proof of our main theorem) has sufficient expressive power. To wit, when 'defining a structured category via functors and natural transformations', one can (i) make free use of products, projections from products, and diagonal functors into products of categories (including the identity functor) (ii) precompose a functor of n variables with up to n functors (iii) precompose a functor of n variables with up to n natural transformations (iv) precompose a natural transformation with functors (v) compose natural transformations (2-cells) $F \xrightarrow{\eta} G$, $G \xrightarrow{\xi} H$ horizontally $F \xrightarrow{\eta \cdot \xi} H$ (vi) compose natural transformations 'vertically', e.g. $F_1 \xrightarrow{\eta} F_2$ and $G_1 \xrightarrow{\xi} G_2$ to $G_1 \circ F_1 \xrightarrow{\eta \star \xi} G_2 \circ F_2$ (where G_i is assumed composable with F_i) and (vii) take the inverse of natural isomorphisms.

Of these, (i), (ii) and (iv) are built into Def. 2.2. That definition also permits the precomposition of a functor by a *single* natural transformation but any instance of (iii) is expressible, by naturality, as a horizontal composition of natural transformations of one variable only, hence as a chain of rewrites, i.e. a path in the graph of terms. By naturality, (vi) is also expressible via (iii), (iv) and (v). Finally, identity morphisms and hence, natural isomorphisms and inverses can be enforced via the convention on paths of lengths zero.

The upshot is that the commutativity of any two-dimensional 'pasting diagram' of functors and natural transformations can be equivalently expressed by the commutativity of a set of pairs of paths in the corresponding graph of terms. Such an equivalent expression is far from unique in general.

3. Coherence and categorification

Recall that an interpretation of the data $\langle S, F, R\rangle$ assigns to each term a functor $\prod_{x_i^{(s)}} \mathcal{C}_s \to \mathcal{C}_r$ and to each edge of the graph of terms a natural transformation. If C is a collection of pairs of paths with the same initial and terminal vertices, an interpretation is called a model of C if each cycle commutes.

Definition 3.1. The axioms $\langle S, F, R, C\rangle$ are *coherent in the sense of MacLane* if, within any model, for any two vertices v, w, the composites of the natural transformations along any two edge-paths $p_1 : \mathsf{v} \to \mathsf{t}_1 \to \ldots \mathsf{t}_i \to \mathsf{w}$, $p_1 : \mathsf{v} \to \mathsf{u}_1 \to \ldots \mathsf{u}_j \to \mathsf{w}$ are equal.

Coherence in this sense is a strong requirement (as commutativity is required to hold even when the path p_1 or p_2 has length 0!) but seems to be in the spirit of MacLane's dictum [Mac71] that coherence theorems state that *all* diagrams well-formed from the data commute — provided the diagrammatic axioms do.

Definition 3.2. Let $\mathcal{U}$ be an equational universal algebra with sorts S and set of function symbols F. The 2-theory $\langle S, F, R, C\rangle$ is a *categorification* of $\mathcal{U}$ if

(i) two terms u, v belong to the same connected component of the graph of terms if and only if $\mathsf{u} = \mathsf{v}$ in the universal algebra $\mathcal{U}$ and
(ii) in any model of $\langle S, F, R, C\rangle$, all edges are interpreted by natural transformations that are natural *isomorphisms*.

In other words, in models of $\langle S, F, R, C\rangle$ the identities that hold between terms in $\mathcal{U}$ are replaced by natural isomorphisms between the corresponding functors; the equivalence relation between terms generated by these natural isomorphisms coincides with equational identity. A universal algebra has infinitely many categorifications in this sense (indeed, every equational axiomatization can be turned into one). The categorification being coherent is the same as saying that the graph of terms, in any model of $\langle S, F, R, C\rangle$, is interpreted by a groupoid whose components are *trivial* (i.e. by a groupoid each of whose hom-sets is either empty or contains precisely one arrow). Any equational universal algebra has coherent categorifications; but the problem that motivated this paper is: given a finitary equational universal algebra, find a *finitary* coherent categorification.

Def. 3.1 is semantic in nature in that it contains quantification over the class of all models. It is possible to give an equivalent syntactic definition expressed directly in terms of the graph of terms, but we shall not need to do so here. Prior to stating and proving the main theorem, though, we need to identify two families of cycles in the graph of terms. The ones from the first family (denoted D_0) are commutative in any interpretation of $\langle S, F, R\rangle$ purely by virtue of category theory. Cycles from the second family (denoted D_1) are commutative in any model of the axioms C.

(D_0) Consequences of the functoriality of natural transformations. Let t be any term, and z a variable of sort s that occurs at least once in t. Let $\mathsf{x} \to \mathsf{y}$ be an edge between terms of sort s. Then let

$$\mathsf{t}[z \mapsto \mathsf{x}] \dashrightarrow \mathsf{t}[z \mapsto \mathsf{y}]$$

denote any directed edge path from $\mathsf{t}[z \mapsto \mathsf{x}]$ to $\mathsf{t}[z \mapsto \mathsf{y}]$ that results from replacing, one by one but in any order, each occurrence of z in t by y instead of x. By convention, if z does not occur in t, let $\mathsf{t}[z \mapsto \mathsf{x}] \dashrightarrow \mathsf{t}[z \mapsto \mathsf{y}]$ stand for the path of length 0 that consists of the term t.

Example 3.3. If $\mathsf{t} = \mathsf{t}_1 z\, \mathsf{t}_2 z\, \mathsf{t}_3 z\, \mathsf{t}_4$, with no other occurrence of z in t, then $\mathsf{t}[z \mapsto \mathsf{x}] \dashrightarrow \mathsf{t}[z \mapsto \mathsf{y}]$ may stand for the path

$$\mathsf{t}_1\mathsf{x}\,\mathsf{t}_2\mathsf{x}\,\mathsf{t}_3\mathsf{x}\,\mathsf{t}_4 \to \mathsf{t}_1\mathsf{y}\,\mathsf{t}_2\mathsf{x}\,\mathsf{t}_3\mathsf{x}\,\mathsf{t}_4 \to \mathsf{t}_1\mathsf{y}\,\mathsf{t}_2\mathsf{x}\,\mathsf{t}_3\mathsf{y}\,\mathsf{t}_4 \to \mathsf{t}_1\mathsf{y}\,\mathsf{t}_2\mathsf{y}\,\mathsf{t}_3\mathsf{y}\,\mathsf{t}_4$$

or

$$\mathsf{t}_1\mathsf{x}\,\mathsf{t}_2\mathsf{x}\,\mathsf{t}_3\mathsf{x}\,\mathsf{t}_4 \to \mathsf{t}_1\mathsf{x}\,\mathsf{t}_2\mathsf{y}\,\mathsf{t}_3\mathsf{x}\,\mathsf{t}_4 \to \mathsf{t}_1\mathsf{y}\,\mathsf{t}_2\mathsf{y}\,\mathsf{t}_3\mathsf{x}\,\mathsf{t}_4 \to \mathsf{t}_1\mathsf{y}\,\mathsf{t}_2\mathsf{y}\,\mathsf{t}_3\mathsf{y}\,\mathsf{t}_4$$

etc. Each of the above edges arises as postcomposition of $\mathsf{x} \to \mathsf{y}$.

Let now $\mathsf{u}_1 \to \mathsf{u}_2$ be an edge, x a variable of sort s, and $\mathsf{v}_1 \to \mathsf{v}_2$ an edge between terms of sort s. That gives rise to cycles of the form

$$\begin{array}{ccc} \mathsf{u}_1[x \mapsto \mathsf{v}_1] & \dashrightarrow & \mathsf{u}_1[x \mapsto \mathsf{v}_2] \\ \downarrow & & \downarrow \\ \mathsf{u}_2[x \mapsto \mathsf{v}_1] & \dashrightarrow & \mathsf{u}_2[x \mapsto \mathsf{v}_2] \end{array}$$

(More precisely, that shape corresponds to x occurring both in u_1 and u_2; if x occurs in u_1 but not in u_2, the cycle 'looks like'

$$\begin{array}{ccc} \mathsf{u}_1[x \mapsto \mathsf{v}_1] & \dashrightarrow & \mathsf{u}_1[x \mapsto \mathsf{v}_2] \\ & \searrow \quad \swarrow & \\ & \mathsf{u}_2 & \end{array}$$

and dually when x occurs in u_2 but not in u_1. If x occurs in neither u_1 nor u_2, the rectangle reduces to a parallel pair of identical edges.)

Let D_0 be the collection of all cycles in the graph of terms arising in this fashion.

Example 3.4. Consider a one-sorted structure with a single binary operation $\mathcal{C} \times \mathcal{C} \xrightarrow{-\circledast-} \mathcal{C}$. Assume it is equipped with a natural transformation of the form $(X \circledast Y) \circledast Z \xrightarrow{\eta_{XYZ}} X \circledast (Y \circledast Z)$. Then the diagram

$$\begin{array}{ccc} (((X \circledast Y) \circledast Z) \circledast V) \circledast W & \longrightarrow & (X \circledast (Y \circledast Z)) \circledast V) \circledast W \\ \downarrow & & \downarrow \\ ((X \circledast Y) \circledast Z) \circledast (V \circledast W) & \longrightarrow & (X \circledast (Y \circledast Z)) \circledast (V \circledast W) \end{array}$$

commutes. It is the case of the above rectangle when $\mathsf{u}_1 \to \mathsf{u}_2 = (U \circledast V) \circledast W \to U \circledast (V \circledast W)$, $\mathsf{v}_1 \to \mathsf{v}_2 = (X \circledast Y) \circledast Z \to X \circledast (Y \circledast Z)$, and the role of x is played by U. In prefix notation, U would be replaced in the literal sense of strings, without the need for occasional extra parentheses.

This quadrangle is one of the faces of the Stasheff polytope K_5. It commutes in the absence of any assumptions about the Stasheff pentagon K_4. There are two more quadrilateral faces in K_5, both of which are naturally commutative; they correspond to analogous replacements of V and W.

(D_1) The commutativities entailed by the paths in C under pre- and post-compositions with terms. First, let us define these operations:

(precomposition) If $\mathsf{x} \to \mathsf{y}$ is an edge and $x_i \mapsto \mathsf{t}_i$ are variable substitutions, then

$$\mathsf{x}[x_i \mapsto \mathsf{t}_i] \to \mathsf{y}[x_i \mapsto \mathsf{t}_i]$$

is also an edge. Indeed, let $\mathsf{x} \to \mathsf{y}$ arise as $\mathsf{x}_1\mathsf{U}\mathsf{x}_2 \to \mathsf{y}_1\mathsf{V}\mathsf{y}_2$ where $\langle \mathsf{U}, \mathsf{V} \rangle$ is a substitution instance of a rewrite rule $\langle \mathsf{u}, \mathsf{v} \rangle$. Then the above edge amounts to performing the substitutions $x_i \mapsto \mathsf{t}_i$ in each of x_1, x_2, y_1, y_2, U and V. (Note that any substitution instance of $\langle \mathsf{U}, \mathsf{V} \rangle$ is also an instance of $\langle \mathsf{u}, \mathsf{v} \rangle$.)

(postcomposition) If $\mathsf{x} \to \mathsf{y}$ is an edge between terms of sort s and $\mathsf{w}_1 z \mathsf{w}_2$ is a well-formed term with z a variable of sort s, then

$$\mathsf{w}_1\mathsf{x}\mathsf{w}_2 \to \mathsf{w}_1\mathsf{y}\mathsf{w}_2$$

is also an edge. Indeed, with x, y as above, this edge arises as $\mathsf{w}_1\mathsf{x}_1\mathsf{U}\mathsf{x}_2\mathsf{w}_2 \to \mathsf{w}_1\mathsf{x}_1\mathsf{V}\mathsf{x}_2\mathsf{w}_2$.

For a tuple of edges between terms all of the same sort (such as an edge path), pre- resp. post-composition with a term is defined by performing these actions for each edge. Close the set of cycles in C under arbitrary precomposition, then under postcomposition to generate a collection of diagrams (necessarily closed under both pre- and post-composition) that we will denote D_1.

To sum up, any interpretation of $\langle S, F, R \rangle$ makes the diagrams in D_0 commute; any model of $\langle S, F, R, C \rangle$ makes in addition the diagrams in D_1 commute. Pasting diagrams of these types side by side, we will be able to deduce the coherence of 2-theory axioms stemming from convergent rewrite systems.

4. COHERENCE OF CONVERGENT PRESENTATIONS

Let us first recall Knuth's *critical pair lemma* that plays a key role in both the Knuth-Bendix algorithm and our coherence proof. A *span* is a diagram $\mathsf{r}_1 \leftarrow \mathsf{t} \to \mathsf{r}_2$ in the graph of terms, i.e. a term with two distinct rewritings. The critical pair lemma says that spans come in one of three types. The first two of these, it turns out, are harmless from the viewpoint of functoriality, since they can be completed by diagrams belonging to the family D_0; and the third one possesses 'templates' which we will use for coherence conditions.

We say that a span $\mathsf{r}_1 \leftarrow \mathsf{t} \to \mathsf{r}_2$ *extends* $\mathsf{r}'_1 \leftarrow \mathsf{t}' \to \mathsf{r}'_2$ if $\mathsf{r}_1 \leftarrow \mathsf{t} \to \mathsf{r}_2$ arises from $\mathsf{r}'_1 \leftarrow \mathsf{t}' \to \mathsf{r}'_2$ by a pre- followed by a post-composition, in the sense of composition of edges by terms introduced above.

Lemma 4.1. (Knuth) *Let a finite set of sorts, function symbols and rewrite rules be given. From this data one can effectively compute a finite set of spans, called* critical pairs*, such that if* $\mathsf{r}_1 \leftarrow \mathsf{t} \rightarrow \mathsf{r}_2$ *is any span, then either*

(1) t *permits a decomposition* $\mathsf{t} = \mathsf{t}_1\mathsf{U}_1\mathsf{t}_2\mathsf{U}_2\mathsf{t}_3$ *and there exist instances* $\langle \mathsf{U}_1, \mathsf{V}_1 \rangle$, $\langle \mathsf{U}_2, \mathsf{V}_2 \rangle$ *of rewrite rules such that one of* $\mathsf{t} \rightarrow \mathsf{r}_1$ *and* $\mathsf{t} \rightarrow \mathsf{r}_2$ *is*

$$\mathsf{t}_1\mathsf{U}_1\mathsf{t}_2\mathsf{U}_2\mathsf{t}_3 \rightarrow \mathsf{t}_1\mathsf{V}_1\mathsf{t}_2\mathsf{U}_2\mathsf{t}_3$$

and the other one is

$$\mathsf{t}_1\mathsf{U}_1\mathsf{t}_2\mathsf{U}_2\mathsf{t}_3 \rightarrow \mathsf{t}_1\mathsf{U}_1\mathsf{t}_2\mathsf{V}_2\mathsf{t}_3,$$

or

(2) *there exist edges* $\mathsf{u}_1 \rightarrow \mathsf{u}_2$, $\mathsf{v}_1 \rightarrow \mathsf{v}_2$, *a variable* x *occurring at least once in* u_1 *and an occurrence of* v_1 *in* $\mathsf{u}_1[x \mapsto \mathsf{v}_1]$ *(corresponding to a replacement instance of* x*) say,* $\mathsf{u}_1[x \mapsto \mathsf{v}_1] = \mathsf{w}_1\mathsf{v}_1\mathsf{w}_2$, *such that one edge of the span can be written*

$$\mathsf{t} = \mathsf{u}_1[x \mapsto \mathsf{v}_1] \rightarrow \mathsf{u}_2[x \mapsto \mathsf{v}_1]$$

while the other one arises as

$$\mathsf{t} = \mathsf{u}_1[x \mapsto \mathsf{v}_1] = \mathsf{w}_1\mathsf{v}_1\mathsf{w}_2 \rightarrow \mathsf{w}_1\mathsf{v}_2\mathsf{w}_2$$

or

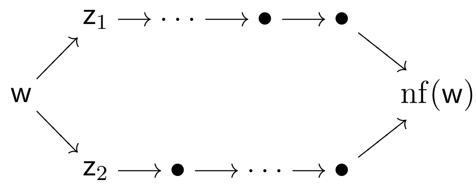
(3) $\mathsf{r}_1 \leftarrow \mathsf{t} \rightarrow \mathsf{r}_2$ *extends a critical pair.*

Sketch of proof. By definition, there exist instances $\langle \mathsf{U}_1, \mathsf{V}_1 \rangle$, $\langle \mathsf{U}_2, \mathsf{V}_2 \rangle$ of rewrite rules $\langle \mathsf{u}_1, \mathsf{v}_1 \rangle$, $\langle \mathsf{u}_2, \mathsf{v}_2 \rangle$ such that $\mathsf{t} \rightarrow \mathsf{r}_1$ is $\mathsf{t}_1\mathsf{U}_1\mathsf{t}_2 \rightarrow \mathsf{t}_1\mathsf{V}_1\mathsf{t}_2$ and $\mathsf{t} \rightarrow \mathsf{r}_2$ is $\mathsf{t}'_1\mathsf{U}_2\mathsf{t}'_2 \rightarrow \mathsf{t}'_1\mathsf{V}_2\mathsf{t}'_2$. The three cases now depend on the relative position of U_1 and U_2 in t. Case (1) is when they are disjoint. If they overlap as substrings then (since they are subterms) one of them, say U_1, must contain the other. Case (2) is when U_2 is substring of a trivial replacement instance of u_1 (U_2 is subterm of a term that is substituted into a variable occurring in u_1). In the remaining case (3), the overlap is non-trivial; under some variable substitution, u_2 becomes a subterm of u_1. That substitution must then factor through the most general unifier of u_2 and the corresponding subterm of u_1. The critical pairs are thus computed as follows: take all pairs $\langle \mathsf{u}_1, \mathsf{v}_1 \rangle$, $\langle \mathsf{u}_2, \mathsf{v}_2 \rangle$ of rules; take all subterms u'_1 of u_1; if u'_1 and u_2 are unifiable, use their most general unifier to define a substitution instance of $\langle \mathsf{u}_1, \mathsf{v}_1 \rangle$ resp. post-composition of a substitution instance of $\langle \mathsf{u}_2, \mathsf{v}_2 \rangle$ with identical left-hand sides. See Knuth–Bendix [KB70] or Baader and Nipkow [BN98] for details.

Remark. Knuth stated his lemma for term rewrite systems in the classical sense of section 1, and that is all we will need. It remains valid for labeled rewrite systems in the sense of section 2, but critical pairs can then also have the form $\mathsf{r} \xleftarrow{\alpha} \mathsf{t} \xrightarrow{\beta} \mathsf{r}$ that would be tautologous for an algebraic term rewrite system.

Now we can state the main result.

Theorem 4.2. *Let* $\langle S, F, R \rangle$ *be a finite convergent axiomatization of an equational universal algebra* $\mathfrak{U}$*, with set of critical pairs* P*. For each* $\mathsf{z}_1 \leftarrow \mathsf{w} \rightarrow \mathsf{z}_2$ *in* P*, choose* some *sequence of rewrites* *that transforms* z_1 *into the normal form* $\mathrm{nf}(\mathsf{w})$ *of* w*, and choose* some *such sequence of rewrites from* z_2 *to* $\mathrm{nf}(\mathsf{w})$ *too. This defines a cycle in the graph of terms:*

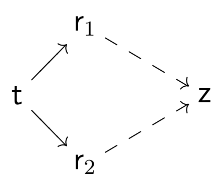


Take C to be the set of cycles obtained in this fashion (one for each critical pair). Then the 2-theory presented by $\langle S, F, R^+ \cup R^-, C \cup I\rangle$ is a finite, MacLane-coherent categorification of $\mathfrak{U}$. Here R^+ is the same as R, the original set of rewrite axioms; R^- is a formal inverse $\mathsf{v} \to \mathsf{u}$ for each rule $\mathsf{u} \to \mathsf{v}$ present in R; and I contains the cycles $\{\mathsf{v} \to \mathsf{u} \to \mathsf{v}; \mathsf{v}\}$, $\{\mathsf{u} \to \mathsf{v} \to \mathsf{u}; \mathsf{u}\}$ for each rule in R.

Proof. Recall from Def. 3.2 that in a categorified algebra, the edges between terms are to be interpreted by natural *isomorphisms*; that is the reason for adding a formal converse to each rewrite rule and the additional diagrams in I. A model of $\langle S, F, R^+ \cup R^-, C \cup I\rangle$ is exactly the same as a model of $\langle S, F, R, C\rangle$ where all rewrites are interpreted by natural isomorphisms. For much of the proof, we will actually work with a model of $\langle S, F, R, C\rangle$ and exploit convergence; it is only at the last step that we use the assumption on isomorphisms.

So, fix an arbitrary model of $\langle S, F, R, C\rangle$. The key is the following lemma, the only one, in fact, that directly exploits the diagrams in C.

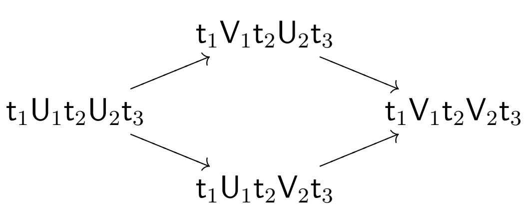
Lemma 4.3. *Let $\mathsf{r}_1 \leftarrow \mathsf{t} \to \mathsf{r}_2$ be any span. There exist a term z and edge-paths $\mathsf{r}_1 \rightsquigarrow \mathsf{z}$, $\mathsf{r}_2 \rightsquigarrow \mathsf{z}$ such that*



commutes.

Proof. By Knuth's theorem, the span $\mathsf{r}_1 \leftarrow \mathsf{t} \to \mathsf{r}_2$ is one of three types. For each of those, we will find a way to appropriately complete the diagram. We retain the notation of Lemma 4.1.

(1) Choose $\mathsf{z} = \mathsf{t}_1\mathsf{V}_1\mathsf{t}_2\mathsf{V}_2\mathsf{t}_3$ and note that the square



completes the span as desired. It commutes since it is an element of D_0.

(2) Assume that x occurs in u_2. There are commutative rectangles belonging to the family D_0

$$\begin{array}{ccc} \mathsf{u}_1[x \mapsto \mathsf{v}_1] & \dashrightarrow & \mathsf{u}_1[x \mapsto \mathsf{v}_2] \\ \downarrow & & \downarrow \\ \mathsf{u}_2[x \mapsto \mathsf{v}_1] & \dashrightarrow & \mathsf{u}_2[x \mapsto \mathsf{v}_2] \end{array}$$

Choose such a sequence of edges for the top row that the first instance of v_1 to be replaced is at the location $\mathsf{w}_1\mathsf{v}_1\mathsf{w}_2$, creating a diagram

$$\begin{array}{ccc} \mathsf{u}_1[x \mapsto \mathsf{v}_1] \longrightarrow \mathsf{w}_1\mathsf{v}_2\mathsf{w}_2 & \dashrightarrow & \mathsf{u}_1[x \mapsto \mathsf{v}_2] \\ \downarrow & & \downarrow \\ \mathsf{u}_2[x \mapsto \mathsf{v}_1] & \dashrightarrow & \mathsf{u}_2[x \mapsto \mathsf{v}_2] \end{array}$$

Set $\mathsf{z} = \mathsf{u}_2[x \mapsto \mathsf{v}_2]$ to obtain a suitable completion of the span. The case when x does not occur in u_2 is similar; the bottom row is replaced by the term u_2, yielding a commutative 'triangle'.

(3) If $\mathsf{r}_1 \leftarrow \mathsf{t} \rightarrow \mathsf{r}_2$ extends the critical pair $\mathsf{z}_1 \leftarrow \mathsf{w} \rightarrow \mathsf{z}_2$, then the diagram

$$\begin{array}{ccccc} & & \mathsf{z}_1 \longrightarrow \cdots \longrightarrow \bullet \longrightarrow \bullet & & \\ & \nearrow & & \searrow & \\ \mathsf{w} & & & & \mathrm{nf}(\mathsf{w}) \\ & \searrow & & \nearrow & \\ & & \mathsf{z}_2 \longrightarrow \bullet \longrightarrow \cdots \longrightarrow \bullet & & \end{array}$$

corresponding to the latter in C will, under the same pre- and post-compositions, extend to a diagram

$$\begin{array}{ccccc} & & \mathsf{r}_1 \longrightarrow \cdots \longrightarrow \bullet \longrightarrow \bullet & & \\ & \nearrow & & \searrow & \\ \mathsf{t} & & & & \mathsf{z} \\ & \searrow & & \nearrow & \\ & & \mathsf{r}_2 \longrightarrow \bullet \longrightarrow \cdots \longrightarrow \bullet & & \end{array}$$

contained in the family D_1.

□

Proposition 4.4. *Let* t *be any term, and consider any two chains of rewrites that convert* t *into its normal form* $\mathrm{nf}(\mathsf{t})$*:*

$$\begin{array}{ccccc} & & \mathsf{r}_1 \longrightarrow \cdots \longrightarrow \bullet \longrightarrow \bullet & & \\ & \nearrow & & \searrow & \\ \mathsf{t} & & & & \mathrm{nf}(\mathsf{t}) \\ & \searrow & & \nearrow & \\ & & \mathsf{r}_2 \longrightarrow \bullet \longrightarrow \cdots \longrightarrow \bullet & & \end{array}$$

The above diagram commutes.

Proof. The following observation follows from the contrapositive of König's lemma on infinite trees, but we include the proof for completeness.

Lemma 4.5. *In a noetherian rewrite system, there exist only finitely many rewrite chains*

$$\mathsf{t}_0 \to \mathsf{t}_1 \to \mathsf{t}_2 \to \ldots \to \mathsf{t}_k$$

starting at any given term t_0.

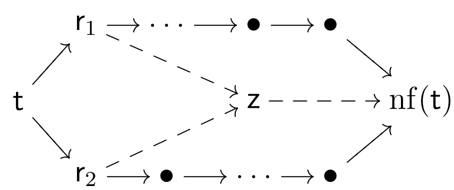
Proof. By contradiction. Suppose that infinitely many rewrite chains have t_0 as their starting term. There are only finitely many terms u such that $\mathsf{t}_0 \to \mathsf{u}$; say, these are $\mathsf{u}_1, \mathsf{u}_2, \ldots, \mathsf{u}_k$. It must be that for some i, u_i is the starting term of infinitely many rewrite chains. Define $\mathsf{t}_1 = \mathsf{u}_i$ for some such i. Continue this way building a non-terminating rewrite chain $\mathsf{t}_0 \to \mathsf{t}_1 \to \mathsf{t}_2 \to \ldots$. But that contradicts the assumption the rewrite system is noetherian. □

Definition 4.6. In a noetherian rewrite system, define the *depth* of a term t by

$$\operatorname{depth}(\mathsf{t}) = \max\left\{k \mid \text{there exists a rewrite chain } \mathsf{t} \to \mathsf{t}_1 \to \mathsf{t}_2 \to \ldots \to \mathsf{t}_k\right\}$$

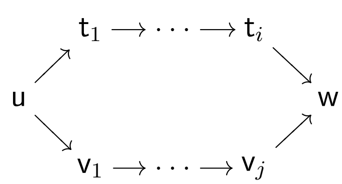
setting $\operatorname{depth}(\mathsf{t}) = 0$ if t permits no rewrites.

The proof of Prop. 4.4 is by induction on $\operatorname{depth}(\mathsf{t})$. When $\operatorname{depth}(\mathsf{t}) = 0$, the statement is vacuously true. We may assume that the edges $\mathsf{t} \to \mathsf{r}_1$ and $\mathsf{t} \to \mathsf{r}_2$ are different, otherwise apply the induction hypothesis to the diagram starting at $\mathsf{r}_1 = \mathsf{r}_2$. Now use Lemma 4.3 to find a term z and paths of arrows $\mathsf{r}_1 \rightsquigarrow \mathsf{z}$, $\mathsf{r}_2 \rightsquigarrow \mathsf{z}$ that make the left-hand diamond in

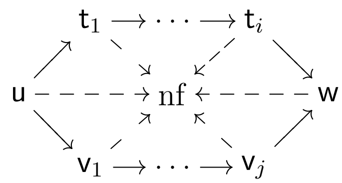


commute. The normal form of z must be $\operatorname{nf}(\mathsf{t})$ as well, so there exists a chain of rewrites from z to $\operatorname{nf}(\mathsf{t})$. Apply the induction hypothesis to the parts of the diagram starting at r_1 resp. r_2, noting that $\operatorname{depth}(\mathsf{r}_1) < \operatorname{depth}(\mathsf{t})$ and $\operatorname{depth}(\mathsf{r}_2) < \operatorname{depth}(\mathsf{t})$, to conclude that the outer cycle commutes as well. □

Now we can finish the proof of Thm. 4.2. The desired conclusion is that in any model of $\langle S, F, R^+ \cup R^-, C \cup I\rangle$, all cycles commute in the graph of terms:



A model of $\langle S, F, R^+ \cup R^-, C \cup I\rangle$ is the same as a model of $\langle S, F, R, C\rangle$ where all R-rewrites happen to be interpreted by natural isomorphisms. All the terms in this diagram belong to the same class modulo the equivalence relation generated by R-rewrites, so must have the same normal form nf. For each vertex, choose *some* sequence of R-rewrites to nf:



Any arrow between adjacent vertices of the original cycle is either an R-rewrite or the inverse of an R-rewrite. Either way, all the interior triangles commute by Prop. 4.4. But this means the outer cycle commutes as well, since all morphisms are isomorphisms. □

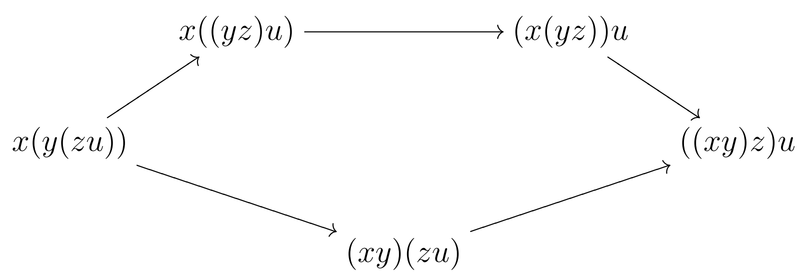
5. Examples and complements

Much of the work in this paper was spent setting up a graphical calculus of commutative diagrams where the notion of critical pair was applicable. (That is the reason, for example, for permitting only *single instances* of rewrites to serve as edges in the graph of terms; if the transitive closure $\to^*$ of elementary rewrites served as edges, the classification of spans would be much more cumbersome.) Having done that work, there are satisfying instances of the main theorem, but also many ways in which it should be extended; we finish with their discussion.

Example 5.1. Take either orientation of the sole axiom $x(yz) = (xy)z$ of semigroups, say, $x(yz) \Rightarrow (xy)z$. There is one critical pair for this system, stemming from the unification of the subterm yz of the left-hand side with the left-hand side itself. The critical pair generated by the most general unifier is

$$x((yz)u) \leftarrow x(y(zu)) \rightarrow (xy)(zu). \tag{5.1}$$

There is a unique chain of rewrites both from $x((yz)u)$ and from $(xy)(zu)$ to $((xy)z)u$. The rewrite system is noetherian as well and hence convergent (without any completion needed). In the notation of Thm. 4.2, the set C of coherence conditions consists of the MacLane pentagon



as advertised in the introduction. The associated 2-theory is that of MacLane monoidal categories (without unit).

Example 5.2. To get from semigroups to monoids, add the rewrite rules $\mathbf{1}x \Rightarrow x$ and $x\mathbf{1} \Rightarrow x$. There are five critical pairs: in addition to (5.1), also

$$(x\mathbf{1})z \leftarrow x(\mathbf{1}z) \rightarrow xz \tag{5.2}$$

$$(xy)\mathbf{1} \leftarrow x(y\mathbf{1}) \rightarrow xy \tag{5.3}$$

$$(\mathbf{1}y)z \leftarrow \mathbf{1}(yz) \rightarrow yz \tag{5.4}$$

$$\mathbf{1} \leftarrow \mathbf{1} \cdot \mathbf{1} \rightarrow \mathbf{1} \tag{5.5}$$

Just as in the previous example, each of these critical pairs can be completed to their normal form uniquely, and the rewrite system is noetherian, so convergent; the associated 2-theory is unital MacLane monoidal categories. Moreover, the five diagrams arising from the application of Thm. 4.2 are exactly the ones originally listed by MacLane [Mac63], predating the discovery of convergent rewrite systems.

Kelly [Kel64] subsequently showed that this set of coherence conditions is redundant: the MacLane pentagon together with the diagram arising from (5.2),

$$\begin{array}{ccc} & x(\mathbf{1}z) & \\ \swarrow & & \searrow \\ (x\mathbf{1})z & \longrightarrow & xz \end{array}$$

imply the other three. (MacLane [Mac71] and many references continue to list the diagrammatic form of (5.5), i.e. that the right and left unit transformations from $\mathbf{1} \cdot \mathbf{1}$ to $\mathbf{1}$ are equal, as an axiom.) Kelly's proof is ingenious and makes heavy use of the assumption that the associativity and unit transformations are natural *isomorphisms*. This shows that the output of Thm. 4.2 need not be a minimal set of coherence conditions. (I am indebted to the referee for this remark.)

Example 5.3. Any convergent axiomatization of groups — cf. Example 1.3 — gives rise to a 2-theory whose models are coherent categorical groups. (Not group objects in categories, but categories with multiplication, unit and inverse functors that are coherent in the sense of MacLane, cf. Def. 3.1.)

A coherent categorical group is necessarily a coherent categorical monoid, so it could equivalently be called a MacLane monoidal category with a coherent inverse for multiplication, at least as long as the underlying signature is the usual (one constant, one binary product, one unary function for the inverse).

Note that Ulbrich [Ulb81], Solian [Sol81] and Laplaza [Lap83] have all introduced notions of coherent categorical groups.

Example 5.4. A homomorphism $f : M \to N$ between semigroups can be considered as a universal algebra with two sorts: source and target, which are semigroups with operation $\star$ and $\cdot$, respectively, and the function symbol $f(-)$ satisfying $f(x \star y) = f(x) \cdot f(y)$. Besides the associativity rewrite rules $x \star (y \star z) \Rightarrow (x \star y) \star z$ and $x \cdot (y \cdot z) \Rightarrow (x \cdot y) \cdot z$, include the rule $f(x) \cdot f(y) \Rightarrow f(x \star y)$. This rewrite system is noetherian, as one easily sees by a suitable term order keeping track of occurrences of f and left parentheses. There are several critical pairs but they are all confluent; so this is a convergent axiomatization. The associated 2-theory is that of coherent monoidal functors (sometimes called 'strong monoidal transformations' since the homomorphism comparison maps were required to be natural isomorphisms). Similarly with units.

Question 5.5. If the universal algebra $\mathcal{U}$ possesses a convergent axiomatization, does the universal algebra $\hom(\mathcal{U})$ possess one too? ($\hom(\mathcal{U})$ has twice as many sorts as $\mathcal{U}$, and its algebras are a pair of $\mathcal{U}$-algebras connected by a $\mathcal{U}$-homomorphism.)

Example 5.6. Coherent monoidal actions. Convergent presentations of monoids M have been extensively investigated, chiefly due to their interaction with properties of the homology groups $H_n(M,\mathbb{Z})$, cf. Squier et al [SOK94]. A monoid presentation

$$\langle g_1, g_2, \ldots, g_n \mid \mathsf{u}_1 \Rightarrow \mathsf{v}_1, \ldots, \mathsf{u}_k \Rightarrow \mathsf{v}_k \rangle$$

(where the u_i, v_i are words in the generators g_j) with *directed* rules is, essentially by definition, a string rewrite system. But it can be considered as an axiomatization of a single-sorted universal algebra where each g_j is a unary function symbol, and the relations $\mathsf{u}_i \Rightarrow \mathsf{v}_i$ express equational identities between composites of these functions (with the single dummy variable suppressed). Models of the associated 2-theory are categories $\mathcal{C}$ equipped with an endofunctor for each of $g_1, g_2, \ldots, g_n$ and natural isomorphisms between the corresponding composites. If one starts with a convergent presentation of the monoid M then Thm. 4.2 applies: a finite number of critical pairs of words and hence coherence diagrams can be found whose validity guarantees (thanks also to the existence of normal forms) that to each element of M one can associate a well-defined endofunctor of $\mathcal{C}$, getting a coherent pseudo-action of M on $\mathcal{C}$ in the usual sense.

Groups with convergent presentations also exist. The case of 'monoids with several objects', i.e. edge rewriting as a case of string rewriting, recalls Jardine's coherent pseudo-simplicial objects [Jar91] and other Reedy diagrams.

Discussion. *Permutative identities.* A severe shortcoming of categorification via convergent rewriting is that it is inapplicable to any universal algebra that contains a permutative (or more precisely *variable-permuting*) axiom. A permutative axiom asserts the equality of the terms t and $\mathsf{t}[x_i \mapsto x_{\sigma(i)}]$, where σ is a non-trivial permutation of variables occurring in t. The best-known example is certainly

$$x \cdot y = y \cdot x$$

but middle-self-interchange

$$\big((a \star b) \star (c \star d)\big) = \big((a \star c) \star (b \star d)\big)$$

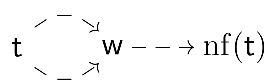
is likewise permutative, as is, say,

$$(x \odot y) \odot (x \circledast x) = (y \odot x) \odot (y \circledast y).$$

The Knuth–Bendix procedure attempts to 'orient' each of the given axioms, hence to include either a rewrite $\mathsf{t} \Rightarrow \mathsf{t}[x_i \mapsto x_{\sigma(i)}]$ or a rewrite $\mathsf{t}[x_i \mapsto x_{\sigma(i)}] \Rightarrow \mathsf{t}$ in the system. But some power of σ is the identity, so either option violates the noetherian property. So no noetherian rewrite system could include, say, the usual axioms of commutative monoids, much less produce MacLane's coherence diagrams for symmetric monoidal categories from them. Moreover, satisfying a permutative identity is an intrinsic property of a universal algebra, independent of its presentation, and one can show that a universal algebra satisfying a non-degenerate permutative identity cannot have a convergent axiomatization. For example, Thm. 4.2 cannot produce any categorification of commutative monoids — not even in a signature or axiomatization different from the usual one.

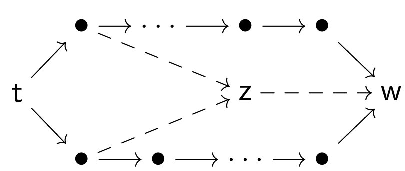
Knuth was aware of the problem with commutativity already in [KB70], and there soon appeared an extensive research effort to develop a theory of rewriting modulo a congruence on terms — such as equivalence induced by a subset of the axioms, for example, by the permutative axioms. See Baader–Nipkow [BN98] for a good overview of the case of theories with commutative-associative operations. The relation between that and categorical coherence — such as symmetric monoidal categories and coherent monoidal functors — has yet to be established. From the viewpoint of category theory, symmetric or braided monoidal axioms arise from the action of an operad on the entire rewrite system. From the viewpoint of term rewriting, what coherence axioms define is a congruence between proofs with respect to which any equational identity possesses a unique equivalence class of equational proofs (i.e. sequence of rewrites from left-hand to right-hand side).

Coherence for natural transformations. Given that the rewrite relation $\mathsf{u} \to \mathsf{v}$ can have *no* symmetric instance in a noetherian system, it is quite ironic that the main result, Thm. 4.2, is applicable only to noetherian systems whose rewrites are interpreted by natural isomorphisms. If one starts with any convergent rewrite system, the proof of Thm. 4.2 remains valid up to and including Prop. 4.4. However, as pointed out above, that proposition would no longer imply the conclusion of Thm. 4.2. Said slightly differently: to go from the fact that all paths from a term t to its normal form $\mathrm{nf}(\mathsf{t})$ commute, to the conclusion that all paths from t to a common target w commute



is automatic if all arrows are isomorphisms (or at least monos), but not in general.

Nonetheless, Laplaza [Lap72] proves that there is a coherent notion of monoidal category whose associator is just a natural transformation; the MacLane pentagon is the only axiom needed. The reason is that in that case, one can sidestep the last part of the proof of Thm. 4.2: given two paths from t to w, some term z *and* path from z to w can always be found to form a diagram



where the inner diamond commutes (by virtue of belonging to D_0 or D_1, as in the proof of Lemma 4.3). The commutativity of the outer diamonds can be assumed by noetherian induction, finishing the proof.

In other words, for the semigroup example 5.1, a finite set of parallel pairs of rewrite paths can be found that (together with the ones commuting by naturality, cf. D_0 above) generate, under pre- and postcomposition and pasting, *all* parallel pairs of rewrite paths. I do not know if that holds in all convergent rewrite systems.

Though Laplaza does not include units in his analysis, I suspect that MacLane's five original axioms for unital monoidal categories — corresponding to the five critical pairs (5.1)–(5.5) —

form a minimal set of coherence conditions for unital monoidal categories where associativity and left and right units are only natural transformations (not necessarily isomorphisms).

2-theories and Lawvere theories. A (finitely presentable) 2-theory in the sense of this paper can be seen to be the same as a (finitely presentable) Lawvere 2-theory over Cat, or the category of models of a finite product sketch enriched over small categories. See Power [Pow99] and Cohen [Coh08]. Hence, models of a 2-theory are algebras of a 2-monad over the category of small categories. These notions may well diverge on the level of morphisms of models (strict vs. pseudo vs. lax homomorphisms of 2-algebras). Analyzing functors between theories (i.e. interpretations of one theory in another) is also likely to be interesting; in particular, developing a notion of coherent equivalence between two coherent categorifications of the same universal algebra.

REFERENCES

[Bek99] Tibor Beke, *Operads from the viewpoint of categorical algebra*, Higher homotopy structures in topology and mathematical physics, Contemp. Math., no. 227, American Mathematical Society, 1999, pp. 29–47.

[BN98] Franz Baader and Tobias Nipkow, *Term rewriting and all that*, Cambridge University Press, 1998.

[Bur93] Albert Burroni, *Higher-dimensional word problems with applications to equational logic*, Theoret. Comput. Sci. **115** (1993), no. 1, 43–62.

[Coh08] Jonathan A. Cohen, *Coherence for rewriting 2-theories*, Thesis, Australian National University, 2008.

[Jar91] J. F. Jardine, *Supercoherence*, Journal of Pure and Applied Algebra **75** (1991), no. 2, 103–194.

[Joh02] P. T. Johnstone, *Sketches of an elephant: a topos theory compendium. Vol. 2.*, Oxford Logic Guides, no. 44, Oxford University Press, 2002.

[KB70] Donald E. Knuth and Peter B. Bendix, *Simple word problems in universal algebras*, Computational Problems in Abstract Algebra (Proc. Conf., Oxford, 1967), Pergamon Press, 1970, pp. 263–297.

[Kel64] G. M. Kelly, *On MacLane's conditions for coherence of natural associativities, commutativities, etc.*, J. Algebra **1** (1964), 397–402.

[Laf95] Yves Lafont, *Equational reasoning with 2-dimensional diagrams*, Term rewriting (Font Romeux, 1993), Lecture Notes in Comput. Sci., no. 909, Springer-Verlag, 1995, pp. 170–195.

[Lap72] Miguel Laplaza, *Coherence for associativity not an isomorphism*, Journal of Pure and Applied Algebra **2** (1972), no. 2, 107–120.

[Lap83] ———, *Coherence for categories with group structure: an alternative approach*, J. Algebra **84** (1983), no. 2, 305–323.

[Mac63] Saunders MacLane, *Natural associativity and commutativity*, vol. 49, Rice Univ. Studies, no. 4, 1963.

[Mac71] ———, *Categories for the working mathematician*, Springer-Verlag, 1971.

[Pow99] John Power, *Enriched lawvere theories*, Theory and Applications of Categories **6** (1999), no. 7, 83–93.

[Sch99] Robert R. Schneck, *Natural deduction and coherence for non-symmetric linearly distributive categories*, Theory and Applications of Categories **6** (1999), no. 9, 105–146.

[SOK94] Craig Squier, Friedrich Otto, and Yuji Kobayashi, *A finiteness condition for rewriting systems*, Theoret. Comput. Sci. **131** (1994), no. 2, 271–294.

[Sol81] Alexandru Solian, *Coherence in categorical groups*, Comm. Algebra **9** (1981), no. 10, 1039–1057.

[Ulb81] K.-H. Ulbrich, *Kohärenz in Kategorien mit Gruppenstruktur*, J. Algebra **72** (1981), no. 2, 279–295.

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF MASSACHUSETTS, LOWELL, ONE UNIVERSITY AVENUE, LOWELL, MA 01854

E-mail address: `tibor_beke@uml.edu`