HERBRAND-CONFLUENCE

STEFAN HETZL \textsuperscript{a} AND LUTZ STRASSBURGER \textsuperscript{b}

\textsuperscript{a} Institute of Discrete Mathematics and Geometry, Vienna University of Technology, Wiedner Hauptstraße 8-10, 1040 Vienna, Austria
\textit{e-mail address:} stefan.hetzl@tuwien.ac.at

\textsuperscript{b} INRIA Saclay – Île-de-France, 1 rue Honoré d’Estienne d’Orves, Bâtiment Alan Turing, Campus de l’École Polytechnique, 91120 Palaiseau, France
\textit{e-mail address:} lutz@lix.polytechnique.fr

Abstract. We consider cut-elimination in the sequent calculus for classical first-order logic. It is well known that this system, in its most general form, is neither confluent nor strongly normalizing. In this work we take a coarser (and mathematically more realistic) look at cut-free proofs. We analyze which witnesses they choose for which quantifiers, or in other words: we only consider the Herbrand-disjunction of a cut-free proof. Our main theorem is a confluence result for a natural class of proofs: all (possibly infinitely many) normal forms of the non-erasing reduction lead to the same Herbrand-disjunction.

1. Introduction

The constructive content of proofs has always been a central topic of proof theory and it is also one of the most important influences that logic has on computer science. Classical logic is widely used and presents interesting challenges when it comes to understanding the constructive content of its proofs. These challenges have therefore attracted considerable attention, see, for example, [Par92, DJS97, CH00], [BB96], [Urb00, UB00], [BBS02], [Koh08], or [BL00] for different investigations in this direction.

A well-known, but not yet well-understood, phenomenon is that a single classical proof usually allows several different constructive readings. From the point of view of applications this means that we have a choice among different programs that can be extracted. In [RT12] the authors show that two different extraction methods applied to the same proof produce two programs, one of polynomial and one of exponential average-case complexity. This phenomenon is further exemplified by case studies in [Urb00, BHL\textsuperscript{05}, BHL\textsuperscript{08}] as well as the asymptotic results [BH11, Het12]. The reason for this behavior is that classical “proofs often leave algorithmic detail underspecified” [Avi10].

On the level of cut-elimination in the sequent calculus this phenomenon is reflected by the fact that the standard proof reduction without imposing any strategy is not confluent.
In this paper we consider cut-elimination in classical first-order logic and treat the question which cut-free proofs one can obtain (by the strategy-free rewriting system) from a single proof with cuts. As our aim is to compare cut-free proofs we need a notion of equivalence of proofs: clearly the syntactic equality makes more differences than those which are mathematically interesting. Being in a system with quantifiers, a natural and more realistic choice is to consider two cut-free proofs equivalent if they choose the same terms for the same quantifiers, in other words: if they have the same Herbrand-disjunction.

A cut-reduction relation will then be called Herbrand-confluent if all its normal forms have the same Herbrand-disjunction. The main result of this paper is that, for a natural class of proofs, the standard reduction without erasing of subproofs is Herbrand-confluent. This result is surprising as this reduction is neither confluent nor strongly normalizing and may produce normal forms of arbitrary size (which—as our result shows—arise only from repetitions of the same instances).

As a central proof technique we use rigid tree languages which have been introduced in [JKV09] with applications in verification (e.g. of cryptographic protocols as in [JKV11]) as their primary purpose. To a proof we will associate a rigid tree grammar whose language is invariant under non-erasing cut-elimination and hence equal to the only obtainable Herbrand-disjunction. This property suggests the new notion of Herbrand-content of a proof, which is defined as the language of the grammar of the proof, and which is a strong invariant. A side effect of this proof technique is a combinatorial description of how the structure of a cut-free proof is related to that of a proof with cut. Such descriptions are important theoretical results which underlie applications such as algorithmic cut-introduction as in [HLW12, HLRW13].

This paper is an extended version of [HS12], where we have worked in a setting that was restricted to proofs of formulas of the shape $\exists x_1 \cdots \exists x_n A$, for $A$ quantifier-free. In this paper we extend the results obtained in [HS12] to proofs of arbitrary end-sequents. For this, we first carry out the central technical work in a setting of skolemized end-sequents, and then extend these results to the general case via deskolemization. This proof strategy is analogous to the proof of the second $\varepsilon$-Theorem from the first $\varepsilon$-Theorem in [HB39].

More precisely, this paper is structured as follows: in Section 2 we briefly review the sequent calculus and cut-elimination for classical first-order logic. In Section 3 we describe regular and rigid tree grammars, which we relate to proofs in Section 4. In Section 5 we prove the main invariance lemma in the skolemized setting, and in Section 6 we establish the necessary techniques and results for lifting the invariance lemma to the general case. This lifting is carried out in Section 7 followed by a discussion of several corollaries such as Herbrand-confluence.

2. Sequent Calculus and Cut-Elimination

For the sake of simplicity, we consider only a one-sided sequent calculus and formulas in negation normal form, but the results can be proved for a two-sided sequent calculus in the same way. Thus, our formulas (denoted by $A$, $B$, ...) are generated from the literals and the constants $\top$ and $\bot$ via the binary connectives $\land$ (and) and $\lor$ (or) and the quantifiers $\exists$ and $\forall$ in the usual way. The negation $\overline{A}$ of a formula $A$ is defined via the usual De Morgan laws. A sequent (denoted by $\Gamma$, $\Delta$, ...) is a multiset of formulas.
Definition 2.1. A proof is a tree of sequents, such that every node forms together with its children an instance of one of the inference rules shown below:

\[
\begin{array}{c}
\frac{\Gamma, A}{\Gamma, A} \quad \frac{\top}{\top} \quad \frac{\Gamma}{\Gamma, A} \quad \frac{\Gamma, A, \neg A}{\Gamma, \neg A} \quad \frac{\Gamma, A, \Delta}{\Gamma, \Delta} \quad \frac{\Gamma, \neg A}{\Gamma, A} \\
\end{array}
\]

where in the \textit{ax}-rule \(A\) has to be a literal, in the \(\forall\)-rule the \(\alpha\) is called \textit{eigenvariable} and does not appear in \(\Gamma\), \(\forall x\, A\), and in the \(\exists\)-rule the term \(t\) does not contain a variable bound in \(A\). We use the notation \([x \setminus \alpha]\) for the substitution that replaces \(x\) by the eigenvariable \(\alpha\). Similarly, \([x \setminus t]\) is the substitution that replaces \(x\) with \(t\).

The explicitly mentioned formula in a conclusion of an inference rule, like \(\neg A\), is called \textit{main formula}. Analogously, the explicitly mentioned formulas in the premises of an inference rule, like \(A\) and \(B\) for \(\lor\), are called \textit{auxiliary formulas}. In the context of a concrete derivation we speak about \textit{main} and \textit{auxiliary formula occurrences} of inferences.

Definition 2.2. A proof is called \textit{regular} if different \(\forall\)-inferences have different eigenvariables.

We use the following convention: We use lowercase Greek letters \(\alpha, \beta, \gamma, \delta, \ldots\) for eigenvariables in proofs, and \(\pi, \psi, \ldots\) for proofs. For a proof \(\pi\), we write \(|\pi|\) for the number of occurrences of inferences in \(\pi\). Furthermore, we write \(\text{EV}(\pi)\) for the set of eigenvariables of \(\forall\)-inferences of \(\pi\).

In a sequent calculus proof, each formula occurrence can be traced downwards via its descendants to either a cut formula or the end-sequent. We write \(\text{EV}_c(\pi)\) for the set of those eigenvariables in \(\pi\) that are introduced by a \(\forall\)-inference whose main formula occurrence can be traced downwards to a cut formula, i.e., is not part of the end-sequent of \(\pi\). The elements of \(\text{EV}_c(\pi)\) will also be called \textit{cut-eigenvariables}.

Definition 2.3. A \textit{weak sequent} is a sequent that does not contain any \(\forall\)-quantifier.

Fact 2.4. If the end-sequent of a proof \(\pi\) is a weak sequent then \(\text{EV}(\pi) = \text{EV}_c(\pi)\).

Remark 2.5. Our results do not depend on technical differences in the definition of the calculus (which in classical logic are inessential) such as the choice between multiplicative and additive rules and the differences in the cut-reduction induced by these choices. However, for the sake of precision, we will formally define the cut-reduction used in this paper.

Definition 2.6. Cut-reduction is defined on regular proofs and consists of the proof rewrite steps shown in Figure 1 (as well as all corresponding symmetric variants), where in the contraction reduction step \(\rho' = [\alpha \setminus \alpha']_{\alpha \in \text{EV}(\psi_2)}\) and \(\rho'' = [\alpha' \setminus \alpha'']_{\alpha \in \text{EV}(\psi_2)}\) are substitutions replacing each eigenvariable \(\alpha\) in \(\psi_2\) by fresh copies, i.e., \(\alpha'\) and \(\alpha''\) are fresh for the whole proof. We write \(\sim\) for the compatible (w.r.t. the inference rules), reflexive and transitive closure of \(\sim\).

The above system for cut-reduction consists of purely local, minimal steps and therefore allows the simulation of many other reduction relations. We chose to work in this system in order to obtain invariance results of maximal strength. Among the systems that can
Axiom reduction:
\[ \psi_1 \quad \psi_2 \]
\[ \frac{\Gamma, A \quad \overline{A}, A}{\overline{\Gamma}, A} \quad \text{cut} \]
\[ \Psi \quad \Gamma, A \]

Quantifier reduction:
\[ \psi_1 \quad \psi_2 \]
\[ \frac{\Delta, \overline{A}[x \setminus t]}{\Delta, \exists x A} \quad \frac{A[x \alpha], \Gamma}{\forall x A, \Gamma} \]
\[ \frac{\Psi}{\Gamma, \Delta} \quad \text{cut} \]

Propositional reduction:
\[ \psi_1 \quad \psi_2 \quad \psi_3 \]
\[ \frac{\Gamma, A \quad \Delta, B \quad \overline{A}, B, \Pi}{\Gamma, \Delta, A \wedge B} \quad \frac{\Pi \lor \overline{\Pi}}{\Gamma, \Delta, \Pi} \]
\[ \frac{\Psi}{\Gamma, \Delta, \Pi} \quad \text{cut} \]

Contraction reduction:
\[ \psi_1 \quad \psi_2 \]
\[ \frac{\Gamma, A, A}{\Gamma, A} \quad \text{cut} \]
\[ \frac{\Gamma, A, A}{\overline{\Gamma}, \Delta} \quad \text{cut} \]

Weakening reduction:
\[ \psi_1 \quad \psi_2 \]
\[ \frac{\Gamma}{\Gamma, A} \quad \frac{\overline{A}, \Delta}{\Gamma, \Delta} \]
\[ \Psi \quad \Gamma, \Delta \quad \text{cut} \]

Unary inference permutation:
\[ \psi_1 \quad \psi_2 \]
\[ \frac{\Gamma', A \quad \overline{A}, \Delta}{\Gamma, A} \quad \text{cut} \]
\[ \frac{\Gamma', A \quad \overline{A}, \Delta}{\Gamma', \Delta} \quad \text{cut} \]

Binary inference permutation:
\[ \psi_1 \quad \psi_2 \quad \psi_3 \]
\[ \frac{\Gamma', \overline{A} \quad \Gamma''}{\Gamma, A} \quad \frac{\overline{A}, \Delta}{\Gamma', \Delta} \quad \frac{\Gamma', \overline{A} \quad \Gamma''}{\Gamma, \Delta} \quad \text{cut} \]
\[ \frac{\Psi}{\Gamma, \Delta} \quad \text{cut} \]

Figure 1: Cut-reduction steps
be simulated literally are for example all color annotations of \cite{DJS97} in the multiplicative version of LK defined there. The real strength of the results in this paper lies however in the general applicability of the used proof techniques: the extraction of a grammar from a proof (that is described in the next sections) is possible in all versions of sequent calculus for classical logic and in principle also in other systems like natural deduction. In particular, our results also apply to inversion-based cut-elimination procedures such as for example in \cite{Sch77}.

3. Regular and Rigid Tree Grammars

Formal language theory constitutes one of the main areas of theoretical computer science. Traditionally, a formal language is defined to be a set of strings but this notion can be generalized in a straightforward way to considering a language to be a set of first-order terms. Such tree languages possess a rich theory and many applications, see e.g. \cite{GS97, CDG+07}. In this section we introduce notions and results from the theory of tree languages that we will use for our proof-theoretic purposes.

A **ranked alphabet** $\Sigma$ is a finite set of symbols which have an associated arity (their **rank**). For $f \in \Sigma$, we sometimes use the notation $f/n$ for saying that $n$ is the arity of $f$. We write $\mathcal{T}_\Sigma$ to denote the set of all finite trees (or terms) over $\Sigma$, and we write $\mathcal{T}_\Sigma(X)$ to denote the set of all trees over $\Sigma$ and a set $X$ of variables (seen as symbols of arity 0). We also use the notion of **position** in a tree, which is a list of natural numbers. We write $\varepsilon$ for the empty list (the root position), and we write $p.q$ for the concatenation of lists $p$ and $q$. We write $p \leq q$ if $p$ is a prefix of $q$ and $p < q$ if $p$ is a proper prefix of $q$. Clearly, $\leq$ is a partial order and $<$ is its strict part. We write $\text{Pos}(t)$ to denote the set of all positions in a tree $t \in \mathcal{T}_\Sigma(X)$. Furthermore, for a given tree or term $t$ and position $p$, we write $t|_p$ to denote the subterm of $t$ that occurs at position $p$.

**Definition 3.1.** A **regular tree grammar** is a tuple $G = \langle N, \Sigma, \theta, P \rangle$, where $N$ is a finite set of **non-terminal symbols**, where $\Sigma$ is a ranked alphabet, such that $N \cap \Sigma = \emptyset$, where $\theta$ is the **start symbol** with $\theta \in N$, and where $P$ is a finite set of production rules of the form $\beta \rightarrow t$ with $\beta \in N$ and $t \in \mathcal{T}_\Sigma(N)$.

The derivation relation $\rightarrow_G$ of a regular tree grammar $G = \langle N, \Sigma, \theta, P \rangle$ is defined as follows. We have $s \rightarrow_G r$ if there is a production rule $\beta \rightarrow t$ in $P$ and a position $p \in \text{Pos}(s)$, such that $s|_p = \beta$ and $r$ is obtained from $s$ by replacing $\beta$ at $p$ by $t$. The **language** of $G$ is then defined as $L(G) = \{ t \in \mathcal{T}_\Sigma \mid \theta \rightarrow_G^* t \}$, where $\rightarrow_G^*$ is the reflexive and transitive closure of $\rightarrow_G$. A derivation $\mathcal{D}$ of a term $t \in L(G)$ is a sequence $t_0 \rightarrow_G t_1 \rightarrow_G \ldots \rightarrow_G t_n$ with $t_0 = \theta$ and $t_n = t$. Note that a term $t$ might have different derivations in $G$.

In \cite{JKV09} the class of rigid tree languages has been introduced with applications in verification (e.g. of cryptographic protocols as in \cite{JKV11}) as primary motivation. It will turn out that this class is appropriate for describing cut-elimination in classical first-order logic. In contrast to \cite{JKV09} we do not use automata but grammars—their equivalence is shown in \cite{Het12a}.

**Definition 3.2.** A **rigid tree grammar** is a tuple $\langle N, N_R, \Sigma, \theta, P \rangle$, where $\langle N, \Sigma, \theta, P \rangle$, is a regular tree grammar and $N_R \subseteq N$ is the set of **rigid non-terminals**. We speak of a **totally rigid tree grammar** if $N_R = N$. In this case we will just write $\langle N_R, \Sigma, \theta, P \rangle$. 
A derivation \( \theta = t_0 \to_G t_1 \to_G \ldots \to_G t_n = t \) of a rigid tree grammar \( G = \langle N, N_R, \Sigma, \theta, P \rangle \) is a derivation in the underlying regular tree grammar satisfying the additional rigid condition: If there are \( i, j < n \), a non-terminal \( \beta \in N_R \), and positions \( p \) and \( q \) such that \( t_i|_p = \beta \) and \( t_j|_q = \beta \) then \( t_i|_p = t_j|_q \). The language \( L(G) \) of the rigid tree grammar \( G \) is the set of all terms \( t \in \mathcal{T}_2 \) which can be derived under the rigidity condition. For a given derivation \( \mathcal{D}: \alpha \to_G \beta \) of a term \( \alpha \) and a non-terminal \( \beta \) we say that \( \beta \) is a \( \beta \)-position in \( \mathcal{D} \) if there is an \( i \leq n \) with \( t_i|_p = \beta \), i.e., either a production rule \( \beta \to s \) has been applied at \( p \) in \( \mathcal{D} \), or \( \beta \) occurs at position \( p \) in \( t \). In the context of a given grammar \( G \), we sometimes write \( \mathcal{D} : \alpha \to^* G t \) to specify that \( \mathcal{D} \) is a derivation starting with \( \alpha \) and ending with the term \( t \).

Example 3.3. Let \( \Sigma = \{0/0, s/1\} \). A simple pumping argument shows that the language \( L = \{f(t,t) \mid t \in \mathcal{T}_2 \} \) is not regular. On the other hand, \( L \) is generated by the rigid tree grammar

\[
G = \langle \{\theta, \alpha, \beta\}, \{\alpha\}, \{0/0, s/1, f/2\}, \theta, P \rangle \quad \text{where} \quad
P = \{ \theta \to f(\alpha, \alpha), \alpha \to 0 | s(\beta), \beta \to 0 \mid s(\beta) \}
\]

Lemma 3.4. Let \( G = \langle N, N_R, \Sigma, \theta, P \rangle \) be a rigid tree grammar and let \( t \in L(G) \). Then there is a derivation \( \alpha \to_G \beta \) which uses at most one \( \beta \)-production for each \( \beta \in N_R \).\hfill \Box

Proof. Given any derivation of \( t \), suppose both \( \beta \to s_1 \) and \( \beta \to s_2 \) are used at positions \( p_1 \) and \( p_2 \) respectively. Then by the rigidity condition \( t|_{p_1} = t|_{p_2} \) and we can replace the derivation at \( p_2 \) by that at \( p_1 \) (or the other way round). This transformation does not violate the rigidity condition because it only copies existing parts of the derivation.

Lemma 3.5. Let \( G = \langle N_R, \Sigma, \theta, P \rangle \) be a totally rigid tree grammar and \( \theta \neq \beta \in N_R \), such that there is exactly one \( t \), with \( \beta \to t \) in \( P \). If \( G' = \langle N_R \setminus \{\beta\}, \Sigma, \theta, (P \setminus \{\beta \to t\})[\beta \to \ell] \rangle \) then \( L(G) = L(G') \).\hfill \Box

Proof. If a \( G \)-derivation of a term \( s \) uses \( \beta \), it must replace \( \beta \) by \( t \) hence \( s \) is derivable using the productions of \( G' \) as well. The rigidity condition is preserved as the equality constraints of the \( G' \)-derivation are a subset of those of the \( G \)-derivation. Conversely, given a \( G' \)-derivation of a term \( s \) we obtain a derivation of \( s \) from the productions of \( G \) by replacing applications of \( \delta \to r[\beta \to \ell] \) by \( \delta \to r \) followed by a copy of \( \beta \to t \) for each occurrence of \( \beta \) in \( r \). Let \( \gamma_1, \ldots, \gamma_n \) be the non-terminals that appear in \( t \). By the rigidity condition for \( i \in \{1, \ldots, n\} \) there is a unique term at all \( \gamma_i \)-positions in the derivation. Hence \( \beta \) fulfills the rigidity condition as well, and we have obtained a \( G \)-derivation of \( s \).

Notation 3.6. For a given non-terminal \( \beta \) and a term \( t \), we will write \( \beta \in t \) or \( t \ni \beta \) for denoting that \( \beta \) occurs in \( t \).

Definition 3.7. Let \( G \) be a tree grammar. A path of \( G \) is a list \( \mathcal{P} \) of productions \( \alpha_1 \to t_1, \ldots, \alpha_n \to t_n \) with \( n \geq 1 \) and \( \alpha_{i+1} \in t_i \) for all \( i \in \{1, \ldots, n-1\} \). The length of a path is \( |\mathcal{P}| = n \). We will also write \( \mathcal{P}: \alpha_1 \to t_1 \ni \alpha_2 \to \ldots \ni \alpha_n \to t_n \) to denote a path.

For a given path \( \mathcal{P}: \alpha_1 \to t_1 \ni \alpha_2 \to \ldots \ni \alpha_n \to t_n \) we say that \( \alpha_1, \ldots, \alpha_n \) are on the path \( \mathcal{P} \) and write \( \alpha_i \in \mathcal{P} \) for that. We also write \( \mathcal{P}: \alpha_1 \to^* t_n \) and \( \mathcal{P}: \alpha_1 \to^\ast \alpha_n \), if we do not want to explicitly mention the intermediate steps. For a fixed grammar \( G \), we write \( \alpha \to \beta \) to denote that there is a path \( \mathcal{P} \) in \( G \) with \( \mathcal{P}: \alpha \to \beta \).
For a set $P$ of production rules, we write $\alpha <_P \beta$ (or simply $\alpha < \beta$, when $P$ is clear from context) if there is a production $\alpha \rightarrow t$ in $P$ with $\beta \in t$. We write $<_+$ for the transitive closure of $<$, and $<^*$ for its reflexive, transitive closure. Note that $\alpha \rightarrow \beta$ implies $\alpha <^+ \beta$, but not the other way around, since $\beta$ could be a non-terminal with no production $\beta \rightarrow s$ in $P$.

**Definition 3.8.** A tree grammar $\langle N, \Sigma, \theta, P \rangle$ is called **cyclic** if $\alpha <^+_P \alpha$ for some $\alpha \in N$, and **acyclic** otherwise.

**Lemma 3.9.** If $G$ is totally rigid and acyclic, then we have that up to renaming of the non-terminals $G = \langle \{\alpha_1, \ldots, \alpha_n\}, \Sigma, \alpha_1, P \rangle$ with $L(G) = \{\alpha_1|\alpha_1|t_1| \cdots |n\alpha_n|t_n| \mid \alpha_i \rightarrow t_i \in P\}$.

**Proof.** Acyclicity permits a renaming of non-terminals, such that $\alpha_i <^+_P \alpha_j$ implies $i < j$. Then $L(G) \supseteq \{\alpha_1|\alpha_1|t_1| \cdots |n\alpha_n|t_n| \mid \alpha_i \rightarrow t_i \in P\}$ is obvious. For the left-to-right inclusion, let $\phi: \alpha_1 = s_1 \rightarrow G \rightarrow G \rightarrow G \Rightarrow s_n = s \in \mathcal{F}_G$ be a derivation in $G$. By Lemma 3.4 we can assume that for each $j$ at most one production whose left-hand side is $\alpha_j$ is applied, say $\alpha_j \rightarrow t_j$. By acyclicity we can rearrange the derivation so that $\alpha_j \rightarrow t_j$ is only applied after $\alpha_i \rightarrow t_i$ for all $i < j$. For those $\alpha_j$ which do not appear in the derivation we can insert any substitution without changing the final term so we obtain $s = \alpha_1|\alpha_1|t_1| \cdots |n\alpha_n|t_n$. \(\square\)

This lemma entails that $|L(G)| \leq \prod_{i=1}^n |\{t \mid \alpha_i \rightarrow t \in P\}|$, in particular we are dealing with a finite language. The central questions in this context are (in contrast to the standard setting in formal language theory) not concerned with representability but with the size of a representation.

### 4. Proofs and Grammars

In this section we will relate sequent calculus proofs to rigid tree grammars. A central tool for establishing this relation is Herbrand’s theorem [Her30 Bus95]. In its simplest form it states that $\exists x A$, for $A$ quantifier-free, is valid iff there are terms $t_1, \ldots, t_n$ such that $\bigwedge_{i=1}^n A[x|t_i]$ is a tautology. Such tautological disjunctions of instances are hence called **Herbrand-disjunctions**. Such a disjunction, or equivalently: the set of terms, can be considered a compact representation of a cut-free proof. The relation to tree grammars is based on the observation that a (finite) set of terms is just a (finite) tree language. While the Herbrand-disjunction of a cut-free proof will be considered a tree language, a proof with cut will give a rise to a grammar and its cut-elimination will be described by the computation of the language of its grammar.

There are different options for extending Herbrand’s theorem to non-prenex formulas, e.g. the Herbrand proofs of [Bus95] or the expansion trees of [Mil87]. For our purposes it will be most useful to follow the approach of [BL94].

**Definition 4.1.** Let $\pi$ be a proof and let $O$ be a formula occurrence in $\pi$. Then we define the **Herbrand-set** $H(O)$ of $O$ inductively as follows:

- If $O$ is the occurrence of a formula $A$ in an axiom, then $H(O) = \{A\}$.
- If $O$ is in the conclusion sequent of an inference rule without being its main occurrence, then $H(O)$ has exactly one ancestor $O'$ in one of the premises, and we let $H(O) = H(O')$.
- If $O$ is the main occurrence in the conclusion of a $\circ$-rule with $\circ \in \{\land, \lor\}$ and with auxiliary occurrences $O_1$ and $O_2$, then $H(O) = \{A \circ B \mid A \in H(O_1), B \in H(O_2)\}$.
- If $O$ is the main occurrence in the conclusion of a $\forall$- or $\exists$-rule with auxiliary occurrence $O_1$ in the premise, then $H(O) = H(O_1)$. 
If \( O \) is the main occurrence in the conclusion of a \( w \)-rule, then \( H(O) = \{ \bot \} \).

If \( O \) is the main occurrence in the conclusion of a \( c \)-rule with auxiliary occurrences \( O_1 \) and \( O_2 \) in the premise, then \( H(O) = H(O_1) \cup H(O_2) \).

Finally, we define
\[
H(\pi) = \bigcup_{P \in \Gamma} H(P)
\]
where \( \Gamma \) is the end-sequent of \( \pi \) and \( P \) ranges over all formula occurrences in \( \Gamma \).

Besides to the Herbrand-set of a formula occurrence, we also need the set of terms associated with an occurrence of an \( \exists \)-formula.

**Definition 4.2.** Let \( Q \) be an occurrence of a formula \( \exists x A \) in a proof. We define the set \( tm(Q) \) of terms associated with \( Q \) as follows: if \( Q \) is introduced as the main formula of a weakening, then \( tm(Q) = \emptyset \). If \( Q \) is introduced by an \( \exists \)-rule
\[
\frac{\Gamma, A[x/t]}{\Gamma, \exists x A}
\]
then \( tm(Q) = \{ t \} \). If \( Q \) is the main formula in the conclusion of a contraction, and \( Q_1 \) and \( Q_2 \) are the two auxiliary occurrences of the same formula in the premise, then \( tm(Q) = tm(Q_1) \cup tm(Q_2) \). In all other cases, an inference with the occurrence \( Q \) in the conclusion has a corresponding occurrence \( Q' \) of the same formula in one of its premises, and we let \( tm(Q) = tm(Q') \).

In the following, we will restrict our attention to a certain class of proofs, that we call simple proofs below.

**Definition 4.3.** A proof \( \pi \) is called simple if

- it is regular (i.e., different \( \forall \)-inferences have different eigenvariables),
- every cut in \( \pi \) is of one of the following forms:

\[
\frac{\Gamma, B \quad \overline{\Delta}}{\Gamma, \overline{\Delta}} \quad \text{cut} \quad \text{or} \quad \frac{\Gamma, \exists x B, \overline{\Delta}}{\Gamma, \overline{\Delta}} \quad \forall
\]

where \( B \) is quantifier-free.

Let us make some remarks on this definition. First, we require regularity which is a necessary assumption in the context of cut-elimination. But since every proof can be trivially transformed into a regular one, this is no real restriction. Second, the requirement of the \( \forall \)-rule being applied directly above the cut is natural as the rule is invertible. Moreover, any proof which does not fulfill this requirement can be pruned to obtain one that does, by simply permuting \( \forall \)-inferences down and identifying their eigenvariables when needed. Thus, the only significant restriction is that of disallowing quantifier alternations in the cut formulas. This corresponds to allowing only \( \Sigma_1 \) (or \( \Pi_1 \)) formulas in cuts.

We conjecture that our central result can be extended to \( \Sigma_n \)-cuts. However, this will require the development of an adequate class of grammars first (see also Section 8).

**Observation 4.4.** Simple proofs have the technically convenient property of exhibiting a 1-1 relationship between cut-eigenvariables and cuts. For an eigenvariable \( \alpha \in EV_c(\pi) \) we will therefore write \( \text{cut}_\alpha \) for the corresponding cut and \( \forall_\alpha \) for the inference introducing \( \alpha \) (when read from bottom to top).
**Definition 4.5.** Let \( \pi \) be a simple proof, let \( \alpha \in EV_c(\pi) \), and let \( Q \) be the occurrence of the existentially quantified cut-formula in the premise of \( \text{cut}_\alpha \). Then we write \( B(\alpha) \) for the set of substitutions \( \{ [\alpha \backslash t] \mid t \in \text{tm}(Q) \} \) and we define
\[
B(\pi) = \bigcup_{\alpha \in EV_c(\pi)} B(\alpha).
\]

Structures similar to the above \( B(\pi) \) have been investigated also in [Hei10] and [McK13], where they form the basis of proof net like formalisms using local reductions for quantifiers in classical first-order logic. Our aim in this work is however quite different: we use these structures for a global analysis of the sequent calculus.

**Definition 4.6.** Let \( \pi \) be a simple proof. Then the **grammar of \( \pi \)** is the totally rigid grammar \( G(\pi) = \langle N_R, \Sigma, \theta, P \rangle \) with
\[
N_R = EV_c(\pi) \cup \{ \theta \}.
\]
\[
\Sigma = \Sigma(\pi) \cup \{ \land, \lor, \top, \bot \}.
\]
\[
P = \{ \theta \to A \mid A \in H(\pi) \} \cup \{ \alpha \to t \mid [\alpha \backslash t] \in B(\pi) \}
\]
where \( \Sigma(\pi) \) is the signature of \( \pi \). The rank of \( \land \) and \( \lor \) is 2, the rank of \( \top \) and \( \bot \) is 0, and \( \theta \) does not occur in \( \pi \).

**Lemma 4.7.** If \( \pi \) is a simple proof, then \( G(\pi) \) is acyclic.

**Proof.** By induction on the number of cuts in \( \pi \). The grammar of a cut-free proof is trivially acyclic. For the induction step, let \( r \) be the lowest binary inference with subproofs \( \pi_1 \) and \( \pi_2 \) such that either (i) \( r \) is a cut or (ii) \( r \) is not a cut but both \( \pi_1 \) and \( \pi_2 \) contain at least one cut. Let \( P, P_1, \) and \( P_2 \) be the set of productions induced by the cuts in \( \pi, \pi_1, \pi_2 \), respectively. In case (iii), \( <_P = <_{P_1} \cup <_{P_2} \), which is acyclic by induction hypothesis (since \( EV_c(\pi_1) \cap EV_c(\pi_2) = \emptyset \)). In case (i), let \( P_r \) be the productions induced by the cut \( r \), then \( <_P = <_{P_1} \cup <_{P_2} \cup <_{P_r} \). By induction hypothesis, \( <_{P_1} \) and \( <_{P_2} \) are acyclic and as the cut-formula in \( r \) contains at most one quantifier, also \( <_{P_r} \) is acyclic. Therefore, a cycle in \( <_{P_r} \) must be of the form \( \alpha_1 <_{P_1} \beta_1 <_{P_1} \alpha_2 <_{P_2} \beta_2 <_{P_1} \alpha_1 \) where \( \alpha_1, \beta_1 \in EV_c(\pi_1) \) and \( \alpha_2, \beta_2 \in EV_c(\pi_2) \). However, \( r \) contains only one quantifier and depending on its polarity all productions in \( P_r \) lead from \( \pi_1 \) to \( \pi_2 \) or from \( \pi_2 \) to \( \pi_1 \) but not both, so \( <_P \) is acyclic.

5. **Grammars and Cut-Elimination**

In this section we will show that the language of the grammar of a proof defined in the previous section is an invariant under cut elimination. Before formulating this invariance result precisely we have to consider the following three aspects of the situation:

First, note that all the reductions shown in Figure 1 preserve simplicity, except the following:

\[
\cdots \quad \vdots \quad \forall_\alpha \quad \cdots \quad \forall_\beta \quad \cdots \\
\hline \hline \text{cut}_\alpha \hline \hline \cdots \quad \forall_\beta \quad \sim \quad \forall_\alpha \quad \cdots \quad \forall_\beta \quad \cdots \\
\hline \hline \cdots \quad \forall_\alpha \quad \hline \hline \text{cut}_\beta \hline \hline \cdots \quad \forall_\alpha \quad \cdots \quad \forall_\beta \quad \cdots \\
\hline \hline \cdots \quad \forall_\beta \quad \sim \quad \forall_\alpha \quad \cdots \quad \forall_\beta \quad \cdots \\
\hline \hline \cdots \quad \forall_\beta \quad \hline \hline \text{cut}_\alpha \hline \hline \cdots \quad \forall_\alpha \quad \cdots \quad \forall_\beta \quad \cdots
\]

\[1\] We consider the eigenvariables in \( EV(\pi) \setminus EV_c(\pi) \) to be part of \( \Sigma(\pi) \).
where \(\text{cut}_\alpha\) is permuted down under \(\text{cut}_\beta\) (using the bottommost reduction in Fig. 1) and the cut formula of \(\text{cut}_\beta\) has its ancestor on the right side of \(\text{cut}_\alpha\). So in the following, when we speak about a reduction sequence of simple proofs we require that the above reduction is immediately followed by permuting \(\forall_\alpha\) down as well, in order to arrive at

\[
\begin{array}{c}
\ldots \\
\ldots \ldots \vdash \forall_\beta \\
\ldots \\
\ldots \ldots \vdash \text{cut}_\beta \\
\ldots \\
\ldots \ldots \vdash \forall_\alpha \\
\ldots \\
\ldots \ldots \vdash \text{cut}_\alpha \\
\ldots \\
\end{array}
\]

which is again simple.

Secondly, observe that there is no mechanism for deletion in the grammar, but there is one in cut-elimination: the reduction of weakening which erases a sub-proof (see Fig. 1). It is hence natural and will turn out to be useful to also consider the reduction relation without this step.

**Definition 5.1.** We define the non-erasing cut-reduction \(\ne\) as \(\sim\) without the reduction rule for weakening.

Note that a \(\ne\)-normal form \(\pi\) is an analytic proof too as \(H(\pi)\) is also a Herbrand-disjunction, i.e. a tautological collection of instances. In contrast to a \(\sim\)-normal form (which might contain implicit redundancy) a \(\ne\)-normal form might also contain explicit redundancy in the form of cuts whose cut-formulas are introduced by weakening on one or on both sides. Non-erasing reduction is also of interest in the context of the \(\lambda\)-calculus where it is often considered in the form of the \(\lambda I\)-calculus and gives rise to the conservation theorem (see Theorem 13.4.12 in [Bar84]). Our situation here is however quite different: neither \(\sim\) nor \(\ne\) is confluent and neither of them is strongly normalizing.

Thirdly, in contrast to the case treated in [HS12] in our more general setting it may happen that the reduction of a weakening deletes sub-formulas of formula instances from the proof. In order to treat this situation adequately, we need to define a generalization of the \(\subseteq\)-relation between sets of formulas. For this reason, we use the symbol \(\perp\) for representing subformulas introduced by weakening, a technique also employed in [BHW12, Wel11] for the purpose of a tighter complexity-analysis.

**Definition 5.2.** The relation \(\preceq\) is defined inductively on quantifier-free formulas as follows:

- for all formulas \(A\) we have \(\perp \preceq A\) and \(A \preceq A\), and
- whenever \(A' \preceq A\) and \(B' \preceq B\) then also \(A' \land B' \preceq A \land B\) and \(A' \lor B' \preceq A \lor B\)

Let \(\mathcal{A}\) and \(\mathcal{B}\) be sets of quantifier-free formulas. Then we define \(\mathcal{A} \preceq \mathcal{B}\) iff for all \(A \in \mathcal{A}\) there is a \(B \in \mathcal{B}\) with \(A \preceq B\).

**Fact 5.3.** The relation \(\preceq\) is transitive on formula sets.

We are now in a position to precisely state our main invariance lemma which connects grammars with cut-elimination for weak sequents.

**Lemma 5.4.** If \(\pi \sim \pi'\) is a reduction sequence of simple proofs of a weak sequent, then \(L(G(\pi)) \geq L(G(\pi'))\). If \(\pi \ne \pi'\) is a reduction sequence of simple proofs of a weak sequent, then \(L(G(\pi)) = L(G(\pi'))\).
The rest of this section is devoted to proving this result. The proof strategy is to carry out an induction on the length of the reduction sequence \( \pi \to \pi' \) (or \( \pi \xrightarrow{\text{ne}} \pi' \) respectively) and to make a case distinction on the type of reduction step. The most difficult step will turn out to be the reduction of contraction which duplicates a sub-proof.

**Lemma 5.5.** Let \( \pi \) be a simple proof, and let \( \pi' \) be obtained from \( \pi \) by the single application of an axiom reduction, or a propositional reduction, or a unary or binary inference permutation (see Figure 7). Then \( L(G(\pi')) = L(G(\pi)) \).

**Proof.** None of these reductions is changing the grammar of the proof, i.e., \( G(\pi') = G(\pi) \) and therefore also \( L(G(\pi')) = L(G(\pi)) \). \( \square \)

**Lemma 5.6.** Let \( \pi \) be a simple proof, and let \( \pi' \) be obtained from \( \pi \) by the single application of a quantifier reduction (see Figure 7). Then \( L(G(\pi')) = L(G(\pi)) \).

**Proof.** Let \( \alpha \) be the eigenvariable of the \( \forall \)-inference and \( t \) be the term of the \( \exists \)-rule directly above the cut that is reduced. Then \( G(\pi') \) can be obtained from \( G(\pi) \) by removing the production rule \( \alpha \to t \) and by applying the substitution \( [\alpha \ \! \! \leftarrow \ t] \) to the right-hand side of all remaining production rules. Thus, \( L(G(\pi')) = L(G(\pi)) \) follows immediately from Lemma 5.5. \( \square \)

**Lemma 5.7.** Let \( \pi \) be a simple proof, and let \( \pi' \) be obtained from \( \pi \) by the single application of a weakening reduction (see Figure 7). Then \( L(G(\pi')) \leq L(G(\pi)) \).

**Proof.** The grammar \( G(\pi') \) is obtained from \( G(\pi) \) via two modifications. First, all productions coming from cuts or \( \exists \)-inferences in \( \psi_2 \) are deleted, and second, the formulas in \( \Delta \) which are ancestors of the end-sequent are replaced by \( \bot \) in \( H(\pi') \). Now let \( A \in L(G(\pi')) \).

Then the derivation of \( A \) in \( G(\pi') \) is also a derivation in \( G(\pi) \), with the difference that some \( \bot \)-subformulas are replaced by other formulas, yielding a formula \( B \in L(G(\pi)) \) with \( B \geq A \). Hence \( L(G(\pi')) \leq L(G(\pi)) \). \( \square \)

It remains to analyze the case of contraction. Surprisingly, also in this case the language of the grammar of a proof remains unchanged. However, the proof of this result is quite technical and requires additional auxiliary results about the relationship between proofs and grammars. Furthermore, this is the case which needs the additional condition that the end-sequent of our proof is weak, i.e., does not contain \( \forall \)-quantifiers.

For simplifying the presentation, we assume in the following (without loss of generality) that the \( \forall \)-side is on the right of a cut and the \( \exists \)-side on the left. Then, a production \( \beta \to t \) in \( G(\pi) \) corresponds to three inferences in \( \pi \): a cut, an instance of the \( \forall \)-rule, and an instance of the \( \exists \)-rule, that we denote by \( \text{cut}\_\beta \), \( \forall\_\beta \), and \( \exists\_i \), respectively, and that are, in general, arranged in \( \pi \) as shown below.

\[
\frac{\Gamma', A[x\!\!\leftarrow \beta]}{\exists x\ A} \quad \frac{\exists x\ A[x\!\!\leftarrow \beta], \Delta'}{\forall x\ A, \Delta'} \quad \forall\_\beta \\
\vdots \\
\frac{\Gamma, \exists x\ A}{\forall x\ A, \Delta} \quad \exists\_i \quad \text{cut}\_\beta \\
\Gamma, \Delta
\] (5.1)

The additional condition that \( \forall\_\beta \) is directly above \( \text{cut}\_\beta \), as indicated in (5.1), is needed because in the following we make extensive use of Observation 4.4: there is a one-to-one correspondence between the cuts and the eigenvariables in \( \text{EV}_c(\pi) \), and thus, the notation \( \text{cut}\_\beta \) makes sense.
Definition 5.8. We say that the instances \( \text{cut}_\beta, \forall_{\beta}, \) and \( \exists_t \) are on a path \( \mathcal{P} \) in \( G(\pi) \) if the production \( \beta \to t \) is in \( \mathcal{P} \).

Definition 5.9. Let \( \pi \) be a proof containing the configuration

\[
\begin{array}{c}
r_1 \\
\vdots \\
r_2 \\
\vdots \\
r_3 \\
\vdots \\
\end{array}
\]

where \( r_1, r_2, \) and \( r_3 \) are arbitrary rule instances, and \( r_3 \) is a branching rule, and \( r_1 \) and \( r_2 \) might or might not be branching. Then we say that \( r_1 \) is on the left above \( r_3 \), denoted by \( r_1 \uparrow r_3 \), and \( r_2 \) is on the right above \( r_3 \), denoted by \( r_3 \uparrow r_2 \), and \( r_1 \) and \( r_2 \) are in parallel, denoted by \( r_1 \uparrow|\uparrow r_2 \).

Lemma 5.10. Let \( \pi \) be a simple proof and \( \mathcal{P} : \alpha_1 \to t_1 \ni \alpha_2 \ldots \to t_n \ni t \) be a path in \( G(\pi) \). Then there is a \( k \in \{1, \ldots, n\} \) such that \( \text{cut}_{\alpha_k} \) is lowermost among all inferences on \( \mathcal{P} \). Furthermore, \( \forall_{\alpha_1} \) is on the right above \( \text{cut}_{\alpha_k} \) and \( \exists_{t_n} \) is on the left above \( \text{cut}_{\alpha_k} \).

Proof. We proceed by induction on \( n \). If \( n = 1 \), then \( n = k = 1 \). For the induction step consider a path \( \alpha_1 \to t_1 \ni \ldots \ni \alpha_n \to t_n \ni \alpha_{n+1} \to t_{n+1} \). By induction hypothesis, there is some \( l \in \{1, \ldots, n\} \) such that we have this configuration

\[
\begin{array}{c}
\exists_{t_n} \\
\vdots \\
\forall_{\alpha_1} \\
\vdots \\
\text{cut}_{\alpha_l} \\
\vdots \\
\end{array}
\]

As \( \alpha_{n+1} \in t_n \) we know that \( \exists_{t_n} \) must be on the right above \( \text{cut}_{\alpha_{n+1}} \). Hence, we are in one of the following two situations

\[
\begin{array}{c}
\exists_{t_n} \\
\vdots \\
\forall_{\alpha_1} \\
\vdots \\
\text{cut}_{\alpha_l} \\
\vdots \\
\end{array}
\]

or

\[
\begin{array}{c}
\exists_{t_{n+1}} \\
\vdots \\
\text{cut}_{\alpha_{n+1}} \\
\vdots \\
\exists_{t_n} \\
\vdots \\
\forall_{\alpha_1} \\
\vdots \\
\text{cut}_{\alpha_l} \\
\vdots \\
\end{array}
\]

In the first case we let \( k = n + 1 \) and in the second we let \( k = l \). In both cases \( \text{cut}_{\alpha_k} \) has the desired properties.

Lemma 5.11. Let \( \pi \) be a simple proof, let \( G(\pi) = \langle N_R, \Sigma, \varphi, P \rangle \), and let \( \beta, \alpha \in \text{EV}_c(\pi) \). If \( \beta \to \alpha \) then either \( \text{cut}_{\alpha} \uparrow \text{cut}_{\beta} \) or \( \text{cut}_{\alpha} \uparrow \text{cut}_{\beta} \) or \( \text{cut}_{\alpha} \uparrow|\uparrow \text{cut}_{\beta} \).

Proof. Since \( \beta \to \alpha \), we have a path \( \beta \to \ldots \ni \alpha \to t \) for some \( t \). By Lemma 5.10 there is a \( \gamma \), such that \( \exists_{\gamma} \uparrow \text{cut}_{\gamma} \) and \( \text{cut}_{\gamma} \uparrow \forall_{\beta} \), and such that \( \text{cut}_{\alpha} \) and \( \text{cut}_{\beta} \) are not below \( \text{cut}_{\gamma} \). Furthermore, \( \text{cut}_{\alpha} \) must be below \( \exists_{\gamma} \), and \( \text{cut}_{\beta} \) below \( \forall_{\beta} \). If \( \gamma = \beta \), then \( \text{cut}_{\alpha} \uparrow \text{cut}_{\beta} \). If \( \gamma = \alpha \), then \( \text{cut}_{\alpha} \uparrow \text{cut}_{\beta} \). And if \( \gamma \neq \beta \) and \( \gamma \neq \alpha \), then \( \text{cut}_{\alpha} \uparrow|\uparrow \text{cut}_{\beta} \).
Lemma 5.12. Let $G(\pi) = \langle N_R, \Sigma, \varphi, P \rangle$ be the grammar of a simple proof $\pi$, such that there are two paths

$$
\beta \rightarrow t \ni \gamma_0 \rightarrow s_0 \ni \gamma_1 \rightarrow s_1 \ni \ldots \rightarrow s_{n-1} \ni \gamma_n = \alpha \rightarrow s_n
$$

$$
\beta \rightarrow t \ni \delta_0 \rightarrow r_0 \ni \delta_1 \rightarrow r_1 \ni \ldots \rightarrow r_{m-1} \ni \delta_m = \alpha \rightarrow r_m
$$

such that $\gamma_0$ and $\delta_0$ occur at two different positions in $t$. Then we have one of the following two cases:

1. we have $\gamma_i = \delta_j$ for some $0 \leq i < n$ and $0 \leq j < m$, or
2. for all $0 \leq i < n$ and $0 \leq j < m$ we have $\text{cut}_\alpha \ni \text{cut}_{\gamma_i}$ and $\text{cut}_\alpha \ni \text{cut}_{\delta_j}$.

Proof. Note that because of acyclicity of $G(\pi)$, we have that $\beta \neq \gamma_i$ for all $i \leq n$ and $\beta \neq \delta_j$ for all $j \leq m$, in particular $\beta \neq \alpha$. Assume, for the moment, that $m, n > 0$; the case of one of them being zero will be treated at the very end of the proof. Then $\gamma_0 \neq \alpha$ and $\delta_0 \neq \alpha$. If $\gamma_0 = \delta_0$, we have case 1. So, assume also $\gamma_0 \neq \delta_0$. As $\beta \rightarrow t$ is a production in $G(\pi)$, the proof $\pi$ contains a formula which contains both $\gamma_0$ and $\delta_0$ hence $\forall_{\gamma_0}$ and $\forall_{\delta_0}$ are not parallel. Since we have $\text{cut}_{\gamma_0} \ni \forall_{\gamma_0}$ and $\text{cut}_{\delta_0} \ni \forall_{\delta_0}$, we also have that $\text{cut}_{\gamma_0}$ and $\text{cut}_{\delta_0}$ are not parallel. Without loss of generality, assume that $\text{cut}_{\delta_0}$ is below $\text{cut}_{\gamma_0}$. Then $\text{cut}_{\delta_0} \ni \text{cut}_{\gamma_0}$ (since $\text{cut}_{\gamma_0} \ni \text{cut}_{\delta_0}$ would entail $\forall_{\gamma_0} \ni \forall_{\delta_0}$). Since we have $\delta_0 \rightarrow \alpha$, we can apply Lemma 5.10 giving us three possibilities:

- If $\text{cut}_\alpha \ni \text{cut}_{\delta_0}$ then we have the situation

  $\vdots$

  \begin{align*}
  \exists_{s_n} & \exists_{s_0} & \forall_{\alpha} & \forall_{\gamma_0} \\
  \vdots & \vdots & \vdots & \vdots \\
  \text{cut}_\alpha & \text{cut}_{\gamma_0} & \text{cut}_{\delta_0} & \vdots \\
  \vdots & \vdots & \vdots & \vdots
  \end{align*}

  By Lemma 5.10 applied to the path $\gamma_0 \rightarrow s_n$ we have that $\text{cut}_{\delta_0}$ must coincide with $\text{cut}_{\gamma_i}$ for some $0 \leq i < n$ (since $\pi$ is a tree), so $\delta_0 = \gamma_i$ (by Observation 4.4), and we are in case 1.

- If $\text{cut}_\alpha \ni \text{cut}_{\delta_0}$ then we are in both of the following two situations:

  $\vdots$

  \begin{align*}
  \forall_{\gamma_0} & \exists_{s_n} & \text{cut}_{\gamma_0} & \text{cut}_{\delta_0} \\
  \vdots & \vdots & \vdots & \vdots \\
  \text{cut}_\alpha & \forall_{\delta_0} & \exists_{r_m} & \text{cut}_{\delta_0} & \text{cut}_\alpha & \vdots \\
  \vdots & \vdots & \vdots & \vdots & \vdots & \vdots
  \end{align*}

  Thus, by Lemma 5.10 applied to the paths $\gamma_0 \rightarrow s_n$ and $\delta_0 \rightarrow r_m$ we know that $\text{cut}_\alpha = \text{cut}_{\gamma_k} = \text{cut}_{\delta_l}$ for some $0 \leq k \leq n$ and $0 \leq l \leq m$ hence $\gamma_k = \alpha = \delta_l$. Furthermore $k = n$ and $l = m$ by acyclicity of $G(\pi)$ and assumption $\gamma_n = \alpha = \delta_m$. Now consider any $\gamma_i$ with $0 \leq i < n$. Since $\gamma_i \rightarrow \alpha$, we can apply Lemma 5.11 and get either $\text{cut}_\alpha \ni \text{cut}_{\gamma_i}$ or $\text{cut}_\alpha \ni \text{cut}_{\gamma_i}$ and $\text{cut}_\alpha \ni \text{cut}_{\gamma_i}$ and $\text{cut}_\alpha \ni \text{cut}_{\gamma_i}$ or $\text{cut}_\alpha \ni \text{cut}_{\gamma_i}$. Since by Lemma 5.10 $\text{cut}_{\gamma_i}$ must be above $\text{cut}_\alpha$, we
conclude \( \text{cut}_\alpha \rightarrow \text{cut}_\gamma \). With the same reasoning we can conclude that \( \text{cut}_\alpha \rightarrow \text{cut}_\delta \) for all \( 0 \leq j < m \). We are therefore in case 2.

- If \( \text{cut}_\alpha \rightarrow \text{cut}_\delta \) then we are in both of the following two situations:

\[
\begin{align*}
\exists_{r_m} & \qquad \forall_{\delta_0} \\
\text{cut}_\alpha & \qquad \text{cut}_\delta \\
\cdots & \qquad \text{r} \\
\end{align*}
\]

and

\[
\begin{align*}
\exists_{s_n} & \qquad \forall_{\gamma_0} \\
\text{cut}_\alpha & \qquad \text{cut}_\gamma \\
\cdots & \qquad \text{r} \\
\end{align*}
\]

By Lemma 5.10 applied to the paths \( \gamma_0 \rightarrow \ldots \rightarrow s_n \) and \( \delta_0 \rightarrow \ldots \rightarrow r_m \), the rule \( r \) coincides with \( \text{cut}_\gamma \) and \( \text{cut}_\delta \), for some \( 0 < i < n \) and \( 0 < j < m \), therefore \( \gamma_i = \delta_j \) (by Observation 4.4), and we are in case 1.

It remains to treat the case \( n = 0 \) or \( m = 0 \). If \( m = n = 0 \) then we are trivially in case 2 (there is no \( 0 \leq i < n \) or \( 0 \leq j < m \)). If \( n = 0 \) and \( m > 0 \), we can apply Lemma 5.10 to the path \( \delta_0 \rightarrow \ldots \rightarrow r_m \) and obtain an \( l \in \{0, \ldots, m\} \) such that we are in the situation

\[
\begin{align*}
\exists_{r_m} & \qquad \forall_\alpha \\
\text{cut}_\alpha & \qquad \forall_{\delta_0} \\
\cdots & \qquad \text{cut}_\delta \\
\end{align*}
\]

But by the same argument as at the beginning of the proof, we also have that \( \forall_\alpha \) and \( \forall_{\delta_0} \) cannot be in parallel (\( \alpha \) and \( \delta_0 \) both appear in \( t \)), and therefore either \( \text{cut}_{\delta_0} \rightarrow \text{cut}_\alpha \) or \( \text{cut}_\alpha \rightarrow \text{cut}_{\delta_0} \). Since \( \delta_0 \rightarrow \alpha \), the only possibility is \( \text{cut}_\alpha \rightarrow \text{cut}_{\delta_0} \), by Lemma 5.11. Thus \( \text{cut}_\alpha = \text{cut}_{\delta_1} \), and therefore \( l = m \) and we are in case 2. The case \( m = 0 \) and \( n > 0 \) is similar.

We have now finally collected together all necessary tools for describing the reduction step for contraction. \qed
Lemma 5.13. Let $\pi$ be a simple proof of a weak sequent such that $\pi$ contains a subproof $\psi$, shown on the left below,

$$
\psi = \frac{\Gamma, A, A}{\Delta, A} \frac{\Gamma, A}{\psi_1} \frac{\psi_2}{\Delta, \Delta} \text{cut}
$$

and let $\pi'$ be the proof obtained from $\pi$ from replacing $\psi$ by $\psi'$ shown on the right above, where $\rho' = [\alpha \backslash \alpha']_{\alpha \in \text{EV}(\psi_2)}$ and $\rho'' = [\alpha \backslash \alpha'']_{\alpha \in \text{EV}(\psi_2)}$ are substitutions that replace all eigenvariables in $\psi_2$ by fresh copies. Then $L(G(\pi')) = L(G(\pi))$.

Proof. Let us first show $L(G(\pi)) \subseteq L(G(\pi'))$. Write $P$ for the productions of $G(\pi)$ and $P'$ for those of $G(\pi')$. Let $F \in L(G(\pi))$ and $\mathcal{D}$ be its derivation. If the duplicated cut is quantifier-free, then $P' = P \rho' \cup P \rho''$, since the substitutions $\rho$ and $\rho'$ do not affect the eigenvariables outside $\psi_2$. Hence $\mathcal{D} \rho'$ (as well as $\mathcal{D} \rho''$) is a derivation of $F$ in $G(\pi')$. If the duplicated cut contains a quantifier, let $\alpha$ be its eigenvariable, let $t_1, \ldots, t_k$ be its terms coming from the left copy of $A$ and $t_{k+1}, \ldots, t_n$ those from the right copy of $A$ and let $Q = \{\alpha \rightarrow t_1, \ldots, \alpha \rightarrow t_n\} \subseteq P$. We then have

$$
P' = (P \setminus Q) \rho' \cup \{\alpha' \rightarrow t_1, \ldots, \alpha' \rightarrow t_k\} \cup (P \setminus Q) \rho'' \cup \{\alpha'' \rightarrow t_{k+1}, \ldots, \alpha'' \rightarrow t_n\}.
$$

If $\mathcal{D}$ does not contain $\alpha$, then $\mathcal{D} \rho'$ (as well as $\mathcal{D} \rho''$) is a derivation of $F$ in $G(\pi')$. If $\mathcal{D}$ does contain $\alpha$, then by Lemma 3.4 we can assume that it uses only one $\alpha$-production, say $\alpha \rightarrow t_1$. If $1 \leq i \leq k$, then $\mathcal{D} \rho'$ is a derivation of $F$ in $G(\pi')$ and if $k < i \leq n$, then $\mathcal{D} \rho''$ is a derivation of $F$ in $G(\pi')$.

Let us now show $L(G(\pi')) \subseteq L(G(\pi))$. Let $F$ be a formula in $L(G(\pi'))$, and let $\mathcal{D}'$ be a derivation of $F$ in $G(\pi')$. We construct $\mathcal{D} = \mathcal{D}'(\rho')^{-1}(\rho'')^{-1}$ by “undoing” the renaming of the variables in $\psi_2$. Then $\mathcal{D}$ is a derivation for $F$, using the production rules of $G(\pi)$, but possibly violating the rigidity condition.

First, recall that $\text{EV}_c(\pi) = \text{EV}(\pi)$ and observe that only non-terminals $\alpha \in \text{EV}(\psi_2)$ can violate the rigidity condition in $\mathcal{D}$: if $\beta \notin \text{EV}(\psi_2)$ violates the rigidity condition then there are $\beta$-positions $p_1, p_2$ in $\mathcal{D}$ with $F|_{p_1} \neq F|_{p_2}$ and as $\beta \rho' \rho'' = \beta$ the positions $p_1, p_2$ are also $\beta$-positions in $\mathcal{D}'$ and they violate the rigidity condition in $\mathcal{D}'$ which is a contradiction to $\mathcal{D}'$ being a $G(\pi')$-derivation.

Now define for each $\alpha \in \text{EV}(\psi_2)$ the value $\mathbf{n}(\mathcal{D}, \alpha)$ to be the number of pairs $(p_1, p_2) \in \text{Pos}(F) \times \text{Pos}(F)$ where $p_1$ and $p_2$ are $\alpha$-positions in $\mathcal{D}$ with $p_1 \neq p_2$ and $F|_{p_1} \neq F|_{p_2}$, and define $\mathbf{n}(\mathcal{D}) = \sum_{\alpha \in \text{EV}(\psi_2)} \mathbf{n}(\mathcal{D}, \alpha)$. We proceed by induction on $\mathbf{n}(\mathcal{D})$ to show that $\mathcal{D}$ can be transformed into a derivation which does no longer violate rigidity. If $\mathbf{n}(\mathcal{D}) = 0$ then $\mathcal{D}$ obeys the rigidity condition, and we are done. Otherwise there is at least one $\alpha \in \text{EV}(\psi_2)$ with $\mathbf{n}(\mathcal{D}, \alpha) > 0$. We now pick one such $\alpha$ which is minimal with respect to $<^*$ (which exists since $G(\pi)$ is acyclic). Let $p_1$ and $p_2$ be $\alpha$-positions in $\mathcal{D}$ with $p_1 \neq p_2$ and $F|_{p_1} \neq F|_{p_2}$, let $p$ be the maximal common prefix of $p_1$ and $p_2$ and let $q$ be the maximal prefix of $p$ where a production rule has been applied in $\mathcal{D}$. Due to the tree structure of $F$, the position $q$ is uniquely defined, and $q$ is a $\beta$-position for some non-terminal $\beta$, and some production rule
\[ \beta \rightarrow t \] has been applied at position \( q \) in \( \mathcal{D} \), and we have two paths:
\[
\beta \rightarrow t \ni \gamma_0 \ni s_0 \ni \gamma_1 \ni s_1 \ni \ldots \ni s_{n-1} \ni \gamma_n = \alpha \ni s_n
\]
\[
\beta \rightarrow t \ni \delta_0 \ni r_0 \ni \delta_1 \ni r_1 \ni \ldots \ni r_{m-1} \ni \delta_m = \alpha \ni r_m
\]
where \( \gamma_0 \) and \( \delta_0 \) occur at two different positions in \( t \). Thus, we can apply Lemma 5.12 giving us the following two cases:

- We have \( \gamma_i = \delta_j \) for some \( 0 \leq i < n \) and \( 0 \leq j < m \). Say \( \eta = \gamma_i = \delta_j \), and let \( p_{\eta} \) and \( p_{\delta} \) be the positions of \( \gamma_i \) and \( \delta_j \) (respectively) in \( \mathcal{D} \). Since \( \eta \ni \alpha \) we know that \( \eta \) does not violate the rigidity condition (we chose \( \alpha \) to be minimal), and therefore \( F | p_{\eta} = F | p_{\delta} = F' \). Let \( \mathcal{D}_\gamma : \gamma_i \rightarrow G(\pi) F' \) and \( \mathcal{D}_\delta : \delta_j \rightarrow G(\pi) F' \) be the two subderivations of \( \mathcal{D} \) starting in positions \( p_{\eta} \) and \( p_{\delta} \), respectively. Without loss of generality, we can assume that \( n(\mathcal{D}_\gamma) \leq n(\mathcal{D}_\delta) \). Then let \( \hat{\mathcal{D}} \) be the derivation obtained from \( \mathcal{D} \) by replacing \( \mathcal{D}_\delta \) by \( \mathcal{D}_\gamma \). Then \( \hat{\mathcal{D}} \) is still a derivation for \( F \), but \( n(\hat{\mathcal{D}}) < n(\mathcal{D}) \).

- For all \( 0 \leq i < n \) and \( 0 \leq j < m \) we have \( \text{cut}_{\eta} \ni \text{cut}_{\gamma_i} \) and \( \text{cut}_{\alpha} \ni \text{cut}_{\delta_j} \). So all inferences of the path \( \gamma_0 \ni \ldots \ni s_{n-1} \) as well as all inferences of \( \delta_0 \ni \ldots \ni r_{m-1} \) are in \( \psi_2 \). Therefore all variables of these paths are in \( \text{EV}(\psi_2) \). As \( \alpha \) violates the rigidity in \( \mathcal{D} \) one of \( p_1, p_2 \) must be a \( \alpha' \)-position and the other a \( \alpha'' \)-position in \( \mathcal{D}' \) because \( \mathcal{D}' \) does satisfy the rigidity condition. Without loss of generality we can assume that \( p_1 \) is the \( \alpha' \)-position and \( p_2 \) the \( \alpha'' \)-position. As the paths are contained completely in \( \psi_2 \) we have \( \gamma_0 \in \text{EV}(\psi_2)\rho' \) and \( \delta_0 \in \text{EV}(\psi_2)\rho'' \) which is a contradiction as no term can contain both a variable from \( \text{EV}(\psi_2)\rho' \) and one from \( \text{EV}(\psi_2)\rho'' \).

**Proof of Lemma 5.4.** By induction on the length of the reduction \( \pi \leadsto \pi' \) or \( \pi \leadsto \pi' \) respectively using one of Lemmas 5.5 5.6 5.7 or 5.8 depending on the current reduction step. \( \square \)

## 6. Skolemization and Deskolemization

In this section we will describe some results that allow one to extend the above invariance lemma to proofs of arbitrary end-sequents (including \( \forall \)-quantifiers). Carrying out the above argument directly for arbitrary end-sequents would require dealing with variable-names on the level of the grammar in order to describe the changes of eigenvariables of the \( \forall \)-quantifiers in the end-sequent. This can be avoided completely by skolemizing proofs to reduce the general case to that of weak sequents and then translating back the results by deskolemization. Skolemization and deskolemization are simple operations on the level of Herbrand-disjunctions or expansion trees [Mil87] and their use in this context suffices for our purposes. In contrast, they have surprising complexity-effects on the level of proofs, see e.g. [BHW12]. The reason why this transfer is possible is that the form of the end-sequent, and in particular the question whether it contains universal quantifiers, does not have an effect on the dynamics of cut-elimination. This observation has been well known for a long time and is apparent already in Gentzen’s consistency proof for Peano Arithmetic [Gen38] which is carried out on a (hypothetical) proof of the empty sequent as well as in the proof of the second \( \varepsilon \)-Theorem from the first \( \varepsilon \)-Theorem by deskolemization [Hil39].

Let us now first define the notion of Herbrand-disjunction precisely. We assume w.l.o.g. that in a formula every variable is bound by at most one quantifier.
**Definition 6.1.** For a given formula $F$, we write $\hat{F}$ for the formula obtained from $F$ by removing all quantifiers. Now let $x_1, \ldots, x_n$ be the existentially bound variables in $F$, and let $y_1, \ldots, y_m$ be the universally bound variables in $F$. Then any formula of the shape

$$\hat{F}[x_1/t_1, \ldots, x_n/t_n, y_1/\alpha_1, \ldots, y_m/\alpha_m]$$

where $\hat{F}$ is an arbitrary formula with $\hat{F} \leq \hat{F}$, where $t_1, \ldots, t_n$ are arbitrary terms, and where $\alpha_1, \ldots, \alpha_m$ are fresh variables, is called an instance of $F$. If $\Gamma$ is a sequent we say that a set $\mathcal{I}$ of formulas is a set of instances of $\Gamma$ if for every $I \in \mathcal{I}$ there is a $F \in \Gamma$, s.t. $I$ is instance of $F$.

Often we will work in the context of a proof $\pi$ of a sequent $\Gamma$ and consider the instances of the formulas in $\Gamma$ that are induced by $\pi$. Then the above fresh variables $\alpha_1, \ldots, \alpha_m$ will be eigenvariables of the proof and their occurrences in terms will be restricted by an acyclicity-condition, see below.

Let $\Gamma = F_1, \ldots, F_n$ be a sequent, let $\mathcal{I}$ be a set of instances of $\Gamma$, let $m_i$ be the number of quantifiers in $F_i$, and let $l_i$ be the number of instances of $F_i$ in $\mathcal{I}$. If we impose an arbitrary linear ordering on the instances of $F_i$ in $\mathcal{I}$, then a tuple $\langle i, j, k \rangle$ for $1 \leq i \leq n$ and $1 \leq j \leq m_i$ and $1 \leq k \leq l_i$ uniquely identifies the term which is substituted for the quantifier $Qx_j$ in the $k$-th instance of the formula $F_i$. We will write $t_{i,j,k}$ for this term (which could just be an eigenvariable if $Qx_j$ happens to be an $\forall$-quantifier). The $k$-th instance of $F_i$ can hence be written as $F_{i,k}[x_1/t_{i,1,k}, \ldots, x_{m_i}/t_{i,m_i,k}]$, where $x_1, \ldots, x_{m_i}$ are the bound variables in $F_i$, and $F_{i,k}$ is some formula with $F_{i,k} \leq F_i$. Such a tuple $\langle i, j, k \rangle$ is called existential position if $x_j$ is bound existentially in $F_i$, and universal position if $x_j$ is bound universally in $F_i$.

A position $\langle i_1, j_1, k_1 \rangle$ is said to dominate another position $\langle i_2, j_2, k_2 \rangle$, if $i_1 = i_2$, and $k_1 = k_2$, and the quantifier $Qx_{j_2}$ is in the scope of the quantifier $Qx_{j_1}$ in $F_i$. A set $\mathcal{I}$ of instances induces a relation $<$ on its existential positions as: $\langle i_1, j_1, k_1 \rangle < \langle i_2, j_2, k_2 \rangle$ if there is a universal position $\langle i_3, j_3, k_3 \rangle$, such that the term $t_{i_2,j_2,k_2}$ contains a variable $\alpha$ with $\alpha = t_{i_3,j_3,k_3}$ and $\langle i_1, j_1, k_1 \rangle$ dominates $\langle i_3, j_3, k_3 \rangle$. Furthermore we define the dependency relation $\ll$ on the existential positions of $\mathcal{I}$ as transitive closure of $\langle$.

**Remark 6.2.** A proof $\pi$ with the property that $H(\pi) = \mathcal{I}$ is sometimes called a sequentialization of $\mathcal{I}$. If $\mathcal{I}$ has positions $\langle i_1, j_1, k_1 \rangle$ and $\langle i_2, j_2, k_2 \rangle$ with $\langle i_1, j_1, k_1 \rangle < \langle i_2, j_2, k_2 \rangle$, then in each sequentialization of $\mathcal{I}$ the inference corresponding to $\langle i_1, j_1, k_1 \rangle$ is below that of $\langle i_2, j_2, k_2 \rangle$. In the literature on proof nets, relations like $<$ are known as jumps.

**Definition 6.3.** A set $\mathcal{I}$ of instances of $\Gamma$ is called Herbrand-disjunction of $\Gamma$ if

- the dependency relation $\ll$ of $\mathcal{I}$ is acyclic, and
- $\bigvee_{I \in \mathcal{I}} I$ is a tautology.

This notion of Herbrand-disjunction is essentially a flat (as opposed to tree-like) formulation of expansion tree proofs [Mil87]. A similar flat formulation can, for instance, be found in [BL94].

**Theorem 6.4.** $\Gamma$ is valid iff it has a Herbrand-disjunction.

**Proof Sketch.** Via translating back and forth with cut-free sequent calculus or alternatively via expansion tree proofs. \qed
Example 6.5. Let $\Gamma = \exists x (P(x) \lor \forall y P(y))$, let $\mathcal{F} = \{ P(c) \lor P(\alpha), \overline{P}(\alpha) \lor P(\beta) \}$ and fix the numbering of quantifiers and instances to be from the left to the right. Then there are the two existential positions $\langle 1, 1, 1 \rangle$ with $t_{1,1,1} = c$ and $\langle 1, 1, 2 \rangle$ with $t_{1,1,2} = \alpha$ and two universal positions $\langle 1, 2, 1 \rangle$ with $t_{1,2,1} = \alpha$ and $\langle 1, 2, 2 \rangle$ with $t_{1,2,2} = \beta$. As $\langle 1, 1, k \rangle$ dominates $\langle 1, 2, k \rangle$, we have $\langle 1, 1, 1 \rangle < \langle 1, 1, 2 \rangle$, but not the other way round because $t_{1,1,1} = c$ is variable-free. Therefore $\ll$ is acyclic. Furthermore $\mathcal{F}$ is a tautology and hence a Herbrand-disjunction.

Note that for a weak sequent $\Gamma$, the induced dependency ordering $\ll$ is empty and hence trivially acyclic. The Herbrand-disjunctions of weak sequents are therefore exactly the tautologies of instances.

Definition 6.6. Let $F[\forall y G]$ be a formula containing a universal quantifier and let $\exists x_1, \ldots, \exists x_n$ be the existential quantifiers in whose scope $\forall y$ is. Then define the Skolemization of this universal quantifier as

$$\text{sk}_1(F[\forall y G]) = F[G[y \backslash g(x_1, \ldots, x_n)]]$$

where $g$ is a fresh $n$-ary function symbol, called a Skolem function symbol. The term $g(x_1, \ldots, x_n)$ is called Skolem-term. For a formula $F$ define its Skolemization $\text{sk}(F)$ to be the iteration of $\text{sk}_1$ until no universal quantifier is left, such that no Skolem function symbol is used for two different universal quantifiers in $F$. For a sequent $\Gamma = F_1, \ldots, F_n$ define its Skolemization $\text{sk}(\Gamma) = \text{sk}(F_1), \ldots, \text{sk}(F_n)$, where no Skolem function symbol is used for two different universal quantifiers in $\Gamma$.

Remark 6.7. Sometimes the above operation on formulas is also called Herbrandization. We prefer to use the name Skolemization due to the simple duality between the satisfiability-preserving replacement of existential quantifiers and the validity-preserving replacement of universal quantifiers by new function symbols. There is no danger of confusion as, in the proof-theoretic context of this work, we are clearly dealing with validity only. This use of terminology is due to [HB39], see in particular Section 3.5.a.

The above side condition on the choice of Skolem function symbols results in a 1-1 mapping between universal quantifiers in the sequent we skolemize and the Skolem function symbols. It could be made formally more precise by equipping the $\text{sk}$-operation with such a bijection as second argument. However, for the sake of notational simplicity we refrain from doing so here.

The Skolemization of formulas and sequents can be extended to a Skolemization of proofs. When skolemizing a proof, all universal quantifiers in the end-sequent are removed and their variables are replaced by Skolem-terms. In contrast, the cut-formulas remain unchanged, more precisely:

Definition 6.8. Let $\pi$ be a proof of a sequent $\Gamma$, and let $y_1, \ldots, y_n$ be the variables that are bound by a $\forall$-quantifier in $\Gamma$. Furthermore, for each $y_i$ let $\alpha_{i,1}, \ldots, \alpha_{i,h_i}$ be the eigen-variables introduced in $\pi$ by an $\forall$-rule whose main formula is of the shape $\forall y_i A$. Then the Skolemization of the proof $\pi$, denoted by $\text{sk}(\pi)$, is the proof with end-sequent $\text{sk}(\Gamma)$ that is obtained from $\pi$ by

(1) removing all $\forall$-quantifiers binding one of $y_1, \ldots, y_n$ everywhere, and
(2) replacing each occurrence of $y_i$ (for $i \in \{1, \ldots, n\}$) and $\alpha_{i,j}$ (for $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, h_i\}$) by the corresponding Skolem-term. This term is in each case uniquely determined if we proceed from the end-sequent of $\pi$ upwards to the axioms and demand
that each rule application remains valid, or, in the case of the $\forall$-rule, becomes void (i.e.,
premise and conclusion coincide), and
(3) removing the void rule instances.

Note that $sk(\pi)$ still can contain $\forall$-quantifiers, namely those coming from a cut.

The Skolemization of a proof $\pi$ also affects the quantifier-free formulas in $\pi$ through the
replacement of eigenvariables by Skolem terms. In the context of proof Skolemization we
hence extend the notation $sk(\cdot)$ to formulas $F$ from which some (or all) $\forall$-quantifiers
have been removed; then $sk(F)$ denotes the formula obtained from skolemizing the remaining
$\forall$-quantifiers and carrying out the replacement of eigenvariables by Skolem-terms. Skolem-
ization of proofs has the following useful commutation properties.

**Lemma 6.9.** If $\pi \rightsquigarrow \pi'$ then $sk(\pi) \rightsquigarrow sk(\pi')$. If $\pi \mathcal{N}\mathcal{E} \pi'$ then $sk(\pi) \mathcal{N}\mathcal{E} sk(\pi')$.

**Proof.** By induction on the number of reductions in $\pi \rightsquigarrow \pi'$ or $\pi \mathcal{N}\mathcal{E} \pi'$, respectively,
making a case distinction on the reduction step. The most interesting case is that of the
permutation of a $\forall$-inference over a cut

\[
\frac{\psi_1}{\psi_2}
\frac{\Gamma, B[x\setminus\alpha], A}{\Gamma, \forall x B, \Delta}
\frac{\psi_1}{\psi_2}
\frac{\psi_1}{\psi_2}
\frac{\Gamma, B[x\setminus\alpha], A, \Delta}{\Gamma, \forall x B, \Delta}
\]

where the main formula of the $\forall$-inference is an ancestor of the end-sequent. This reduction
step is translated to an identity-step as Skolemization maps both of the above proofs to

\[
\frac{sk(\psi_1), sk(\forall x B), sk(A)}{sk(\psi_2), sk(\psi_2), sk(\Delta)}
\frac{sk(\psi_1), sk(\psi_2), sk(\psi_2), sk(\Delta)}{sk(\psi_1), sk(\psi_2), sk(\psi_2), sk(\Delta)}
\]

Each of the other reduction steps translates directly into exactly one reduction step in the
skolemized sequence.

**Lemma 6.10.** $L(G(sk(\pi))) = sk(L(G(\pi)))$.

**Proof.** First note that $EV_c(\pi) = EV_c(sk(\pi))$ hence $G(\pi)$ and $G(sk(\pi))$ have the same non-
terminals. Furthermore, to each $\alpha \in EV(\pi) \setminus EV_c(\pi)$ corresponds a unique Skolem-term in
$sk(\pi)$, hence to each $F \in H(\pi)$ and $\sigma \in B(\pi)$ corresponds a unique $F' \in H(sk(\pi))$ and
$\sigma' \in B(\pi)$ and therefore to each production $\alpha \rightarrow t$ in $G(\pi)$ corresponds a unique production
$\alpha \rightarrow t'$ in $G(sk(\pi))$ that is obtained from replacing eigenvariables by their respective
Skolem-terms. If $I \in sk(L(G(\pi)))$ then by Lemma 3.9 we have $I = sk(F[\alpha_1 \setminus s_1] \cdots [\alpha_n \setminus s_n])$.
Now for $\theta \rightarrow F, \alpha_1 \rightarrow s_1, \ldots, \alpha_n \rightarrow s_n$ being the productions in $G(\pi)$, letting $\theta \rightarrow F', \alpha_1 \rightarrow s'_1, \ldots, \alpha_n \rightarrow s'_n$ be the corresponding productions in $G(sk(\pi))$ we obtain
$F'[\alpha_1 \setminus s'_1] \cdots [\alpha_n \setminus s'_n] = sk(F[\alpha_1 \setminus s_1] \cdots [\alpha_n \setminus s_n])$. Thus, $sk(L(G(\pi))) \subseteq L(G(sk(\pi)))$.
For the other direction, note that every Skolem-term has at least one corresponding $\alpha \in EV(\pi) \setminus EV_c(\pi)$, and as before, this relation translates to productions. So, if $J \in L(G(sk(\pi)))$ then by Lemma 3.9 we have $J = G[\alpha_1 \setminus t_1] \cdots [\alpha_n \setminus t_n]$ for $\theta \rightarrow G, \alpha_1 \rightarrow t_1, \ldots, \alpha_n \rightarrow t_n$ being the productions in $G(sk(\pi))$. By choosing one corresponding set of productions $\theta \rightarrow G', \alpha_1 \rightarrow t'_1, \ldots, \alpha_n \rightarrow t'_n$ where Skolem-terms are replaced by the eigenvariables from which they originate we obtain $sk(G'[\alpha_1 \setminus t'_1] \cdots [\alpha_n \setminus t'_n]) = G[\alpha_1 \setminus t_1] \cdots [\alpha_n \setminus t_n]$.
As we have seen in the above proof, Skolemization can identify instances that differ only in their variable names. The reason for this ability lies in the use of variable names which can be chosen in a redundant way. These superfluous instances can also be removed by an appropriate variable renaming as shown in the following example.

**Example 6.11.** Let \( \Gamma = \exists x \forall y (\overline{P}(x, y) \lor \overline{Q}(x, y)), \exists x P(c, x) \land \exists x Q(c, x) \). Then the set of instances obtained from a sequent calculus proof that ends with an \( \land \) -inference is
\[
\mathcal{I} = \{ \overline{P}(c, \alpha) \lor \overline{Q}(c, \alpha), \overline{P}(c, \beta) \lor \overline{Q}(c, \beta), P(c, \alpha) \land Q(c, \beta) \}.
\]

Skolemizing would produce the following set of instances
\[
sk(\mathcal{I}) = \{ \overline{P}(c, f(c)) \lor \overline{Q}(c, f(c)), P(c, f(c)) \land Q(c, f(c)) \}
\]
by implicitly identifying the two formulas that become equal. A similar effect (but without using Skolemization) can be achieved by directly identifying \( \alpha \) and \( \beta \) as in
\[
\mathcal{I}[\beta \mid \alpha] = \{ \overline{P}(c, \alpha) \lor \overline{Q}(c, \alpha), P(c, \alpha) \land Q(c, \alpha) \}.
\]

We now generalize the observations made in the above example. For every Herbrand-disjunction \( \mathcal{I} \) there is a substitution \( \rho \), such that \( \mathcal{I} \rho \) is a Herbrand-disjunction having the following property: If two universal positions \( \langle i, j, k_1 \rangle \) and \( \langle i, j, k_2 \rangle \) have different variables then there is a \( j' \), such that the quantifier \( \exists x_{j'} \) dominates \( \forall x_j \) in \( F_i \) and \( t_{i, j', k_1} \neq t_{i, j', k_2} \). This follows for example from the formulation of expansion trees in [CHM12a, CHM12b] which use sets of terms for the \( \exists \) -quantifier and a single variable for the \( \forall \) -quantifier. A Herbrand-disjunction with this property is \( \alpha \) -equivalent to one with canonical variable names in the following sense.

**Definition 6.12.** Let \( \mathcal{I} \) be a set of instances. The **canonical name** of the eigenvariable of the universal position \( \langle i, j, k \rangle \) is \( \alpha_{i, j, t_1, \ldots, t_m} \) where \( t_1, \ldots, t_m \) are the terms of the existential positions that dominate \( \langle i, j, k \rangle \). The **canonical variable renaming** \( \rho_c \) of \( \mathcal{I} \) is the substitution which replaces all variable names by their canonical names.

**Remark 6.13.** Note that this relationship is significantly more complex than \( \alpha \) -equivalence, as differently named variables are identified according to certain criteria external to variable names. In particular, for some fixed \( \mathcal{I} \), there are \( \mathcal{I}_n \) of unbounded size such that \( \mathcal{I}_n \rho_c = \mathcal{I} \). This can be seen, for example, by continuing Example 6.5, take \( \mathcal{I}_n = \{ \overline{P}(c) \lor P(\alpha_1), \overline{P}(\alpha_1) \lor P(\beta_1) \mid 1 \leq i \leq n \} \).

We now turn to **deskolemization**, the inverse operation of Skolemization. In our setting, we only consider deskolemization of sequents and their instances, but not of proofs. Furthermore, we always assume that the original sequent with \( \forall \) -quantifiers is known. Hence the deskolemization of a sequent trivially replaces it by the original sequent. More interesting is the deskolemization of instances which will consist of replacing Skolem-terms by (canonically named) variables.

**Definition 6.14.** Let \( \Gamma = F_1, \ldots, F_n \) be a sequent with Skolem function symbol \( f_{i,j} \) for the universal quantifier \( \forall x_j \) in \( F_i \). Let \( \mathcal{I} \) be a set of instances of \( \Gamma \) and define its **deskolemization** \( sk^{-1}(\mathcal{I}) \) by repeating the replacement
\[
f_{i,j}(t_1, \ldots, t_m) \mapsto \alpha_{i, j, t_1, \ldots, t_m}
\]
on maximal Skolem-terms (w.r.t. the subterm ordering).
In the deskolemization of a Herbrand-disjunction, the acyclicity of the dependency relation is obtained from the acyclicity of the subterm ordering on the Skolem-terms. Conversely, during Skolemization, the Skolem-terms are well-defined due to the acyclicity of the dependency relation (see e.g. Mil87, Wel11, BHW12 for more details). We hence obtain the following properties:

**Lemma 6.15.** Let $\Gamma$ be a sequent and $\Gamma'$ be a weak sequent with $\Gamma' = sk(\Gamma)$.

1. If $\mathcal{J}$ is a Herbrand-disjunction of $\Gamma$, then $sk(\mathcal{J})$ is a Herbrand-disjunction of $sk(\Gamma)$.
2. If $\mathcal{J}'$ is a Herbrand-disjunction of $\Gamma'$, then $sk^{-1}(\mathcal{J}')$ is a Herbrand-disjunction of $\Gamma$.
3. If $\mathcal{J}$ is a Herbrand-disjunction of $\Gamma$, then $sk^{-1}(sk(\mathcal{J})) = \mathcal{J}_c$.

### 7. Herbrand-Content

**Definition 7.1.** For a simple proof $\pi$, we define its Herbrand-content as $[\pi] = L(G(\pi))\rho_c$.

Note that for a cut-free proof $\pi$ we have $[\pi] = H(\pi)\rho_c$, i.e. the Herbrand-content is nothing other than the Herbrand-disjunction of the proof after variable normalization. Also note that for a proof $\pi$ of a weak sequent we have $[\pi] = L(G(\pi))$, and hence, for a cut-free proof of a weak sequent we have $[\pi] = H(\pi)$. We can now lift the main invariance lemma, Lemma 5.4, to proofs of arbitrary end-sequents and formulate this result in terms of the Herbrand-content.

**Theorem 7.2.** If $\pi \rightsquigarrow \pi'$ is a reduction sequence of simple proofs, then $[\pi] \preceq [\pi']$. If $\pi \rightsquigarrow \pi'$ is a reduction sequence of simple proofs, then $[\pi] = [\pi']$.

**Proof.** If $\pi \rightsquigarrow \pi'$ then $sk(\pi) \rightsquigarrow sk(\pi')$ by Lemma 6.9. So, by Lemma 5.4 we have $L(G(sk(\pi))) \preceq L(G(sk(\pi')))$. By Lemma 6.10 we get $sk(L(G(\pi))) \preceq sk(L(G(\pi')))$. Using Lemma 6.15 and the observation that $sk^{-1}$ commutes with $\preceq$ we see that

$$[\pi] = L(G(\pi))\rho_c = sk^{-1}(sk(L(G(\pi)))) \preceq sk^{-1}(sk(L(G(\pi')))) = L(G(\pi'))\rho_c = [\pi']$$

The proof for $\pi \rightsquigarrow \pi'$ is step-by-step the same, replacing $\preceq$ by $\succeq$.

**Corollary 7.3.** If $\pi \rightsquigarrow \pi'$ is a reduction sequence of simple proofs and $\pi'$ is cut-free, then $H(\pi')\rho_c \preceq [\pi]$.

**Proof.** This is a direct consequence of Theorem 7.2.

This corollary shows that $[\pi]$ is an upper bound on the Herbrand-disjunctions obtainable by cut-elimination from $\pi$. Let us now compare this result with another upper bound that has previously been obtained in Het10. To that aim let $G_0(\pi)$ denote the regular tree grammar underlying $G(\pi)$ which can be obtained by setting all non-terminals to non-rigid. In this notation, a central result of Het10, adapted to this paper’s setting is

**Theorem 7.4.** Let $\pi$ be a proof of a formula of the shape $\exists x_1 \ldots \exists x_n A$ with $A$ quantifier-free, and let $\pi \rightsquigarrow \pi'$ with $\pi'$ cut-free. Then $H(\pi') \subseteq L(G_0(\pi))$.

While the Theorem 7.4 applies also to non-simple proofs, Corollary 7.3 is stronger in several respects:

First, the size of the Herbrand-content is by an exponential smaller than the size of the bound given by Theorem 7.4. Indeed, it is a straightforward consequence of Lemma 8.9.
that the language of a totally rigid acyclic tree grammar with \( n \) production rules is bound by \( n^n \) but on the other hand:

**Proposition 7.5.** There is an acyclic regular tree grammar \( G \) with \( 2n \) productions and \( |L(G)| = n^n \).

**Proof.** Let \( f \) be an \( n \)-ary function symbol, then the productions \( \alpha_0 \to f(\alpha_1, \ldots, \alpha_1), \ldots, \alpha_{n-1} \to f(\alpha_n, \ldots, \alpha_n) \) create a tree with \( n^n \) leaves. Let \( c_1, \ldots, c_n \) be terminal symbols, then by adding the productions \( \alpha_n \to c_1, \ldots, \alpha_n \to c_n \) we obtain the desired grammar \( G \).

Secondly, the class of totally rigid acyclic tree grammars can be shown to be in exact correspondence with the class of simple proofs in the following sense. Not only can we use a totally rigid acyclic tree grammar to simulate the process of cut-elimination, we can also—in the other direction—use cut-elimination to simulate the process of calculating the language of a grammar. It is shown in [Het12a] how to transform an arbitrary acyclic totally rigid tree grammar \( G \) into a simple proof that has a \( \lll \) normal form whose Herbrand-disjunction is essentially the language of \( G \).

The third and—for the purposes of this paper—most important difference is that the bound of Corollary 7.3 is **tight** in the sense that it can actually be reached by a cut-elimination strategy, namely \( \lll \). In fact, an even stronger statement is true: not only is there a normal form of \( \lll \) that reaches the bound but all of them do. This property leads naturally to the following confluence result for classical logic.

**Definition 7.6** (Herbrand-confluence). A relation \( \lll \) on a set of proofs is called **Herbrand-confluent** if \( \pi \lll \pi_1 \lll \pi_2 \) with \( \pi_1 \lll \pi_2 \) being normal forms for \( \lll \) implies that \( H(p) = H(p_1) = H(p_2) \).

**Corollary 7.7.** The relation \( \lll \) is Herbrand-confluent on the set of simple proofs.

**Proof.** This is a direct consequence of Theorem 7.2.

How does this result fit together with \( \lll \) being neither confluent nor strongly normalizing? In fact, note that it is possible to construct a simple proof which permits an infinite \( \lll \) reduction sequence from which one can obtain normal forms of arbitrary size by bailing out from time to time. This can be done by building on the propositional double-contraction example found e.g. in [DJS97, Gal93, Urb00] and in a similar form in [Zuc74]. While these infinitely many normal forms do have pairwise different Herbrand-disjunctions when regarded as multisets, Corollary 7.7 shows that as sets they are all the same. This set-character of Herbrand-disjunctions is assured by using canonical variable names (or equivalently: Skolemization) and thus identifying repeated instances. This observation shows that the lack of strong normalization is taken care of by using sets instead of multisets as data structure. But what about the lack of confluence? Results like [BH11] and [Het12b] show that the number of \( \lll \) normal forms with different Herbrand-disjunctions can be enormous. On the other hand we have just seen that \( \lll \) induces only a **single** Herbrand-disjunction: \( \lll p \). The relation between \( \lll p \) and the many Herbrand-disjunctions induced by \( \lll \) is explained by Corollary [7.3]. \( \lll p \) contains them all.
8. Conclusion

We have shown that non-erasing cut-elimination for the class of simple proofs is Herbrand-confluent. While there are different and possibly infinitely many normal forms, they all induce the same Herbrand-disjunction. This result motivates the definition of this unique Herbrand-disjunction as Herbrand-content of the proof with cut.

As future work, the authors plan to extend this result to arbitrary first-order proofs. The treatment of blocks of quantifiers is straightforward: the rigidity condition must be changed to apply to vectors of non-terminals. Treating quantifier alternations is more difficult: the current results suggest to use a stack of totally rigid tree grammars, each layer of which corresponds to one layer of quantifiers (and is hence acyclic). Concerning further generalizations, note that the method of describing a cut-free proof by a tree language is applicable to any proof system with quantifiers that has a Herbrand-like theorem, e.g., even full higher-order logic as in [Mil87]. The difficulty consists in finding an appropriate type of grammars.

Given the wealth of different methods for the extraction of constructive content from classical proofs, what we learn from our work about the class of simple proofs is this: the first-order structure possesses (in contrast to the propositional structure) a unique and canonical unfolding. The various extraction methods hence do not differ in the choice of how to unfold the first-order structure but only in choosing which part of it to unfold. We therefore see that the effect of the underspecification of algorithmic detail in classical logic is redundancy.

Acknowledgments

The authors would like to thank Paul-André Melliès for helpful comments on this work. The first author was supported by a Marie Curie Intra European Fellowship within the 7th European Community Framework Programme, by the projects I603, P22028 and P25160 of the Austrian Science Fund (FWF) and the WWTF Vienna Research Group 12-04.

References


