CS@UH Department of Computer Science
University of Houston

Uniqueness of Normal Forms and Modularity of Confluence for Term Rewriting Systems

Rakesh M. Verma

Department of Computer Science
University of Houston
Houston, TX, 77204, USA
http://www.cs.uh.edu

Technical Report Number UH-CS-11-06

July 6, 2011

Keywords: Term Rewriting Systems, Uniqueness of Normal Forms, Confluence, Combination of Rewrite Systems

Abstract

Programming language interpreters, proving theorems of the form A = B, abstract data types and program optimization can all be represented by a finite set of rules called a rewrite system. In this paper, we study two fundamental concepts, uniqueness of normal forms and confluence, for nonlinear systems in the absence of termination. This is a difficult topic with only a few results so far. Through a novel approach, we show that every system with no overlaps and in which every nonoverlap is an inhomogeneity nonoverlap has unique normal forms. We also prove the confluence of the union (function symbols can be shared) of a nonlinear system with a left-linear system under fairly general conditions. Persistence plays a key role in this proof. We are not aware of any confluence result that allows the same level of function symbol sharing.

Research supported in part by NSF grant CCF 0306475

Uniqueness of Normal Forms and Modularity of Confluence for Term Rewriting Systems

Rakesh M. Verma

Abstract

Programming language interpreters, proving theorems of the form A = B, abstract data types and program optimization can all be represented by a finite set of rules called a rewrite system. In this paper, we study two fundamental concepts, uniqueness of normal forms and confluence, for nonlinear systems in the absence of termination. This is a difficult topic with only a few results so far. Through a novel approach, we show that every system with no overlaps and in which every nonoverlap is an inhomogeneity nonoverlap has unique normal forms. We also prove the confluence of the union (function symbols can be shared) of a nonlinear system with a left-linear system under fairly general conditions. Persistence plays a key role in this proof. We are not aware of any confluence result that allows the same level of function symbol sharing.

Index Terms

Term Rewriting Systems, Uniqueness of Normal Forms, Confluence, Combination of Rewrite Systems

I. Introduction

Two of the most challenging and important problems in rewriting are proving the Unique-Normal-Form and Church-Rosser (also called confluence) properties for non-left-linear (nonlinear, for short) systems, particularly in the absence of termination. There is considerable progress on proving Church-Rosser theorems for left-linear systems (systems in which the left-hand sides (lhs's) of the rules contain at most one occurrence of any variable) [1], [2], [3], [4]. In contrast, for nonlinear systems there are only a handful of general results and almost all of them require termination [2], [5], [6], [7], [8] or left-linearity [9], [10].

In this paper, we attack these two fundamental problems and prove the following results:

- We classify nonoverlaps into two classes: nonoverlaps due to inhomogeneity (I-nonoverlaps) and nonoverlaps due to occurs-check (O-nonoverlaps). We then show that every system with no overlaps in which all nonoverlaps are I-nonoverlaps has the unique normal form property. This result is a generalization of Chew's 1981 result [11] in one direction since it allows more general kinds of nonoverlaps. However, to keep the technical details understandable we do not permit root overlaps here, which Chew does allow. The approach used in proving this result is also novel and should be outlined.

We introduce the idea of constraints and their satisfiability in a rewrite system. We then characterize nonoverlaps as certain kinds of unsatisfiable constraints. We then prove that these kinds of constraints remain unsatisfiable even when certain kinds of rules are added to a system and exploit this fact to prove the unique normal form property (UN). We also clarify the relationship between persistence and the unique normal form property. We show that persistence alone is not sufficient to imply uniqueness of normal forms (correcting a claim in [12]).

Comparison with related work on uniqueness of normal forms. To our knowledge, the following works are more closely related to the uniqueness of normal forms work presented here, do not require termination, and are applicable to non-left-linear systems [11], [13], [14], [15], [16]. There is also some work by Gomi, Oyamaguchi and Ohta in Transactions IPSJ of Japan, which we have not been able to access.

In 1980, Klop [13] proved the Church-Rosser property for the disjoint sum of an orthogonal (i.e., left-linear and nonoverlapping; see next section for precise definitions) combinatory reduction system and a single nonlinear rule of various specific forms (e.g., $D(x, x) \to x$ and $D(x, x) \to E(x)$). In 1987, Toyama [15] proved that the disjoint-sum of two Church-Rosser rewrite systems is Church-Rosser. In 1992, Oyamaguchi and Ohta [14] considered the Church-Rosser property for non-E-overlapping right-ground (i.e., right-hand sides contain no variables) rewrite systems. A weaker result than Church-Rosser, viz., uniqueness of normal forms for strongly nonoverlapping, compatible

Research supported in part by NSF grant CCF 0306475

Chew allows root overlaps provided they are compatible, e.g., $x + 0 \to x$ and $0 + x \to x$ root overlap in $0 + 0$.

systems was claimed by Chew in the 1981 STOC [11] (see also [17] for some unique-normal-form results for λ-calculus + specific rules). A strongly nonoverlapping system is one that remains nonoverlapping even when the variables in the lhs's are renamed to make the rules left-linear. More recently, [16] (see also references cited therein) gave a new proof of Chew's theorem using a more sophisticated version of Chew's approach in which they establish the confluence of a specialized conditional linearization of the given non-left-linear system. For obvious reasons, we refer to this approach as *indirect*. In contrast to the indirect approach, our approach is shorter, direct and uses proof orderings on equational proofs that do not require termination of the underlying rewrite system. It appears to us that direct approaches are more promising for subsequent generalization since they do not require a stronger property than uniqueness of normal forms, viz., confluence.

• We prove that the union (generalization of disjoint sum, function symbols can be shared) of a system R_2 with a left-linear system R_1 is confluent provided that the union is semi-terminating (no sequence containing infinite R_2 reductions), persistent and rhs's of rules in R_1 do not share any function symbols with lhs's of rules in R_2. Finally, we give several sufficient conditions that can be checked syntactically, which ensure that the union has the properties we need.

Comparison with related work on confluence of non-disjoint combinations. We are not aware of any confluence result which allows this much function sharing. The closest result is that of Klop's on CRS's. However, Klop's proof cannot be used directly, since it uses postponement of certain kinds of reductions, which does not hold for us. Moreover, Klop gets persistence for free because of the specificity of rules in R_2. Note that Toyama's proof technique cannot be used since it uses the non-increasing nature of ranks of terms, which does not hold for non-disjoint sums. Rao [7] generalizing a result of [8] proved a confluence result for terminating systems that allows some sharing provided that the union is a hierarchical combination and constructor-based. In particular, no sharing of defined symbols is allowed in the lhs's and only constructors can be shared between lhs of the higher system with rhs's of the lower. We note that Rao's proof is somewhat easier since his conditions ensure that the union is also terminating. See also [9], [10] for combination results that either do not allow as much function sharing, or require one of the two properties: termination/left-linearity.

II. Preliminaries

We assume familiarity with basic notions of rewriting (see [18], [19] for excellent surveys). Let V be a countable set of elements called *variables* and Σ be a countable set of function symbols with $\Sigma \cap V = \emptyset$. $\mathcal{T}$ is the set of all terms of a first-order language constructed from V and Σ. It is convenient to think of terms as ordered rooted trees. $\mathcal{T}(S)$ denotes that the terms are constructed from function symbols in S (the set V of variables is implicit). The size of a term s is denoted by $|s|$. The *height* (resp. *size*) of a term s is 0 (resp. 1) if s is a variable or a constant, and $1 + max_i height(s_i)$ (resp. $1 + \sum_{i=1}^{m} |s_i|$) if $s = f(s_1, \ldots, s_m)$. The root symbol of a term is: $root(t) = f$ if $t = f(t_1, \ldots, t_n)$, and $root(t) = t$ if $t \in V$. Consider an extra constant $\Box$ called a hole and the set $\mathcal{T}' = \mathcal{T}(\Sigma \cup \{\Box\})$. Then $C \in \mathcal{T}'$ is called a *context* on Σ. We use the notation $C[, \ldots,]$ for the context containing n holes ($n \geq 0$). A is a *subterm* of B if $B = C[A]$ for some context C.

The notion of a *path* or *occurrence* is used to refer to subterms in a term as follows. A path is either the empty string λ that reaches the root or $o.i$ (o is a path and i an integer) which reaches the ith argument of the root of the subterm reached by o. t/o refers to the subterm of t reached by o and $t[o \leftarrow s]$ denotes the term obtained by replacing the subterm t/o by s. We define the prefix relation (notation: $\leq$) on occurrences $o \leq q$ whenever $\exists p\ o.p = q$; if $p \neq \lambda$ also, then $o < q$ (proper prefix). For any term t its set of occurrences is denoted $O(t)$.

A *substitution* maps variables to terms. An *instance* $\sigma(s)$ of a term s is obtained by substituting $\sigma(x)$ for every variable x in s. A *rule* is a pair of terms $l \rightarrow r$, such that $l \notin V$ and every variable occurring in r also appears in l (variables are implicitly universally quantified). A *system* R is a finite set of rules. We say that a rule $A \rightarrow B$ is a *rule instance* of the rule $l \rightarrow r$ if we can substitute terms for the variables in $l \rightarrow r$ to get $A \rightarrow B$. As the variables of each rule are universally quantified we assume hereafter that any two distinct rules do not share any variable. Terms s and t are *unifiable* if and only if there is a ground term C which is an instance of both s and t. We say s *overlaps* t if and only if a non-variable proper subterm u of one of the two terms unifies with the other term. (When checking for overlaps it is best to relabel the variables in s and t so that they do not share any variables.) A set $S \subseteq \mathcal{T}$ is *nonoverlapping* if and only if for all s, $t \in S$, *not*(s overlaps t). (Since s and t could be equal, the definition of nonoverlapping does not allow self-overlapping rules like associativity.) We say that s and t *root overlap* if and only if they are left-hand sides of two distinct rules in the system and they unify.

Note that our definition of nonoverlapping allows root overlaps. Therefore, whenever necessary we will use the phrase *no overlaps* to forbid both root and nonroot overlaps. A rewrite system with root overlaps is *consistent* if $\sigma(r) = \sigma'(R)$ whenever $\sigma(l) = \sigma'(L)$ for two distinct rules $l \rightarrow r$ and $L \rightarrow R$.

We say that s rewrites to t in one step at position p (by R), denoted by $s \rightarrow_{R,p} t$, if $s|_p = l\sigma$ and $t = s[r\sigma]_p$, for some $l \rightarrow r \in R$ and substitution σ. If $p = \lambda$, then the rewrite step is said to be applied *at the topmost position* (at the root) and is denoted by $s \overset{r}{\rightarrow}_R t$; it is denoted by $s \overset{nr}{\rightarrow}_R t$ otherwise.

Notation. The letters v, x, y, z denote variables, and f, g, h, etc., denote function symbols of nonzero arity. We use $=_R$ to represent the least equivalence relation containing $\rightarrow_R$. We also say that $a =_R b$ is an equational proof. When the set of rules R is clear from the context, we drop the subscripts from $\rightarrow$. The reflexive-transitive closure of $\rightarrow$ is denoted by $\xrightarrow{*}$, the transitive closure by $\xrightarrow{+}$, and symmetric closure by $\leftrightarrow$. We use $p : A \xrightarrow{*} B$ to give the name p to the reduction sequence ($|p|$ denotes its length). We use $\xrightarrow{r}$ to indicate root reduction (i.e., reduction of the entire term) and $\xrightarrow{nr}$ to indicate nonroot reduction (i.e., reduction at a proper subterm). Similarly, $\xrightarrow{nr*}$ represents a sequence of zero or more nonroot reductions, etc. For every natural number n, $[n]$ denotes $\{1, 2, \ldots, n\}$, $[0] = \emptyset$. If s is any term, then $Var(s)$ denotes the set of variables in s. We say that an equational proof $q : a =_R b$ contains a *root* reduction if there are a_1 and a_2 in the proof such that $q : a =_R a_1 \overset{r}{\longleftrightarrow} a_2 =_R b$. Strictly speaking every equational proof has its associated multiset of occurrences at which the individual proof steps are made but we often omit it for convenience. Let $q : a = a_0 \leftrightarrow_R a_1 \leftrightarrow_R \ldots \leftrightarrow_R a_n = b$ $(n \geq 0)$ be an equational proof, then n the number of steps is its length (also denoted by $|q|$), and equational proof p is a *subproof* of q provided $p : a_i =_R a_j$ for some $0 \leq i \leq j \leq n$. We use [ ] to denote multisets.

Proposition 1 (Basic facts of equational proofs): (i) $s =_R t$ implies $\sigma(s) =_R \sigma(t)$ for any substitution σ (stability under substitution). (ii) $s =_R t$ implies $C[s] =_R C[t]$ for any context C (stability under contexts). (iii) Let $s =_R t$ and let O be the set of occurrences at which reductions are carried out in this proof. If $o \in O$ is any minimal occurrence, then $s/o =_R t/o$ (projection property).

Proof: For (i) a simple induction on the length of the equational proof suffices, using the fact that the rewrite relation is stable under substitutions. For (ii) again an induction using the stability of the rewrite relation under contexts. The proof of (iii) is obvious. □

Remark. Note that these facts can be used together with the subproof operation and the reflexive, symmetric and transitive properties of equational proofs to obtain equational proofs from a given equational proof. For example if $R = \{a \rightarrow b\}$ and the given proof is $f(a) =_R f(b)$, then we can apply projection to get the proof $a =_R b$ and stability under contexts to get the proof $f(f(a)) = f(f(b))$, etc. Similarly, if $R = \{a \rightarrow b, b \rightarrow c\}$, then the proof $a =_R c$ also yields the proofs $a =_R b$ and $b =_R c$.

Definition 2: We say that an equational proof q *yields* an equational proof p (or that p can be *obtained* from q), if p can be obtained by applying some sequence of steps to q, where each step is either a projection, a subproof operation, or use of an equivalence property of equational proofs.

This definition can be made even more precise as a sequence just as the concept of proof is defined in any book on formal logic. Note that our notion of yields only gives proofs of terms that are subterms of terms appearing in the given proof since we do not use the "expansive" stability under contexts operation. We say that relation $\rightarrow$ is *confluent* (CR) if and only if $\forall A, B, C$ $A \xrightarrow{*} B$ and $A \xrightarrow{*} C$ implies $\exists$ D such that $B \xrightarrow{*} D$ and $C \xrightarrow{*} D$.

Definition 3: Let R be a rewrite system. A term t is a *normal form* if there is no term u such that $t \rightarrow u$. A term t *has a normal form* if there is a normal form u such that $t \xrightarrow{*} u$. R (or $\rightarrow_R$) is uniquely normalizing (is $UN^{\rightarrow}$) if for all terms A, B, C such that $A \xrightarrow{*} B$ and $A \xrightarrow{*} C$ and B, C are normal forms we have $B = C$. R (or $\rightarrow_R$) has unique normal forms (is UN) if for all normal forms A, B with $A =_R B$ we have $A = B$. R is terminating if there are no infinite reduction sequences $t_0 \rightarrow t_1 \rightarrow \ldots$.

Lemma 4: For every system R: $CR \Rightarrow UN \Rightarrow UN^{\rightarrow}$. Reverse implications generally do not hold.

Proof: The proofs of both statements are standard (see, for example [13]). We include examples refuting the reverse implications. $UN^{\rightarrow} \not\Rightarrow UN$. Let $R = \{a \rightarrow b, a \rightarrow c, c \rightarrow c, d \rightarrow c, d \rightarrow e\}$. R is $UN^{\rightarrow}$, but not UN since $b =_R e$ and b, e are distinct normal forms. $UN \not\Rightarrow CR$. Let $R = \{a \rightarrow b, b \rightarrow b, a \rightarrow c, c \rightarrow c\}$. R is UN (a, b and c are not normal forms) but R is not CR since b and c do not have a common reduct. □

III. Nonoverlaps, Constraints and Occurrence Sets

Let R be a rewrite system and let $l \rightarrow r$, $l' \rightarrow r'$ be two rules (not necessarily distinct) in R. Assume, for simplicity, that the rules do not share any variables (make a copy with new variables if necessary). Let $Unif$

denote the unification algorithm given below and $Unif^\infty$ denote the algorithm "without the occurs-check". We say that a nonoverlap of l with a non-variable subterm u of l' (proper subterm, if the two rules are the same) is due to *occurs-check* (O-nonoverlap) if l and u unify with $Unif^\infty$ (but not with $Unif$). We say that a nonoverlap is due to *inhomogeneity* (I-nonoverlap), if l and u do not unify even with $Unif^\infty$.

Example. The nonoverlap of $f(x, g(x))$ and $f(x, x)$ at the root is due to the occurs-check, whereas the non-overlaps of all non-variable subterms of $h(z, g(z), a)$ with non-variable subterms of $h(x, x, b)$ are due to inhomogeneity.

We now formalize the concept of constraints introduced by unification of two terms and *the set of occurrences* through which the Unification algorithm generates a constraint. For this purpose, we modify a "naive" unification algorithm given in [20], which is presented below. To present this algorithm we introduce first some notation from [20].

In contrast to the algorithm of [20], the input to the algorithm below are two labeled, rooted, directed trees G and H with two distinguished nodes (roots) u and v, which represent the two given terms to be unified. The algorithm constructs and maintains a relation $\mathcal{R}$ as undirected edges in $G \cup H$. The relation $\mathcal{R}$ is symmetric and reflexive by its representation. In order to make $\mathcal{R}$ a c-e relation, both correspondence and equivalence must be satisfied. Setting children equivalent, when their parents are equivalent is called *propagation*. For $\mathcal{R}$ to be an equivalence relation one must also enforce *transitivity*. Since we do not allow any sharing in the input, ensuring transitivity in our algorithm takes the slightly modified form given below, which we also call *consistency*. After creating the minimal c-e relation $\mathcal{R}$ for which $u\mathcal{R}v$ the algorithm tests for homogeneity. The relation $\mathcal{R}$ is said to be *homogeneous* if for any two nodes $p\mathcal{R}q$, the labels of p and q are not different function symbols. If the check succeeds a new labeled graph G' is constructed by coalescing classes of nodes in G. If G' is acyclic the input is unifiable. If G' is not acyclic, the input is $Unif^\infty$ unifiable. For an example of the algorithm, see [20].

Notation: let x be any node in a rooted tree T. The term represented by x is denoted t_x. Further, o_x denotes the edge labels on the unique directed path from the root to x in T.

Proc Naive-unification(U,V) /* G and H are implicit */

```
set $U\mathcal{R}V$; $O(U, V) = \{\lambda\}$
while ($\mathcal{R}$ is not a c-e relation) do
propagation: while (there exist $p\mathcal{R}q$ having corresponding
    $i$th children $p_i$, $q_i$ not related by $\mathcal{R}$) do
    set $p_i\mathcal{R}q_i$; $O(p_i, q_i) = O(q_i, p_i) = O(p, q) \cup \{o_p.i, o_q.i\}$ od;
transitivity: while (there exist $m\mathcal{R}n$ and $p\mathcal{R}q$ such that $m$ and $p$
    are labeled by the same variable, but $n$, $q$ are not related by $\mathcal{R}$) do
    set $n\mathcal{R}q$; $O(q, n) = O(n, q) = O(m, n) \cup O(p, q) \cup \{o_m, o_p\}$ od od ;
if $\mathcal{R}$ not homogeneous then print Ununifiable
else {coalesce equivalence classes to produce labeled graph $G'$}
  Occurs check: if $G'$ has a cycle then print Ununifiable but $Unif^\infty$ unifiable
  else print Unifiable
```

This process may be viewed pictorially on the labeled trees G and H by starting with an undirected edge between the two roots and then adding undirected edges between nodes based on propagation and transitivity. If the two terms t_u, t_v do not unify and this root nonoverlap is an I-nonoverlap, then (by definition) the algorithm fails in the check for homogeneity. We can then extract all pairs of nodes (p, q) from the input such that t_p and t_q are maximal subterms with root(t_p) $\neq$ root(t_q) and $p\mathcal{R}q$. We call such a pair, an inhomogeneity witness and the pair (t_p, t_q) a constraint introduced by unification. Associated with $p\mathcal{R}q$ is the occurrence set $O(p, q)$ that gives the "unification proof" of $p\mathcal{R}q$. It is clear that any equational proof between instances of t_u and t_v is also a way of adding edges between the nodes of G and H where the edges are equational proofs. Finally, it follows that any equational proof between instances of t_u and t_v must "obstruct" all unification proofs of inhomogeneity witnesses, or more formally the occurrence set of any equational proof must intersect with $O(p, q)$ for every inhomogeneity witness (p, q).

IV. Normal Form Uniqueness Results

In this section we first define persistence and then give sufficient conditions that imply the UN property. Finally, we discuss the relationship of persistence and the UN property. Intuitively, persistence requires that the template of the lhs in a redex is untouched by nonroot reductions, and any root reduction after a sequence of nonroot reductions starting from a redex of l ($l \to r \in R$) can only be applied with the same rule. Hence all the terms in the sequence of nonroot reductions are instances of the template of l. (This property is the inspiration for the term *persistence*.)

Definition 5: R (or $\to_R$) is persistent if for every term A such that (i) $A \overset{r}{\to}_R B$ via $l \to r \in R$, (ii) $A \xrightarrow{nr*}_R A'$, and (iii) $A' \to_R B'$ via $l' \to r' \in R$ applied at $o \in O(A')$, either (1) $A' \overset{r}{\to}_R B'$ and $l' \to r' = l \to r$, or (2) there is a $u \in O(l), u \leq o$, and l/u is a variable.

Note that persistence implies that the system must be nonoverlapping, but the converse is false, e.g., let $R = \{f(x,x) \to a, f(x,g(x)) \to a, b \to g(b)\}$. The above definition of persistence is a slight modification of the definition in [21] to avoid root overlaps.

A. Sufficient Condition for UN

We begin with some definitions. Let $lhs(R)$ (resp. $rhs(R)$) denote the set of distinct (distinct means distinct even after variable renaming) lhs's (resp. rhs's) of rules in R and let $SL(R)$ (resp. $SR(R)$) denote the set of all distinct *non-variable subterms* of the terms in $lhs(R)$ (resp. $rhs(R)$).

We now give our first sufficient condition for uniqueness of normal forms and prove the following characterization theorem, which is used later in the proof.

Notation. For any equational proof q, $O(q)$ is the set of occurrences at which reductions are made in q.

Definition 6: We say that an equational proof between instances of two terms $l, L, p : \sigma(l) =_R \sigma'(L)$ is *harmless* if l and L unify and for every reduction at occurrence $o \in O(p)$ either (i) there is a variable in l at an occurrence which is a prefix of o, or (ii) there is a variable in L at an occurrence which is a prefix of o.

Definition 7: An *interval* is a harmless equational proof $A =_R B$ satisfying $A = \sigma(s)$ and $B = \sigma'(s)$ for some non-variable term $s \in lhs(R) \cup rhs(R)$.

Definition 8: A reduction in an equational proof q is *covered by an interval p of q* if both of its endpoints are part of p. A term in an equational proof q is *in the middle of an interval p of q* if the up to two reductions the term is involved in q are covered by p.

Observe that any root reduction of an equational proof q cannot be covered by a single interval because of the two conditions on every rule of the rewrite system (cf. the Preliminaries section) except for rules of the form $l \to r$ where l and r unify. We assume that each such rule has been replaced by the two rules $l \to h(x_1, \dots, x_n)$ and $h(x_1, \dots, x_n) \to r$, where h is a new function symbol and the x_i's are all the variables in l. Note that the unique normal form property is (trivially) preserved by such a transformation. Note also that we cannot simply eliminate such rules since deleting such rules expands the set of normal forms on the original signature of the rewrite system and hence can destroy the uniqueness of normal forms property.

Definition 9: Let q be an equational proof. The set (resp. multiset) of all the rule instances applied in q are called q's *set (resp. multiset) of associated rule instances*, denoted by $RI(q)$ (resp. $RI[q]$). Let q be a non-null equational proof of the form $q : a = a_1 \leftrightarrow_R a_2 \leftrightarrow_R \dots \leftrightarrow_R a_n = b$ for $n > 1$. We call the set $\{a_2, \dots, a_{n-1}\}$ the set of *inner terms* of q.

For example, if rewrite system R contains the rules $a \to b, f(x) \to x$ and q is the proof $f(a) \to f(b) \to b$, then $RI[q] = RI(q) = \{a \to b, f(b) \to b\}$. Note how the context of the first rule, $a \to b$, applied in q is “lost” in the process.

Definition 10: Given an equational proof q we define a covering of a proof by intervals as follows:

1) If q is an interval, then a possible covering of q is q itself.
2) If $q : C[s_1, \dots s_n] =_R C[t_1, \dots, t_n]$ for a non-empty context C, $n \geq 1$ and $q_i : s_i =_R t_i$ for all i, then either we may cover each q_i separately (called *lower intervals*), or we may cover q itself by intervals, then cover the remaining parts of q by lower intervals.
3) If q has at least one root reduction, we split q at the root reductions and then cover each part separately using (1) or (2).
4) The intervals in any covering must be edge-disjoint, i.e., no two intervals can share a step of the proof.

We are only interested in certain kinds of coverings, called maximum coverings, which we define in two steps. A *maximal covering* of q is any covering in which no interval can be added, nor can any existing interval be extended to an interval that includes one more step of q without violating condition (4). A *maximum covering* of q is a maximal covering that covers the greatest number of reductions of q. There may be more than one maximum covering. Next we define an interval forest structure of a proof.

Definition 11: Given an equational proof q we define an *interval forest structure* by first finding *all*, not necessarily edge-disjoint, intervals in q. Then a DAG structure is defined by considering all the intervals that are just below an interval (immediately nested) and intersect with it as the children of an interval. Doing this for each interval gives rise to a forest of DAGs of intervals. The *height* of an interval I, notation $h(I)$ is the length in number of nodes of a longest path from I to a leaf interval that is I's descendant in the forest. The height of a leaf interval is 1 since we count nodes not edges.

With these definitions, we can define the complexity of a proof.

Definition 12: Let $q : \sigma(l) =_R \sigma'(L)$ be an equational proof. We define the complexity of q with respect to a set S of intervals, denoted $C(q, S)$ and the *complexity* of q, denoted $N(q)$, as follows. Let S be any maximum covering of q, $S = \{p_1, \ldots, p_k\}$, $k \geq 0$. Let $C(q, \emptyset)$ be the multiset $[|t| \mid t \in RI[q]]$. Next suppose S is nonempty. For each interval $p_i : a_i =_R b_i \in S$, let $N(p_i) = [h(p_i) + |t| \mid t = max\{|s| \mid s \to \ldots \in RI(p_i)\}]$. Now, let $C(q, S) = \cup_{i=1}^{k} N(p_i) \ \cup \ [|t| \mid t \to \ldots \in RI[q] - \cup_{p_i \in S} RI[p_i]]$. Finally, $N(q) = min_S\{C(q, S)\}$.

We use the multiset extension of the the usual ordering on the numbers to compare the complexities of proofs. For multisets, the $\cup$ symbol denotes multiset union.

Example: Let R contain the rules $a \to b$ and $f(x) \to x$ and let q be the proof $a \leftarrow f(a) \to f(b) \to b$. Then, $RI[q] = RI(q) = \{f(a) \to a, f(b) \to b, a \to b\}$ and $N(q) = [2, 2, 2]$. One 2 is contributed by the interval ($2 = 1 + |a|$) and the other two are the sizes of lhs's $f(a)$ and $f(b)$ of rules in $RI[q]$.

Fact 13: If p is a subproof of q, then $N(p) \leq N(q)$. Further, if p is a proper subproof of q that is missing at least one reduction of q not covered by any interval, then $N(p) < N(q)$.

Proof: Let S be the set of intervals that minimizes $N(q)$ and $S' = \{I \in S \mid$ Both endpoints of I are in the proof $p\}$. We show that $C(p, S') \leq N(q)$. The result then follows from the fact that $N(p) \leq C(p, S')$ by definition of N. To see that $C(p, S') \leq N(q)$, note that $RI[p] \subset RI[q]$ and that any interval $I \in S$ of q that is cut by p and no longer an interval contributes more to $N(q)$ than to $C(p, S')$ (at least [1+ max lhs size in $RI(I)$] versus multiset of lhs sizes in $RI(I)$).

When p is a proper subproof that is missing at least one reduction, say $s \to t$, of q not covered by any interval, then there is one more occurrence of $|s|$ in $N(q)$ than in $N(p)$ and combining this with $N(p) \leq N(q)$ it follows that $N(p) < N(q)$. □

The following lemma is used in the proof of the characterization theorem.

Lemma 14: Let $l, L \in SL(R)$ and let $q : \sigma(l) =_R \sigma'(L)$ be any equational proof containing at least one root reduction.

(1) Suppose that there is one end, let it be $a = \sigma(l)$, of q such that the root reduction in q closest to a is directed away from a. Consider the subproof p of q, where p is the proof from a to the first root reduction of q. Then, $N(p) < N(q)$, where

$$q : a = \sigma(l) \ \overset{nr*}{\longleftrightarrow} \ s \ \overset{r}{\to} \ t =_R \sigma'(L), \ p : a = \sigma(l) \ \overset{nr*}{\longleftrightarrow} \ s$$

(2) Suppose that q contains two consecutive root reductions using the same rule that are directed away from each other towards its ends, i.e.,

$$q : \sigma(l) =_R t_1 \ \overset{r}{\leftarrow} \ s_1 \ \overset{nr*}{\longleftrightarrow} \ s_2 \ \overset{r}{\to} \ \ t_2 =_R \sigma'(L)$$

Let Q denote the equational proof obtained from q by replacing the subproof $p : t_1 =_R t_2$ of q by a harmless subproof $p' : t_1 =_R t_2$ with $N(p') < N(p)$. Then, $N(Q) < N(q)$. Further if p is a proper subproof of q, then $N(p) < N(q)$.

Proof: For (1) we note that q has at least one root reduction, and no root reduction can be covered by a single interval. So, let S be any set of intervals for which $N(q) = C(q, S)$. Then, each interval in S is either an interval of p also or completely disjoint from p, i.e., no interval spans over a subproof of p and the subproof of q disjoint from p. Hence we may partition S into S_p and $S - S_p$, where S_p are the intervals in S that are contained in p. It is now easily seen that if S minimizes $C(q, S)$ then S_p minimizes $C(p, S_p)$ simultaneously. Moreover, $N(p) \cup [|s|] \subseteq N(q)$. Hence $N(p) < N(q)$.

For (2) we reason as follows. The root reductions of q cannot be covered by single intervals for computing $N(q)$. The intervals that cover the subproof q_1 of q from $\sigma(l)$ to t_1 can also cover the corresponding subproof of Q. Similarly, for the subproof q_2 of q from t_2 to $\sigma'(L)$. Therefore, $N(q) = N(p) \cup N(q_1) \cup N(q_2)$ and $N(Q) \leq N(p') \cup N(q_1) \cup N(q_2)$. From the first equation we get $N(p) \leq N(q)$. and also $N(p) < N(q)$ when p is a proper subproof of q since in that case one of the $N(q_i)$'s must be nonempty. From the two relations for $N(q)$ and $N(Q)$ we get $N(Q) < N(q)$ since $N(p') < N(p)$. □

Definition 15: Define an ordering on equational proofs $q : C =_R D$ and $p : A =_R B$ as follows: $p < q$ if either (i) $N(p) < N(q)$ or (ii) $N(p) = N(q)$ and the multiset $[A, B] < [C, D]$, where $<$ is the multiset-extension of the usual subterm ordering, or (ii) $N(p) = N(q)$, $[A, B] = [C, D]$, and $|p| < |q|$. In other words, a lexicographic combination of N value, multiset of endpoints, and length, in order as stated.

Clearly, the ordering $<$ on equational proofs is well-founded.

Notation: From now on, when we use the phrase *minimal proof*, it means minimal in the proof ordering above.

Fact 16: If equational proof p is a shortest proof obtained from an equational proof $q : \sigma(l) =_R \sigma'(L)$, where l and L do not unify, by steps that mimic a unification algorithm on the ends of q (i.e., by projection and equivalence operations on proofs extracted from q after projection), then $N(p) \leq N(q)$.

Proof: Obvious from definitions of N, projection and equivalence operations since projection does not increase $RI[q]$, and potentially decreases it, nor can equivalence operations on subterms of the ends of q increase $RI[q]$. To see this, let $p : s =_R t$ and $p' : t =_R u$ be two proofs on subterms of the ends of q. Then $RI[p] \cup RI[p'] \subseteq RI[q]$, so an equivalence operation yields a proof $p'' : s =_R u$ with $RI[p''] \subseteq RI[q]$. Note that projection of an interval in q may give rise to proofs with no intervals or smaller intervals. This does not create any problems since the union of the complexities of a collection of smaller intervals is by definition smaller than the complexity of a large interval that spans all of them because the nesting depth of the larger interval is at least one more than its descendant intervals. Moreover, since we consider only maximum coverings in the definition of complexity of a proof, it is not possible to obtain intervals after projection and equivalence operations on a proof that had no intervals to begin with in its maximum covering (i.e., its only maximum covering was empty). Since p is a shortest proof it cannot contain any redundant steps, which can potentially increase the complexity. □

Lemma 17: Let R be any rewrite system with no overlaps with every nonoverlap an I-nonoverlap. Let q be any minimal equational proof in the proof ordering above such that $q : \sigma(l) =_R \sigma'(L)$ between two terms $\sigma(l) \neq \sigma'(L)$ where l and L are two terms in $SL(R)$ that do not unify. Then q contains a root reduction.

Proof: Let $O = O(q)$. Since l and L do not unify (and their nonoverlap is an I-nonoverlap), there are non-variable maximal subterms s and S of $\{l, L\}$ such that $root(s) \neq root(S)$ and $Unif^{\infty}(l, L)$ generates the constraint $s = S$. Without loss of generality, we assume that there is only one such constraint. Let $O_1, \ldots, O_k$, be the different sets of occurrences through which $Unif^{\infty}(l, L)$ generates this constraint. Here k is the number of different ways in which this constraint can be generated. Any equational proof of l, L either interferes with all of these proofs, where by interfere we mean that for each O_i there must be an occurrence $o_i \in O$ such that o_i is a prefix of some occurrence in O_i, or yields an equational proof between instances of s, S.

Hence, in either case we can obtain using one of the O_i's or O an equational proof $q' = b_1 \leftrightarrow_R \ldots \leftrightarrow_R b_m$ from q of an instance of a non-variable subterm, say s', of l or L that contains an instance of s, and a non-variable subterm, say S', of l or L that contains an instance of S, which contains a root reduction. We let q' denote a shortest such proof. A root reduction can be ensured by repeatedly applying projection and equivalence properties on proofs extracted from q by projection (mimicking the steps of a unification algorithm). Now since $root(s) \neq root(S)$ and s and S are non-variable subterms, s and S do not unify, which also means that the superterms of s and S do not unify as well. Further, if q does not contain a root reduction and q' does, then q' is obtained by applying at least one projection operation from q, which means s' and S' are proper subterms and $N(q') \leq N(q)$ (Fact 16). Thus, q' also violates the conditions of the theorem and $q' < q$ since $[s', S'] < [a, b]$, which contradicts the choice (minimality in $<$ ordering) of q. Thus the claim is proved. □

Theorem 18: Let R be any rewrite system with no overlaps and in which every nonoverlap is an I-nonoverlap. Then, there is no (non-null) equational proof $q : \sigma(l) =_R \sigma'(L)$ between two terms $\sigma(l) \neq \sigma'(L)$ where l and L are two terms in $SL(R)$, unless either:

1) at least one of l or L is in $lhs(R)$ and q contains a root reduction, or
2) l and L unify and there exists a harmless proof $p : \sigma(l) =_R \sigma'(L)$ with $N(p) \leq N(q)$.

Proof: Suppose there is a non-null equational proof between two instances of terms in $SL(R)$, $a = \sigma(l)$ and $b = \sigma'(L)$ such that $a \neq b$, which violates both conditions. Choose q, a, b such that $q : a =_R b$ is a minimal such proof. Let $O = O(q)$.

If proof p is obtained by applying a projection to a minimal equational proof q with an occurrence not equal to λ, then $p < q$.

Case A: l and L do not unify. This case also covers the case when at least one of l or L is in $lhs(R)$ with $l \neq L$ because of the no-overlap requirement. By Lemma 17, since q is a minimal proof, q contains a root reduction, $\lambda \in O$. Since $\lambda \in O$ and q violates the theorem, neither l nor L is in $lhs(R)$, so $l, L \in SL(R) - lhs(R)$. There are two cases.

(i) There is at least one end of q such that the first root reduction from it is directed towards the opposite end of q. Wlog, let this end be $a = \sigma(l)$. Consider the (non-null) subproof q' between a and the term closest to a to which a root reduction is applied in q. The proof q' is between an instance of an lhs l' and an instance of $l \in SL(R) - lhs(R)$. Because of the no overlap requirement, l' does not unify with l. Thus, the proof q' also violates the theorem and $N(q') < N(q)$ (Lemma 14), which contradicts the minimality of q.

(ii) Otherwise there are two consecutive root reductions (i.e., there is no root reduction in between the two) in q that are directed towards the ends of q and away from each other. So let t_1 and t_2 be the corresponding inner terms of q such that $t_1 = a_i$ and $t_2 = a_j$ for some $i, j \in [m-1]$ and $i \leq j$. Let $R_1 : \sigma_1(l_1) = t_1 \xrightarrow{r} t_1' = \sigma_1(r_1)$ and $R_2 : \sigma_2(l_2) = t_2 \xrightarrow{r} t_2' = \sigma_2(r_2)$ be the root reductions applied to terms t_1 and t_2 in q. If $t_1 = t_2$, then because of the no-overlap requirement $l_1 = l_2$ and $t_1' = t_2'$. Hence we may delete from q: t_1, the associated root reductions R_1 and R_2 and also one of the terms $t_1' = t_2'$, getting a shorter proof p from q, which contradicts the choice of q. Hence we must have $t_1 \neq t_2$. Let q' be the non-null proof between t_1 and t_2. If lhs's l_1 and l_2 are distinct, then the proof q' satisfies $N(q') < N(q)$ (by definition of complexity since $N(q') \cup [|t_1|, |t_2|] \subseteq N(q)$) and is between instances of distinct lhs's (and distinct lhs's do not unify since R has no overlaps). Thus its existence contradicts the choice of q. If $l_1 = l_2$, then by the minimality of q, q' cannot violate the theorem. This implies that there is a harmless proof $p' : t_1 =_R t_2$ corresponding to q' such that $N(p') \leq N(q')$ (p' may be the same as q').

Since p' is harmless and $l_1 = l_2$ there is a harmless proof $P : \sigma_1(r_1) =_R \sigma_2(r_2)$ ($r_1 = r_2$) with $N(P) < N(q'')$, where $q'' : t_1' =_R t_2'$ is the subproof of q. It is obtained by removing the terms t_1 and t_2, the root reductions R_1 and R_2, and applying the steps of the harmless proof p' (which are between instances of variables in the lhs l_1) between the instances $\sigma_1(x)$ and $\sigma_2(x)$ for each variable x in $r_1 = r_2$ (since every variable in the rhs also appears in the lhs). Note that by construction of P and definition of $N(P)$, $N(p')$ we get $N(P) \leq N(p')$ since the associated rule instances of P are a subset of the associated rule instances of p' and either both are intervals or only P is not. Also since $N(q'') = N(q') \cup [|t_1|, |t_2|]$ we get $N(P) < N(q'')$ since $N(p') \leq N(q')$. Let Q be the proof obtained from q by replacing the subproof between t_1' and t_2' in q by P. By Lemma 14, $N(Q) < N(q)$ and Q also violates the theorem, which contradicts the choice of q.

Case B: Terms a and b are instances of terms l, L in $SL(R)$ that unify and there is no harmless proof $p : a =_R b$ with $N(p) \leq N(q)$. In particular, this means that q itself is not harmless, i.e., there is at least one reduction in q at an occurrence $o \in O$ such that l/o and L/o are nonvariable subterms of l and L respectively and $\sigma(l/o) \neq \sigma'(L/o)$. Further, every harmless proof $p : a =_R b$ must satisfy $N(p) > N(q)$. Note that in Case B, if either $l \in lhs(R)$ or $L \in lhs(R)$, then so must the other term, and therefore $l = L$ by the no overlap requirement. Since q violates the theorem, then $\lambda \notin O$.

Let o' be a minimal occurrence in O such that l/o' and L/o' are maximal subterms with a (non-null) equational subproof $q' : \sigma(l/o') =_R \sigma'(L/o')$ induced by q, $\sigma(l/o') \neq \sigma'(L/o')$, and there is no harmless proof p corresponding to q' with $N(p) \leq N(q')$. There must exist such an occurrence for, if not, then there is a harmless proof corresponding to q and no bigger than q - a contradiction. It is obtained by using the harmless proofs corresponding to the q''s. Now, by our choice of o', q' must contain a root reduction. Since q does not contain a root reduction and q' does, q' is obtained by applying projection with an occurrence not equal to λ and so $l/o', L/o' \in SL(R) - lhs(R)$ and also $N(q') \leq N(q)$. The rest of the argument for this subcase is similar to that of Case A starting at "There are two cases," the contradiction is that we get a proof q'' such that $N(q'') < N(q') \leq N(q)$ that also violates the theorem.

The other subcase is when $l, L \in SL(R) - lhs(R)$. Let o' be a minimal occurrence in O such that l/o' and L/o' are maximal subterms with a (non-null) equational subproof $q' : \sigma(l/o') =_R \sigma'(L/o')$ induced by q, $\sigma(l/o') \neq \sigma'(L/o')$, and there is no harmless proof p corresponding to q' with $N(p) \leq N(q')$. There must exist such an occurrence for, if not, then there is a harmless proof corresponding to q and no bigger than q - a contradiction. It is obtained by

When $r_1 = r_2$ is a variable

Note that we do not exploit the non-unification of l and L in Case A except to argue for a root reduction in the proof, hence the unification of l/o' and L/o' does not affect the rest of the argument, which only uses the fact that l and L are in $SL(R) - lhs(R)$.

using the harmless proofs corresponding to the q''s. Now, by our choice of o', q' must contain a root reduction, since otherwise the chosen q' satisfies $q' < q$ and also violates the theorem, contradicting the choice of q. The rest of the proof is similar to that of Case A and hence omitted. □

B. Unique Normal Form Property

We assume that for every rewrite systems R, we may extend the signature of R by two new constants and a new unary function symbol.

Theorem 19: Every system R with no overlaps and which contains only I-nonoverlaps has the unique normal form property UN.

Proof: (i) Suppose that R does not have the unique normal form property. Then, there are R-normal forms A and B such that $A =_R B$ and $A \neq B$. Let σ be any substitution such that $\sigma(A) \neq \sigma(B)$ and both $\sigma(A)$ and $\sigma(B)$ are ground R normal forms. Then, $R' = R \cup \{h(\sigma(A)) \to a, h(\sigma(B)) \to b\}$ also has no overlaps and furthermore has only I-nonoverlaps, but violates Theorem 18 since $h(\sigma(A)) =_{R'} h(\sigma(B))$ and $h(\sigma(A))$, $h(\sigma(B))$ do not unify since they are distinct ground terms, which is a contradiction. □

Remark: The reason for insisting on distinct *ground* normal forms A and B in the above proof and using a ground substitution in the above proof is to avoid the possibility of adding overlapping rules to R. For example, if $f(x) =_R f(y)$ and we add the rules $h(f(x)) \to a$, $h(f(y)) \to b$, then we create an overlap. A simple example of a rewrite system R with this property is $\{g(x,y) \to f(x), g(x,y) \to f(y)\}$.

C. The relationship of persistence and UN

We now give an example to show that persistence does not imply $UN^{\to}$, hence it also does not imply UN.

Theorem 20: Persistence does not imply uniquely normalizing ($UN^{\to}$).

Proof (by counterexample): Consider the following nonoverlapping system which contains only one O-nonoverlap:

$$\begin{align}
c &\to g(c) \tag{1}\\
g(x) &\to f(x, g(x)) \tag{2}\\
f(x,x) &\to e \tag{3}\\
F(e,e) &\to G(H(e, g(e))) \tag{4}\\
H(e,e) &\to F(e, g(e)) \tag{5}\\
I(x,x) &\to a \tag{6}\\
I(x, G(x)) &\to b \tag{7}
\end{align}$$

Consider the term $A = I(H(c,c), F(c,c))$. Now $A \xrightarrow{nr+} I(F(e,g(e)), F(e,g(e)) \xrightarrow{r} a$ and $A \xrightarrow{nr+} I(H(e,g(e)), G(H(e,g(e)))) \xrightarrow{r} b$, hence the system is not uniquely normalizing. However, it is not hard to show that the system is persistent, since e and $g(e)$ are not joinable. □

The above example can be extended to show that the following condition is not sufficient for persistence.

Theorem 21: The condition: R is nonoverlapping and for every O-nonoverlap in R at least one constraint of the form $(x, C[...x...])$ introduced by $Unif$ is not left-reducible (see [12] for definition), does not imply that R is persistent.

Proof (by counterexample): Consider the previous system extended by two rules. This system contains only one O-nonoverlap and the constraint $(x, G(x))$ is clearly not left-reducible.

$$\begin{align}
c &\to g(c) \tag{8}\\
g(x) &\to f(x, g(x)) \tag{9}\\
f(x,x) &\to e \tag{10}\\
F(e,e) &\to G(H(e, g(e))) \tag{11}\\
H(e,e) &\to F(e, g(e)) \tag{12}\\
I(x,x) &\to a \tag{13}\\
I(x, G(x)) &\to b \tag{14}
\end{align}$$

$$J(x,x) \quad \rightarrow \quad d_1 \tag{15}$$

$$J(a,b) \quad \rightarrow \quad d_2 \tag{16}$$

Consider the term $B = J(A,A)$, where A is the term of Theorem 20. Now B is an instance of the lhs of Rule 15 and $B \xrightarrow{nr+} J(a,b)$ (since $A \xrightarrow{*} a$ and $A \xrightarrow{*} b$), which is an instance of the lhs of Rule 16. Hence the system is not persistent. □

V. The Church-Rosser property of the Union

For brevity's sake the following hypothesis are assumed in all the lemmas of this section: Let R_1 be any left-linear system. Let R_2 be any system such that every function symbol appearing in the lhs of any rule in R_2 does not appear on the rhs of any rule in R_1 (we say that R_2 and R_1 are lr-disjoint; similarly one can define ll-disjoint, etc). Further, assume that $R = R_1 \cup R_2$ is persistent and satisfies the following finiteness condition called semi-termination. (F) There is no sequence of R-reductions from any term t containing an infinite number of R_2 reduction steps.

Note that (F) immediately implies that R_2 is terminating, but termination of R_2 is not sufficient for semi-termination of R as is easily seen by the following example.

Example. Let $R_1 = \{a \rightarrow b\}$, $R_2 = \{h(x,x) \rightarrow h(a,b)\}$. Now R_1 and R_2 are both terminating (in fact, simply terminating, i.e., their termination can be established by simplification orderings; see [22] for a definition), but R has the following cyclic derivation: $h(a,b) \rightarrow_R h(b,b) \rightarrow h(a,b)$. The example can be easily modified so that R is not even quasiterminating. Note that all conditions except (F) are satisfied.

We now prove that R is confluent and then give sufficient conditions that ensure persistence and finiteness of R. Observe that Toyama's [15] technique cannot be used since it depends on the non-increasing property of ranks w.r.t. reductions, which does not hold for us. We need the following lemmas for the proof of confluence.

Lemma 22: For all A, B, C such that $A \overset{*}{\rightarrow}_{R_1} B$ and $A \overset{*}{\rightarrow}_R C$, there exists D such that $C \overset{*}{\rightarrow}_{R_1} D$ and $B \overset{*}{\rightarrow}_R D$.

Proof: An easy argument using the confluence of R_1 shows that it is sufficient to prove the following statement: for all A, B, C such that $A \overset{*}{\rightarrow}_{R_1} B$ and $A \rightarrow_{R_2} C$, there exists D such that $C \overset{*}{\rightarrow}_{R_1} D$ and $B \overset{*}{\rightarrow}_R D$. Let $A = \mathcal{C}[u]$, where u is the redex contracted in the reduction $A \rightarrow C$, and let $l \rightarrow r \in R_2$ be the rule used. So, there exists a substitution σ such that $u = \sigma(l)$ and $C = \mathcal{C}[\sigma(r)]$.

Let $u = C'[A_1, A_2, \ldots, A_m]$, where C' is the template of l, $l = C'[x_1, x_2, \ldots, x_m]$ for some variables x_j (not necessarily distinct) and $\alpha = \langle x_1, \ldots, x_m \rangle \propto \beta = \langle A_1, \ldots, A_m \rangle$ (i.e., $x_i = x_j$ implies $A_i = A_j$, $i, j \in [m]$; see [15] for formal definition of $\propto$). Let $r = C_r[x_{i_1}, \ldots, x_{i_k}]$, where $i_j \in [m]$. Then, $\sigma(r) = C_r[A_{i_1}, \ldots, A_{i_k}]$. Now, we mark the redex u in A and all its descendants (see [3] for a formal definition) in the reduction sequence q: $A \overset{*}{\rightarrow}_{R_1} B$. Therefore, B contains $n \geq 0$ marked terms of the form $C'[B_1^i, B_2^i, \ldots B_m^i]$ for $i \in [n]$. Note that the template of l is unchanged in all the marked terms in B because of the persistence of R. Also, observe that all the marked terms in B are disjoint (because they are all descendants of the same redex) so we can write $B = C''[C'[B_1^1, B_2^1, \ldots B_m^1], \ldots, C'[B_1^n, B_2^n, \ldots B_m^n]]$ for some context C''. The reduction sequence q can be divided into an inner part and an outer part with respect to the marked subterms. Call a step in q inner if it takes place inside one of the marked terms of the form $C'[\ldots]$ and outer otherwise. Let Q_o be the reduction sequence obtained from q by replacing every marked term in it by some new variable, say v. Let $\mathcal{Q}_o$ be the reduction sequence $C \overset{*}{\rightarrow}_{R_1} E$ obtained from Q_o, where v is replaced by $\sigma(r)$. Then, $E = C''[\sigma(r), \ldots, \sigma(r)]$.

Now note that the inner part of q consists of reductions that change the A_i's in A. Let γ_i denote $\langle B_1^i, \ldots, B_m^i \rangle$ for $i \in [n]$. Then, more precisely in the inner part of q we have $\beta \overset{*}{\rightarrow}_{R_1} \gamma_i$ for each $i \in [n]$. Because R_1 is confluent, we have $CR(A_i)$ for all $i \in [n]$. Therefore, we can find δ_i such that $\beta \overset{*}{\rightarrow}_{R_1} \gamma_i \overset{*}{\rightarrow}_{R_1} \delta_i$ and $\beta \propto \delta_i$. Therefore, let $F = C''[C'[\delta_1], \ldots, C'[\delta_n]]$. Then, $B \overset{*}{\rightarrow}_{R_1} F$. Now, since $\beta \propto \delta_i$ and $\alpha \propto \beta$, we have $\alpha \propto \delta_i$ (by transitivity), so each $C'[\delta_i]$ is an instance of l. So let $\delta_i = \sigma_i(l)$ and let $D = C''[\sigma_1(r), \ldots, \sigma_n(r)]$. Obviously, $F \overset{*}{\rightarrow}_{R_2} D$ and $E \overset{*}{\rightarrow}_{R_1} D$ and the lemma is proved. □

We classify the set of function symbols that appear in R into *linear* and *nonlinear* as follows: if a function symbol f appears in the lhs of *any* rule in R_2, then f is nonlinear and linear otherwise.

Note the use of the fact that a confluent rewrite system remains confluent even if the signature is expanded. The A_i's may contain function symbols not appearing in the rules of R_1, but this cannot affect the confluence of R_1.

Definition 23: The *nonlinear height* of a term t (notation $|t|_n$) is the maximum number of nonlinear symbols on any path from the root of t to a leaf.

Lemma 24: If $A \rightarrow_{R_2} B$, $A \xrightarrow{*}_{R_1} C'$ and $C' \rightarrow_{R_2} C$, then there is a D such that $B \xrightarrow{*}_R D$ and $C \xrightarrow{*}_R D$.

Proof (sketch): The full proof of this lemma is somewhat long, but not difficult. The reader can easily fill in the details. By Lemma 22 we have D' and E such that $C' \xrightarrow{*}_{R_1} D'$, $D' \xrightarrow{*}_{R_2} E$ and $B \xrightarrow{*}_{R_1} E$. Also, all the reduction steps from C' to D' are covered by the steps from D' to E. We consider three cases: (i) the reduction step Rs from C' to C is independent (this makes sense since the steps from D' to E are performed at occurrences in $O(C')$) of all the steps from D' to E, (ii) Rs is covered by a single step (single because the steps from D' to E are disjoint) from D' to E, and (iii) Rs covers some (possibly all) of the reductions steps from D' to E. Case (i) is easy. For cases (ii) and (iii) we use persistence of R and the confluence of R_1 to find the desired term D. For case (iii), we also use the fact that the same rule is applied at disjoint occurrences in the reduction sequence from D' to E. □

Remark. The above lemma uses the lr-disjointness condition in an essential way to control the interference of R_2 steps after an R_1 reduction step has been applied. We now define the nonlinear derivation height (notation DH_n) of a term.

Definition 25: $DH_n(t) = max\{n \mid \exists u, t \xrightarrow{*}_R u$ with n R_2-reductions $\}$.

Lemma 26: (1) $DH_n(t)$ is finite for every t. (2) If $t \rightarrow_R u$, then $DH_n(t) \geq DH_n(u)$. If $t \rightarrow_{R_2} u$, then $DH_n(t) > DH_n(u)$.

Proof: Use the definition of DH_n, transitivity of $\xrightarrow{*}$, and the finiteness (F) conditions. □

Theorem 27: If R_1 is left-linear, $R = R_1 \cup R_2$ is lr-disjoint, persistent and semi-terminating, then R is Church-Rosser.

Proof: We prove that CR(A) by induction on $DH_n(A)$. The base case is $DH_n(A) = 0$. In this case, the only derivations possible from A consist solely of R_1-reductions. Since R is persistent so is R_1 and since every left-linear persistent system is confluent [3], [4], [2], the claim holds for the base case. Assume $CR(A)$ for $DH_n(A) < m$ ($m > 0$). We show the following claim:

Claim. $A \rightarrow_{R_2} B$, $A \xrightarrow{*}_{R_1} C'$, and $C' \xrightarrow{*}_{R_2} C$ implies there is a D such that $B \xrightarrow{*}_R D$ and $C \xrightarrow{*}_R D$ for the case $DH_n(A) \leq m$.

Proof of claim: If there are zero reductions from C' to C, then we use Lemma 22 to get the desired term D. Otherwise, let $C' \rightarrow_{R_2} C''$ be the first R_2-reduction step. By Lemma 24 we have a D' such that $B \xrightarrow{*} D'$ and $C'' \xrightarrow{*} D'$. Now, $DH_n(C'') < DH_n(C') \leq DH_n(A)$ by Lemma 26, therefore by the induction hypothesis for the theorem, i.e., $CR(C'')$, we have a D such that $D' \xrightarrow{*} D$ and $C \xrightarrow{*} D$. □

Induction Step: We now prove $CR(A)$ when $DH_n(A) = m$. So suppose that $A \xrightarrow{*}_R B$ and $A \xrightarrow{*}_R C$. If all the reductions in either $A \xrightarrow{*}_R B$ or $A \xrightarrow{*}_R C$ are R_1-reductions, then we are done by Lemma 22. Otherwise we have the following situation: $A \xrightarrow{*}_{R_1} C' \rightarrow_{R_2} C'' \xrightarrow{*} C$, and $A \xrightarrow{*}_{R_1} B'$, $B' \rightarrow_{R_2} B''$ and $B'' \xrightarrow{*} B$, where $C' \rightarrow C''$ and $B' \rightarrow B''$ are the first reductions from R_2 on the respective derivations. Now, we have the following derivations. By confluence of R_1 (see base case) we have a D' such that $B' \xrightarrow{*}_{R_1} D'$ and $C' \xrightarrow{*}_{R_1} D'$. By Lemma 22 we have a D'' such that $D' \xrightarrow{*}_{R_1} \cdot \xrightarrow{*}_{R_2} D''$ and $C'' \xrightarrow{*}_{R_1} D''$. By the above claim, we have an E such that $B'' \xrightarrow{*}_R E$ and $D'' \xrightarrow{*}_R E$. By Lemma 26, $DH_n(B'') < DH_n(B') \leq DH_n(A)$ and $DH_n(C'') < DH_n(C') \leq DH_n(A)$. Therefore, by the induction hypothesis of the theorem, i.e., $CR(B'')$ and $CR(C'')$, we have a D such that $B \xrightarrow{*}_R D$ and $C \xrightarrow{*}_R D$ and the proof is complete. □

Remarks. The above result can be generalized in several different ways, we omit proofs of the generalizations for lack of space. First, we can drop the finiteness requirement and prove CR(A) for only those terms A for which $DH_n(A)$ is finite. Second, we do not really need full persistence of R, a slightly weaker form is sufficient. This is important because it permits some kinds of harmless root and nonroot overlaps in R. Finally, note that this proof shows some similarity to Klop's proof. However, as noted earlier Klop's proof cannot be used since it uses postponement of nonlinear reductions, which does not hold for us and also persistence is immediate there.

We now give sufficient conditions that ensure persistence and semitermination of the union. First, we note that nonoverlapping and semitermination imply persistence.

Lemma 28: If the lr-disjoint union R of a left-linear system R_1 and any system R_2 containing only I-nonoverlaps is nonoverlapping and semi-terminating, then R is persistent.

Proof: A left-linear rule can have only I-nonoverlap with another (not necessarily left-linear) rule. Therefore, there are no O-nonoverlaps in R. Hence Theorem 18 applies. □

We can drop the requirement that R_2 should have only I-nonoverlaps provided that there are no collapsing rules in R_1 by generalizing the proof of Lemma 14. It seems that even the collapsing requirement can be dropped, but this has been difficult to prove.

Theorem 29: The following conditions are sufficient for the semi-termination of a nonoverlapping system R. R_1 is linear (i.e., left-linear & right-linear), and (1) No function symbol that appears in the lhs of any rule in R_2 appears in the rhs of any rule, or (2) All R_2 rules are height decreasing, i.e. $ht(l) > ht(r)$ for all $l \rightarrow r \in R_2$, or (3) All R_2 rules are nonlinear-height decreasing, i.e, $|l|_n > |r|_n$ for every rule $l \rightarrow r \in R_2$.
Proof: Straightforward. □

We give an example to show that the finiteness requirement cannot be dropped completely. Let $R_1 = \{a \rightarrow b\}$ and $R_2 = \{f(x, x) \rightarrow e, g(x) \rightarrow f(x, g(x)), h(x, x) \rightarrow g(h(a, b))\}$. Obviously the lr-disjoint conditions and persistence are satisfied. But, the union is not confluent since $h(b, b) \overset{*}{\rightarrow} e$ and $h(b, b) \overset{*}{\rightarrow} g(e)$ but e and $g(e)$ do not have a common reduct. An example due to Klop also shows that the lr-disjoint condition cannot be completely dropped. Let $R_2 = \{f(x, x) \rightarrow a, f(x, g(x)) \rightarrow b\}$ and $R_1 = \{c \rightarrow g(c)\}$. Since the R_2 rules need an f and strictly decrease the number of f's and the R_1 rules cannot increase this number the finiteness condition is obvious. The union is of course not confluent as is easily seen.

VI. Conclusion

In this paper we have studied two fundamental concepts, uniqueness of normal forms and confluence, for nonlinear systems in the absence of termination. This is a difficult topic with only a few results so far. We classified nonoverlaps into two classes: nonoverlaps due to inhomogeneity (I-nonoverlaps) and nonoverlaps due to occurs-check (O-nonoverlaps). Through a novel approach, we then proved that every nonoverlapping system in which all nonoverlaps are I-nonoverlaps has the unique normal form property. This result is tight and a substantial generalization of previous work. We also proved the confluence of the union of a nonlinear system with a left-linear system under fairly general conditions. Persistence plays a key role in this proof. There are several promising directions for future work. First, we note that the finiteness requirement can be weakened somewhat although it cannot be dropped completely. The proof of this is likely to be difficult but fruitful since it might lead to new techniques for dealing with unions (or decompositions) rather than disjoint sums. Second, our work here suggests some natural generalizations to deal with non-persistent systems. Any progress along these two lines will obviously be of considerable importance to rewriting and its applications.

References

[1] A. Church and J. Rosser, "Some properties of conversion," *Transactions of the AMS*, vol. 39, pp. 472–482, 1936.

[2] G. Huet, "Confluent reductions: Abstract properties and applications to term rewriting systems," *Journal of the ACM*, vol. 27, no. 4, pp. 797–821, 1980, also in 18th IEEE FOCS, 1977.

[3] M. O'Donnell, *Computing in Systems Described by Equations*, ser. Lecture Notes in Computer Science. Springer-Verlag, 1977, vol. 58.

[4] B. Rosen, "Tree manipulating systems and church-rosser theorems," *Journal of the ACM*, vol. 20, pp. 160–187, 1973, also in the 2nd ACM Symposium on Theory of Computing.

[5] D. Knuth and P. Bendix, "Simple word problems in universal algebra," in *Computational Problems in Abstract Algebra*, J. Leech, Ed., Oxford. Pergammon Press, 1970, pp. 263–297.

[6] M. Newman, "On theories with a combinatorial definition of equivalence," *Ann. Math.*, vol. 43, pp. 223–243, 1942.

[7] M. K. Rao, "Completeness of hierarchical combinations of term rewriting systems," in *Proc. Conf. on Foundations of Software Technology & Theoretical Comp. Sci.*, 1993.

[8] A. Middeldorp and Y. Toyama, "Completeness of combinations of constructor systems," in *Proc. Conf. on Rewriting Techniques & Applications*, 1991, pp. 188–199.

[9] F. Baader and T. Nipkow, *Term Rewriting and All That.* Cambridge University Press, 1998.

[10] E. Ohlebusch, *Advanced Topics in Term Rewriting.* Springer, 2002.

[11] P. Chew, "Unique normal forms in term rewriting systems with repeated variables," in *Proc. ACM Symp. on Theory of Computing*, vol. 13, 1981, pp. 7–18.

[12] R. M. Verma, "Unique normal forms and confluence for rewrite systems," *Proc. Int'l Joint Conf. on Artificial Intelligence*, pp. 362–368, 1995.

[13] J. Klop, "Combinatory reduction systems," Ph.D. dissertation, Mathematisch Centrum, Amsterdam, 1980.

[14] M. Oyamaguchi and Y. Ohta, "Church rosser property of right-ground systems," *Trans. of the IEICE*, vol. J76-D-I, 1992, in Japanese.

[15] Y. Toyama, "On the church-rosser property for the direct sum of term rewriting systems," *Journal of the ACM*, vol. 34, no. 1, pp. 128–143, 1987.

[16] K. Mano and M. Ogawa, "Unique normal form property of compatible term rewriting systems: a new proof of Chew's theorem," *Theoretical Computer Science*, vol. 258, no. 1–2, pp. 169–208, 2001.

[17] J. Klop and R. de Vrijer, "Unique normal forms for lambda calculus with surjective pairing," *Information and Control*, vol. 80, pp. 97–113, 1989.

[18] N. Dershowitz and D. Plaisted, "Rewriting," in *Handbook of Automated Reasoning*, J. A. Robinson and A. Voronkov, Eds. Elsevier Science, 2001, vol. 1, ch. 9, pp. 535–610.

[19] J. Klop, "Rewrite systems," in *Handbook of Logic in Computer Science*. Oxford, 1992.

[20] C. Dwork, P. Kanellakis, and J. Mitchell, "On the sequential nature of unification," *Journal of Logic Programming*, vol. 1, pp. 35–50, 1984.

[21] R. M. Verma, "A theory of using history for equational systems with applications," *Journal of the ACM*, vol. 42, no. 5, pp. 984–1020, 1995, also in the 32nd IEEE FOCS Symposium, 1991.

[22] N. Dershowitz, "Termination of rewriting," *Journal of Symbolic Computation*, vol. 3, pp. 69–116, 1987.