A Verified Prover Based on Ordered Resolution
Anders Schlichtkrull, Jasmin Christian Blanchette, Dmitriy Traytel

To cite this version:
Anders Schlichtkrull, Jasmin Christian Blanchette, Dmitriy Traytel. A Verified Prover Based on Ordered Resolution. CPP 2019 - The 8th ACM SIGPLAN International Conference on Certified Programs and Proofs, Jan 2019, Cascais, Portugal. CPP 2019 - The 8th ACM SIGPLAN International Conference on Certified Programs and Proofs. <10.1145/3293880.3294100>. <hal-01937141>

HAL Id: hal-01937141
https://hal.archives-ouvertes.fr/hal-01937141
Submitted on 27 Nov 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
A Verified Prover Based on Ordered Resolution

Anders Schlichtkrull
DTU Compute
Technical University of Denmark
Kongens Lyngby, Denmark
andschl@dtu.dk

Jasmin Christian Blanchette
Department of Computer Science
Vrije Universiteit Amsterdam
Amsterdam, the Netherlands
j.c.blanchette@vu.nl

Dmitriy Traytel
Department of Computer Science
ETH Zürich
Zürich, Switzerland
traytel@inf.ethz.ch

Abstract
The superposition calculus, which underlies first-order theorem provers such as E, SPASS, and Vampire, combines ordered resolution and equality reasoning. As a step towards verifying modern provers, we specify, using Isabelle/HOL, a purely functional first-order ordered resolution prover and establish its soundness and refutational completeness. Methodologically, we apply stepwise refinement to obtain, from an abstract nondeterministic specification, a verified deterministic program, written in a subset of Isabelle/HOL from which we extract purely functional Standard ML code that constitutes a semidecision procedure for first-order logic.

CCS Concepts • Theory of computation → Logic and verification; Automated reasoning;

Keywords automatic theorem provers, proof assistants, first-order logic, stepwise refinement

ACM Reference Format:

1 Introduction
Automatic theorem provers based on superposition, such as E [42], SPASS [33], and Vampire [21], are often employed as backends in proof assistants and program verification tools [7, 20, 32]. Superposition is a highly successful calculus for first-order logic with equality, which generalizes both ordered resolution [2] and ordered completion [1].

Resolution operates on sets of clauses. A clause is an n-ary disjunction of literals \( L_1 \lor \cdots \lor L_n \) whose variables are interpreted universally. Each literal is either an atom \( A \) or its negation \( \neg A \). An atom is a symbol applied to a tuple of terms—e.g., \( \text{prime}(n) \). The empty clause is denoted by \( \bot \).

Resolution works by refutation: Conceptually, the calculus proves a conjecture \( \forall x. C \) from a set of axioms \( D \) by deriving \( \bot \) from \( D \cup \{ \exists x. \neg C \} \), indicating its unsatisfiability. As an optimization, it uses a redundancy criterion to discard tautologies, subsumed clauses, and other unnecessary clauses; for example, \( p(x) \lor q(x) \) and \( p(5) \) are both subsumed by \( p(x) \). Compared with plain resolution, ordered resolution relies on an order on the atoms to further prune the search space.

Modern superposition provers are highly optimized programs that rely on sophisticated calculi, with a rich metatheory. In this paper, we propose to verify, using Isabelle/HOL [30], a purely functional prover based on ordered resolution. Although our primary interest is in metatheory per se, there are of course applications for verified provers [50].

The verification relies on stepwise refinement [55]. Four layers are connected by three refinement steps.

Our starting point, layer 1 (Section 3), is an abstract Prolog-style nondeterministic resolution prover in a highly general form, as presented by Bachmair and Ganzinger [2] and as formalized in our earlier work [39, 40]. It operates on possibly infinite sets of clauses. Its soundness and refutational completeness are inherited by the other layers.

Layer 2 (Section 4) operates on finite multisets of clauses and introduces a priority queue to ensure that inferences are performed in a fair manner, guaranteeing completeness: Given a valid conjecture, the prover will eventually derive \( \bot \).

Layer 3 (Section 5) is a deterministic program that works on finite lists, committing to a strategy for assigning priorities to clauses. However, it is not fully executable: It abstracts over operations on atoms and employs logical specifications instead of executable functions for auxiliary notions.

Finally, layer 4 (Section 6) is a fully executable program. It provides a concrete datatype for atoms and executable definitions for all auxiliary notions, including unifiers, clause subsumption, and the order on atoms.

From layer 4, we can extract Standard ML code by invoking Isabelle's code generator [11]. The resulting prover constitutes a proof of concept: It uses an efficient calculus (layer 1) and a reasonable strategy to ensure fairness (layers 2 and 3), but depends on inefficient list-based data structures. Further refinement steps will be required to obtain a prover that is competitive with the state of the art.
The refinement steps connect vastly different levels of abstraction. The most abstract level is occupied by an infinitary logical calculus and the semantics of first-order logic. Soundness and completeness relate these two notions. At the functional programming level, soundness amounts to a safety property: Whenever the program terminates normally, its outcome is correct, whether it is a proof or a finite saturation witnessing unprovability. Correspondingly, refutational completeness is a liveness property: If the conjecture is valid, the program will always terminate normally. We find that, far from being academic exercises, Bachmair and Ganzinger’s framework [2] and its formalization [39, 40] adequately capture the metatheory of actual provers.

To our knowledge, our program is the first verified prover for first-order logic implementing an optimized calculus. It is also the first example of the application of refinement in a first-order context. This methodology has been used to verify SAT solvers [6, 29], which decide the satisfiability of propositional formulas, but first-order logic is semidecidable—sound and complete provers are guaranteed to terminate only for unsatisfiable (i.e., provable) clause sets. This complicates the transfer of completeness results across refinement layers.

The present work is part of the IsaFoL (Isabelle Formalization of Logic) project,1 which aims at developing a library of results about logic and automated reasoning [3]. The Isabelle files are available in the Archive of Formal Proofs [38, 39] and in the IsaFoL repository.2 The parts specific to the functional prover refinement amount to about 4000 lines of source text. A convenient way to study the files is to open them in Isabelle/jEdit [54], as explained in the repository’s readme file. The files were created using Isabelle version 2018, but the repositories will be updated to follow Isabelle’s evolution.

## 2 Atoms and Substitutions

The first three refinement layers are based on an abstract library of first-order atoms and substitutions. In the fourth and final layer, the abstract framework is instantiated with concrete datatypes and functions.

We start from IsaFoL’s library of clausal logic [6], which is parameterized by a type ‘a of logical atoms. Literals L are defined as an inductive datatype: ‘a literal = Pos ‘a | Neg ‘a. The type of clauses C, D, E is introduced as the alias ‘a clause = ‘a literal multiset, where multiset is the type constructor of finite multisets. Thus, the clause A ∨ B, where A and B are arbitrary atoms, is represented by the multiset {Neg A, Pos B}, and the empty clause ⊥ is represented by the empty multiset Ø. The complement operation is defined as −Neg A = Pos A and −Pos A = Neg A for any atom A.

In automated reasoning, it is customary to view clauses as multisets of literals rather than as sets. One reason is that multisets behave more naturally under substitution. For example, applying {y → x} to the two-literal clause p(x) ∨ p(y) results in p(x) ∨ p(x), which preserves the clause’s structure.

The truth value of ground (i.e., variable-free) atoms is given by a Herbrand interpretation: a set I, of type ‘a set, of all true ground atoms. The “models” predicate |= is defined as I |= A ⇔ A ∈ I. This definition is lifted to literals, clauses, and sets of clauses in the usual way. A set of clauses D is satisfiable if there exists an interpretation I such that I |= D.

Resolution depends on a notion of substitution and of most general unifier (MGU). These auxiliary concepts are provided by a third-party library, IsaFoR (Isabelle Formalization of Rewriting) [51]. To reduce our dependency on external libraries, we hide them behind abstract locales parameterized by a type of atoms ‘a and a type of substitutions ‘s.

We start by defining a locale substitution_ops that declares application (·), identity (id), and composition (◦):

```isabelle
locale substitution Ops = fixes id :: ‘s and o :: ‘s ⇒ ‘s and ◦ :: ‘a ⇒ ‘s ⇒ ‘a
```

Within the locale’s scope, we introduce a number of derived concepts. Ground atoms are defined as those atoms that are left unchanged by substitutions: is_ground A ⇔ ∀σ. A = A · σ. A ground substitution is a substitution whose application always results in ground atoms. Nonstrict and strict generalization are defined as generalizes A B ⇔ ∃σ. A · σ = B
strictly_generalizes A B ⇔ generalizes A B ∧ ¬ generalizes B A

The operators on atoms are lifted to literals, clauses, and sets of clauses. The grounding of a clause is defined as

```isabelle
grounding_of C = {C · σ | is_ground σ}
```

The operator is lifted to sets of clauses in the obvious way. Clause subsumption is defined as

```isabelle
subsumes C D ⇔ ∃σ. C · σ ⊆ D
```

with strictly_subsumes as its strict counterpart.

The next locale, substitution, characterizes the operations defined by substitution_ops. A separate locale is necessary because we cannot interleave assumptions and definitions in a single locale. In addition, substitution fixes a function for renaming clauses apart (so that they do not share any variables) and a function that, given a list of atoms, constructs an atom with these as subterms:

```isabelle
locale substitution = substitution_ops + fixes renamings apart :: ‘a clause list ⇒ ‘s list and atm_of_atms :: ‘a list ⇒ ‘a assumes A · id = A A · (σ o τ) = (A · σ) · τ (∀A. A · σ = A · τ) ⇒ σ = τ is_ground (C · σ) ⇒ ∃τ. is_ground τ ∧ C · τ = C · σ
```

1https://bitbucket.org/isafol/isafol/wiki/Home
The above definition is presented to give a flavor of our development. We refer to the Isabelle files for the exact definitions. Inside the locale, we prove further properties of the substitution_ops operations. Notably, we prove well-foundedness of the strictly_subsumes predicate based on the well-foundedness of strictly_generalizes, which is stated as an assumption. The atm_of_atms operation is used to encode a clause as a single atom in this well-foundedness proof.

Finally, a third locale, mgu, extends substitution by fixing a function mgu :: "a set set ⇒ 's option" that computes an MGU σ given a set of unification constraints.

### 3 Bachmair and Ganzinger’s Prover

Our earlier formalization [39, 40] of a nondeterministic ordered resolution prover presented by Bachmair and Ganzinger [2] forms layer 1 of our refinement. In this paper, we restrict our focus to binary resolution, which can be implemented efficiently and forms the basis of modern provers.

The ordered resolution calculus is parameterized by a total order \(>\) ("larger than") on ground atoms. For first-order logic, the order \(>\) is extended to an order \(\succ\) on nonground atoms so that \(B \succ A\) if and only if for all ground substitutions \(\sigma\), we have \(B \cdot \sigma > A \cdot \sigma\). The calculus consists of the single rule

\[
\frac{C ∨ A_1 ∨ ⋯ ∨ A_k \succ A ∨ D}{(C ∨ D) ∖ σ}
\]

where \(σ\) is the (canonical) MGU that solves the unification problem \(A_1 \equiv ⋯ \equiv A_k \equiv A\). In addition, each \(A_i ∖ σ\) must be strictly \(\succ\)-maximal with respect to the atoms in \(C ∖ σ\) (meaning that \(A_i\) is not \(\leq\) any atom in \(C ∖ σ\), and \(A ∖ σ\) is \(\succ\)-maximal with respect to the atoms in \(D ∖ σ\). To achieve completeness, the rule must be adapted slightly to rename apart the variables occurring in different premises.

A set of clauses \(D\) is saturated if any conclusion from premises in \(D\) is already in \(D\). The ordered resolution calculus is refutational complete, meaning that any unsatisfiable saturated set of clauses necessarily contains \(⊥\).

Resolution provers start with a finite set of initial clauses—the input problem—and successively add conclusions from premises in the set. If the inference rule is applied in a fair fashion, the set reaches saturation at the limit; if the set is unsatisfiable, this means \(⊥\) is eventually derived.

Crucially, not only do efficient provers add clauses to their working set, they also remove clauses that are deemed redundant. This requires a refined notion of saturation. We call a set of clauses \(D\) saturated up to redundancy, written \(\text{saturated}_\text{up to} D\), if any inference from nonredundant clauses in \(D\) yields a redundant conclusion.

Bachmair and Ganzinger’s nondeterministic first-order prover, called RP, captures the "dynamic" aspects of saturation. RP is defined as an inductive predicate \(\sim\) on states, which are triples \(S = (N, \mathcal{P}, \mathcal{O})\) of new clauses \(N\), processed clauses \(\mathcal{P}\), and old clauses \(\mathcal{O}\). Initially, \(N\) is the input problem, and \(\mathcal{P} \cup \mathcal{O}\) is empty. Clauses can be removed if they are tautological or subsumed or after subsumption resolution has been applied. When all clauses in \(N\) have been processed (either removed entirely or moved to \(\mathcal{P}\), a clause \(C\) from \(\mathcal{P}\) can be chosen for inference computation; \(C\) is then moved from \(\mathcal{P}\) to \(\mathcal{O}\), and all its conclusions with premises from the other old clauses form the new \(N\). Formally:

\[
\text{inductive}\ \sim : \ 'a\ state \Rightarrow \ 'a\ state \Rightarrow \ bool\ \ \text{where}
\]

\[
\begin{align*}
&\neg A \in C ∩ Pos A \in C \Rightarrow \\
&(N ∪ \{C\}, \mathcal{P}, \mathcal{O}) \sim_1 (N ∪ \{P\}, \mathcal{P}, \mathcal{O}) \\
&D \in \mathcal{P} ∪ \mathcal{O} \text{ \& subsumes } D \mathcal{C} \Rightarrow \\
&(N ∪ \{C\}, \mathcal{P}, \mathcal{O}) \sim_2 (N ∪ \{P\}, \mathcal{P}, \mathcal{O}) \\
&D \in N ∩ \text{strictly subsumes } D \mathcal{C} \Rightarrow \\
&(N ∪ \{C\}, \mathcal{P}, \mathcal{O}) \sim_3 (N ∪ \{P\}, \mathcal{P}, \mathcal{O}) \\
&D \in N ∩ \text{strictly subsumes } D \mathcal{C} \Rightarrow \\
&(N ∪ \{C\}, \mathcal{P}, \mathcal{O}) \sim_4 (N ∪ \{P\}, \mathcal{P}, \mathcal{O}) \\
&D \in \mathcal{P} ∪ \mathcal{O} \text{ \& reduces } D \mathcal{C} \mathcal{L} \Rightarrow \\
&(N ∪ \{C\}, \mathcal{P} ∪ \{L\}, \mathcal{O}) \sim_5 (N ∪ \{C\}, \mathcal{P}, \mathcal{O}) \\
&D \in N ∩ \text{reduces } D \mathcal{C} \mathcal{L} \Rightarrow \\
&(N ∪ \{C\}, \mathcal{P} ∪ \{L\}, \mathcal{O}) \sim_6 (N ∪ \{C\}, \mathcal{P}, \mathcal{O}) \\
&(0, \mathcal{P} ∪ \{C\}, \mathcal{O}) \sim_8 \ (\text{concl_of } \text{`infers between} \ \mathcal{O} \ \mathcal{C}, \mathcal{P} ∪ \{C\})
\end{align*}
\]

Subscripts on \(\sim\) identify the rules. The notation \(f ∖ X\) stands for the image of \(X\) under \(f\), \(\text{infers between} \ \mathcal{O} \ \mathcal{C}\) calculates all the ordered resolution inferences whose premises are a subset of \(\mathcal{O} \cup \mathcal{C}\) that contains \(C\), and reduces \(D \mathcal{C} \mathcal{L}\) is defined as \(∃L' ∖ σ, D = D' ∪ \{L'\} ∩ L = L' ∖ σ ∩ D' ∖ σ \subseteq C\).

The following derivation shows that RP can diverge even on unsatisfiable clause sets:

\[
\begin{align*}
&\sim_8^* (0, \{\neg p(a, a), p(x, y) \vee p(x, y)\}, \emptyset) \\
&\sim_9 (\emptyset, \{\neg p(a, a), p(x, y) \vee p(x, y)\}) \\
&\sim_9 (\{p(x, f(f(x)))\}, \{\neg p(a, a)\}) \\
&\sim_8 (0, \{\neg p(a, a), p(x, f(x))\}, \{p(x, y) \vee p(x, y), p(x, x)\}) \\
&\sim_8 (\{p(x, f(f(x)))\}, \{\neg p(a, a)\}) \\
&\sim_8 \cdots
\end{align*}
\]

We can leave \(\neg p(a, a)\) in \(\mathcal{P}\) forever and always generate more clauses of the form \(p(x, f^i(x))\), for increasing values of \(i\). This emphasizes the importance of a fair strategy for selecting clauses to move from \(\mathcal{P}\) to \(\mathcal{O}\) using rule 9.

Formally, a derivation is a possibly infinite sequence of states \(S_0 \sim S_1 \sim S_2 \sim \cdots\). In Isabelle, this is expressed by the codatatype of lazy lists:
The canonical way of expressing the unsatisfiability of a set appears in the \( \bot \) states that the limit of a fair derivation \( S \) or multiset of first-order clauses with respect to Herbrand \( \textit{els} \) before and after a transition: from a set of clauses, that set must be unsatisfiable: \( \text{Liminf} \) is defined as the set of all positions of \( S \) that are the projections of the \( N \), \( \mathcal{P} \), and \( O \) components of \( S \).

The soundness theorem states that if \( \text{RP} \) derives \( \bot \) (i.e., \( 0 \)) from a set of clauses, that set must be unsatisfiable:

**Theorem RP_sound:**

\[ \theta \in \text{clss} \_\text{of} \ (\text{Liminf} \ S) \Rightarrow \\
\neg \ \text{satisfiable} \ (\text{grounding} \_\text{of} \ (\text{lhd} \ S)) \]

In the above, \( \text{clss} \_\text{of} \ (N, \mathcal{P}, O) = N \cup \mathcal{P} \cup O \).

A stronger, finer-grained notion of soundness relates models before and after a transition:

**Theorem RP_model:**

\[ S \sim S' \Rightarrow \\
(\ i \models \text{grounding} \_\text{of} \ S' \iff i \models \text{grounding} \_\text{of} \ S) \]

The canonical way of expressing the unsatisfiability of a set or multiset of first-order clauses with respect to Herbrand interpretations is as the unsatisfiability of its grounding. Completeness of the prover can only be guaranteed when its rules are executed in a fair order, such that clauses do not get stuck forever in \( N \) or \( \mathcal{P} \). Accordingly, fairness is defined as \( \text{Liminf} \ N_S = \text{Liminf} \ P_S = 0 \). The completeness theorem states that the limit of a fair derivation \( S \) is saturated:

**Theorem RP_saturated_if_fair:**

\[ \text{fair} \ S \Rightarrow \text{saturated}_\text{upto} \ (\text{Liminf} \ (\text{grounding} \_\text{of} \ S)) \]

In particular, if the initial problem is unsatisfiable, \( \bot \) must appear in the \( O \) component of the limit of any fair derivation:

**Corollary RP_complete_if_fair:**

\[ \text{fair} \ S \land \neg \text{satisfiable} \ (\text{grounding} \_\text{of} \ (\text{lhd} \ S)) \Rightarrow \\
\theta \in \text{O} \_\text{of} \ (\text{Liminf} \ S) \]

### 4 Ensuring Fairness

The second refinement layer is the prover \( \text{RP}_w \), which ensures fairness by assigning a weight to every clause and by organizing the set of processed clauses—the \( \mathcal{P} \) state component—as a priority queue, where lighter clauses are chosen first. By assigning somewhat heavier weights to newer clauses, we can guarantee that all derivations are fair.

Another necessary ingredient for fairness is that derivations must be complete. For example, the incomplete derivation consisting of the single state \( (\{\}, 0, \emptyset) \) is not fair. This requirement is expressed formally as \( \text{full}_\text{chain} \ (\sim_w) \ S \).

For the rest of this section, we fix a full chain \( S \) such that \( \mathcal{P} \_\text{of} \ (\text{lhd} \ S) = \mathcal{O} \_\text{of} \ (\text{lhd} \ S) = 0 \).

Because each \( \text{RP}_w \) rule corresponds to an \( \text{RP} \) rule, it is straightforward to lift the soundness and completeness results from \( \text{RP} \) to \( \text{RP}_w \). The main difficulty is to show that the priority queue ensures fairness of full derivations, which is needed to obtain an unconditional completeness theorem for \( \text{RP}_w \), without the assumption fair \( S \).

**Definition.** The weight of a clause \( C \), which defines its priority in the queue, may depend both on the clause itself and on when it was generated. To reflect this, the \( \text{RP}_w \) prover uses a weight function \( w(C, i) \), where \( i \) is the timestamp. The larger the timestamp, the newer the clause. A state is now a quadruple \( S = (N, \mathcal{P}, O, I) \), where the first three components are finite multisets and \( I \) is the timestamp to assign to the next generation of clauses. Formally, we have the following type abbreviations:

- **Type synonym** \( \text{wclause} = ' \text{a clause} \times \text{nat} \)
- **Type synonym** \( \text{wstate} = ' \text{a wclause multiset} \times ' \text{a wclause multiset} \times ' \text{a wclause multiset} \times \text{nat} \)

We extend the \( \text{FO_resolution_prover} \) locale, in which \( \text{RP} \) is defined, with a weight function that, for any given clause, is strictly monotone with respect to the timestamp, so that older copies of a clause are preferred to newer ones:

**locale** \( \text{weighted_FO_resolution_prover} = \\
\text{FO_resolution_prover} + \\
\text{fixes} \ \text{weight} :: 'a \text{wclause} \Rightarrow \text{nat} \\
\text{assumes} \ i < j \Rightarrow \text{weight} (C, i) < \text{weight} (C, j) \)

The weight function is otherwise arbitrary. This gives nearly unlimited freedom when selecting clauses, which is possibly the most crucial heuristic in modern provers [43]. For example, breadth-first search corresponds to the instance where weight \( (C, i) \) is defined as \( i \).

The \( \text{RP}_w \) prover uses \( 'a \text{wclause} \) for clauses. It is defined inductively as follows:

**Inductive** \( \sim_w :: 'a \text{wstate} \Rightarrow 'a \text{wstate} \Rightarrow 'a \text{wstate} \Rightarrow bool \ where \\
\text{Neg} \ A \in C \land \text{Pos} \ A \in C \\
(\ N \cup \{C,t\}, \mathcal{P}, O, t) \sim_{w1} (N, \mathcal{P}, O, t) \\
| \ D \in \text{fst} \ (\mathcal{P} \cup O) \land \text{subsumes} \ D \ C \Rightarrow \\
(\ N \cup \{C,t\}, \mathcal{P}, O, t) \sim_{w2} (N, \mathcal{P}, O, t) \)
To prevent this, the \( \text{RP} \) rules are formulated so that whenever they remove the earliest copy of any clause \( C \in \mathcal{P} \), they also remove all its copies from \( \mathcal{P} \). This property is captured by the following lemma:

**Lemma** \( \text{preserve_min}_{\mathcal{P}} \):

\[
\mathcal{S} \sim_{\omega} \mathcal{S}' \land (C, i) \in \mathcal{P}_{\mathcal{S}} \land C \in \text{fst} \cdot \mathcal{P} \iff (C, i) \in \mathcal{P}_{\mathcal{S}'} \land (C, i) \in \text{fst} \cdot \mathcal{P}
\]

This completes our review of \( \text{RP}_w \). As an intermediate step towards a more concrete prover, we restrict the weight function to be a linear equation that considers both timestamps and clause sizes:

**locale** \( \text{weighted}_F O_\text{resolution_prover_with_size}_{\text{timestamp_factors}} = \)**

\[
\text{FO_resolution_prover} + \text{fixes} \text{ size_factor :: nat and timestamp_factor :: nat}
\]

**assumes** timestamp_factor \( > 0 \)

**begin**

\[
\text{fun} \quad \text{weight} :: \text{ a wclause } \Rightarrow \text{ nat where}
\]

\[
\text{weight} (C, i) = \text{ size_factor } \cdot |C| + \text{ timestamp_factor } \cdot i
\]

**end**

where \( |C| = \sum_{A : \text{AEC \sim AEC}} |A| \). It is easy to prove that this definition of weight is strictly monotone and hence that this locale is a sublocale of \( \text{weighted}_F O_\text{resolution_prover} \). This gives us a correspondingly specialized version of \( \text{RP}_w \) that will form the basis of further refinement steps.

The idea of organizing \( \mathcal{P} \) as a priority queue is well known in the automated reasoning community. Bachmair and Ganzinger [2, p. 44] mention it in a footnote, but they require the weight to be monotone not only in the timestamp but also in the clause size, claiming that this is necessary to ensure fairness. Our proof reveals that clause size is irrelevant, even in the presence of reductions. This demonstrates how working out the details and making all assumptions explicit using a proof assistant can help clarify fine theoretical points.

**Refinement Proofs.** To lift the soundness and completeness results about \( \text{RP} \) to \( \text{RP}_w \), we must first show that any possible behavior of \( \text{RP}_w \) on states of type \( \text{wstate} \) is a possible behavior of \( \text{RP} \) on the corresponding values of type \( \text{state} \):

**Lemma** \( \text{weighted}_R \text{P}_w_{\text{imp}}_{\text{RP}} : \)**

\[
\mathcal{S} \sim_{\omega} \mathcal{S}' \Rightarrow \text{state_of} \mathcal{S} \sim \text{state_of} \mathcal{S}'
\]

The proof is by induction on the rules of \( \text{RP}_w \), with one difficult case. Inference computation (rule 9) converts a set to a finite multiset using \( \text{mset_set} \), which is undefined for infinite sets. Thus, we must show only a finite set of inferences may be performed from a finite clause set:

**Lemma** \( \text{finite}_o F O_\text{_resolution_inferences_between} : \)**

\[
\text{finite} \mathcal{D} \Rightarrow \text{finite (infers_between} \mathcal{D} \mathcal{C})
\]

A binary resolution inference takes two premises, of the form \( \text{CAA} = C \lor A_1 \lor \cdots \lor A_k \) and \( \text{DA} = \neg A \lor D \), and produces a conclusion \( E = (C \lor D) \lor \). It can be represented compactly by a tuple of the form \( (\text{CAA}, \text{DA}, \text{AA}, A, E) \), where \( \text{AA} = A_1 \lor \cdots \lor A_k \). We must show that the set of such
tuples produced by \texttt{infers\_between} is finite, assuming \( \mathcal{D} \) is finite. First, observe that the last component \( E \) of a tuple is determined by the other four. Hence it suffices to consider quadruples \((\text{CAA}, \text{DA}, \text{AA}, A)\). Let \( \mathcal{DC} = \mathcal{D} \cup \{ C \} \), and let \( n \) be the length of the longest clause in \( \mathcal{DC} \). Moreover, let \( \mathcal{A} = \bigcup_{D \in \mathcal{DC}} \text{ atms\_of } D \) and \( \mathcal{AAA} = \{ B \mid \text{ set\_mset } B \subseteq \mathcal{A} \wedge |B| \leq n \} \). Then all inferences between \( D \) and \( C \) belong to \( \mathcal{DC} \times \mathcal{DC} \times \mathcal{AAA} \times \mathcal{A} \), a cartesian product of finite sets.

**Soundness and Completeness Proofs.** Using the refinement lemma \texttt{weighted\_RP\_imp\_RP}, it is easy to lift the \texttt{RP\_model} theorem (Section 3) to \texttt{RP\_w}:

\begin{align*}
\text{theorem \texttt{weighted\_RP\_model}}: \\
S \rightsquigarrow_w S' \implies \\
(1 \models \text{ grounding\_of } S' \iff 1 \models \text{ grounding\_of } S)
\end{align*}

Completeness is considerably more difficult. We first show that the use of timestamps ensures that all full \texttt{RP\_w} derivations are fair. In principle, a full derivation could be unfair by virtue of being finite and ending in a state such as \( \mathcal{N} \) or \( \mathcal{P} \) is nonempty. However, this is impossible because a transition of rule 8 or 9 could then be taken from the last state, contradicting the hypothesis that the derivation is full. Hence, finite full derivations are necessarily fair:

\begin{align*}
\text{lemma \texttt{fair\_if\_finite}}: \\
\text{finite } S \iff \text{ fair \texttt{Imap \_state\_of } S}
\end{align*}

There are two ways in which an infinite derivation \( S \) in \texttt{RP\_w} could be unfair: A clause could get stuck forever in \( \mathcal{N} \) or in \( \mathcal{P} \). We show that the \( \mathcal{N} \) case is impossible by defining a measure on states that decreases with respect to the lexicographic extension of \( > \) on \texttt{nat} to pairs:

\begin{align*}
\text{abbreviation \texttt{RP\_basic\_measure}}: & \quad \text{’} a \text{ wstate } \Rightarrow \text{ nat}^2 \\
\text{where}& \\
\text{RP\_basic\_measure} (\mathcal{N}, \mathcal{P}, O, t) \equiv \\
(\text{sum} ((\lambda (C, _.). \mid |C| + 1) \cdot (\mathcal{N} \cup \mathcal{P} \cup O)), \mid \mathcal{N}\mid)
\end{align*}

The first component of the pair is the total size of all the clauses in the state, plus 1 for each clause to ensure that empty clauses are also counted. The second component is the number of clauses in \( \mathcal{N} \). It is easy to see why the measure is decreasing. Rule 9, inference computation, is not applicable due to our assumption that a clause remains stuck in \( \mathcal{N} \). Rule 8, which moves a clause from \( \mathcal{N} \) to \( \mathcal{P} \), decreases the measure’s second component while leaving the first component unchanged. The other rules decrease the first component since they remove clauses or literals. Formally:

\begin{align*}
\text{lemma \texttt{weighted\_RP\_basic\_measure\_decreasing\_N}}: \\
S \rightsquigarrow_w S' \wedge (C, \_.) \in \mathcal{N} \cap S \implies \\
(\text{RP\_basic\_measure } S', \text{RP\_basic\_measure } S) \\
\in \text{ RP\_basic\_rel}
\end{align*}

where \texttt{RP\_basic\_rel} = \texttt{natLess <lex} \texttt{natLess and natLess = (m, n) \mid m < n}.

What if a clause \( C \) is stuck in \( \mathcal{P} \)? Lemma \texttt{preserve\_min\_P} states that in any step, either all copies of \( C \) are removed or the one with the lowest timestamp is kept. Hence, \( C \)'s timestamp will either remain stable or decrease over time. Since \( > \) is well founded on natural numbers, eventually a fixed \( i \) will be reached and will belong to the limit:

\begin{align*}
\text{lemma \texttt{persistent\_wclause\_in\_P\_if\_persistent\_clause}}: \\
C \in \text{ Liminf } (\text{ Imap } \_p\_of (\text{ Imap \_state\_of } S)) \implies \\
\exists i. (C, i) \in \text{ Liminf } (\text{ Imap } (\text{ set\_mset } \_p\_of ) S)
\end{align*}

Again, we define a measure, but it must also decrease when inferences are computed and new clauses appear in \( \mathcal{N} \). In this case, \texttt{RP\_basic\_measure} may increase.) Our new measure is parameterized by a predicate \( p \) that can be used to filter out undesirable clauses:

\begin{align*}
\text{abbreviation \texttt{RP\_filtered\_measure}}: & \quad (\lambda \wclause \Rightarrow \text{ bool}) \Rightarrow (\text{’} a \text{ wstate } \Rightarrow \text{ nat}^2 \\
\text{where}& \\
\text{RP\_filtered\_measure} (\lambda, \mathcal{P}, O, t) \equiv \\
(\text{sum} ((\lambda (C, _.). \mid |C| + 1) \cdot (\mathcal{N} \cup \mathcal{P} \cup O | |p\_Di|), \mid \mathcal{N} \mid)
\end{align*}

Notice that the case \texttt{RP\_filtered\_measure} (\( \lambda \), True) essentially amounts to \texttt{RP\_basic\_measure}. In the formalization, we use \texttt{RP\_filtered\_measure} (\( \lambda \), True) to avoid duplication.

Suppose the clause \( C \) that is stuck in \( \mathcal{P} \) has weight \( w \) in the limit, and suppose that a clause \( D \) is moved from \( \mathcal{P} \) to \( \mathcal{O} \) by rule 9. That clause’s weight must be at most \( w \); otherwise, it would not have been preferred to \( C \). Thus, infinite derivations necessarily consist of segments each consisting of finitely many applications of rules other than rule 9 followed by an application of rule 9: \( (\rightsquigarrow_{\text{w}\_\_}, \rightsquigarrow_{\text{w}\_0})^m \). Since each application of rule 9 increases the \( i \) component of the state, eventually we reach a state in which \( t > w \). As a consequence of strict monotonicity of weight, any clauses generated by inference computation from that point on will have weights above \( C \)'s, and if \( C \) remains stuck, then so must these clauses. Thus, we can ignore these clauses altogether, by using \( \lambda (C, i) \). \( i \leq w \) as the filter \( p \). We adapt the corresponding relation to consider the extra argument:

\begin{align*}
\text{abbreviation \texttt{RP\_filtered\_rel}}: (\text{nat}^2)^2 \text{ set where} \\
\text{RP\_filtered\_rel} = \\
\text{natLess <lex} \text{ natLess <lex} \text{ natLess}
\end{align*}

The measure \texttt{RP\_filtered\_measure} (\( \lambda (\_, i) \). \( i \leq w \)) decreases for steps occurring between inference computations and for all steps once we have reached a state where \( t > w \) (at which point all inference computations are blocked by \( C \)). To obtain a measure that also decreases on inference computation, we add a component \( w + 1 - t \) of the measure. We also add a component \texttt{RP\_basic\_measure} \( S \) to ensure that the measure decreases when a clause \( (C, i) \) such that \( i > w \) is simplified. This yields the combined measure

\begin{align*}
\text{abbreviation \texttt{RP\_combined\_measure}}: & \quad \text{nat } \Rightarrow (\text{’} a \text{ wstate } \Rightarrow \text{ nat} \times \text{ nat}^2 \times \text{ nat}^3 \\
\text{where}& \\
\text{RP\_combined\_measure } w S \equiv \\
w + 1 - t\_of S, \\
\text{RP\_filtered\_measure} (\lambda (\_, i). i \leq w) S, \\
\text{RP\_basic\_measure } S)
\end{align*}
This measure is indeed decreasing with respect to a lexicographic order:

**lemma** weighted\_RP\_basic\_measure\_decreasing\_P:
\[ S \leadsto \_w S' \land Ci \in P \rightarrow (RP\_combined\_measure (weight Ci) S',
RP\_combined\_measure (weight Ci) S) \in \text{natLess } \langle\text{lex}\rangle\] RP\_filtered\_rel \langle\text{lex}\rangle\] RP\_basic\_rel

By combining the two lemmas weighted\_RP\_basic\_measure\_decreasing\_N and weighted\_RP\_basic\_measure\_decreasing\_P, we can prove all derivations starting with \( P = O = \emptyset \) fair:

**theorem** weighted\_RP\_fair:
fair (limap state of Ss)

Since all derivations are fair and \( RP_w \) derivations correspond to \( RP \) derivations, it is trivial to lift \( RP \)'s saturation and completeness theorems:

**corollary** weighted\_RP\_saturated:
saturated\_upto (Liminf (limap grounding of Ss))

**corollary** weighted\_RP\_complete:
\( \neg \) satisfiable (grounding\_of (lhd Ss)) \Rightarrow \emptyset \in O\_of (Liminf (limap state of Ss))

5 Eliminating Nondeterminism

The third refinement layer defines a functional program \( RP_d \) that embodies a specific rule application strategy, thereby eliminating \( RP_w \)'s nondeterminism.Clauses are represented by lists, and multisets of clauses by lists of lists.

**Definition.** Our prover corresponds roughly to the following pseudocode:

**function** RP\_d\( (N, P, O, t) \) is
repeat forever
| if \( \bot \in P \uplus O \) then
return \( P \uplus O \)
else if \( N = P = O = t \) then
return \( O \)
else if \( N = 0 \) then
let \( C \) be a minimal-weight clause in \( P \);
\( N := \) conclusions of all inferences from \( O \uplus \{C\} \)
involving \( C \), with timestamp \( t \);
move \( C \) from \( P \) to \( O \);
\( t := t + 1 \)
else
remove an arbitrary clause \( C \) from \( N \);
reduce \( C \) using \( P \uplus O \);
if \( C = \bot \) then
return \( \{\bot\} \)
else if \( C \) is neither a tautology nor subsumed by a clause in \( P \uplus O \) then
reduce \( P \) using \( C \);
reduce \( O \) using \( C \), moving any reduced clauses from \( O \) to \( P \);
remove all clauses from \( P \) and \( O \) that are strictly subsumed by \( C \);
add \( C \) to \( P \)

The function should be invoked with \( N \) as the input problem, \( P = O = \emptyset \), and an arbitrary timestamp \( t \) (e.g., 0). The loop is loosely modeled after Vampire’s proof procedure [52].

In Isabelle, the list-based representations compel us to introduce the following type abbreviations:

**type synonym** ‘a lclause = ‘a literal list
type synonym ‘a delclause = ‘a lclause × nat
type synonym ‘a dstate =
‘a delclause list × ‘a delclause list × ‘a delclause list × nat

The prover is defined inside a locale that inherits weighted\_FO\_resolution\_prover\_with\_size\_timestamp\_factors. The core function, \( RP_d\_step \), performs a single iteration of the main loop. Here is the definition, excluding auxiliary functions:

**fun** RP\_d\_step :: \( ‘a dstate \Rightarrow ‘a dstate \) where
RP\_d\_step \((N, P, O, t)\) =
| if \( \exists Ci \in P \at O . \) lst Ci = [] then
(\([\],[\] , remdups P \at O , t + |remdups P|) )
else case \( N \) of
\([[\] ] ⇒ (case P of
\([\] ] ⇒ (N, P, O, t)
\) P0 \# P \) \Rightarrow
let
\( (C, i) = \) select\_min\_weight\_clause \( P0 \) \( P \);
\( N = \) map \((\lambda D . (D, t))\) \( \) remdups
(resolve\_rename \( C \@ C \) \at \( (\) map
(resolve\_rename\_both\_ways \( C \at \) \at \) \( O \))\);
\( P = \) filter \((\lambda(D, i). \) mset \( D \neq \) mset \( C \) \at \( P \);
\( O = (C, i) \# O \);
\( t = t + 1 \)
in
(N, P, O, t)
\) \)
| \( (\{\} ) \Rightarrow \)
let 
\( C = \) reduce \( (\) map \( \) lst \( (P \at O)\)\) \( \) \( C \) in
\( (\) \)
if \( C = [] \) then
(\([\],[\] , \( (\) \), \( i \), \( t + 1 \) \)
else if is\_tautology \( C \)
\) \at
\) \)
in
(N, P, O, t)
\) \)
else let
\( P = \) reduce\_all \( C \) \at \( P \);
\( (\) \)
\( (\) \)
\( (\) \)
in
(N, P, O, t)
The \# operator abbreviates the Cons constructor, and @ is the append operator.

The existential quantifier above is unproblematic because it ranges over a finite set, but some of the auxiliary functions use infinite quantification. Notably, subsumption of a clause
$D$ by another clause $C$ is defined as $\exists \sigma. \ C \cdot \sigma \subseteq D$ (Section 2), where $\sigma$ ranges over substitutions. Nonexecutable constructs are acceptable if we know that we can replace them by equivalent executable constructs further down the refinement chain; for example, an implementation of subsumption can compute a witness $\sigma$ using matching, instead of blindly enumerating all possible substitutions.

The main program is a tail-recursive function that repeatedly calls $\text{RP}_d$ step until a final state $([[],[],O,t])$ is reached, at which point it returns the set $O$ stripped of its timestamps:

\[
\text{partial\_function\ (option)} \\
\quad \text{RP}_d :: 'a \text{ dstate }\Rightarrow 'a \text{ lclause\ list\ option} \\
\quad \text{where} \\
\quad \text{RP}_d\ S = \text{if is\_final } S \text{ then Some\ (map\ fst\ (O\_of\ S))} \text{ else RP}_d\ \text{(RP}_d\ \text{step}\ S)
\]

Since the recursion may diverge, we cannot introduce the function using the $\text{fun}$ command [22]. Instead, we use $\text{partial\_function\ (option)}$ [23], which puts the computation in an option monad. The function’s result is of the form Some $R$ if the recursion terminates and None otherwise.

**Refinement Proofs.** Using refinement, we connect the $\text{RP}_d$ step function to the $\text{RP}_w$ predicate. $\text{RP}_d$ step has a coarser granularity than $\text{RP}_w$: A single invocation on a nonfinal state $S$ can amount to a chain of $\text{RP}_w$ transitions. This is captured by the following (weak)-refinement property:

\[
\text{lemma\ nonfinal\_deterministic\_RP\_step:} \\
\quad \neg\ \text{is\_final } S \Rightarrow \\
\quad \text{wstate\_of } S \xrightarrow{w} \text{wstate\_of } (\text{RP}_d\ \text{step}\ S)
\]

where $\text{wstate\_of}$ converts $\text{RP}_d$ states to $\text{RP}_w$ states. The entire proof, including key lemmas, is about 1300 lines long. It follows the case distinctions in $\text{RP}_d\ \text{step}$’s definition:

**case** $\exists C \in P \cap O.\ \text{fst } CI = [1]:$

By induction on $|\text{remdups } P|$ (where remdups removes duplicates), there must exist a derivation of the form

\[
\begin{align*}
\text{wstate\_of } (N', P, O, t) \\
\xrightarrow{w_2} \text{wstate\_of } (P, O, t) \\
\xrightarrow{w_9} \text{wstate\_of } (N', (C, i) \# O, t + 1) \\
\xrightarrow{w} \text{wstate\_of } ([][], O', t + |\text{remdups } P'|)
\end{align*}
\]

for $P'$ = filter ($\lambda(D,j)$. mset $D \neq \text{mset } C) P, O' = \text{remdups } P' \cap O$, and suitable $N'$ and $(C, i) \in P$. The last step is justified by the induction hypothesis.

**case** $N = P = [1]:$

Contradiction with the assumption that $(N, P, O, t)$ is a nonfinal state.

**case** $N = [\_]:$

It suffices to show that the transition

\[
\text{wstate\_of } ([], P, O, t) \\
\xrightarrow{w_5} \text{wstate\_of } (N', P', (C, i) \# O, t + 1)
\]

is possible, where $(C, i) \in P$ is a minimal-weight clause, $N' = \text{map } (\lambda D. (D, t)) (\text{remdups } (\text{resolve\_rename } C C \\text{@ \ concat } (\text{map\ resolve\_rename\_both\_ways } C \circ \text{fst}) \ O)))$, and $P' = \text{filter } (\lambda(D,j)$. mset $D \neq \text{mset } C) P$. The main proof obligation is that $N'$, converted to multisets, equals the multiset $\text{mset\_set } ((\text{Ind}. (D, t)) ' \\text{conc\_of } \text{in\_fers\_between } (\text{set\_mset } (\text{fst } O)) C)$ specified in rule $\sim w_9$. The distance between the functional program and its mathematical specification is at its greatest here. The proof is tedious but straightforward.

**otherwise:**

Let $C' = \text{reduce } (\text{map\ fst } P @ \text{map\ fst } O) [\_]. C$. If $C' = [\_]$, then

\[
\begin{align*}
\text{wstate\_of } ((C, i) \# N', P, O, t) \\
\text{wstate\_of } ((([], i) \# N', P, O, t) \\
\text{wstate\_of } ((([], i) \# N', P, O, t) \\
\text{wstate\_of } ((([], i) \# N', P, O, t) \\
\text{wstate\_of } ((([], i) \# N', P, O, t) \\
\text{wstate\_of } ((([], i) \# N', P, O, t) \\
\text{wstate\_of } ((([], i) \# N', P, O, t)
\end{align*}
\]

Otherwise, if $\text{is\_tautology } C' \lor \text{subsume } (\text{map\ fst } (P@O)) C'$, then

\[
\begin{align*}
\text{wstate\_of } ((C, i) \# N', P, O, t) \\
\text{wstate\_of } ((C', i) \# N', P, O, t) \\
\text{wstate\_of } ((C', i) \# N', P, O, t) \\
\text{wstate\_of } ((C', i) \# N', P, O, t) \\
\text{wstate\_of } ((C', i) \# N', P, O, t) \\
\text{wstate\_of } ((C', i) \# N', P, O, t)
\end{align*}
\]

for suitable lists $P'$, $\text{back\_to } P, O', O'', P''$.}

**Soundness and Completeness Proofs.** Let $S_0 = (N_0, [], [], t_0)$ be an arbitrary initial state. Soundness means that whenever $\text{RP}_d S_0$ terminates with some clause set $R$, then $R$ is a saturation that satisfies the same models as $N_0$. Moreover, if $N_0$ is unsatisfiable, then $R$ contains $\perp$, providing a simple syntactic check for unsatisfiability. Completeness means that divergence is possible only if $N_0$ is satisfiable. For satisfiable clause sets $N_0$, both termination and divergence are possible.

To lift soundness and completeness results from $\text{RP}_w$ to $\text{RP}_d$, we first define $S_0$ as a full chain of nontrivial $\text{RP}_d$ steps starting from $S_0$. We let $S_0 = \text{derivation\_from } S_0$, with $\text{primcorec\ deriv\_from :: 'a \text{ dstate }\Rightarrow 'a \text{ dstate\ list\ where}}$

\[
\text{deriv\_from } S = \text{LCons } S \text{ (if is\_final } S \text{ then LNil\ else derivation\_from } (\text{RP}_d\ \text{step}\ S))
\]

Based on $S_0$, we let $wS_0 = \text{lmap\ wstate\_of } S_0$ and note that $wS_0$ is a full chain of "big" $\sim w_9$ steps. Using a lemma that will be proved below, we obtain a full chain $sswS_0$ of "small"
The proof is by contradiction:

Assume that $\neg$ satisfiable (grounding_of $N_0$). Hence, by weighted_RP_complete we have $0 \in \mathcal{O}_w$ of sswoS$. It is easy to show that sswoS's limit is a subset of wS's limit; hence $0 \in \mathcal{O}_w$ of wS$. This implies the existence of a natural number $k$ such that $0 \in \mathcal{O}_w$ of (lhd wS $k$). Hence $0 \in \mathcal{O}_w$ of (RPd_step$k$ S0). However, by an induction on $k$, we can show that RPd must terminate after at most $k$ iterations, contradicting the assumption that RPd diverges.

A Coinductive Puzzle. A single “big” step of the deterministic prover RPd may correspond to many “small” steps of the weighted prover RPw. To transfer the results from RPw to RPd, we must expand the big steps. The core of the expansion is an abstract property of chains and transitive closure:

Let $R$ be a relation and xs a chain of $R^*$ transitions. There exists a chain of $R$ transitions that embeds xs—i.e., that contains all elements of xs in the same order and with only finitely many elements inserted between each pair of consecutive elements of xs.

On finite chains, this property can be proved by straightforward induction. But the completeness proof must also consider infinite chains. Coinduction and corecursion up-to techniques are useful for such tasks.

The desired property is stated formally as follows:

**Lemma chain_trancpl_imp_exists_chain:**

chain R' xs $\Rightarrow$

$\exists ys. \text{chain } R \text{ ys } \land \text{ lhd xs } = \text{lhd ys} \land \text{llast xs } = \text{llast ys}$

where the embedding $\subseteq$ of lazy lists is defined coinductively using $\vdash$, which preends a finite list to a lazy list:

**coinductive $\subseteq$:** a list $\Rightarrow$ a list $\Rightarrow$ bool where

finite xs $\Rightarrow$ LNil $\subseteq$ xs

| xs $\subseteq$ ys $\Rightarrow$ LCons x xs $\subseteq$ zs $\vdash$ LCons x ys

**fun $\vdash$:** a list $\Rightarrow$ a list $\Rightarrow$ a list where

| [] $\vdash$ xs = xs

| (z # zs) $\vdash$ xs = LCons z (zs $\vdash$ xs)

The definition of $\subseteq$ ensures that infinite lazy lists only embed other infinite lazy lists, but not the finite ones: xs $\subseteq$ ys $\Rightarrow$ (finite xs $\Rightarrow$ infinite ys). The unguarded calls to llast may seem worrying, but the function is conveniently defined to always return the same unspecified element for infinite lists.

To prove the lemma above, we instantiate the existential quantifier by the following corecursively defined witness:

**corec wit ((a $\Rightarrow$ a $\Rightarrow$ bool) $\Rightarrow$ a list $\Rightarrow$ a list where**

wit R xs = (case xs of

LCons x (LCons y ys) $\Rightarrow$

LCons x (pick R x y $\vdash$ wit R (LCons y ys))

$\_ = \Rightarrow$ xs)

Here pick R x y returns an arbitrary finite list of R-related states connecting the R*-related x and y. Its definition is pick R x y = (SOME zs. chain R (llist_of (x # zs @ [y]))), where llist_of converts finite lists into lazy lists and SOME is Hilbert’s choice operator. Thus, pick satisfies the characteristic property R' x y $\Rightarrow$ chain R (llist_of (x # pick R x y @ [y])). The nonexecutability entailed by the use of Hilbert choice is unproblematic because the wit function is used only in the proofs and not in the prover’s code.

The definition of wit is not primitively corecursive. Although there is a guarding LCons constructor, the corecursive call occurs under $\vdash$, which makes the productivity of this function nontrivial. This syntactic structure of the definition is called corecursive up to $\vdash$. Ultimately, wit is productive because $\vdash$ does not remove any LCons constructors from its second arguments. A slightly weaker requirement, called friendliness, is supported by Isabelle’s corec command [5]. For the above definition to be accepted $\vdash$ must be registered as a “friend.” This involves a one-line proof.

The four conjuncts in chain_trancpl_imp_exists_chain are discharged in turn under the assumption chain R' xs. In order of increasing difficulty: lhd (wit R xs) = lhd xs follows by simple rewriting. Next, llast (wit R xs) = llast xs requires an induction in the case of finite chains xs. For any infinite chain zs, llast zs is defined as a fixed unspecified ‘a value. The properties xs $\subseteq$ wit R xs and chain R (wit R xs) require a coinduction on $\subseteq$ and chain, respectively. In keeping with the
definition, plain coinduction on $\sqsubseteq$ and chain does not suffice, and we must use coinduction up to $\sqsubseteq$ on $\sqsubseteq$ and chain.

The property $\text{chain_transcl_imp_exists_chain}$ easily extends to full chains.

6 Obtaining Executable Code

Our deterministic prover $\text{RP}_d$ is already quite close to being an executable program. The fourth refinement, the prover $\text{RP}_n$, adds the missing ingredients: a concrete representation of terms and an executable algorithm for clause subsumption.

First-Order Terms. We instantiate our abstract notion of term using a particularly comprehensive formalization of terms developed as part of the IsaFoR library [51]. This rewriting-independent part of IsaFoR has recently moved to the Archive of Formal Proofs [47].

IsaFoR terms are defined as the following datatype:

\[
\text{ datatype } ('f,'v)\text{ term } = \text{ Var } 'v | \text{ Fun } 'f ((('f,'v) \text{ term}\text{ list}) \Rightarrow ('f,'v) \text{ term} )
\]

To simplify notation, in this paper we fix $'f = 'v = \text{ nat}$ and abbreviate $('f,'v) \text{ term}$ by $\text{ term}$. In the formalization, polymorphic types are used whenever possible. IsaFoR also defines the standard monadic term substitution $:: \text{ term } \Rightarrow (('v \Rightarrow \text{ term}) \Rightarrow \text{ term})$ and a unify $:: (\text{ term } \times \text{ term}) \Rightarrow \text{ subst} \Rightarrow \text{ Isubst}$ function, where $\text{Isubst} = (('v \times \text{ term}) \text{ list} )$ is the list-based representation of a finite substitution. The function unify computes the MGU for a list of unification constraints that is compatible with a given substitution. IsaFoR includes a wealth of theorems, including the correctness of unify and the well-foundedness of strict term generalization, defined as $(\exists \sigma. s \cdot \sigma = t) \land (\exists \sigma. t \cdot \sigma = s)$.

This infrastructure allows us to conveniently instantiate our locales $\text{ substitution_ops}$, $\text{ substitution}$, and $\text{ mgu}$. We instantiate the type $'a$ of atoms with $\text{ term}$ and the type $'s$ of substitutions with $'v \Rightarrow \text{ term}$ and the constants $\text{id}$, $\text{id}$, and $\text{ atm_of }$, $\text{ atms }$, $\text{ Var}$, $\text{ Var}$, $\text{ Atom}$, $\text{ Atom}$, and (arbitrarily) $\text{ Fun}$ 0. For the MGU computation, there is a slight type mismatch: IsaFoR offers a list-based unifier, whereas our locale requires the type $\text{ term } \text{ set } \Rightarrow (\text{ term } \Rightarrow \text{ term})$ option. It is easy to translate a finite set of finite sets of terms into a finite list of pairs of constraints. To be executable, the translation requires us to sort the terms belonging to set with respect to an arbitrary (but executable) linear order.

Only the function $\text{renamings}_\text{apart}$ was not present in IsaFoR. We supply a definition:

\[
\begin{align*}
\text{fun } & \text{renamings}_\text{apart} :: \text{ term clause list } \Rightarrow (\text{'}v \Rightarrow \text{ term}) \text{ list} \\
\text{where } & \text{renamings}_\text{apart} [] = [] \\
& \text{renamings}_\text{apart} (C \# Cs) = \\
& \quad (\lambda v. v + \max ((\emptyset) \cup \text{ vars clause list } (Cs \cdot \text{ as})) + 1) \# \text{ as} \\
\text{where } & \text{vars clause list } :: \text{ term clause list } \Rightarrow \text{ 'v set} \text{ returns the variables contained in a list of clauses. The creation of fresh variable names relies on 'v = \text{ nat}.}
\end{align*}
\]

Finally, the $\text{FO_resolution_prover}$ locale requires that the type of atoms supports two comparison operators: a well-order $>$ and a comparison $\leq$ that is stable under substitution (i.e., $B > A \Rightarrow B \cdot \sigma > A \cdot \sigma$). Moreover, $>$ and $\leq$ must coincide on ground atoms. We instantiate $>$ with the Knuth-Bendix order [19] on terms, provided by IsaFoR [46]. This order is executable, stable under substitution, well-founded, and total on ground terms. The well-order $>$, which must be total on all terms, is then defined as an arbitrary extension of a partial well-founded order $\rightarrow$ to a well-order, using Hilbert choice. This makes $>$ nonexecutable, but this is acceptable since it is $>$, not $\rightarrow$, that is used in the prover’s code.

Clause Subsumption. The second hurdle concerns clause subsumption. Its mathematical definition, subsumes $C D \Leftrightarrow \exists \sigma. C \cdot \sigma \subseteq D$, involves an infinite quantification.

The problem of deciding whether such a substitution exists is NP-complete [18]. We start with the following naive code. In contrast to the mathematical definition, which operates on multisets of literals, our function operates on lists:

\[
\begin{align*}
\text{fun } & \text{subsumes list :: term literal list } \Rightarrow \\
& \text{term literal list } \Rightarrow (\text{'}v \Rightarrow \text{ term option}) \Rightarrow \text{ bool}
\end{align*}
\]

where

\[
\begin{align*}
\text{subsumes list } [\] Ks \sigma = \text{ True} \\
\text{subsumes list } (L \# Ls) Ks \sigma = \\
(\exists K \in \text{ set } Ks. \text{ is_pos } K = \text{ is_pos } L \land \text{ case match_term list } [(\text{ atm_of } L, \text{ atm_of } K)] \sigma \text{ of } \\
\text{ None } \Rightarrow \text{ False} \\
\text{ Some } \rho \Rightarrow \text{ subsumes list } Ls (\text{ remove1 } K Ks) \rho)
\end{align*}
\]

In the Cons case, we must consider all possible matching literals for $L$ from $Ks$ compatible with the substitution $\sigma$. The bounded existential quantification that expresses this non-determinism can be executed by iterating over the finite list $Ks$. The functions $\text{is_pos}$ and $\text{atm_of}$ are the discriminator and selector for literals. The function $\text{match_term list}$ is provided by IsaFoR. It attempts to extend a given substitution into Some matcher for a list of matching constraints, given as term pairs. If the extension is impossible, $\text{match_term list}$ returns None. This substitution-passing style is typical of purely functional implementations of matching.

It is easy to prove that the above executable function implements clause subsumption: subsumes (mset $Ls$) (mset $Ks$) = subsumes list $Ls$ $Ks$ $\lambda x. \text{ None}$, where mset converts lists to multisets by forgetting the order of the elements. After the registration of this equation, Isabelle’s code generator will rewrite any code that contains the nonexecutable left-hand side to use the executable right-hand side instead.

Clause subsumption is a hot spot in a resolution prover [41]. Following Tammet [49], we implement a heuristic that often reduces the number of calls to $\text{match_term list}$, which is linear in the size of the input terms, by first performing a simpler, imprecise comparison. For example, terms with different root symbols will never match, and these can be compared in
A Verified Prover Based on Ordered Resolution

The lemma of the art, we compare our prover’s performance with that
of three other provers on a benchmark suite. TPTP (Thousands of Problems for Theorem Provers) [48] is the de facto standard library for benchmarking automatic provers. We extended RPx with the trusted TPTP parser from Metis [15]. We benchmarked E 2.1, Vampire 4.2.2, Metis 2.4, and RPx on 1000 randomly selected equality-free problems from the TPTP’s FOF (first-order formulas) and CNF (first-order formulas in conjunctive normal form) categories. We converted all FOF problems to CNF using E’s classifier. Each prover was run on each problem for 60 s on an Intel Core i9-7900X (3.3 GHz 10-Core) with 128 GB of RAM.

The results are summarized in the following table, showing for each prover how many unsatisfiable and satisfiable problems were solved and how many seconds were needed on average by each prover on the problems that were solved by all four:

<table>
<thead>
<tr>
<th></th>
<th>Vampire</th>
<th>E</th>
<th>Metis</th>
<th>RPx</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unsatisfiable</td>
<td>675</td>
<td>635</td>
<td>436</td>
<td>331</td>
</tr>
<tr>
<td>Satisfiable</td>
<td>158</td>
<td>135</td>
<td>91</td>
<td>22</td>
</tr>
<tr>
<td>Average time (s)</td>
<td>0.032</td>
<td>0.014</td>
<td>0.637</td>
<td>3.126</td>
</tr>
</tbody>
</table>

The detailed results of the evaluation are available online, together with instructions for reproducing them.3

As expected, RPx is not competitive. A prover’s performance comes from its calculus, its heuristics, and its indexing data structures. RPx employs an excellent calculus but mediocre heuristics and data structures. Better performance could be achieved by working on these last two aspects. Heuristics are often easy to verify, because their input–output specifications are permissive, but formalizing optimized data structures can be very laborious [10].

Nevertheless, sometimes the calculus is all that matters. Benchmark MSC015 from the TPTP library is a particularly challenging family Φn of first-order problems, each consisting of the following n + 2 clauses:

\[
\neg p(b, \ldots, b) \lor (p(a, \ldots, a)) \\
\neg p(a, b, \ldots, b) \lor (p(a, b, \ldots, a)) \\
\neg p(x_1, a, b, \ldots, b) \lor (p(x_1, b, a, \ldots, a)) \\
\vdots \\
\neg p(x_1, \ldots, x_{n-2}, a, b) \lor (p(x_1, \ldots, x_{n-2}, b, a)) \\
\neg p(x_1, \ldots, x_{n-2}, x_{n-1}, a) \lor (p(x_1, \ldots, x_{n-2}, x_{n-1}, b))
\]

A comment in the benchmark warns us that back in 2007, no prover could solve the Φ33 within an hour. Even in 2018, only one prover solves Φ33 within 300 s, and four provers solve Φ20 within 300 s. RPx solves Φ20 in 100 s and Φ22 in 200 s. Presumably, the reason for this success is that RPx fortuitously chooses an instance of the Knuth–Bendix order that is well suited to this benchmark.

3http://matroyshka.gforge.inria.fr/pubs/fun_rp_data.tar.gz

constant time. Similarly, literals with opposite polarities cannot match. We sort our (list-represented) clauses with respect to a literal quasi-order (i.e., a transitive and reflexive relation)

\[\text{leq}_\text{lit} \; L \; K \; L \; K \; \text{if} \; \text{is} \_ \text{pos} \; L \; = \; \text{is} \_ \text{pos} \; K \; \text{and} \; \text{match} \_ \text{term} \_ \text{list} \; [\; \text{atom} \_ \text{of} \; L \; , \; \text{atom} \_ \text{of} \; K \; ] \; \sigma \; = \; \text{Some} \; \rho \; \text{for} \; \text{some} \; \sigma \; \text{and} \; \rho. \; \text{Any} \; \text{quasi-order} \; \text{satisfying} \; \text{this} \; \text{property} \; \text{can be used in a refinement of \text{subsumes} \_ \text{list} \; \text{to remove} \; \text{too small} \; \text{literals} \; \text{(with respect to \text{leq} \_ \text{lit})}, \; \text{as highlighted in gray below:}

\[
\text{fun} \; \text{subsumes} \_ \text{list} \; : \; \text{term} \; \text{literal} \; \text{list} \; \Rightarrow \; \text{term} \; \text{literal} \; \text{list} \; \Rightarrow \; (\; \prime \; \Rightarrow \; \text{term} \; \text{option} \; \Rightarrow \; \text{bool}
\]

\[
\text{where}
\]

\[
\text{subsumes} \_ \text{list}' \; [\; ] \; \text{Ks} \; \sigma \; = \; \text{True}
\]

\[
| \; \text{subsumes} \_ \text{list}' \; (\; L \; \# \; \text{Ls} \; ) \; \text{Ks} \; \sigma \; =
\]

\[
\text{let} \; \text{Ks} \; = \; \text{filter} \; (\; \text{leq} \_ \text{lit} \; \text{L}) \; \text{Ks} \; \text{in}
\]

\[
(\; \exists \; \text{K} \; \in \; \text{set} \; \text{Ks} . \; \text{is} \_ \text{pos} \; \text{K} \; = \; \text{is} \_ \text{pos} \; \text{L} \; \land
\]

\[
\text{case} \; \text{match} \_ \text{term} \_ \text{list} \; (\; \text{atom} \_ \text{of} \; \text{L} \; , \; \text{atom} \_ \text{of} \; \text{K} \; ] \; \sigma \; \Rightarrow
\]

\[
\text{None} \; \Rightarrow \; \text{False}
\]

\[
| \; \text{Some} \; \rho \; \Rightarrow \; \text{subsumes} \_ \text{list}' \; \text{Ls} \; (\; \text{remove1} \; \text{K} \; \text{Ks} \; \rho \; )
\]

The lemma \text{subsumes} \_ \text{list} \; \text{Ls} \; \text{Ks} \; \rho \; = \; \text{subsumes} \_ \text{list}' \; (\; \text{sort} \; \text{leq} \_ \text{lit} \; \text{Ls} \; ) \; \text{Ks} \; \rho \; \text{allows the code generator to refine the original version. In \text{RPx}, we let \text{leq} \_ \text{lit} \; \text{be a quasi-order that} \; \text{(1) considers negative literals smaller than positive ones; (2) considers variables smaller than nonvariables; and (3) sorts atoms according to a total order on their root symbols. This refinement is a local optimization: It requires us to explicitly sort one of the input clauses. A potentially more efficient refinement would be to ensure that all clauses in the prover’s state are sorted with respect to \text{leq} \_ \text{lit}. Sorting \text{Ls} \; \text{at each invocation of subsumption could then be avoided, and filtering \text{Ks} \; \text{could be performed more efficiently. However, maintaining the invariant would require changes throughout the prover’s code.}

The End Result. Finally, Isabelle can export our prover to Standard ML, Haskell, OCaml, or Scala. The command

\[
\text{export} \_ \text{code} \; \text{RPx} \; \in \; \text{SML} \; \text{module} \_ \text{name} \; \text{RP}
\]

generates an ML module containing the implementation of our prover in about 1000 lines of code, including dependencies. The generated module exports the function \text{RPx} \; : \; (\; \text{term} \; \text{literal} \_ \text{list} \; \ast \; \text{nat} \; ) \; \text{list} \; \Rightarrow \; \text{bool}. The input is the \text{N} component of an initial state, which consists of pairs of clauses and arbitrary timestamps (e.g., 0).

Even though in Isabelle we have proved that for any unsatisfiable input \text{RPx}, will terminate and return False, the code generator guarantees only partial correctness of its output: If the ML program terminates on the ML input generated from the Isabelle term \text{t} \; \text{and evaluates to the Boolean result} \; \text{b}, the proposition \text{RPx} \; \text{t} \; = \; \text{b} \; \text{is provable in Isabelle; by soundness,} \; \text{b} \; \text{indicates the satisfiability of the input clause set. There is recent work towards providing stronger guarantees and}

Empirical Evaluation. To measure the gap with the state of the art, we compare our prover’s performance with that

\url{http://matroyshka.gforge.inria.fr/pubs/fun_rp_data.tar.gz}
7 Discussion and Related Work

We found Bachmair and Ganzinger’s [2] chapter and its formalization [39, 40] suitable as a starting point for a verified prover. Nonetheless, we faced some difficulties, notably concerning the identification of suitable refinement layers. We developed layers 2, 3, and 4 largely in parallel, with each of the authors working on a separate layer. Bringing layer 2 into a state such that it both ensures fairness and could be refined further by layer 3 required several iterations.

Stepwise refinement helped us achieve separation of concerns: fairness, determinism, and executability were achieved successively. Another strength of this methodology is that it allows us to prove results at a high level of abstraction; for example, fairness is established at layer 2 already and is inherited by subsequent layers. The main difficulty with refinement is that some nontrivial machinery is necessary to lift results from one layer to the next. We believe the gain in modularity makes this worthwhile.

It took us quite some time to design a suitable measure to prove the fairness of the layer 2 prover RPₚ. Our solution amounts to advancing to a state carrying a suitably high timestamp and filtering out all overly heavy clauses. Initially, our proof consisted of two steps—advancing and filtering—each with its own measure. This proof gave us the assurance that RPₚ was fair, but we found that combining the measures is both more succinct and more intelligible.

Our main objective was not to reach qed as quickly as possible but rather to investigate how to harness a modern proof assistant to formalize the metatheory of automatic theorem provers. We found Isabelle suitable for this verification task. The Isar proof language allows us to state key intermediate steps, as in a paper proof. Standard tactics, including Isabelle’s simplifier, can be used to discharge proof obligations. The Sledgehammer tool [32] employs superposition provers and SMT (satisfiability modulo theories) solvers to swiftly identify which lemmas can be used to prove a goal; standard Isabelle tactics are then used to reconstruct the proof. Isabelle’s support for coinductive methods, including the coinductive, codatatype, and corec commands, helps us reason about infinite processes. Locales are a useful abstraction for defining the refinement layers. And Isabelle’s libraries, the Archive of Formal Proofs, and IsaFoR certainly saved us months of labor.

The Archive also includes a refinement framework [25], which has been used in a separate effort to connect the imperative code of an efficient SAT solver to an abstract calculus [6]. The framework is helpful in a variety of situations, including when the refinement relation between a concrete and an abstract data representation is not a function. But since converting a list to a multiset (between our levels 3 and 2) or a multiset to a set (between levels 2 and 1) is a function, we did not see a need to employ it. Moreover, the framework is currently not designed for refining semidecision procedures, as acknowledged privately by its developer. We conjecture that its support for separation logic could be useful if we were to refine the prover further to obtain imperative code.

Thanks to the verification, we can trust to a very high extent that our ordered resolution prover is sound and complete. To make the prover’s performance competitive with E, SPASS, and Vampire, we would need to extend the current work along two axes. First, we should use superposition, together with its extensive simplification machinery, as the base calculus. A good starting point would be to apply our methodology to Peltier’s [33] formalization of superposition. Given that a large part of a modern superposition prover’s code consists of heuristics, which are easy to verify, the full verification of a competitive superposition prover appears to be a realistic objective for a forthcoming Ph.D. thesis. Second, the refinement chain should be continued to cover optimized algorithms and data structures. These could be specified by refining layer 4 further, along the lines of Fleury et al.’s [10] refinement of an imperative SAT solver.

In computer science, a metatheory may inspire an implementation, or vice versa, but the connection is seldom made explicit. By formalizing the metatheory, the implementation, and their connection, we can demonstrate not only the implementation’s correctness but also the metatheory’s adequacy for describing potential implementations. In particular, we have now confirmed that Bachmair and Ganzinger [2] accurately describe the abstract principles of an executable functional prover (with a few exceptions [40]), even though they provide few details beyond layer 1.

We built the prover on our earlier formalization [39, 40] of first-order resolution. Related efforts developed using Isabelle/HOL include Peltier’s [33] formalization of superposition and Schlichtkrull’s [37] formalization of unordered resolution. These developments cover only logical calculi; had we started with any of them, the first step would have been to define an abstract prover in the style of layer 1 and prove basic properties about it. Another related effort is Hirokawa et al.’s [13] formalization of ordered completion, which (like ordered resolution) can be regarded as a special case of superposition.

Formalizing a theorem proving tool using a theorem proving tool is a thrilling (if self-referential) prospect for many researchers. An early result is Ridge and Margarson’s [35] verified first-order prover, based on a sequent calculus for first-order logic without full first-order terms but only variables. Kumar et al. [24] formalized the soundness of a proof assistant for higher-order logic. Jensen et al. [16] verified the soundness of a kernel for a proof assistant for first-order logic that includes a tableau prover. There are several verified SAT solvers [6, 27–29, 31, 44]. SAT being a decidable problem, termination has been proved for most solvers. First-order logic, on the other hand, is semidecidable.

A pragmatic approach to combining the efficiency of unverified code with the trustworthiness of verified code involves checking certificates produced by reasoning tools—
e.g., proofs produced by SAT solvers [9, 26]. Researchers from the first-order theorem proving community are now advocating this approach for their systems [34]. An ad hoc version of this approach is used in Sledgehammer and HOLy-Hammer to reconstruct proofs found by external automatic provers [4, 17].

8 Conclusion

Starting from an abstract description of an ordered resolution prover [39, 40], we verified, through a refinement chain, a purely functional prover that uses lists as its main data structure. The resulting program is interesting in its own right and could be refined further to obtain an implementation that is competitive with the state of the art.

Stepwise refinement is a keystone of our methodology, and we found it adequate. Each refinement step cleanly isolates concerns, yielding intelligible proof obligations. Refinement also helped us identify an unnecessary assumption in Bachmair and Ganzinger’s [2] chapter and clarify the argument. Lifting results from one layer to another required some thought, especially the completeness results, which correspond to liveness properties.

Having now established a methodology and built basic formal libraries, we expect that verifying other provers, using Isabelle or other systems, will be substantially easier. Because it is based on Bachmair and Ganzinger’s framework, our approach generally applies to all saturation-based provers, with or without redundancy. This includes resolution, paramodulation, ordered rewriting, superposition, and variants thereof, covering many of the most successful provers for equational [8, 12], first-order [21, 42, 53], and higher-order logic [45].

Acknowledgments

Johannes Hölzl gave us some useful advice on how to specify and reason about possibly nonterminating functions in Isabelle/HOL. Alexander Bentkamp, Mathias Fleury, Andreas Halkjær From, Carsten Fuhs, Peter Lammich, Mark Summerfield, Jørgen Villadsen, and the anonymous reviewers suggested many textual improvements. Eugene Kotelnikov, Stephan Schulz, and Geoff Sutcliffe graciously answered our questions about Vampire, E, and TPTP.

Schlichtkrull has received funding from a Ph.D. scholarship in the Algoritms, Logic and Graphs section of DTU Compute and from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program (grant agreement No. 700321, LIGHTTest). Blanchette has received funding from the ERC under the European Union’s Horizon 2020 research and innovation program (grant agreement No. 713999, Matryoshka). An earlier version of this paper was included as a chapter of Schlichtkrull’s Ph.D. thesis [36] with the same authors’ list.

References
