Introducing Quantified Cuts in Logic with Equality

Stefan Hetzl[1], Alexander Leitsch[2], Giselle Reis[2], Janos Tapolczai[1], and Daniel Weller[1]

[1] Institut für Diskrete Mathematik und Geometrie, Technische Universität Wien
[2] Institut für Computersprachen, Technische Universität Wien

Abstract. Cut-introduction is a technique for structuring and compressing formal proofs. In this paper we generalize our cut-introduction method for the introduction of quantified lemmas of the form $\forall x.A$ (for quantifier-free A) to a method generating lemmas of the form $\forall x_1 \ldots \forall x_n.A$. Moreover, we extend the original method to predicate logic with equality. The new method was implemented and applied to the TSTP proof database. It is shown that the extension of the method to handle equality and quantifier-blocks leads to a substantial improvement of the old algorithm.

1 Introduction

Computer-generated proofs are typically analytic, i.e., they only contain logical material that also appears in the statement of the theorem. This is due to the fact that analytic proof systems have a considerably smaller search space which makes proof-search practically feasible. In the case of sequent calculus, proof-search procedures typically work on the cut-free fragment. But also resolution is essentially analytic as resolution proofs satisfy the subformula property of first-order logic. One interesting property of non-analytic proofs is their considerably smaller length. The exact difference depends on the logic (or theory) under consideration, but it is typically enormous. In (classical and intuitionistic) first-order logic there are proofs with cut of length n whose theorems have only cut-free proofs of length 2_n (where $2_0 = 1$ and $2_{n+1} = 2^{2_n}$) (see [16] and [12]). The length of a proof plays an important role in many situations such as human readability, space requirements and time requirements for proof checking. For most of these situations general-purpose data compression methods cannot be used as the compressed representation is not a proof in a standard calculus anymore and hence does not allow fast processing, e.g. linear time proof checking. It is therefore of high practical interest to develop methods of proof transformation which produce non-analytic and hence potentially much shorter proofs.

Work on cut-introduction can be found at a number of different places in the literature. Closest to our work are other approaches which aim to abbreviate or structure a given input proof. In [20] an algorithm for the introduction of atomic cuts that is capable of exponential proof compression is presented.

S. Demri, D. Kapur, and C. Weidenbach (Eds.): IJCAR 2014, LNAI 8562, pp. 240–254, 2014.

The method [5] for propositional logic is shown to never increase the size of proofs more than polynomially. Another approach to the compression of first-order proofs by introduction of definitions for abbreviating terms is [19]. There is a large body of work on the generation of non-analytic formulas carried out by numerous researchers in various communities. Methods for lemma generation are of crucial importance in inductive theorem proving which frequently requires generalization [1], see e.g. [10] for a method in the context of rippling [2] which is based on failed proof attempts. In automated theory formation [3,4], an eager approach to lemma generation is adopted. This work has, for example, led to automated classification results of isomorphism classes [14] and isotopy classes [15] in finite algebra. See also [11] for an approach to inductive theory formation.

Our method of *algorithmic cut-introduction*, based on the inversion of Gentzen's cut-elimination method, has been defined in [8] and [7]. The method in [8] works on a cut-free **LK**-proof φ of a prenex skolemized end-sequent S and consists of the following steps: (1) extraction of a set of terms T from φ, (2) computation of a compressed representation of T, (3) construction of the cut formula, (4) improvement of the solution by computation of smaller cut-formulas, and (5) construction of an **LK**-proof with the universal cut formula obtained in (4) and instantiation of the quantifiers with the terms obtained in (2). It has been shown in [8] that the method is capable of compressing cut-free proofs quadratically. The paper [7] generalized the method to the introduction of arbitrarily many universal cut formulas, where the steps defined above are roughly the same, though the improvement of the solution (step 4) and the final construction of the proof with cuts (step 5) are much more difficult. The method of introducing arbitrarily many universal cuts in [7] leads even to an exponential compression of proof length. Still the methods described above were mainly designed for a theoretical analysis of the cut-introduction problem rather than for practical applications. In particular, they lacked efficient handling of equality (as they were defined for predicate logic without equality) and the introduction of several universal quantifiers in cut formulas (all cut formulas constructed in [7] are of the form $\forall x.A$ for a single variable x and a quantifier-free formula A).

In this paper we generalize our cut-introduction method to predicate logic with equality and to the construction of a (single) quantified cut containing blocks of universal quantifiers. The efficient compression of the terms (step 2) and the improvement of the solution (step 4) require new and non-trivial techniques. Moreover, we applied the new method in large-scale experiments to proofs generated by prover9 on the TPTP library. This empirical evaluation demonstrates the feasibility of our method on realistic examples.

2 Proofs and Herbrand Sequents

Throughout this paper we consider predicate logic with equality. For practical reasons equality will not be axiomatized but handled via substitution rules. We extend the sequent calculus **LK** to the calculus $\mathbf{LK}_{=}$ by allowing sequents of the form $\rightarrow t = t$ as initial sequents and adding the following rules:

$$\frac{\Gamma \to \Delta, s = t \quad A[s], \Pi \to \Lambda}{A[t], \Gamma, \Pi \to \Delta, \Lambda}\ \text{El1} \qquad \frac{\Gamma \to \Delta, t = s \quad A[s], \Pi \to \Lambda}{A[t], \Gamma, \Pi \to \Delta, \Lambda}\ \text{El2}$$

$$\frac{\Gamma \to \Delta, s = t \quad \Pi \to A[s], \Lambda}{\Gamma, \Pi \to A[t], \Delta, \Lambda}\ \text{Er1} \qquad \frac{\Gamma \to \Delta, t = s \quad \Pi \to A[s], \Lambda}{\Gamma, \Pi \to A[t], \Delta, \Lambda}\ \text{Er2}$$

$\mathbf{LK}_=$ is sound and complete for predicate logic with equality.

For convenience we write a substitution $[x_1 \backslash t_1, \ldots, x_n \backslash t_n]$ in the form $[\bar{x} \backslash \bar{t}]$ for $\bar{x} = (x_1, \ldots, x_n)$ and $\bar{t} = (t_1, \ldots, t_n)$. A *strong quantifier* is a $\forall$ ($\exists$) quantifier with positive (negative) polarity. We restrict our investigations to end-sequents in prenex form without strong quantifiers.

Definition 1. *A* Σ_1-sequent *is a sequent of the form*

$$\forall x_1 \cdots \forall x_{k_1} F_1, \ldots, \forall x_1 \cdots \forall x_{k_p} F_p \to \exists x_1 \cdots \exists x_{k_{p+1}} F_{p+1}, \ldots, \exists x_1 \cdots \exists x_{k_q} F_q.$$

for quantifier free F_i.

Note that the restriction to Σ_1-sequents does not constitute a substantial restriction as one can transform every sequent into a validity-equivalent Σ_1-sequent by skolemisation and prenexing.

Definition 2. *A sequent* S *is called* E-valid *if it is valid in predicate logic with equality;* S *is called a* quasi-tautology *[13] if* S *is quantifier-free and E-valid.*

Definition 3. *The length of a proof* φ*, denoted by* $|\varphi|$*, is defined as the number of inferences in* φ*. The quantifier-complexity of* φ*, written as* $|\varphi|_q$*, is the number of weak quantifier-block introductions in* φ*.*

2.1 Extraction of Terms

Herbrand sequents of a sequent S are sequents consisting of instantiations of S which are quasi-tautologies. The formal definition is:

Definition 4. *Let* S *be a* Σ_1*-sequent as in Definition 1 and let* H_i *be a finite set of* k_i*-vectors of terms for every* $i \in \{1, \ldots, q\}$*. We define* $\mathcal{F}_i = \{F_i[\bar{x}_i \backslash \bar{t}] \mid \bar{t} \in H_i\}$ *if* $H_i \neq \emptyset$ *and* $\mathcal{F}_i = \{F_i\}$ *if* $H_i = \emptyset$*. Let*

$$S^* : \quad \mathcal{F}_1 \cup \cdots \cup \mathcal{F}_p \to \mathcal{F}_{p+1} \cup \ldots \cup \mathcal{F}_q.$$

If S^* *is a quasi-tautology then it is called a* Herbrand sequent *of* S *and* $H\colon (H_1, \ldots, H_q)$ *is called a* Herbrand structure *of* S*. We define the* instantiation complexity *of* S^* *as* $|S^*|_i = \sum_{i=1}^{q} |H_i|$.

Note that, in the instantiation complexity of a Herbrand sequent, we only count the formulas obtained by instantiation.

Example 1. Consider the language containing a constant symbol a, unary function symbols f, s, a binary predicate symbol P, and the sequent S defined below. We write f^n, s^n for n-fold iterations of f and s and omit parentheses around the argument of a unary symbol when convenient. Let

$$S : P(f^4a, a), \forall x.fx = s^2x, \forall xy(P(sx, y) \supset P(x, sy)) \to P(a, f^4a)$$

and $H = (H_1, H_2, H_3, H_4)$ for

$$H_1 = \emptyset,\ H_4 = \emptyset,\ H_2 = \{a, fa, f^2a, f^3a\},$$
$$H_3 = \{(s^3f^2a, a), (s^2f^2a, sa), (sf^2a, s^2a), (f^2a, s^3a), (s^3a, f^2a), (s^2a, sf^2a), (sa, s^2f^2a), (a, s^3f^2a)\}.$$

Then

$$\mathcal{F}_1 = \{P(f^4a, a)\},\ \mathcal{F}_4 = \{P(a, f^4a)\},$$
$$\mathcal{F}_2 = \{fa = s^2a,\ f^2a = s^2fa,\ f^3a = s^2f^2a,\ f^4a = s^2f^3a\}$$
$$\begin{aligned}\mathcal{F}_3 = \{&P(s^4f^2a, a) \supset P(s^3f^2a, sa), P(s^3f^2a, sa) \supset P(s^2f^2a, s^2a),\\ &P(s^2f^2a, s^2a) \supset P(sf^2a, s^3a), P(sf^2a, s^3a) \supset P(f^2a, s^4a),\\ &P(s^4a, f^2a) \supset P(s^3a, sf^2a), P(s^3a, sf^2a) \supset P(s^2a, s^2f^2a),\\ &P(s^2a, s^2f^2a) \supset P(sa, s^3f^2a), P(sa, s^3f^2a) \supset P(a, s^4f^2a)\}.\end{aligned}$$

A Herbrand-sequent S^* corresponding to H is then $\mathcal{F}_1 \cup \mathcal{F}_2 \cup \mathcal{F}_3 \to \mathcal{F}_4$. Note that $fa = s^2a,\ f^2a = s^2fa,\ f^3a = s^2f^2a,\ f^4a = s^2f^3a \models f^4a = s^8a$.
The instantiation complexity of S^* is 12. S^* is a quasi-tautology but not a tautology.

Theorem 1 (mid-sequent theorem). *Let S be a Σ_1-sequent and π a cut-free proof of S. Then there is a Herbrand-sequent S^* of S s.t. $|S^*|_i \leq |\pi|_q$.*

Proof. This result is proven in [6] (section IV, theorem 2.1) for **LK**, but the proof for $\mathbf{LK}_=$ is basically the same. By permuting the inference rules, one obtains a proof π' from π which has an upper part containing only propositional inferences and the equality rules (which can be shifted upwards until they are applied to atoms only) and a lower part containing only quantifier inferences. The sequent between these parts is called *mid-sequent* and has the desired properties.

S^* can be obtained by tracing the introduction of quantifier-blocks in the proof, which for every formula $Q\bar{x}_i.F_i$ in the sequent (where $Q \in \{\forall, \exists\}$) yields a set of term tuples H_i, and then computing the sets of formulas $\mathcal{F}_i$.

The algorithm for introducing cuts described here relies on computing a compressed representation of the Herbrand structure, which is explained in Section 3. Note, though, that the Herbrand structure $(H_1, \ldots, H_q)$ is a list of sets of term tuples (i.e. each H_i is a set of tuples $\overline{t}$ used to instantiate the formula F_i). In order to facilitate computation and representation, we will add to the language fresh function symbols $f_1, \ldots, f_q$. Each f_i will be applied to the tuples of the set H_i, therefore transforming a list of sets of tuples into a set of terms. In this new set, each term will have an f_k as its head symbol, that indicates to which formula the arguments of f_k belong.

Example 2. Using this new notation, the Herbrand structure H of the previous example is now represented as the set of terms:

$$T : \{f_2(a), f_2(fa), f_2(f^2a), f_2(f^3a), f_3(s^3f^2a, a), f_3(s^2f^2a, sa), \ldots, f_3(a, s^3f^2a)\}.$$

Henceforth we will refer to the transformed Herbrand structure as the *term set* of a proof.

3 Computing a Decomposition

We shall now describe an algorithm for computing a compressed representation of a term set T. Term sets will be represented by decompositions which are defined as follows:

Definition 5. *Let $T = \{t_1, \ldots, t_n\}$ be a set of ground terms. A decomposition D of T is a pair, written as $U \circ_{\bar{\alpha}} W$, where U is a set of terms containing the variables $\alpha_1, \ldots, \alpha_m$, and $W = \left\{ \bar{w}_1 = \begin{pmatrix} w_{1,1} \\ \vdots \\ w_{1,m} \end{pmatrix}, \ldots, \bar{w}_q = \begin{pmatrix} w_{q,1} \\ \vdots \\ w_{q,m} \end{pmatrix} \right\}$ is a set of vectors of ground terms s.t. $T = U \circ_{\bar{\alpha}} W = \{u[\bar{\alpha} \backslash \bar{w}] \mid u \in U, \bar{w} \in W\}$. The size of a decomposition $U \circ_{\bar{\alpha}} W$ is $|U| + |W|$. When it is clear that the variables in question are $\alpha_1, \ldots, \alpha_m$, we just write $U \circ W$.*

In [7], we gave an algorithm that treated the special case where $m = 1$. Here, we will extend that approach with a *generalized Δ-vector* Δ_G, which, together with a so-called Δ-table, can compute decompositions with an arbitrary m. Δ_G, given in Algorithm 1, computes a *simple decomposition*, i.e. a decomposition with only one term in U. The Δ-table stores such decompositions and builds more complex ones out of them. Due to space reasons, the algorithm can only be sketched here, for details the interested reader is referred to the technical report [18].

Definition 6. *Let T be a set of ground terms. The Δ-table for T is a set of key/value-entries, where each entry is of the form $W \Rightarrow U'$, where U' is a set $\{(u_1, T_1), \ldots, (u_q, T_q)\}$, W is a set of ground term vectors, u_i is a term containing variables, q is some a priori unknown bound different for each line, and T_i is a subset of T s.t. the following two conditions are satisfied:*

1. *For every entry $W \Rightarrow \{(u_1, T_1), \ldots, (u_q, T_q)\}$, $\{u_i\} \circ W$ is a decomposition of T_i (for $1 \leq i \leq q$).*
2. *For every $T' \subseteq T$, there is a pair $W \Rightarrow U'$ in the Δ-table s.t. $(u, T') \in U'$.*

In practice, condition 2 is relaxed in order to improve performance. Whenever a subset T' of T has only a *trivial* decomposition, i.e. $\Delta_G(T') = (\alpha_i, T')$, its information is not added to the Δ-table. Moreover, no superset of T' is considered from this point on, since we know that these will also have trivial decompositions.

Algorithm 1. Generalized Δ-vector Δ_G

function $\Delta_G(t_1, \ldots, t_n$: a list of terms)
  return transposeW(Δ_G'$(t_1, \ldots, t_n)$)
end function
function Δ_G'$(t_1, \ldots, t_n$: a list of terms)
  if $t_1 = t_2 = \ldots = t_n \wedge n > 0$ **then** ▷ case 1: all terms identical
    return $(t_1, ())$
  else if $t_i = f(t_1^i, \ldots, t_m^i)$ for $1 \leq i \leq n$ **then** ▷ case 2: recurse
    $(\bar{w}_1, \ldots, \bar{w}_q) \leftarrow \bigsqcup_{1 \leq j \leq m} \pi_2(\Delta_G(t_j^1, \ldots, t_j^n))$ ▷ $\bigsqcup \equiv$ concatenation
    $u_j \leftarrow \pi_1(\Delta_G(t_j^1, \ldots, t_j^n))$ for all $j \in \{1, \ldots, m\}$
    return merge$(f(u_1, \ldots, u_m), (\bar{w}_1, \ldots, \bar{w}_q))$ ▷ merge all α_i, α_j where $\bar{w}_i = \bar{w}_j$
  else ▷ case 3: introduce new α
    return $(\alpha_{\text{FRESH}}, (t_1, \ldots, t_n))$
  end if
end function

Note that this implies that $\Delta_G(T)$ is almost never computed. An actual decomposition for T is found by iterating over the Δ-table and, for each entry, trying to find a set of pairs $\{(u_{i_1}, T_{i_1}), \ldots, (u_{i_q}, T_{i_q})\} \subseteq U'$ such that $\{u_{i_1}, \ldots, u_{i_n}\} \circ W$ generates T.

Theorem 2 (Soundness). *Let T be a term set. If $U \circ W$ is extracted from iterating over the Δ-table, then $U \circ W$ is a decomposition of T.*

Proof. See [18].

In fact, a stronger result holds: for every decomposition, there exists a unique normal form, and iterating over the Δ-table will only return decompositions in such a normal form. For details, see [18]. To illustrate the algorithm, we compute a decomposition of the term set of Example 2. We remark that, for our cut-introduction method, we are interested in a *decomposition* $(U_1, \ldots, U_q) \circ W$ *of a Herbrand structure* $H = (H_1, \ldots, H_q)$ which has the property that $H_j = \{u[\bar{\alpha} \backslash \bar{w}] \mid u \in U_j, \bar{w} \in W\}$. This is trivially obtained from a decomposition $U \circ W$ of the term set of a Herbrand structure by setting $U_j = \{u \mid f_j(u) \in U\}$.

Example 3. Let $T = T_2 \cup T_3$ with

$$T_2 = \{t_1 : f_2(a),\ \ t_2 : f_2(fa),\ \ t_3 : f_2(f^2a),\ \ t_4 : f_2(f^3a)\}$$
$$T_3 = \{t_5 : f_3(s^3f^2a, a),\ \ t_6 : f_3(s^2f^2a, sa),\ \ \ldots,\ \ t_{12} : f_3(a, s^3f^2a)\}$$

be a term set corresponding to the Herbrand structure $H = (H_1, H_2, H_3, H_4)$:

$$H_1 = \emptyset,\ H_4 = \emptyset, H_2 = \{a, fa, f^2a, f^3a\},$$
$$H_3 = \{(s^3f^2a, a), (s^2f^2a, sa), \ldots, (a, s^3f^2a)\}$$

We now compute Δ_G for every subset of T — consider for instance the subset $T' = \{f_3(s^3f^2a, a), f_3(s^2f^2a, sa)\} \subseteq T$:

$$\Delta_G(f_3(s^3f^2a, a), f_3(s^2f^2a, sa)) = (f_3(s^2\alpha_1, \alpha_2), \left\{ \begin{pmatrix} sf^2a \\ a \end{pmatrix}, \begin{pmatrix} f^2a \\ sa \end{pmatrix} \right\}) = (u, W).$$

If the Δ-table already has an entry $W \Rightarrow U'$, we add (u, T') to U'. If not, we insert a new entry $W \Rightarrow \{(u, T')\}$. After Δ_G has been computed for all subsets, we iterate through it, looking for simple decompositions that can be composed into a decomposition of T. We find the entry

$$\begin{aligned}
W \Rightarrow\ & U'_1 \cup U'_2 \cup U'_3 \cup U'_4 \\
U'_1 =\ & \{\} \\
U'_2 =\ & \{(f_2(\alpha_1), \{t_1, t_3\}),\ (f_2(f\alpha_1), \{t_2, t_4\}),\ (f_2(\alpha_2), \{t_1, t_3\})\ (f_2(f\alpha_2), \{t_2, t_4\})\} \\
U'_3 =\ & \{(f_3(s^3\alpha_1, \alpha_2), \{t_5, t_9\}),\ (f_3(s^2\alpha_1, s\alpha_2), \{t_6, t_{10}\}), \\
& \quad (f_3(s\alpha_1, s^2\alpha_2), \{t_7, t_{11}\}),\ (f_3(\alpha_1, s^3\alpha_2), \{t_8, t_{12}\})\} \\
U'_4 =\ & \{\} \\
W =\ & \left\{ \begin{pmatrix} f^2a \\ a \end{pmatrix}, \begin{pmatrix} a \\ f^2a \end{pmatrix} \right\}
\end{aligned}$$

and can see that $U'_i \circ W = T_i$ for $1 \leq i \leq 4$. Therefore, $(U'_1 \cup U'_2 \cup U'_3 \cup U'_4) \circ W$ is a decomposition of T. We then translate this decomposition T back into a decomposition of H by removing the function symbols f_2 and f_3 from U'_2 & U'_3 (the empty sets U'_1 and U'_4 can be disregarded):

$$\begin{aligned}
U =\ & (U_2, U_3), \\
U_2 =\ & \{\alpha_1,\ f\alpha_1,\ \alpha_2,\ f\alpha_2\}, \\
U_3 =\ & \{(s^3\alpha_1, \alpha_2),\ (s^2\alpha_1, s\alpha_2),\ (s\alpha_1, s^2\alpha_2),\ (\alpha_1, s^3\alpha_2)\}, \\
W =\ & \left\{ \begin{pmatrix} f^2a \\ a \end{pmatrix}, \begin{pmatrix} a \\ f^2a \end{pmatrix} \right\}.
\end{aligned}$$

4 Computing a Cut-Formula

After having computed a decomposition as described in Section 3, the next step consists in computing a cut-formula based on that decomposition. A decomposition D specifies the instances of quantifier blocks in a proof with a $\forall$-cut, but does not contain information about the propositional structure of the cut formula to be constructed. The problem to find the appropriate propositional structure is reflected in the following definition.

Definition 7. *Let S be a Σ_1-sequent and F_i, k_i as in Definition 1, H be a Herbrand structure for S, and $D{:}\, U \circ W$ a decomposition of H with $V(D) = \{\alpha_1, \ldots, \alpha_n\}$. Let $U = (U_1, \ldots, U_q)$ and $W = \{\bar{w}_1, \ldots, \bar{w}_k\}$, where the $\bar{w}_j$ are n-vectors of terms not containing variables in $V(D)$, and $\mathcal{F}'_i = \{F_i[\bar{x}_i \backslash \bar{t}] \mid \bar{t} \in U_i\}$ for $U_i \neq \emptyset$ and $\mathcal{F}'_i = \{F_i\}$ for $U_i = \emptyset$. Furthermore let X be an n-place predicate variable. Then the sequent*

$$S^{\sim}:\quad X\bar{\alpha} \supset \bigwedge\nolimits_{i=1}^{k} X\bar{w}_i, \mathcal{F}'_1, \ldots, \mathcal{F}'_p \to \mathcal{F}'_{p+1}, \ldots, \mathcal{F}'_q.$$

is called a schematic extended Herbrand sequent *of S w.r.t. D. The* instantiation complexity *of $S^{\sim}$, denoted by $|S^{\sim}|_i$, is defined as $k + \sum_{i=1}^{q} |U_i|$.*

Definition 8. *Let $S^\sim$ be a schematic extended Herbrand sequent of S w.r.t. a decomposition D as in Definition 7 and A be a formula with $V(A) \subseteq \{\alpha_1, \ldots, \alpha_n\}$. Then the second-order substitution $\sigma{:}[X\backslash\lambda\bar{\alpha}.A]$ is a solution of $S^\sim$ if $S^\sim\sigma$ is a quasi-tautology; in this case $S^\sim\sigma$ is called an* extended Herbrand sequent. *The instantiation complexity of $S^\sim\sigma$ is defined as $|S^\sim|_i$.*

Theorem 5 in Section 4.2 shows that, from a solution of a schematic extended Herbrand sequent $S^\sim$ of S, we can define a proof ψ of S with a $\forall$-cut and $|\psi|_q = |S^\sim|_i$. The question remains whether every schematic extended Herbrand sequent is solvable. We show below that this is indeed the case.

Let $S^\sim$ as in Definition 7. We define

$$F[l] = \bigwedge \bigcup_{i=1}^{p} \mathcal{F}'_i \text{ and } F[r] = \bigvee \bigcup_{i=p+1}^{q} \mathcal{F}'_i.$$

Definition 9. *Let $S^\sim$ be a schematic extended Herbrand sequent of S as in Definition 7. We define the* canonical formula $C(S^\sim)$ *of $S^\sim$ as $F[l] \wedge \neg F[r]$. The substitution $[X\backslash\lambda\bar{\alpha}.C(S^\sim)]$ is called the* canonical substitution *of $(S, S^\sim)$.*

Theorem 3. *Let S be a Σ_1-sequent, and $S^\sim$ be a schematic extended Herbrand sequent of S. Then the canonical substitution is a solution of $S^\sim$.*

Proof. Let $S^\sim$ be a schematic extended Herbrand sequent as in Definition 7 and $C(S^\sim)$ be the canonical formula of $S^\sim$. We have to prove that

$$S_1 : \;\; C(S^\sim)(\bar{\alpha}) \supset \bigwedge_{i=1}^{k} C(S^\sim)(\bar{w}_i),\;\; F[l] \to F[r]$$

is a quasi-tautology. But, by definition of $C(S^\sim)$, S_1 is equivalent to

$$S_2 : \;\; (F[l] \wedge \neg F[r]) \supset \bigwedge_{i=1}^{k} (F[l] \wedge \neg F[r])(\bar{w}_i), (F[l] \wedge \neg F[r]) \to .$$

Clearly S_2 is a quasi-tautology if the sequent S_3, defined as

$$S_3 : \;\; \bigwedge_{i=1}^{k} (F[l] \wedge \neg F[r])(\bar{w}_i) \to$$

is a quasi-tautology. But, by $D = U \circ W$ being a decomposition of H, S_3 is logically equivalent to the Herbrand sequent S^* defined over H, which (by definition) is a quasi-tautology.

Example 4. Let

$$S : \;\; P(f^4a, a), \forall x.fx = s^2x, \forall xy(P(sx, y) \supset P(x, sy)) \to P(a, f^4a)$$

like in Example 1 and D be the decomposition $U \circ W$ of H constructed in Example 3. We have

$$\begin{aligned} U &= (U_2, U_3), \\ U_2 &= \{\alpha_1,\ f\alpha_1,\ \alpha_2,\ f\alpha_2\}, \\ U_3 &= \{(s^3\alpha_1, \alpha_2),\ (s^2\alpha_1, s\alpha_2),\ (s\alpha_1, s^2\alpha_2),\ (\alpha_1, s^3\alpha_2)\}, \\ W &= \left\{ \begin{pmatrix} f^2a \\ a \end{pmatrix}, \begin{pmatrix} a \\ f^2a \end{pmatrix} \right\}. \end{aligned}$$

The corresponding schematic extended Herbrand sequent $S^\sim$ is

$$X(\alpha_1,\alpha_2) \supset (X(f^2a,a) \wedge X(a,f^2a)),$$
$$f\alpha_1 = s^2\alpha_1,\ f^2\alpha_1 = s^2f\alpha_1,\ f\alpha_2 = s^2\alpha_2,\ f^2\alpha_2 = s^2f\alpha_2,$$
$$P(s^4\alpha_1,\alpha_2) \supset P(s^3\alpha_1,s\alpha_2),\ P(s^3\alpha_1,s\alpha_2) \supset P(s^2\alpha_1,s^2\alpha_2),$$
$$P(s^2\alpha_1,s^2\alpha_2) \supset P(s\alpha_1,s^3\alpha_2), P(s\alpha_1,s^3\alpha_2) \supset P(\alpha_1,s^4\alpha_2), P(f^4a,a) \to P(a,f^4a).$$

Its canonical formula $C(S^\sim)$ which we write as $A(\alpha_1,\alpha_2)$ is

$$\textstyle\bigwedge_{i=1}^{2}(f\alpha_i = s^2\alpha_i \wedge\ f^2\alpha_i = s^2f\alpha_i) \wedge$$
$$\textstyle\bigwedge_{i=0}^{3}(P(s^{4-i}\alpha_1,s^i\alpha_2) \supset P(s^{4-i-1}\alpha_1,s^{i+1}\alpha_2)) \wedge P(f^4a,a) \wedge \neg P(a,f^4a).$$

The canonical solution is $[X\backslash\lambda\alpha_1\alpha_2.A(\alpha_1,\alpha_2)]$ and the corresponding extended Herbrand sequent S' is like $S^\sim$ with $X(\alpha_1,\alpha_2) \supset (X(f^2a,a) \wedge X(a,f^2a))$ replaced by $A(\alpha_1,\alpha_2) \supset (A(f^2a,a) \wedge A(a,f^2a))$. Note that $|S'|_i = 10$, while $|S^*|_i = 12$. So we obtained a compression of quantifier complexity.

4.1 Improving the Solution

In the last section, we have shown that, given a decomposition D of the termset of a cut-free proof of a Σ_1-sequent S, there exists a canonical solution to the schematic extended Herbrand sequent induced by S, D, which gives rise to a proof with a $\forall$-cut. Furthermore, as we will show in Theorem 5, all solutions are equivalent from the point of view of the $|\cdot|_q$ measure. On the other hand, the quality of solutions can be distinguished by other properties, for example by the length of the proof with cut they induce. Since the best known general upper bound on the length of proofs in propositional logic (which corresponds to our setting once a decomposition is fixed) depends on the size of the theorem to be proven, our approach is to search for solutions that have *smaller size than the canonical solution.*

We will consider E-validity of quantifier-free formulas F containing free variables; by "F is E-valid" we mean to say "the universal closure of F is E-valid". Throughout this section, we consider a fixed Σ_1-sequent S using the notation of Definition 1, a fixed decomposition $D = (U_1,\ldots,U_q) \circ W$, with $W = \{\bar{w}_i \mid 1 \leq i \leq k\}$, of a Herbrand structure H of S, along with the schematic extended Herbrand sequent $S^\sim$ induced by S, H, D, using the notation of Definition 7. We will abbreviate $\mathcal{F}'_1 \cup \cdots \cup \mathcal{F}'_p$ by Γ and $\mathcal{F}'_{p+1} \cup \cdots \cup \mathcal{F}'_q$ by Δ, and write "A is a solution" for "$[X\backslash\lambda\bar{x}.A]$ is a solution for $S^\sim$" (note that we will consider the names $\bar{x}$ fixed). In this section, we will focus our attention on solutions in conjunctive normal form (CNF), which always exist since the solution property is semantic (if A is a solution and $A \Leftrightarrow B$ is E-valid, then B is a solution). A clause C is said to be $\bar{x}$-free if it contains no symbol from $\bar{x}$.

The algorithm we will present will involve generating E-consequences of formulas. Although in principle an abstract analysis of our algorithm based on a notion of *E-consequence generator* can be performed, we have chosen, for lack of space, to present only the concrete E-consequence generator used in our implementation.

Our E-consequence generator is based on *forgetful reasoning*. Let C_1, C_2 be two clauses, then denote the set of propositional resolvents of C_1, C_2 by $\mathrm{res}(C_1, C_2)$ and the set of clauses that can be obtained from C_1, C_2 by ground paramodulation by $\mathrm{para}(C_1, C_2)$. Letting F be a formula with CNF $\{C_i\}_{i\in I}$ we define

$$\mathcal{F}(F) = \{C \wedge \bigwedge\nolimits_{i\in I\setminus\{j,k\}} C_i \mid C \in \mathrm{res}(C_j, C_k) \cup \mathrm{para}(C_j, C_k)\}.$$

Using $\mathcal{F}$, we can now present Algorithm 2: the solution-finding algorithm $\mathrm{SF}_{\mathcal{F}}$. It prunes a solution A of $\bar{x}$-free clauses, then recurses upon those consequences of the pruned A generated by $\mathcal{F}$ which pass a certain E-validity check, finally returning a set of formulas (which will all be solutions).

Algorithm 2. $\mathrm{SF}_{\mathcal{F}}$

```
function SF_F(A: solution in CNF)
    A ← A without x̄-free clauses
    S ← {A}
    for B ∈ F(A) do
        if B[x̄\w̄_1], ..., B[x̄\w̄_k], Γ → Δ is E-valid then        ▷ B is a solution
            S ← S ∪ SF_F(B)
        end if
    end for
    return S
end function
```

Theorem 4 (Soundness & Termination). *Let A be any solution in CNF. Then $\mathrm{SF}_{\mathcal{F}}$ terminates on A and, for all $B \in \mathrm{SF}_{\mathcal{F}}(A)$, B is a solution.*

Proof. Termination is trivial since CNFs get smaller under $\mathcal{F}$. If $C(S^{\sim}) \supset B$ is E-valid then B is a solution iff $B[\bar{x}\backslash\bar{w}_1], \ldots, B[\bar{x}\backslash\bar{w}_k], \Gamma \to \Delta$ is E-valid. Hence the E-validity check in $\mathrm{SF}_{\mathcal{F}}$ suffices. Let A be a solution in CNF and A' be A without $\bar{x}$-free clauses. The fact that A' is a solution follows from the fact that $A \supset C[\bar{x}\backslash\bar{w}_k]$ is E-valid for all $\bar{x}$-free clauses C of A, together with the assumption that A is a solution.

The algorithm $\mathrm{SF}_{\mathcal{F}}$ can be used to lift a compression in quantifier-complexity to a compression w.r.t. proof length. Indeed, in the setting of first-order logic without equality, $\mathrm{SF}_{\mathcal{F}}$ has been applied in [7] to obtain an exponential speed-up result w.r.t. proof length.

Example 5. Consider the canonical formula $C(S^{\sim})$ of Example 4. Then $\mathrm{SF}_{\mathcal{F}}$ generates the CNF

$$F(\alpha_1, \alpha_2) : \; f^2\alpha_1 = s^4\alpha_1 \wedge f^2\alpha_2 = s^4\alpha_2 \wedge (\neg P(s^4\alpha_1, \alpha_2) \vee P(\alpha_1, s^4\alpha_2))$$

for the CNF of $C(S^{\sim})$ by applying paramodulation twice to equational atoms and resolution thrice to the clauses corresponding to the implications between the P-atoms. It can be checked that $\lambda\alpha_1\alpha_2.F(\alpha_1, \alpha_2)$ is a solution for $S^{\sim}$ which is smaller than the canonical solution.

4.2 Proof with Cut

Theorem 5. *Let $S^{\sim}$ be an extended Herbrand sequent of a Σ_1-sequent S. Then S has a proof φ with a $\forall$-cut s.t. $|\varphi|_q = |S^{\sim}|_i$.*

Proof. As in [7] (page nr. 12, Theorem 7). Note that the quantifier-blocks in the cut and the equality rules do not change the measured number of weak quantifier-block introductions analyzed in the paper above. The main steps in the proof are the following ones: let S' be an extended Herbrand sequent obtained by the solution $[X \backslash \lambda \bar{\alpha}.A]$. Then a proof with cut formula $\forall \bar{x}.A[\bar{\alpha} \backslash \bar{x}]$ can be constructed where the quantifier substitution blocks for the cut formula on the right-hand-side are $[\bar{x} \backslash \bar{w}]$ for $\bar{w} \in W$ while the cut formula on the left-hand-side gets the substitution $[\bar{x} \backslash \bar{\alpha}]$. The substitutions $[\bar{x_i} \backslash \bar{t}]$ for $\bar{t} \in U_i$ are inserted to introduce the quantifiers of the formula F_i in the end-sequent.

Example 6. Let $\Gamma = P(f^4a, a), \forall x.fx = s^2x, \forall xy(P(sx, y) \supset P(x, sy))$ be the left-hand-side of S. Then, to the canonical solution corresponds an **LK**-proof ψ of the form

$$\dfrac{\dfrac{\dfrac{(\psi_1)}{\Gamma \to P(a, f^4a), A(\alpha_1, \alpha_2)}}{\Gamma \to P(a, f^4a), \forall xy.A(x, y)} \forall^*_{\mathrm{r}} \quad \dfrac{\dfrac{(\psi_2)}{\Gamma, A(f^2a, a), A(a, f^2a) \to P(a, f^4a)}}{\Gamma, \forall xy.A(x, y) \to P(a, f^4a)} \forall^*_{\mathrm{l}}}{\Gamma \to P(a, f^4a)} \mathrm{cut} + \mathrm{c}^*$$

where ψ_1 and ψ_2 are cut-free and ψ_2 contains only structural and propositional inferences (in ψ_2 only $P(f^4a, a)$ is needed from Γ). The quantifier inferences in ψ_1 use exactly the 8 substitutions encoded in U_1 and U_2, ψ_2 uses the 2 substitutions represented by W. So we have $|\psi|_q = 10$.

5 Implementation and Experiments

Summing up the previous sections, the structure of our cut-introduction algorithm is the following:

Algorithm 3. Cut-Introduction

Require: π: cut-free proof
  $T \leftarrow \mathrm{extractTermSet}(\pi)$
  $D \leftarrow \mathrm{getMinimalDecomposition}(T)$
  $C(\bar{x}) \leftarrow \mathrm{getCanonicalSolution}(D)$
  $F(\bar{x}) \leftarrow \mathrm{improveSolution}(C(\bar{x}))$
  return $\mathrm{constructProof}(F(\bar{x}))$

Depending on whether the input proof π contains equality reasoning or not we either work modulo quasi-tautologies as described in this paper or modulo tautologies (as described in [8,7]) in improveSolution and constructProof.

In getMinimalDecomposition we can either compute decompositions with a single variable as in [8,7] or with an unbounded number of variables as described in Section 3. We denote these two variants with CI^1 and CI^* respectively.

These algorithms have been implemented in the gapt-system[1] which is a framework for transforming and analyzing formal proofs. It is implemented in Scala and contains data structures such as formulas, sequents, resolution and sequent calculus proofs and algorithms like unification, skolemization, cut-elimination as well as backends for several external solvers and provers. For deciding whether a quantifier-free formula is a tautology we use MiniSat[2]. We use veriT[3] for deciding whether a quantifier-free formula is a quasi-tautology and prover9[4] for the actual proof construction based on the import described in [9].

We have conducted experiments on the prover9-part of the TSTP-library (Thousands of Solutions of Theorem Provers, see [17]). The choice of prover9 was motivated by the simple and clean proof output format Ivy which makes proof import (comparatively) easy. This library contains 6341 resolution proofs. Of those, 5254 can be parsed and transformed into a sequent calculus proof using the transformation described in [9]. Of those, 2849 have non-trivial termsets (we call a term set trivial if every quantified formula in the end-sequent is instantiated at most once).

The input data we have used for our experiments is this collection of proofs with non-trivial term sets. In this collection 66% use equality reasoning and hence must be treated with the method introduced in this paper. The average term set size is 37,1 but 46% have a term set of size ≤ 10. The experiments have been conducted with version 1.6 of gapt on an Intel i5 QuadCore with 3,33GHz with an allocation of 2GB heap space and a timeout of 60 seconds for the cut-introduction algorithm.

On 19% of the input proofs our algorithm terminates with finding a compression, i.e. a non-trivial decomposition (of size at most that of the original termset) and a proof with cut that realizes this decomposition. On 49% it terminates determining that the proof is uncompressible, more precisely: that there is no proof with a single $\forall$-cut which (by cut-elimination) reduces to the given input term set and is of smaller quantifier complexity, see [7]. Figure 1 depicts the return status (in percent) depending on the size of the term set. When reading this figure one should keep in mind the relatively high number of small proofs (see above). One can observe that proofs with term sets up to a size of around 50 can be treated well by our current implementation, beyond that the percentage of timeouts is very large. Small proofs – unsurprisingly – tend to be uncompressible.

In Figure 2 we restrict our attention to runs terminating with a compression. As one can see from the diagram on the left, a significant reduction of quantifier-complexity can be achieved by our method. The diagram on the right

[1] Generic Architecture for Proof Transformations, `http://www.logic.at/gapt/`
[2] `http://minisat.se/`
[3] `http://www.verit-solver.org/`
[4] `http://www.cs.unm.edu/~mccune/prover9/`

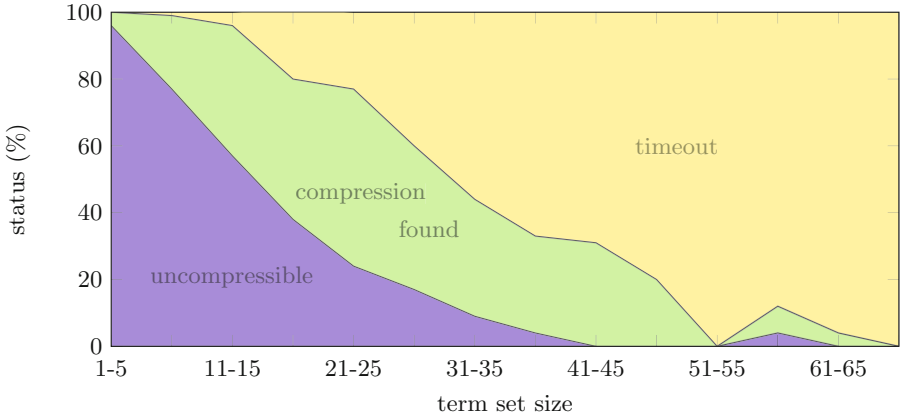

Fig. 1. CI*: return status by term set size

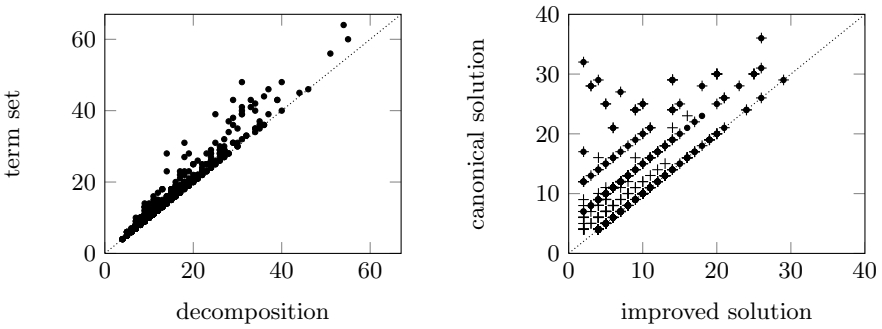

Fig. 2. Size Comparison

demonstrates that forgetful reasoning is highly useful for improving the canonical solution. The points plotted as • are the result after using forgetful resolution only, the points plotted as + are the result after forgetful resolution and paramodulation.

Our experiments also show that the generalization to the introduction of a block of quantifiers introduced in this paper has a strong effect: of the 548 proofs on which CI* finds a compression, 22% are found to be uncompressible by CI[1].

6 Conclusion

We have introduced a cut-introduction method that works modulo equality and is capable of generating cut-formulas containing a block of quantifiers. We have implemented our new method and have conducted a large-scale empirical evaluation which demonstrates its feasibility on realistic examples. Lessons learned from these experiments include that blocks of quantifiers allow for significantly

more proofs to be compressed and that forgetful reasoning methods, while rough in theory, are highly useful for our application in practice.

As future work we plan to extend our method to work modulo (suitably specified) equational theories. We also plan to evaluate our method on proofs produced by Tableaux-provers and SMT-solvers. Another important, and non-trivial, extension will be to cope with cuts that contain quantifier-alternations.

Acknowledgements. The authors would like to thank Pascal Fontaine for help with the veriT-solver and Geoff Sutcliffe for providing the prover9-TSTP test set.

References

1. Bundy, A.: The Automation of Proof by Mathematical Induction. In: Voronkov, A., Robinson, J.A. (eds.) Handbook of Automated Reasoning, vol. 1, pp. 845–911. Elsevier (2001)
2. Bundy, A., Basin, D., Hutter, D., Ireland, A.: Rippling: Meta-Level Guidance for Mathematical Reasoning. Cambridge Tracts in Theoretical Computer Science. Cambridge University Press (2005)
3. Colton, S.: Automated Theory Formation in Pure Mathematics. Ph.D. thesis, University of Edinburgh (2001)
4. Colton, S.: Automated Theory Formation in Pure Mathematics. Springer (2002)
5. Finger, M., Gabbay, D.: Equal Rights for the Cut: Computable Non-analytic Cuts in Cut-based Proofs. Logic Journal of the IGPL 15(5-6), 553–575 (2007)
6. Gentzen, G.: Untersuchungen über das logische Schließen. Mathematische Zeitschrift 39, 176–210, 405–431 (1934-1935)
7. Hetzl, S., Leitsch, A., Reis, G., Weller, D.: Algorithmic Introduction of Quantified Cuts (2013), `http://arxiv.org/abs/1401.4330` (submitted)
8. Hetzl, S., Leitsch, A., Weller, D.: Towards Algorithmic Cut-Introduction. In: Bjørner, N., Voronkov, A. (eds.) LPAR-18 2012. LNCS, vol. 7180, pp. 228–242. Springer, Heidelberg (2012)
9. Hetzl, S., Libal, T., Riener, M., Rukhaia, M.: Understanding Resolution Proofs through Herbrand's Theorem. In: Galmiche, D., Larchey-Wendling, D. (eds.) TABLEAUX 2013. LNCS, vol. 8123, pp. 157–171. Springer, Heidelberg (2013)
10. Ireland, A., Bundy, A.: Productive Use of Failure in Inductive Proof. Journal of Automated Reasoning 16(1-2), 79–111 (1996)
11. Johansson, M., Dixon, L., Bundy, A.: Conjecture synthesis for inductive theories. Journal of Automated Reasoning 47(3), 251–289 (2011)
12. Orevkov, V.: Lower bounds for increasing complexity of derivations after cut elimination. Zapiski Nauchnykh Seminarov Leningradskogo Otdeleniya Matematicheskogo Instituta 88, 137–161 (1979)
13. Shoenfield, J.R.: Mathematical Logic, 2nd edn. Addison Wesley (1973)
14. Sorge, V., Colton, S., McCasland, R., Meier, A.: Classification results in quasigroup and loop theory via a combination of automated reasoning tools. Commentationes Mathematicae Universitatis Carolinae 49(2), 319–339 (2008)
15. Sorge, V., Meier, A., McCasland, R., Colton, S.: Automatic Construction and Verification of Isotopy Invariants. Journal of Automated Reasoning 40(2-3), 221–243 (2008)

16. Statman, R.: Lower bounds on Herbrand's theorem. Proceedings of the American Mathematical Society 75, 104–107 (1979)
17. Sutcliffe, G.: The TPTP World - Infrastructure for Automated Reasoning. In: Clarke, E.M., Voronkov, A. (eds.) LPAR-16 2010. LNCS (LNAI), vol. 6355, pp. 1–12. Springer, Heidelberg (2010)
18. Tapolczai, J.: Cut-Introduction with Multiple Universal Quantifiers. Technical report, `http://www.logic.at/staff/hetzl/deltavector.pdf`
19. Vyskočil, J., Stanovský, D., Urban, J.: Automated Proof Compression by Invention of New Definitions. In: Clarke, E.M., Voronkov, A. (eds.) LPAR-16 2010. LNCS (LNAI), vol. 6355, pp. 447–462. Springer, Heidelberg (2010)
20. Woltzenlogel Paleo, B.: Atomic Cut Introduction by Resolution: Proof Structuring and Compression. In: Clarke, E.M., Voronkov, A. (eds.) LPAR-16 2010. LNCS (LNAI), vol. 6355, pp. 463–480. Springer, Heidelberg (2010)