THE LOGIC IN COMPUTER SCIENCE COLUMN
by
Yuri GUREVICH

Newman's Lemma – a Case Study in Proof Automation and Geometric Logic *

Marc Bezem† Thierry Coquand‡

Abstract

Newman's Lemma states that a graph is confluent if it is locally confluent and has no infinite paths. Newman's Lemma requires essentially higher-order logic. We show that the induction step in Huet's inductive proof of Newman's Lemma is completely first-order and can be automated easily with the help of a resolution theorem prover. The resolution proof uses classical logic. For the automation of a constructive proof we consider Newman's Lemma in geometric logic and sketch an algorithm leading to a constructive proof. We formalize the proof procedure and prove its completeness for geometric logic by a direct topological argument.

*Bull. Eur. Assoc. Theor. Comput. Sci., Feb. 2003

†Department of Computer Science, Bergen University, Postboks 7800, N-5020 Bergen, Norway, `bezem@ii.uib.no`

‡Department of Computer Science, Chalmers University of Technology and Gothenburg University, SE-412 96 Göteborg, Sweden, `coquand@cs.chalmers.se`

1 Introduction

A directed graph G is called *confluent* if any two diverging paths having the same start node can be extended to paths leading to the same end node. Directed graphs are ubiquitous, but often go by different names:

- abstract rewriting systems (underlying functional languages, etc.)
- labelled transition systems (underpinning operational semantics)
- Kripke frames (the basis of the semantics of modal logic)

In the respective cases, confluence is a key notion for:

- uniqueness of normal forms (independence of the evaluation strategy)
- certain forms of determinism (in operational semantics)
- quantifier shifts (in dynamic logic)

The graph G is called *locally confluent* if any two different *steps* (paths of length one) departing from the same node can be extended into paths leading to the same end node. As we will see in the next section, local confluence is strictly weaker than confluence. Newman's Lemma [10] (NL hereafter) states that local confluence implies confluence provided there are no infinite paths. We have automated the combinatorial core of Huet's [5] proof of NL using the resolution theorem prover Otter [11]. Otter is not the fastest theorem prover anymore, but is by far the best documented, see [13, 14]. We discuss the various sensitivities regarding the precise formulation in relation to the theorem proving procedure in Section 3. Geometric logic is introduced in Section 4, its proof theory in Sections 5 and 6, and the semantics and the completeness results in Section 7.

2 Local confluence is weaker than confluence

Consider the following directed graph attributed by Hindley to Kleene.

$$\begin{tikzcd} a & b \arrow[l] \arrow[r, bend left] & c \arrow[l, bend left] \arrow[r] & d \end{tikzcd}$$

Obviously, we have local confluence here since there are no diverging steps from a nor d, and diverging steps from b and c can be extended to paths ending in a and d, respectively, by using the cycle. However, the graph is not confluent, since the diverging paths $b \to c \to d$ and $b \to a$ cannot be extended to join. Attempts to deny this fact by applying a tiling analysis lead to the diagram below, where the cycle has been unwound into an infinite path. Here the dotted lines from a to a and from d to d represent paths of length zero for the sole purpose of keeping things orthogonal.

$$\begin{tikzcd}
b \arrow[r] \arrow[dddd] & c \arrow[rrr] \arrow[d] & & & d \arrow[d, dotted, no head] \\
& b \arrow[r] \arrow[ddd] & c \arrow[rr] \arrow[d] & & d \arrow[d, dotted, no head] \\
& & b \arrow[r] \arrow[dd] & c \arrow[r] & d \\
& & & \vdots & \cdots \\
a \arrow[r, dotted, no head] & a \arrow[r, dotted, no head] & a & &
\end{tikzcd}$$

3 Proving Newman's Lemma

NL provides an interesting test case for several reasons. First, it consists of a mix of first-order and higher-order aspects. The higher-order aspects are the transitive closure and the finiteness of the paths. This makes the identification of the first-order combinatorial core of the proof non-trivial. Second, the proof of NL is not completely trivial, as experienced by everybody seeing it for the first time. In particular the automation of the constructive proof seems to provide a challenge to many theorem proving systems. For a precise analysis we have to introduce some notions from rewriting theory.

Definition 3.1 Let $\to$ be a binary relation on a set S and let $\twoheadrightarrow$ be the reflexive–transitive closure of $\to$. In other words, we consider a directed graph $(S, \to)$ with path relation $\twoheadrightarrow$.

1. We say that $x \in S$ is *confluent* if, for all $x_1, x_2 \in S$, $x \twoheadrightarrow x_1$ and $x \twoheadrightarrow x_2$ implies that $x_1 \twoheadrightarrow y$ and $x_2 \twoheadrightarrow y$ for some $y \in S$. Such a y is called a *common reduct* of x_1 and x_2. We say that $\to$ is *confluent* if every $x \in S$ is confluent.

2. We say that $x \in S$ is *locally confluent* if, for all x_1, x_2, $x \to x_1$ and $x \to x_2$ implies that $x_1 \twoheadrightarrow y$ and $x_2 \twoheadrightarrow y$ for some $y \in S$. Note $x \to x_i$ here and $x \twoheadrightarrow x_i$ above. We say that $\to$ is *locally confluent* if every $x \in S$ is locally confluent.

3. We say that $\to$ is *terminating* if there is no infinite sequence $x_0 \to x_1 \to x_2 \to \cdots$ in S.

One of the classical proofs of NL is by contradiction. Let $\to$ be locally confluent and terminating. Assume there is an x which is not confluent, that is, there exist $x_1, x_2 \in S$ such that $x \twoheadrightarrow x_1$ and $x \twoheadrightarrow x_2$ and x_1 and x_2 have no common reduct. Since $\to$ is terminating, we may assume without loss of generality that x is an $\to$-maximal[1] non-confluent element. If not, there would be a non-confluent x' with $x \to x'$, and if that x' is not $\to$-maximal, then there would be a non-confluent x'' with $x' \to x''$ and so on, leading to a sequence contradicting the termination of $\to$. (This part is difficult to explain, it actually uses the Axiom of Dependent Choice, DC.) Hence we have that every $y \in S$ with $x \to y$ is confluent. From the fact that x_1 and x_2 have no common reduct, it follows that we do not have $x = x_1$ or $x = x_2$, so there must exist intermediate points i_1, i_2 such that $x \to i_1 \twoheadrightarrow x_1$ and $x \to i_2 \twoheadrightarrow x_2$. To x and these intermediate points we can apply local confluence to obtain a common reduct of the intermediate points. By the maximality of x we can then complete the diagram below. This is a contradiction and hence NL has been proved.

$$\begin{tikzcd}
x \arrow[r] \arrow[d] \arrow[dr, phantom, "\text{local confluence}"] & i_1 \arrow[r, two heads] \arrow[d, two heads] & x_1 \arrow[d, two heads] \\
i_2 \arrow[r, two heads] \arrow[d, two heads] & \cdot \arrow[r, two heads] & \cdot \arrow[d, two heads] \\
x_2 \arrow[rr, two heads] & & \cdot
\end{tikzcd}$$

The formalization of the above argument requires higher-order logic (to express transitive closure), DC and (most naturally) three-sorted first-order logic: one sort for the set S, one for the natural numbers and one for infinite sequences of elements of S. There exist many alternative proofs of NL, for example, based on the multiset extension of $\to$. All those proofs are a bit complicated by the fact that the absence of infinite paths has to be turned into an inductive argument, which again requires DC.

An important improvement, essentially due to Huet, is obtained by taking a constructive reformulation of NL as point of departure. In this formulation the infinite sequences such as used in the definition of termination and in DC are avoided by using an inductively defined predicate called accessibility (which can be traced back to Gentzen [4] 'erreichbar').

[1]If the transitive closure of $\to$ is viewed as a *greater than* ordering, then it would be natural to speak of $\to$-*minimal* instead.

Definition 3.2 Let $\rightarrow$ be a binary relation on a set S. The unary predicate $Acc_{\rightarrow}$ is inductively defined by the closure property $(*)$ if $Acc_{\rightarrow}(y)$ for all $y \in S$ such that $x \rightarrow y$, then $Acc_{\rightarrow}(x)$. By $Acc_{\rightarrow}(S)$ we express that $Acc_{\rightarrow}(x)$ for all $x \in S$.

In other words, all $\rightarrow$-maximal elements are accessible, as well as all elements whose successors are all $\rightarrow$-maximal, and so on. The inductive definition stipulates $Acc_{\rightarrow}$ to be the *smallest* unary predicate closed under $(*)$. For this to be well-defined, one easily checks that the intersection of an arbitrary set of unary predicates which are closed under $(*)$ is itself closed under $(*)$. Furthermore, any infinite sequence $x_0 \rightarrow x_1 \rightarrow x_2 \rightarrow \cdots$ consists of elements that are not accessible. The reason is that they can be left out without violating the closure property $(*)$. On the other hand, if there exists a non-accessible $x_0 \in S$, then there must be a non-accessible $x_1 \in S$ with $x_0 \rightarrow x_1$, and hence a non-accessible $x_2 \in S$ with $x_1 \rightarrow x_2$, and so on. Thus one proves using DC that $\rightarrow$ is terminating if and only if all elements of S are accessible, that is, if $Acc_{\rightarrow}(S)$.

The point is now that we avoid the above equivalence proof by adopting $Acc_{\rightarrow}(S)$ straightaway as the 'positive' formulation of the absence of infinite paths. The advantages of using $Acc_{\rightarrow}(S)$ instead of the traditional formulation of termination are three-fold.

- DC is not needed anymore in the proof of NL.
- The sorts for the natural numbers and for infinite sequences are not needed.
- We can reason by induction[2] on $Acc_{\rightarrow}(x)$, the induction step being first-order.

These reasons motivate the following reformulation of NL.

Lemma 3.3 *If $Acc_{\rightarrow}(S)$, then confluence of $\rightarrow$ follows from local confluence.*

We could have added a fourth advantage to the three advantages above, namely that the proof of NL in the formulation with the accessibility predicate can be done constructively, as shown in the next paragraph. The constructive proof is not more complicated than the classical one, it is actually shorter, but the relevant point here is that we do not know a theorem prover that is able to generate the constructive proof. Instead we had to use a resolution theorem prover, which first clausifies the negation of the statement and then searches for a resolution refutation of the resulting clauses. The automation of a constructive proof will be explored later, in the sections on geometric logic.

We will sketch the constructive argument. By induction one proves that every accessible $x \in S$ is confluent. By $Acc_{\rightarrow}(S)$ we then obtain confluence. The

[2]Induction is a very practical method. You can, for example, also cook by (magnetic) induction.

induction step we have to prove is that confluence is preserved under the closure property $(*)$ in the inductive definition of $Acc_{\rightarrow}$. In other words, we have to prove that x is confluent if the induction hypothesis (IH) holds, that is, if every $y \in S$ such that $x \rightarrow y$ is confluent. Assume IH and let $x_1, x_2 \in S$ such that $x \twoheadrightarrow x_1$ and $x \twoheadrightarrow x_2$. If $x = x_1$ or $x = x_2$ then x_2 or x_1 is a common reduct of x_1, x_2. Otherwise, actually appealing to the inductive definition of the reflexive–transitive closure, there exist intermediate points as in the classical proof above. Now a common reduct can be obtained in exactly the same way as in the classical proof, with IH replacing the $\rightarrow$-maximality of x. This proves the induction step.

The above proof of the induction step is completely first-order and can be cast in the form of a refutation problem, namely the inconsistency of the conjunction below. Note that $=, \rightarrow, \twoheadrightarrow$ are uninterpreted binary relations here, with (local) confluence shorthand and $\rightarrow\cdot\twoheadrightarrow$ the composition of (first) $\rightarrow$ and (last) $\twoheadrightarrow$.

$$\bigwedge \begin{cases} \text{1. } = \text{ is reflexive and symmetric} \\ \text{2. } \twoheadrightarrow \text{ includes } = \text{ and } \rightarrow \text{ and is transitive} \\ \text{3. } \twoheadrightarrow \text{ is included in the union of } = \text{ and } \rightarrow\cdot\twoheadrightarrow \\ \text{4. } \rightarrow \text{ is locally confluent} \\ \text{5. there exists a non-confluent } x \text{ such that all } y \leftarrow x \text{ are confluent} \end{cases}$$

Some remarks on the conjuncts 1–3 and 5 above are in order here.

1. Transitivity of $=$ is not needed, which was obvious from the human proof.

2. Taking $\twoheadrightarrow$ to include $=$ is a convenient way of stating that $\twoheadrightarrow$ is reflexive. One can then conclude $a \twoheadrightarrow b$ from $a = b$ without any appeal to congruence axioms. Otter didn't use the fact that $\twoheadrightarrow$ includes $\rightarrow$. This came, admittedly, as a surprise, although it also follows from a close analysis of the human proof. It implies that NL can be strengthened to a (diamond) result for any reflexive–transitive relation satisfying conjunct 3. This strengthening is in itself not very interesting, but results in a surprising speed-up of the proof search by a factor 5.

3. It is well-known that transitive closure is not first-order definable. The conjunction of 2 and 3 is weaker than the reflexive–transitive closure. However, one can prove by induction on $Acc_{\rightarrow}(x)$ that conjunct 3 implies that $\twoheadrightarrow$ is a subset of the reflexive–transitive closure of $\rightarrow$. So in the presence of $Acc_{\rightarrow}(S)$ the conjunction of 2 and 3 exactly axiomatizes the reflexive–transitive closure. The observation that transitive closure is first-order definable for terminating relations may be of use in other applications as well.

(5) This is the negation of the induction step.

Otter found a refutation within some 15 minutes (on a Sun Ultra 10), with a combination of positive hyperresolution and unit-resulting resolution. The Otter input and output files match `www.ii.uib.no/~bezem/NL/NL1.*`. The refutation found by Otter is quite close to a human proof, see the diagram chase below. The arrows are labelled by numbers indicating the clause representing that arrow. A negative number means that the clause refutes the arrow. An ill-understood phenomenon is why it took so long after clause 477 to infer 24727 by transitivity from 45 and 70. Otter can produce many other refutations of the same set of clauses, but all seem to be inessential variations of the same diagram chase, due to the many symmetries.

$$\begin{tikzcd}
x \arrow[r, "27"] \arrow[d, "28"'] \arrow[dr, phantom, "\text{local confluence}"] & i_1 \arrow[r, ->>, "31"] \arrow[d, ->>, "45"'] \arrow[dd, ->>, bend left, "24727"] & x_1 \arrow[dd, ->>, "25185"] \\
i_2 \arrow[r, ->>, "48"] \arrow[d, ->>, "32"'] & \cdot \arrow[d, ->>, "70"'] & \\
x_2 \arrow[r, ->>, "477"] \arrow[rr, ->>, bend right, "-25536"'] & \cdot \arrow[r, ->>, "25291/-25719"] & \cdot
\end{tikzcd}$$

An obvious difficulty for the proof search is the symmetry of the formulation of NL. For example, by applying local confluence, any step $x \to y$ gives rise to useless common reducts of y and y. It is well-known that it is possible to distinguish between 'horizontal' and 'vertical' steps in the formulation of both confluence and local confluence. This leads to an asymmetrical version of Newman's Lemma (aNL), which can be proved by the same proof with all the steps properly labelled as either 'horizontal' or 'vertical'. NL can easily be recovered from aNL by removing the distinction. The asymmetrical analogues of confluence and local confluence are known in the literature as commutativity and weak commutativity, respectively.

Definition 3.4 Let $\to_h$ and $\to_v$ be binary relations on a set S, with reflexive–transitive closures $\twoheadrightarrow_h$ and $\twoheadrightarrow_v$, respectively.

1. We say that x is *commutative* if, for all $x_1, x_2 \in S$, $x \twoheadrightarrow_h x_1$ and $x \twoheadrightarrow_v x_2$ implies that $x_1 \twoheadrightarrow_v y$ and $x_2 \twoheadrightarrow_h y$ for some $y \in S$. We say that $\to_h$ and $\to_v$ *commute* if every $x \in S$ is commutative.

2. We say that x is *weakly commutative* if, for all $x_1, x_2 \in S$, $x \to_h x_1$ and $x \to_v x_2$ implies that $x_1 \twoheadrightarrow_v y$ and $x_2 \twoheadrightarrow_h y$ for some $y \in S$. We say that $\to_h$ and $\to_v$ *commute weakly* if every $x \in S$ is weakly commutative.

The precise statement of aNL is that $\to_h$ and $\to_v$ commute if they commute weakly, provided $Acc_{\to_{hv}}(S)$. Here $\to_{hv}$ is the union of $\to_h$ and $\to_v$. A glance at the diagram below tells us that we need the induction hypothesis both for i_1 with $x \to_h i_1$ and for i_2 with $x \to_v i_2$.

$$
\begin{tikzcd}
x \arrow[r, "h"] \arrow[d, "v"'] \arrow[dr, phantom, "\text{weak commutativity}"] & i_1 \arrow[r, two heads, "h"] \arrow[d, two heads, "v"] & x_1 \arrow[d, two heads, "v"] \\
i_2 \arrow[r, two heads, "h"] \arrow[d, two heads, "v"'] & \cdot \arrow[r, two heads, "h"] & \cdot \arrow[d, two heads, "v"] \\
x_2 \arrow[rr, two heads, "h"] & & \cdot
\end{tikzcd}
$$

The proof of aNL follows the pattern of the proof of NL, but is based on the inconsistency of the following conjunction:

$$
\bigwedge \left\{ \begin{array}{l}
= \text{ is reflexive and symmetric} \\
\twoheadrightarrow_h \text{ and } \twoheadrightarrow_v \text{ include } = \text{ and are transitive} \\
\twoheadrightarrow_h \text{ is included in the union of } = \text{ and } \to_h \cdot \twoheadrightarrow_h \\
\twoheadrightarrow_v \text{ is included in the union of } = \text{ and } \to_v \cdot \twoheadrightarrow_v \\
\to_h \text{ and } \to_v \text{ are weakly commutative} \\
\text{there exists a non-commutative } x \text{ such that all } y \leftarrow_{hv} x \text{ are commutative}
\end{array} \right.
$$

Proof search in the asymmetrical case is two orders of magnitude faster than in the symmetrical case: less than a second. The relevant Otter files match `~bezem/NL/NLa.*`.

4 Geometric Logic

An example of geometric logic is the formalization of projective geometry[3] by Skolem [12] (1920). There are two sorts: lines $l, m, \ldots$ and points $P, B, \ldots$ The atomic relations are (with Skolem's notation): equality of lines (lm), equality of points (PQ) and coincidence (Pl), expressing that a point P is on a line l. The theory is then (all axioms are implicitly universally closed):

$$(PP),\ (PQ) \to (QP),\ (PQ) \wedge (QR) \to (PR) \text{ (equality axioms for points)}$$

[3]However, the qualifier 'geometric' does not come from this example, but refers to the origin of geometric logic in algebraic geometry [2].

(ll), $(lm) \to (ml)$, $(lm) \wedge (mn) \to (ln)$ (equality axioms for lines)

$(PQ) \wedge (Ql) \to (Pl)$, $(Pl) \wedge (lm) \to (Pm)$ (congruence axioms)

$(Pl) \wedge (Ql) \wedge (Pm) \wedge (Qm) \to (PQ) \vee (lm)$ (projective uniqueness axiom)

$(\exists l)((Pl) \wedge (Ql))$, $(\exists P)((Pl) \wedge (Pm))$ (projective axioms of incidence)

Geometric logic abstracts from the above example. A *geometric theory* is a set of geometric formulae. For simplicity we suppose that we have only predicate symbols and no function symbols. Terms are either variables $x, y, z, \ldots$, or parameters $a_0, a_1, \ldots$ Atomic formulae $A, A', \ldots$ are of the form $P\, t_1 \, \ldots \, t_n$, with $t_1, \ldots, t_n$ terms and P a predicate symbol of arity n. A *geometric formula* is a formula of the form $C \to D$ where C and D are given by the following grammar:

$$C ::= 1 \mid C \wedge A \qquad D ::= 0 \mid D \vee E \qquad E ::= (\exists \vec{x})C$$

Here 0 and 1 represent the empty disjunction and conjuntion, respectively. A *closed* formula contains no free variables, only parameters and bound variables. A closed atom is also called a *fact*, denoted by $F, F' \ldots$ To sum up, we use A, C, D, E, F as stems of syntactical variables for atoms, conjunctions, disjunctions, existential formulae and facts, respectively.

We may write D for $1 \to D$, A for $1 \wedge A$ and so on, economizing on empty conjunctions, disjunctions, existential quantifications and brackets as much as possible. As an example, we prefer to write (and read) (ll) instead of $1 \to (0 \vee (\exists)(ll))$.

Note that Skolem's theory above is actually two-sorted. It would be possible to consider as well *multisorted* geometric logic. This is not strictly necessary, since the usual reduction to one sort, using unary predicates *Line* and *Point* and relativizing the quantifiers, results in a one-sorted geometric theory.

In [3], the notion of geometric theory is extended by function symbols and equations on these symbols. A typical axiom is the theory of fields, with function symbols for additions and products, with the usual equational axioms for rings and the geometric axiom

$$x = 0 \vee (\exists y)(x \cdot y = 1)$$

The first-order formulation of Newman's Lemma can naturally be expressed in geometric form. The language has three binary predicates: E (for equality), R and S (approximating the reflexive–transitive closure of R), and one parameter a_0. As in Section 3, we formalise the induction step in the proof that accessibility implies confluence. However, whereas resolution logic (or rather the existing implementations) forces us to negate the conclusion and use a refutation procedure, geometric logic makes it possible to formulate the conclusion positively.

The theory NLg is the following collection of geometric formulae:

(1) $E\,x\,x$

(2) $E\,x\,y \to E\,y\,x$

(3) $E\,x\,y \wedge E\,y\,z \to E\,x\,z$

(4) $E\,x\,y \to S\,x\,y$

(5) $R\,x\,y \to S\,x\,y$

(6) $S\,x\,y \wedge S\,y\,z \to S\,x\,z$

(7) $S\,x\,y \to E\,x\,y \vee (\exists z)(R\,x\,z \wedge S\,z\,y)$ (S included in the union of E and $R{\cdot}S$)

(8) $R\,x\,y \wedge R\,x\,z \to (\exists u)(S\,y\,u \wedge S\,z\,u)$ (local confluence)

(9) $R\,a_0\,x \wedge S\,x\,y \wedge S\,x\,z \to (\exists u)(S\,y\,u \wedge S\,z\,u)$ (induction hypothesis)

The formulae (1)–(8) correspond with the conjuncts 1–4 on page 6 and (9) with the part of conjunct 5 expressing that $R\,a_0\,x$ implies confluence of x. The conclusion that a_0 is confluent can here be stated in a positive way:

(10) $S\,a_0\,x \wedge S\,a_0\,y \to (\exists z)(S\,x\,z \wedge S\,y\,z)$

The reader can now check that the geometric formula (10) is a first-order consequence of the theory NLg. Otter proves this much in the same way as in Section 3, the relevant files match `~bezem/NL/NLg.*`.

5 Proof procedure for geometric theories

The following proof procedure can be seen as a specialization of the proof of the completeness theorem as presented for instance in [7]. We present it informally first, and make it precise in the next section. We divide a geometric theory in two parts: one part collects the axioms $C \to D$ with D essentially one single atom; the remaining part collects the axioms where D is 0, a proper disjunction or an existential formula. Let us call the first part the (definite) *Horn* part and the second the *disjunctive* part[4]. For the theory NLg, the axioms (1)–(6) form the Horn part. Proving conclusion (10) starts with instantiating (10) with two new parameters, say a_1 and a_2, and adding the facts $S\,a_0\,a_1$, $S\,a_0\,a_2$. The proof obligation is then

$$(\exists x)(S\ a_1\ x \wedge S\ a_2\ x)$$

At each step in the proof procedure, only a finite number of parameters are alive (as in [7]). We then collect all facts that we can derive using the Horn part

[4]Such a separation of the theory appears both in [12] and [3]. In [12] this is used to prove that Desargues' Theorem is independent of the theory of projective geometry presented above.

of the theory. In the example of the theory NL, only a_0, a_1, a_2 are alive at first and from the facts $S\,a_0\,a_1$, $S\,a_0\,a_2$ we can only derive

$$S\,a_0\,a_0,\ E\,a_0\,a_0,\ S\,a_1\,a_1,\ E\,a_1\,a_1,\ S\,a_2\,a_2,\ E\,a_2\,a_2$$

Next, we look for the possible disjunctions and existential statements that we can derive from these facts. In the example of the theory NL, we can only derive from (7)

$$E\,a_0\,a_1 \vee (\exists z)(R\,a_0\,z \wedge S\,z\,a_1) \qquad E\,a_0\,a_2 \vee (\exists z)(R\,a_0\,z \wedge S\,z\,a_2)$$

We branch then on all such disjunctions and statements[5]. For instance, in the case above we get four branches. Abbreviate $E\,a_0\,a_1$ by F_1 and $E\,a_0\,a_2$ by F_2. Instantiating the two existential statements with new parameters, say a_3 and a_4, respectively, results in four facts

$$R\,a_0\,a_3,\ S\,a_3\,a_1,\ R\,a_0\,a_4,\ S\,a_4\,a_2$$

which we abbreviate by F_3, F_3', F_4, F_4', respectively. In one branch we add the facts F_1, F_2, in the second F_1, F_4, F_4', in the third F_3, F_3', F_2 and in the fourth branch we add F_3, F_3', F_4, F_4'. In all branches we then close under the Horn part of NLg and continue.

In this way, we build a finitely branching search tree, to each node of which is associated a finite set of facts which is closed under the Horn part of the theory. For the present search tree, the formula

$$(\exists x)(S\ a_1\ x \wedge S\ a_2\ x)$$

will eventually be true on every branch in the finite model defined by the associated collection of facts. This completes the proof that (10) is a a geometric consequence of the theory NLg.

A similar proof procedure is used in [3] to analyse theories such as algebraically closed fields or real closed fields.

6 Proof system for geometric theories

The relevance of geometric logic for computation has been stressed already (see for instance [2]). The previous section shows that it might have some interest for automatic proof search as well. The goal of this section is to give a precise definition of the proof procedure, and to state some remarkable theorems, that show in particular that, for this fragment, intuitionistic and classical derivability coincide. We present a possible short proof of these theorems in the next section.

[5]An optimisation is to first delete all such statements that hold already in the finite model defined by the facts we have so far.

Definition 6.1 Let T be a geometric theory, X a finite collection of facts and D a closed geometric disjunction. As usual, we write $X, F_1, \ldots, F_m$ for $X \cup \{F_1, \ldots, F_m\}$. We define inductively $X \vdash_T D$, which expresses that D is a geometric consequence in T of the facts in X. Here and below we simply write $\vdash$ instead of $\vdash_T$ since T is clear from the context.

- (base case) $X \vdash D$ if there is a disjunct $(\exists \vec{x})C$ of D and a closed instance C_0 of C all whose conjuncts are in X, and
- (induction step) $X \vdash D$ if there is a close instance $C_0 \to D_0$ of an axiom of T such that all conjuncts of C_0 are in X and for every disjunct

  $$(\exists \vec{x})(A_1 \wedge \ldots \wedge A_m)$$

  of D_0 we have
  $$X, A_1[\vec{x} := \vec{a}], \ldots, A_m[\vec{x} := \vec{a}] \vdash D$$
  where $\vec{a}$ consists of new parameters not occurring in $X \vdash D$, nor in D_0.

As a simple example, since the understanding of the induction step above may be hampered by the many abbreviations in geometric formulae, we infer $A \vdash D$ for arbitrary D from the geometric theory $A \to B, B \to 0$. Using the induction step with axiom $A \to B$, $A \vdash D$ follows from $A, B \vdash D$ since A is the only conjunct of $1 \wedge A$ and B is the only disjunct of $0 \vee (\exists)B$ and there is nothing to instantiate. Again using the induction step, but now with axiom $B \to 0$, since B occurs in A, B and 0 is the empty disjunction, $A, B \vdash D$ follows without any further proof obligation.

A more interesting example can be obtained by formalising the proof procedure described in the previous section along the lines of the above inductive definition.

Theorem 6.2 *If D is a consequence of X using higher-order intuitionistic logic*[6]*, then $X \vdash D$. Here and below $X \vdash D$ is as in Definition 6.1.*

The next result is also known as Barr's Theorem [1, 2, 9].

Theorem 6.3 *If D is a consequence of X using higher-order classical logic, then $X \vdash D$.*

One reading of this theorem is that higher-order logic and geometric logic give the same consequence relation on geometric formulae.

Classically, given Gödel's completness theorem for first-order logic, one possible formulation of this theorem is

[6] By higher-order we mean that we allow function and relation variables, and quantification over those variables.

Theorem 6.4 *If D holds in all models of X, then $X \vdash D$.*

A proof along the lines of [7] is direct. However, this would not be constructively valid, in contrast to the formulation of Theorem 6.3 and to the proof we present in the next section. The reference [9] presents a formulation of classical first-order sequent calculus with admissible cut-rules, in which derivations of geometric formulae are automatically geometric.

7 Topological model

In order to prove Theorem 6.2, we build a special topological model of the theory T. The *domain* M of this model is the set $\{a_0, a_1, \ldots\}$ of all parameters. The truth values will form a *complete Heyting algebra* (or *frame*) H [6]. The model is defined by choosing an interpretation $[\![F]\!] \in H$ of each fact. Each closed conjunction $C = F_1 \wedge \cdots \wedge F_m$ is then interpreted as usual by a finite meet

$$[\![C]\!] = [\![F_1]\!] \wedge \cdots \wedge [\![F_m]\!]$$

each closed existential formula $E = (\exists \vec{x})C$ by an infinite join

$$[\![E]\!] = \bigvee_{\vec{a} \subseteq M} [\![C[\vec{x} := \vec{a}]]\!]$$

and finally each closed disjunction $D = E_1 \vee \cdots \vee E_n$ by a finite join

$$[\![D]\!] = [\![E_1]\!] \vee \cdots \vee [\![E_n]\!]$$

The Heyting algebra H and the interpretation $[\![\_]\!]$ will be built in such a way that we have $F_1, \ldots, F_m \vdash D$ iff

$$[\![F_1]\!] \wedge \cdots \wedge [\![F_m]\!] \leq [\![D]\!]$$

in H. Given that higher-order intuitionistic logic can be interpreted in a topological model, see [6], Theorem 6.2 will follow at once.

Following [6], we can define H by generators and relations. We take as generators the facts themselves, so that each closed conjunction C can be seen as an element $C \in H$, and for the relations we take

$$[\![C_0]\!] \leq [\![D_0]\!]$$

for each closed instance $C_0 \to D_0$ of an axiom $C \to D$ in the theory T.

The completeness of this model w.r.t. geometric rules of inference follows from the remark (cf. [8], page 86, for a similar remark used to give a constructive proof of Gödel's completeness theorem):

Lemma 7.1 *If we have*

$$X, F_1(a_{k+1}, \ldots, a_{k+n}), \ldots, F_m(a_{k+1}, \ldots, a_{k+n}) \vdash D$$

where $a_k, \ldots, a_{k+n}$ do not appear in X and D, then we have for all $u_1, \ldots, u_n \in M$

$$X, F_1(u_1, \ldots, u_n), \ldots, F_m(u_1, \ldots, u_n) \vdash D$$

A slight variation on the well-known negative translation gives a proof of Barr's Theorem 6.3. Given X and D, we consider the element $G = (\wedge_{F \in X} [\![F]\!]) \rightarrow [\![D]\!] \in H$ and we transform the frame H into a complete Boolean algebra H_0, by taking H_0 to be the frame defined by the nucleus [6] $j(p) = (p \rightarrow G) \rightarrow G$ on H. This model is now a model of all facts in X, and the fact that D holds in this model is equivalent to $G = 1 \in H$ which is equivalent, by completeness, to $X \vdash D$ as desired.

A more sophisticated argument, which uses the notion of *site* instead of complete Heyting algebra, can be given. This argument would reflect the fact that in the proof procedure we consider only a finite number of parameters at each step.

Acknowledgement

The authors are indebted to Andreas Blass, Wilfried Buchholz, Dimitri Hendriks, Gérard Huet and Tobias Nipkow for valuable comments. The second author would like to thank Jan von Plato for his explanation of Skolem's presentation of projective geometry.

References

[1] M. Barr, Toposes without points, *Journal of Pure and Applied Algebra* **5**:265–280, 1974.

[2] A. Blass, Topoi and computation, *Bulletin of the EATCS* **36**:57–65, October 1998.

[3] M. Coste, H. Lombardi, and M.F. Roy, Dynamical methods in algebra: effective Nullstellensätze, *Annals of Pure and Applied Logic* **111**(3):203–256, 2001.

[4] G. Gentzen, Die Widerspruchsfreiheit der reinen Zahlentheorie, *Mathematische Annalen* **112**:493–565, 1936.

[5] G. Huet, Confluent reductions: Abstract properties and applications to term rewriting systems, *Journal of the ACM* **27**(4):797–821, 1980.

[6] P. Johnstone, *Stone Spaces*, Cambridge Studies in Advanced Mathematics **3**, Cambridge University Press, 1982.

[7] S.C. Kleene, *Mathematical logic,* John Wiley & Sons, Inc., New York-London-Sydney, 1967.

[8] P. Martin-Löf, *Notes on Constructive Mathematics.* Almqvist & Wiksell, Stockholm, 1970.

[9] S. Negri, Contraction-free sequent calculi for geometric theories, with an application to Barr's theorem, to appear in *Archive for Mathematical Logic*.

[10] M.H.A. Newman, On theories with a combinatorial definition of 'equivalence', *Annals of Mathematics* **43**(2):223–243, 1942.

[11] Otter, `www-unix.mcs.anl.gov/AR/otter/`

[12] T. Skolem, *Selected Works in Logic,* edited by J.E. Fenstad, Universitetsforlaget, Oslo, 1970.

[13] L. Wos, *The Automation of Reasoning: an Experimenter's Notebook with Otter Tutorial*, Academic Press, 1996.

[14] L. Wos and G. Pieper, *A Fascinating Country in the World of Computing: Your Guide to Automated Reasoning*, World Scientific, 2000.