Journal of Automata, Languages and Combinatorics **u** (v) w, x–y

ON THE NUMBER OF BROKEN DERIVED TERMS OF A RATIONAL EXPRESSION

Pierre-Yves Angrand
Télécom ParisTech
46 rue Barrault, 75634 Paris Cedex 13, France
e-mail: `angrand@enst.fr@telecom-paristech.fr`

Sylvain Lombardy
Institut Gaspard-Monge, Université Paris-Est Marne-la-Vallée
77454 Marne-la-Vallée Cedex 2, France
e-mail: `Sylvain.Lombardy@univ-paris-est.fr`

and

Jacques Sakarovitch
LTCI, CNRS / Télécom ParisTech
46 rue Barrault, 75634 Paris Cedex 13, France
e-mail: `sakarovitch@telecom-paristech.fr`

ABSTRACT

Bounds are given on the number of broken derived terms (a variant of Antimirov's 'partial derivatives') of a rational expression E. It is shown that this number is less than or equal to $2\ell(\mathsf{E})+1$ in the general case, where $\ell(\mathsf{E})$ is the literal length of the expression E, and that the classical bound $\ell(\mathsf{E})+1$ which holds for partial derivatives also holds for broken derived terms if E is in star normal form.

In a second part of the paper, the influence of the bracketing of an expression on the number of its derived terms is also discussed.

Keywords: regular expressions, rational expressions, derivatives, derivation of expressions.

1. Introduction

The transformation of a rational (regular) expression into an automaton is as old as automata and regular language theory. In [6], Glushkov built an automaton, the *position* automaton of an expression E, with exactly $\ell(\mathsf{E})+1$ states, where $\ell(\mathsf{E})$ is the literal length of the expression E. With the notion of *derivatives*, Brzozowski

[0] Full version of the paper published in the Proceedings of the 11th International Workshop on Descriptional Complexity of Formal Systems, held in Magdeburg, Germany, July 6-9, 2009.

succeeded in lifting the Myhill-Nerode theorem to the level of expressions [3]. Since the result is a deterministic automaton the same kind of bound cannot hold. In [1], Antimirov proposed to compute *partial derivatives* — which we call *derived terms* here — rather than derivatives and built an automaton which is at most as big as the position automaton. It has even been shown that this automaton is not only smaller than, but also a *quotient* of the position automaton [4].

In [7], the notion of *breaking derivation* was introduced as a possible variant of Antimirov's derivation. Like Antimirov's derivation, the breaking derivation is ultimately intended for building an automaton that recognises the language denoted by the expression. And as its name indicates, the breaking derivation will break the derived terms into pieces, leading thus to an automaton which is likely to have more states than the one built with Antimirov's procedure; it is easy to produce examples where this effectively does happen. The motivation for being interested in an apparently less efficient construction comes from an earlier work of the authors (*cf.* [8, 9]) where it was shown that broken derived terms give very good results when the expression under derivation is obtained by the state elimination method applied to an automaton. Somehow, and to some extent, the structure of the automaton is coded by the the state elimination method into the expression it produces, and the breaking derivation is a good candidate for decoding the structure. These considerations are motivated by the search of a method that would be reversible, in the sense it could find an expression from an automaton and then recover the automaton from the expression.

The natural bound that holds for the derived terms no longer holds in general for broken derived terms and might lead to the computation of an automaton bigger than the automaton of derived terms – and it is not even a quotient of the position automaton. The question of finding a bound for the size of the set of broken derived terms was left open [7, Remark 13]. This paper addresses this question and proves the following two results, the second one giving for a natural class of expressions the same bound as for derived terms. Moreover, examples are given which show that both bounds are tight.

Theorem 1 *The number of broken derived terms of an expression* E *whose starred subexpressions are not constant is bounded by* $2\,\ell(\mathsf{E})+1$.

Theorem 2 *The number of broken derived terms of an expression* E *in star normal form is bounded by* $\ell(\mathsf{E})+1$.

The star normal form is defined in [2] by Brüggemann-Klein in order to compute the position automaton efficiently in quadratic time. It amounts to avoiding the star operator on expressions whose constant term is not 1. It is remarkable that this classical notion pops up again in our problem. For the purpose of the proof of Theorem 2, we give a recursive definition of star normal form which is slightly different from the original one, and which is more suitable for proofs by induction on the depth of expressions.

In Section 2 we recall the definition of broken derived terms. The bounds are established in Section 3. In the last section, we address another problem related to the number of (broken) derived terms. We show that the results of the derivation and of the broken derivation depend on the bracketing of the rational expression. More precisely, the left bracketing (of concatenation) in an expression yields a number of derived and broken derived terms that is always less than or equal to the one obtained by the right bracketing.

The proof of two already known propositions, in particular the one of Brüggemann-Klein, are given for sake of completeness but put into an appendix.

2. Definition of the broken derived terms

Let us first briefly recall basic definitions of rational expressions and some of their properties that we use in the following sections. We also recall the definition of the derivation as Antimirov defined it in [1] and finally we give the definition of the broken derived terms (and of the automaton which they allow to build).

2.1. Basic notions

In the sequel, A is a finite alphabet and A^* the free monoid generated by A. The empty word, denoted by 1_{A^*}, is the identity of A^*. We denote by A^+ the set of non-empty words: $A^+ = A^* \setminus \{1_{A^*}\}$.

For any subset L, *i.e.,* language, of A^*, and any word f in A^*, the *(left) quotient* of L by f is defined as $f^{-1}L = \{g \in A^* \mid f\,g \in L\}$. The *constant term* of a language L, is the Boolean value $\mathsf{c}\,(L)$ which is equal to 1 if 1_{A^*} belongs to L and to 0 if it does not.

The set of *rational expressions over* A, denoted by $\mathsf{RatE}(A)$, is the set of well-formed formulas built inductively from the *constants* $\mathsf{0}$ and $\mathsf{1}$ and the letters a in A as *atomic formulas* and with two binary operators $+$ and $\cdot$ and one unary operator $*$: if E and F are rational expressions, so are $(\mathsf{E}+\mathsf{F})$, $(\mathsf{E}\cdot\mathsf{F})$, and (E^*). We often write simply *expression* instead of *rational expression.*

With every rational expression E is associated a language of A^* which is called the *language denoted by* E and we write it $L\,(\mathsf{E})$. Two expressions are *equivalent* if they denote the same language.

It is very common to introduce a *precedence relation* between operators: '$\,* > \cdot > +\,$' which allows to save parentheses in the writing of expressions — *e.g.* $\mathsf{E}+\mathsf{F}\cdot\mathsf{G}^*$ is an unambiguous writing for the expression $(\mathsf{E}+(\mathsf{F}\cdot(\mathsf{G}^*)))$ — but one should be aware that for instance $(\mathsf{E}\cdot(\mathsf{F}\cdot\mathsf{G}))$ and $((\mathsf{E}\cdot\mathsf{F})\cdot\mathsf{G})$ are two equivalent *but distinct* expressions. The operation that we shall study in this paper gives indeed different results on these two expressions, as we shall see in the last section.

Even if for the sake of our purpose and the correct definition of the derivation, we distinguish between expressions that seem to be so obviously equivalent, all the computations on expressions that will be defined below are performed modulo a set of seven identities, that we call *trivial identities*:

$$\mathsf{E}+\mathsf{0}\equiv\mathsf{E},\quad \mathsf{0}+\mathsf{E}\equiv\mathsf{E},\quad \mathsf{E}\cdot\mathsf{0}\equiv\mathsf{0},\quad \mathsf{0}\cdot\mathsf{E}\equiv\mathsf{0},\quad \mathsf{E}\cdot\mathsf{1}\equiv\mathsf{E},\quad \mathsf{1}\cdot\mathsf{E}\equiv\mathsf{E},\quad \mathsf{0}^*\equiv\mathsf{1}. \tag{\textbf{T}}$$

An expression is said to be *reduced* if it contains no subexpression which is a left-hand side of one of the above identities; in particular, $\mathsf{0}$ does not appear in a non-zero reduced expression. It is not necessary to set up a full theory of rational identities in order to understand that any expression H can be rewritten in an equivalent reduced expression H', and that this H' is unique and independant from the way the rewriting is conducted.

In the sequel, most of the operators defined on expressions are implicitely, or explicitely, extended additively to sets of expressions. For instance, we have:

$$\forall X \subseteq \mathsf{RatE}(A) \qquad L(X) = \bigcup_{\mathsf{E}\in X} L(\mathsf{E})\,. \tag{1}$$

Like any formula, a rational expression E can be canonically represented by a tree, which is called *the syntactic tree* of E. Let us denote by $\ell(\mathsf{E})$ the *literal length* of the expression E (*i.e.*, the number of all occurences of letters from A in E) and by $\mathsf{d}(\mathsf{E})$ the *depth* of E which is defined as the depth – or height – of the syntactic tree of the expression. We call an expression of zero literal length a *constant expression.*

Definition 1 *Let* E *be a rational expression. The* constant term *of* E*, written* $\mathsf{c}(\mathsf{E})$*, is the Boolean value defined as follows:*

$$\mathsf{c}(\mathsf{0}) = 0, \quad \mathsf{c}(\mathsf{1}) = 1, \quad \forall a \in A \quad \mathsf{c}(a) = 0,$$
$$\mathsf{c}(\mathsf{F}+\mathsf{G}) = \mathsf{c}(\mathsf{F}) \vee \mathsf{c}(\mathsf{G})\,, \quad \mathsf{c}(\mathsf{F}\cdot\mathsf{G}) = \mathsf{c}(\mathsf{F}) \wedge \mathsf{c}(\mathsf{G})\,, \quad \mathsf{c}(\mathsf{F}^*) = 1.$$

The operator ‘constant term’ is extended to any set X of expressions:

$$\mathsf{c}(X) = \bigvee_{\mathsf{E}\in X} \mathsf{c}(\mathsf{E})\,. \tag{2}$$

The notions of constant term of expressions and of languages are consistent, as it is stated in the following property.

Property 1 *The constant term of an expression* E *is equal to* 1 *if, and only if, the empty word* 1_{A^*} *is in the language* $L(\mathsf{E})$*, that is,* $\mathsf{c}(\mathsf{E}) = \mathsf{c}(L(\mathsf{E}))$.

2.2. Derived terms

We now define what we call the derived terms of an expression, which are the *partial derivatives* of [1]. We also gives in this subsection an inductive procedure to compute directly the set of derived terms.

Definition 2 ([1]) *Let* E *be a rational expression over* A *and let* a *be a letter in* A*. The* $\mathbb{B}$-derivation[1] *of* E *with respect to* a*, denoted* $\frac{\partial}{\partial a}\mathsf{E}$*, is* a set of rational expressions

[1] We call it ‘$\mathbb{B}$-derivation’ for two reasons. First in order to avoid confusion with the derivatives defined by Brzozowski, and second because the formulae depend on the semiring of weights and can be defined for other semirings (*cf.* [7]). Here the weight semiring is the Boolean semiring.

over A, *recursively defined by*

$$\frac{\partial}{\partial a}0 \;=\; \frac{\partial}{\partial a}1 \;=\; \emptyset, \qquad \forall b \in A \quad \frac{\partial}{\partial a}b \;=\; \begin{cases} \{1\} & \textit{if} \quad b = a\,, \\ \emptyset & \textit{otherwise}, \end{cases},$$

$$\frac{\partial}{\partial a}(\mathsf{F}+\mathsf{G}) \;=\; \frac{\partial}{\partial a}\mathsf{F} \,\cup\, \frac{\partial}{\partial a}\mathsf{G}, \tag{3}$$

$$\frac{\partial}{\partial a}(\mathsf{F}\cdot\mathsf{G}) \;=\; \left(\frac{\partial}{\partial a}\mathsf{F}\right)\cdot\mathsf{G} \,\cup\, \mathsf{c}\,(\mathsf{F})\,\frac{\partial}{\partial a}\mathsf{G}, \tag{4}$$

$$\frac{\partial}{\partial a}(\mathsf{F}^*) \;=\; \left(\frac{\partial}{\partial a}\mathsf{F}\right)\cdot\mathsf{F}^*. \tag{5}$$

Equation (4) should be understood as

$$\frac{\partial}{\partial a}(\mathsf{F}\cdot\mathsf{G}) \;=\; \begin{cases} \left(\frac{\partial}{\partial a}\mathsf{F}\right)\cdot\mathsf{G} \,\cup\, \frac{\partial}{\partial a}\mathsf{G}, & \text{if} \quad \mathsf{c}\,(\mathsf{F}) = 1\,, \\ \left(\frac{\partial}{\partial a}\mathsf{F}\right)\cdot\mathsf{G}, & \text{if} \quad \mathsf{c}\,(\mathsf{F}) = 0\,. \end{cases} \tag{6}$$

That is, the product xX of a set X by a Boolean value x is $xX = X$ if $x = 1$ and $xX = \emptyset$ if $x = 0$.

The induction implied by Equations (3–5) should be interpreted by extending derivation additively (as are always derivation operators) and by distributing (on the right) the $\cdot$ operator over sets as well:

$$\frac{\partial}{\partial a}X = \bigcup_{\mathsf{E}\in X}\frac{\partial}{\partial a}\mathsf{E}, \qquad\qquad (X)\cdot\mathsf{F} = \bigcup_{\mathsf{E}\in X}(\mathsf{E}\cdot\mathsf{F}). \tag{7}$$

Finally, every operation on rational expressions is computed modulo the trivial identities (**T**).

Example 1 Let $\mathsf{E}_1 = (a^* + b^*)\cdot(a\cdot(a^*+b^*)) = \mathsf{F}_1\cdot(a\cdot\mathsf{F}_1)$, with $\mathsf{F}_1 = (a^*+b^*)$. The $\mathbb{B}$-derivation of E_1 by a and b yields:

$$\frac{\partial}{\partial a}\mathsf{E}_1 = \left(\frac{\partial}{\partial a}\mathsf{F}_1\right)\cdot(a\cdot\mathsf{F}_1) \,\cup\, \mathsf{c}\,(\mathsf{F}_1)\,\frac{\partial}{\partial a}(a\cdot\mathsf{F}_1) = \{a^*\cdot(a\cdot\mathsf{F}_1)\,,\,\mathsf{F}_1\}.$$

$$\frac{\partial}{\partial b}\mathsf{E}_1 = \left(\frac{\partial}{\partial b}\mathsf{F}_1\right)\cdot(a\cdot\mathsf{F}_1) \,\cup\, \mathsf{c}\,(\mathsf{F}_1)\,\frac{\partial}{\partial b}(a\cdot\mathsf{F}_1) = \{b^*\cdot(a\cdot\mathsf{F}_1)\}.$$

The $\mathbb{B}$-derivation of an expression E over A with respect to a non-empty word f of A^+ is defined by induction on the length of f: either $f = a$ is a letter of A and $\frac{\partial}{\partial f}\mathsf{E}$ is defined above, or f is of length greater than 1: $f = g\,a$ with g in A^+ and a in A and $\frac{\partial}{\partial f}\mathsf{E}$ is defined by

$$\frac{\partial}{\partial ga}\mathsf{E} = \frac{\partial}{\partial a}\left(\frac{\partial}{\partial g}\mathsf{E}\right). \tag{8}$$

Definition 3 *Let* E *be a rational expression over* A*. We call every expression that belongs to* $\frac{\partial}{\partial f}\,\mathsf{E}$ *for some word* f *of* A^+*, a* true derived term *and we write* $\mathrm{TD}\,(\mathsf{E})$ *for the set of true derived terms of* E*:*

$$\mathrm{TD}\,(\mathsf{E}) = \bigcup_{f \in A^+} \frac{\partial}{\partial f}\,\mathsf{E}. \tag{9}$$

The set $\mathrm{D}\,(\mathsf{E}) = \mathrm{TD}\,(\mathsf{E}) \cup \{\mathsf{E}\}$ *is the set of* derived terms *of* E*.*[2]

Example 2 (*Ex. 1 continued*) We have:

$$\begin{aligned}
\frac{\partial}{\partial aa}\,\mathsf{E}_1 &= \frac{\partial}{\partial a}\,(\{a^* \cdot (a \cdot \mathsf{F}_1)\,,\, \mathsf{F}_1\}) = \{a^* \cdot (a \cdot \mathsf{F}_1)\,,\, \mathsf{F}_1\} \cup \{a^*\},\\
\frac{\partial}{\partial ab}\,\mathsf{E}_1 &= \frac{\partial}{\partial b}\,(\{a^* \cdot (a \cdot \mathsf{F}_1)\,,\, \mathsf{F}_1\}) = \{b^*\},\\
\frac{\partial}{\partial ba}\,\mathsf{E}_1 &= \frac{\partial}{\partial a}\,(\{b^* \cdot (a \cdot \mathsf{F}_1)\}) = \{\mathsf{F}_1\},\\
\frac{\partial}{\partial bb}\,\mathsf{E}_1 &= \frac{\partial}{\partial b}\,(\{b^* \cdot (a \cdot \mathsf{F}_1)\}) = \{b^* \cdot (a \cdot \mathsf{F}_1)\}.
\end{aligned}$$

No new terms are found by derivation with respect to any longer word and the process terminates. Hence we have $\mathrm{TD}\,(\mathsf{E}_1) = \{a^* \cdot (a \cdot \mathsf{F}_1)\,,\, b^* \cdot (a \cdot \mathsf{F}_1)\,,\, \mathsf{F}_1\,,\, a^*\,,\, b^*\}$ and then:

$$\mathrm{D}\,(\mathsf{E}_1) = \{\mathsf{E}_1\,,\, a^* \cdot (a \cdot \mathsf{F}_1)\,,\, b^* \cdot (a \cdot \mathsf{F}_1)\,,\, \mathsf{F}_1\,,\, a^*\,,\, b^*\}.$$

Although we shall not make explicit use of it, it is worth mentioning that the derived terms of an expression E allow to build an automaton $\mathcal{A}_{\mathsf{E}}$ that recognises $L\,(\mathsf{E})$: its states are the derived terms of E, the initial state is the expression E itself, the final states are those derived terms whose constant term is 1 and there is a transition labeled by a between the derived terms F and G if, and only if, G belongs to $\frac{\partial}{\partial a}\,\mathsf{F}$.

The key to the proof that $\mathcal{A}_{\mathsf{E}}$ recognises $L\,(\mathsf{E})$ is the fact that the *derivation* of an expression with respect to a word corresponds to the quotient of the denoted language by the same word, in the sense that the following holds:

$$\forall f \in A^+\,,\ \forall \mathsf{E} \in \mathsf{RatE}(A) \qquad L\left(\frac{\partial}{\partial f}\,\mathsf{E}\right) = f^{-1} L\,(\mathsf{E})\,. \tag{10}$$

And as the language denoted by a non constant rational expression contains at least one non-empty word, we obtain the following property.

Property 2 *Every non-constant rational expression has at least one true derived term whose constant term is* 1*, that is:*

$$\forall \mathsf{E} \in \mathsf{RatE}(A) \qquad \ell(\mathsf{E}) \neq 0 \quad \Longrightarrow \quad \mathsf{c}\,(\mathrm{TD}\,(\mathsf{E})) = 1. \tag{11}$$

[2] Derived terms were called *partial derivatives* by Antimirov in [1]. We have already explained in [10] the reason for this renaming: first the (Brzozowski) derivative expressions of E were already 'partial' inasmuch as they are the result of a derivation with respect to *one* of the letters, and because 'partial derivatives' further overloads an established mathematical term.

Finally, and to make the picture complete, let us state the result that motivated the construction of $\mathcal{A}_{\mathsf{E}}$ by Antimirov: to build a hopefully small automaton that recognises $L(\mathsf{E})$, at least one not bigger than the position automaton.

Theorem 3 ([1]) *Let* E *be a rational expression. The number of derived terms of* E*, and thus the number of states of* $\mathcal{A}_{\mathsf{E}}$*, is finite and less than, or equal to,* $\ell(\mathsf{E})+1$.

Example 3 (*Ex. 1 continued*) Figure 1 shows the derived term automaton of E_1.

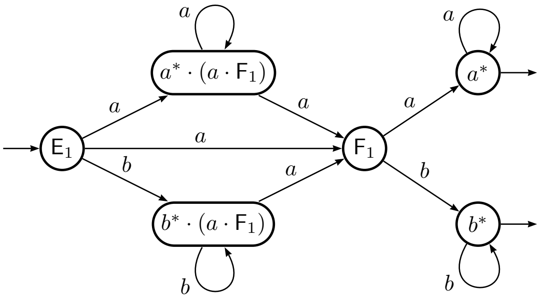
a
$a^* \cdot (a \cdot \mathsf{F}_1)$
a
a
a
a^*
a
E_1
a
F_1
b
a
b
$b^* \cdot (a \cdot \mathsf{F}_1)$
b^*
b
b

Figure 1: The derived term automaton of E_1

A very important property of derived terms, on which all the results of this paper are based, is that the set $\mathrm{D}(\mathsf{E})$ — the set $\mathrm{TD}(\mathsf{E})$ indeed — can be computed by a direct induction on the depth of E, without reference to the derivation operator.[3]

Proposition 4 ([7]) *Let* F *and* G *be two rational expressions. Then, the following holds.*

$$\mathrm{TD}((\mathsf{F}+\mathsf{G})) = \mathrm{TD}(\mathsf{F}) \cup \mathrm{TD}(\mathsf{G}), \tag{12}$$

$$\mathrm{TD}((\mathsf{F}\cdot\mathsf{G})) = (\mathrm{TD}(\mathsf{F}))\cdot\mathsf{G} \cup \mathrm{TD}(\mathsf{G}), \tag{13}$$

$$\mathrm{TD}((\mathsf{F}^*)) = (\mathrm{TD}(\mathsf{F}))\cdot\mathsf{F}^*. \tag{14}$$

These three equations follow from the application of the inductive definition of the derivation by words (Equation (8)) to the derivation of a sum, a product, and a star, of an expression. The complete computations, somewhat tedious, not in line with the rest of the paper, and which can be infered from the more general formulas given in [7], are postponed to the appendix.

Corollary 5 *The set of derived terms* $\mathrm{D}((\mathsf{E}))$ *of any rational expression* E *over* A *can be computed by induction on the depth of* E *by using Equations (12)–(14) and the base clauses:*

$$\mathrm{TD}(\mathsf{0}) = \emptyset, \quad \mathrm{TD}(\mathsf{1}) = \emptyset, \quad \forall a \in A \quad \mathrm{TD}(a) = \{\mathsf{1}\}. \tag{15}$$

It follows in particular that for every constant expression E, $\mathrm{TD}(\mathsf{E}) = \emptyset$.

[3] It is even the way the set $\mathrm{D}(\mathsf{E})$ *should be defined* for weighted rational expressions (*cf.* [7]).

2.3. Broken derived terms

The broken derived terms, which are the subject of this paper, were first defined by the authors (in [7]) for *weighted rational expressions* as a possible variant of the derived terms. We have already explained in the introduction that we have used them, for *Boolean expressions*, in order to set up a method that would be the converse of the state elimination method. We have not been able to solve completely this last problem, but we have shown, in [8], that the broken derived terms will most probably be part of the solution. We have then discovered in this paper a flaw whose origin was an inadequate definition of the breaking derivation; this has been corrected in [9].[4] The definition of the breaking derived terms makes use of the following notations.

Let X be a *set of expressions*. We denote by δ_X the Boolean value equal to 1 if the expression $\mathsf{1}$ belongs to X and to 0 otherwise. We denote by X_p the set $X_\mathsf{p} = X \setminus \{\mathsf{1}\}$. For instance, $(X \cup Y)_\mathsf{p} = X_\mathsf{p} \cup Y_\mathsf{p}$ and $\delta_{X_\mathsf{p}} = \mathsf{0}$ for any sets X and Y. It is also immediate to verify the following three identities.

$$\forall X \subseteq \mathsf{RatE}(A) \qquad \frac{\partial}{\partial a} X_\mathsf{p} = \frac{\partial}{\partial a} X, \quad \mathsf{c}\,(X_\mathsf{p}) \vee \delta_X = \mathsf{c}\,(X)\,, \tag{16}$$

$$\text{and} \qquad X_\mathsf{p} \cup \delta_X\{\mathsf{1}\} = X. \tag{17}$$

We then define a new operation on rational expressions, which we denote by $\mathrm{B}\,()$, and which, roughly speaking, consists in decomposing an expression into a set of expressions whose left factor is not a sum.

Definition 4 *The set* $\mathrm{B}\,(\mathsf{E})$ *of* broken terms *of a rational expression* E *over* A *is the* set of expressions *inductively defined as follows:*

$$\mathrm{B}\,(\mathsf{0}) = \{\mathsf{0}\}, \quad \mathrm{B}\,(\mathsf{1}) = \{\mathsf{1}\}, \quad \forall a \in A \quad \mathrm{B}\,(a) = \{a\},$$

$$\mathrm{B}\,(\mathsf{F}+\mathsf{G}) = \mathrm{B}\,(\mathsf{F}) \cup \mathrm{B}\,(\mathsf{G})\,, \tag{18}$$

$$\mathrm{B}\,(\mathsf{F}\cdot\mathsf{G}) = (\mathrm{B}\,(\mathsf{F}))_\mathsf{p}\cdot\mathsf{G} \cup \delta_{\mathrm{B}(\mathsf{F})}\mathrm{B}\,(\mathsf{G})\,, \tag{19}$$

$$\mathrm{B}\,(\mathsf{F}^*) = \{\mathsf{F}^*\}. \tag{20}$$

By definition, the breaking operator is additive:

$$\forall X \subseteq \mathsf{RatE}(A) \qquad \mathrm{B}\,(X) = \bigcup_{\mathsf{E}\in X} \mathrm{B}\,(\mathsf{E})\,. \tag{21}$$

And it is immediate to check that it is idempotent.

Example 4 (*Ex. 1 continued*) The breaking of E_1 and F_1 gives:

$$\mathrm{B}\,(\mathsf{E}_1) = \{a^*\cdot(a\cdot\mathsf{F}_1)\,,\, b^*\cdot(a\cdot\mathsf{F}_1)\}, \quad \mathrm{B}\,(\mathsf{F}_1) = \{a^*\,,\, b^*\}.$$

[4] The publication of [9] has been much delayed for several reasons, and the correct definition of broken derived terms appears for the first time in print in the present paper.

Definition 5 *The* breaking $\mathbb{B}$-derivation *of a rational expression* E *over* A *with respect to a letter* a *in* A *is defined as:*

$$\frac{\partial_{\mathsf{b}}}{\partial a}\,\mathsf{E} = \mathrm{B}\left(\frac{\partial}{\partial a}\,\mathsf{E}\right). \tag{22}$$

The breaking $\mathbb{B}$*-derivation with respect to an non-empty word is defined by induction on the length of the words:*

$$\frac{\partial_{\mathsf{b}}}{\partial fa}\,\mathsf{E} = \frac{\partial_{\mathsf{b}}}{\partial a}\left(\frac{\partial_{\mathsf{b}}}{\partial f}\,\mathsf{E}\right). \tag{23}$$

We call every rational expression that belongs to $\dfrac{\partial_{\mathsf{b}}}{\partial f}\,\mathsf{E}$*, for some word* f *in* A^+*, a* true broken derived term *of* E *and we write* $\mathrm{TBD}\left(\mathsf{E}\right)$ *for the set of true broken derived terms of* E*:*

$$\mathrm{TBD}\left(\mathsf{E}\right) = \bigcup_{f\in A^+} \frac{\partial_{\mathsf{b}}}{\partial f}\,\mathsf{E}. \tag{24}$$

The set of broken derived terms*,* $\mathrm{BD}\left(\mathsf{E}\right)$*, is defined by:* $\mathrm{BD}\left(\mathsf{E}\right) = \mathrm{TBD}\left(\mathsf{E}\right) \,\cup\, \mathrm{B}\left(\mathsf{E}\right).$

It is easy to check (using (16)) that for any rational expression E over A we have:

$$\forall a \in A \quad \frac{\partial}{\partial a}\,\mathrm{B}\left(\mathsf{E}\right) = \frac{\partial}{\partial a}\,\mathsf{E}, \qquad \text{and thus} \qquad \forall f \in A^+ \quad \frac{\partial_{\mathsf{b}}}{\partial f}\,\mathsf{E} = \mathrm{B}\left(\frac{\partial}{\partial f}\,\mathsf{E}\right), \tag{25}$$

which in turn implies the following property.

Property 3 *The broken derived terms and true broken derived terms of an expression* E *are obtained by 'breaking' the derived terms and true derived terms of* E*, that is:*

$$\forall \mathsf{E} \in \mathsf{RatE}(A) \qquad \mathrm{BD}\left(\mathsf{E}\right) = \bigcup_{\mathsf{K}\in\mathrm{D}(\mathsf{E})} \mathrm{B}\left(\mathsf{K}\right), \quad \mathrm{TBD}\left(\mathsf{E}\right) = \bigcup_{\mathsf{K}\in\mathrm{TD}(\mathsf{E})} \mathrm{B}\left(\mathsf{K}\right). \tag{26}$$

It follows in particular that for every constant expression E, $\mathrm{TBD}\left(\mathsf{E}\right) = \emptyset$.

Example 5 (*Ex. 1 continued*) The set of true broken derived terms of E_1 is:

$$\begin{aligned}\mathrm{TBD}\left(\mathsf{E}_1\right) &= \mathrm{B}\left(\{a^*\cdot(a\cdot\mathsf{F}_1)\,,\, b^*\cdot(a\cdot\mathsf{F}_1)\,,\, \mathsf{F}_1\,,\, a^*\,,\, b^*\}\right),\\ &= \{a^*\cdot(a\cdot\mathsf{F}_1)\,,\, b^*\cdot(a\cdot\mathsf{F}_1)\,,\, a^*\,,\, b^*\}.\end{aligned}$$

Hence the set of broken derived terms is:

$$\mathrm{BD}\left(\mathsf{E}_1\right) = \mathrm{TBD}\left(\mathsf{E}_1\right) \,\cup\, \mathrm{B}\left(\mathsf{E}_1\right) = \{a^*\cdot(a\cdot\mathsf{F}_1)\,,\, b^*\cdot(a\cdot\mathsf{F}_1)\,,\, a^*\,,\, b^*\}.$$

As above, it is also possible to define the *broken derived term automaton* of an expression E: its states are the broken derived terms of E, the initial states are the broken terms of E, the final states are those broken derived terms whose constant term is 1, and there is a transition between two broken derived terms, F and G, labeled by a if, and only if, G belongs to $\frac{\partial_{\mathsf{b}}}{\partial a}\,\mathsf{F}$. Of course, the automaton obtained in this way recognises $L\left(\mathsf{E}\right)$. And as above, the present paper does not develop along this line.

Example 6 (*Ex. 1 continued*) Figure 2 shows the broken derived term automaton of E_1.

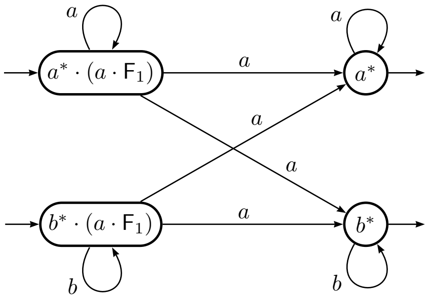
a
a
$a^* \cdot (a \cdot \mathsf{F}_1)$
a
a^*
a
a
$b^* \cdot (a \cdot \mathsf{F}_1)$
a
b^*
b
b

Figure 2: The broken derived term automaton of E_1

.

As we proved for derived terms, with Proposition 4, and as a consequence of that result, we now establish that the sets of broken and true broken derived terms of an expression can be computed by induction on the depth of the expression, and this definition, without any reference to the derivation operator, happens to be more suited for the forthcoming proofs.

Proposition 6 *Let* F *and* G *be two rational expressions. Then, the following holds.*

$$\mathrm{TBD}\left(\mathsf{F}+\mathsf{G}\right) = \mathrm{TBD}\left(\mathsf{F}\right) \,\cup\, \mathrm{TBD}\left(\mathsf{G}\right), \tag{27}$$

$$\mathrm{TBD}\left(\mathsf{F}\cdot\mathsf{G}\right) = \left(\mathrm{TBD}\left(\mathsf{F}\right)\right)_{\mathsf{p}}\cdot\mathsf{G} \,\cup\, \mathrm{TBD}\left(\mathsf{G}\right) \,\cup\, \delta_{\mathrm{TBD}(\mathsf{F})}\mathrm{B}\left(\mathsf{G}\right), \tag{28}$$

$$\mathrm{TBD}\left(\mathsf{F}^*\right) = \left(\mathrm{TBD}\left(\mathsf{F}\right)\right)\cdot\mathsf{F}^*. \tag{29}$$

Proof. This statement is a straightforward consequence of the application of (26) to the inductive computation of derived terms (Proposition 4, Equations (12)–(14)). Since Equation (27) is obvious, let us develop (28).

$$\begin{aligned}\mathrm{TBD}\left(\mathsf{F}\cdot\mathsf{G}\right) &= \mathrm{B}\left(\mathrm{TD}\left(\mathsf{F}\cdot\mathsf{G}\right)\right) = \mathrm{B}\left(\left(\mathrm{TD}\left(\mathsf{F}\right)\right)\cdot\mathsf{G} \,\cup\, \mathrm{TD}\left(\mathsf{G}\right)\right)\\ &= \mathrm{B}\left(\left(\mathrm{TD}\left(\mathsf{F}\right)\right)\cdot\mathsf{G}\right) \,\cup\, \mathrm{B}\left(\mathrm{TD}\left(\mathsf{G}\right)\right)\\ &= \left(\mathrm{TBD}\left(\mathsf{F}\right)\right)_{\mathsf{p}}\cdot\mathsf{G} \,\cup\, \delta_{\mathrm{TBD}(\mathsf{F})}\mathrm{B}\left(\mathsf{G}\right) \,\cup\, \mathrm{TBD}\left(\mathsf{G}\right).\end{aligned}$$

For establishing (29), let us first note that, for any set X of expressions, $\left(\mathrm{B}\left(X\right)\right)_{\mathsf{p}} \subseteq \mathrm{B}\left(\left(X\right)_{\mathsf{p}}\right)$ and thus, by (17), $\mathrm{B}\left(\left(X\right)_{\mathsf{p}}\right) \,\cup\, \delta_{\mathrm{B}(X)}\{1\} = \mathrm{B}\left(X\right)$. Then:

$$\begin{aligned}\mathrm{TBD}\left(\mathsf{F}^*\right) &= \mathrm{B}\left(\mathrm{TD}\left(\mathsf{F}^*\right)\right) = \mathrm{B}\left(\left(\mathrm{TD}\left(\mathsf{F}\right)\right)\cdot\mathsf{F}^*\right)\\ &= \left(\mathrm{B}\left(\left(\mathrm{TD}\left(\mathsf{F}\right)\right)_{\mathsf{p}}\right)\right)\cdot\mathsf{F}^* \,\cup\, \delta_{\mathrm{B}(\mathrm{TD}(\mathsf{F}))}\mathrm{B}\left(\mathsf{F}^*\right)\\ &= \mathrm{B}\left(\left(\left(\mathrm{TD}\left(\mathsf{F}\right)\right)_{\mathsf{p}} \,\cup\, \{1\}\right)\cdot\mathsf{F}^*\right) = \left(\mathrm{TBD}\left(\mathsf{F}\right)\right)\cdot\mathsf{F}^*.\end{aligned}$$

□

Proposition 7 *Let* F *and* G *be two rational expressions. Then, the following holds.*

$$\mathrm{BD}\left(\mathsf{F}+\mathsf{G}\right) = \mathrm{BD}\left(\mathsf{F}\right) \,\cup\, \mathrm{BD}\left(\mathsf{G}\right), \tag{30}$$

$$\mathrm{BD}\left(\mathsf{F}\cdot\mathsf{G}\right) = \left(\mathrm{BD}\left(\mathsf{F}\right)\right)_{\mathsf{p}} \cdot \mathsf{G} \,\cup\, \mathrm{TBD}\left(\mathsf{G}\right) \,\cup\, \delta_{\mathrm{BD}(\mathsf{F})}\mathrm{B}\left(\mathsf{G}\right), \tag{31}$$

$$\mathrm{BD}\left(\mathsf{F}^*\right) = \left(\mathrm{TBD}\left(\mathsf{F}\right)\right)_{\mathsf{p}} \cdot \mathsf{F}^* \,\cup\, \{\mathsf{F}^*\}. \tag{32}$$

Proof. As above, (30) is obvious, so let us first develop (31):

$$\begin{aligned}\mathrm{BD}\left(\mathsf{F}\cdot\mathsf{G}\right) &= \mathrm{TBD}\left(\mathsf{F}\cdot\mathsf{G}\right) \,\cup\, \mathrm{B}\left(\mathsf{F}\cdot\mathsf{G}\right)\\ &= \left(\mathrm{TBD}\left(\mathsf{F}\right)\right)_{\mathsf{p}} \cdot \mathsf{G} \,\cup\, \left(\mathrm{B}\left(\mathsf{F}\right)\right)_{\mathsf{p}} \cdot \mathsf{G} \,\cup\, \mathrm{TBD}\left(\mathsf{G}\right) \,\cup\, \delta_{\mathrm{TBD}(\mathsf{F})}\mathrm{B}\left(\mathsf{G}\right) \,\cup\, \delta_{\mathrm{B}(\mathsf{F})}\mathrm{B}\left(\mathsf{G}\right)\\ &= \left(\mathrm{BD}\left(\mathsf{F}\right)\right)_{\mathsf{p}} \cdot \mathsf{G} \,\cup\, \mathrm{TBD}\left(\mathsf{G}\right) \,\cup\, \delta_{\mathrm{BD}(\mathsf{F})}\mathrm{B}\left(\mathsf{G}\right).\end{aligned}$$

Equation (32) is obvious as well, with just a writing trick that will be useful later. □

Corollary 8 *The sets of broken derived terms* $\mathrm{BD}\left(\mathsf{E}\right)$ *and of true broken derived terms* $\mathrm{TBD}\left(\mathsf{E}\right)$ *of any rational expression* E *over* A *can be computed by induction on the depth of* E *by using Equations (27)–(32) and the base clauses:*

$$\begin{array}{lll}\mathrm{TBD}\left(0\right) = \emptyset, & \mathrm{TBD}\left(1\right) = \emptyset, & \forall a \in A \quad \mathrm{TBD}\left(a\right) = \{1\},\\ \mathrm{BD}\left(0\right) = \{0\}, & \mathrm{BD}\left(1\right) = \{1\}, & \forall a \in A \quad \mathrm{BD}\left(a\right) = \{a, 1\}.\end{array}$$

We are now ready to establish the two main results stated in the introduction.

3. Bounds on the size of the broken derived term set

A remarkable feature of Antimirov's derived terms is that their number, for a given expression E, is bounded by $\ell(\mathsf{E})+1$, that is, the automaton they allow to build is not bigger than the position automaton. Although such a bound seems, at first sight, to be out of reach for the number of broken derived terms, we shall see that twice the same bound, or even the same bound, indeed hold under reasonable, and sensible, hypotheses. One must say that the situation seems to be rather desesperate when one consider expressions such as: $\mathsf{K}_4 = 1 + 1^* + (1^*)^* + ((1^*)^*)^*$. The broken terms of K_4 are:

$$\mathrm{B}\left(\mathsf{K}_4\right) = \{1, 1^*, (1^*)^*, ((1^*)^*)^*\}. \tag{33}$$

The expression K_4 does not contain any letter, and from this example it is easily seen how to construct an expression without letters and with an arbitrary number of broken derived terms, making impossible a bound which is a function of the literal length of the expression. At this point, one should notice that $\mathrm{TBD}\left(\mathsf{K}_4\right) = \emptyset$ and that the blowing phenomenon comes from the breaking of such expressions only, *not from their derivation.* It seems thus reasonable to rule out the playing with 1 and $*$.

3.1. A bound for arbitrary (reasonable) expressions

In order to describe 'reasonable' expressions, let us call a constant expression which contains at least one star operator a *starred constant expression.* Modulo the trivial identities, the only non-starred constant expressions are the sums of 1's. Even if we rule out the starred constant expressions, the bound $\ell(\mathsf{E})+1$ does not hold for the number of broken derived terms of an arbitrary expression, as the following example shows.

Example 7 Let $\mathsf{G}_k = ((a^*)^*+1)^k$, for every positive integer k. More precisely, as bracketing matters (as we show below), let $\mathsf{G}_1 = ((a^*)^*+1)$ and $\mathsf{G}_{k+1} = (\mathsf{G}_1 \cdot \mathsf{G}_k)$, for every positive integer k. Clearly, $\ell(\mathsf{G}_k) = k$. We then have:

$$\mathrm{B}(\mathsf{G}_1) = \{(a^*)^*,\, 1\}, \qquad \text{and} \qquad \mathrm{BD}(\mathsf{G}_1) = \{(a^*)^*, a^*(a^*)^*, 1\}, \tag{34}$$

and, for every k, $\mathrm{B}(\mathsf{G}_{k+1}) = (a^*)^* \cdot \mathsf{G}_k \,\cup\, \mathrm{B}(\mathsf{G}_k)$. By (31) and since $\delta_{\mathrm{BD}(\mathsf{G}_1)} = 1$, we have:

$$\begin{aligned}\mathrm{BD}(\mathsf{G}_{k+1}) &= \mathrm{BD}(\mathsf{G}_1 \cdot \mathsf{G}_k) = (\mathrm{BD}(\mathsf{G}_1))_{\mathsf{p}} \cdot \mathsf{G}_k \,\cup\, \mathrm{BD}(\mathsf{G}_k)\\ &= (a^*)^* \cdot \mathsf{G}_k \,\cup\, (a^*(a^*)^*) \cdot \mathsf{G}_k \,\cup\, \mathrm{BD}(\mathsf{G}_k)\,.\end{aligned}$$

Since the expressions G_k are pairwise distinct, the same holds for the $(a^*)^* \cdot \mathsf{G}_k$ and the $(a^*(a^*)^*) \cdot \mathsf{G}_k$. Hence $\mathrm{BD}(\mathsf{G}_k) \cap \big((a^*)^* \cdot \mathsf{G}_k \,\cup\, (a^*(a^*)^*) \cdot \mathsf{G}_k\big) = \emptyset$, and then:

$$\mathrm{card}\,(\mathrm{BD}(\mathsf{G}_k)) = 2\,\ell(\mathsf{G}_k) + 1. \tag{35}$$

The expressions given in this example reach indeed the bound for the generic case of what we consider reasonable expressions.

Theorem 1 *Let* E *be a rational expression which contains no starred constant subexpression. Then, the following holds:*

$$\mathrm{card}\,(\mathrm{BD}(\mathsf{E})) \leqslant 2\,\ell(\mathsf{E}) + 1.$$

Theorem 1 is a direct consequence of the following slightly more technical statement whose detailed form is necessary for the induction:

Proposition 9 *Let* E *be a rational expression that contains no starred constant subexpression. Then, the following holds:*

$$\mathrm{card}\Big((\mathrm{BD}(\mathsf{E}))_{\mathsf{p}}\Big) \leqslant 2\,\ell(\mathsf{E}), \tag{36}$$

$$\textit{if } \ell(\mathsf{E}) \geqslant 1, \quad \mathrm{card}\,(\mathrm{TBD}(\mathsf{E})) \leqslant 2\,\ell(\mathsf{E}) - 1. \tag{37}$$

As $\mathrm{card}\,(X) \leqslant \mathrm{card}\Big((X)_{\mathsf{p}}\Big) + 1$ for any set X of expressions, (36) directly implies Theorem 1. The necessity of establishing (37) in the course of proving (36) by induction comes from the fact that $\mathrm{TBD}(\mathsf{F})$ appears in the inductive computation of $\mathrm{BD}(\mathsf{F}^*)$ as described by Proposition 7.

Proof. By induction on the depth of E, that is, $\mathsf{E} = \mathsf{F} + \mathsf{G}$, $\mathsf{E} = \mathsf{F} \cdot \mathsf{G}$ or $\mathsf{E} = \mathsf{F}^*$. In the first two cases, $\ell(\mathsf{E}) = \ell(\mathsf{F}) + \ell(\mathsf{G})$, in the third, $\ell(\mathsf{E}) = \ell(\mathsf{F})$. If E contains no starred constant subexpressions, then the same is true of F and G. We thus have by induction:

$$\mathrm{card}\left((\mathrm{BD}\,(\mathsf{F}))_{\mathsf{p}}\right) \leqslant 2\,\ell(\mathsf{F}) \qquad \text{and} \qquad \mathrm{card}\left((\mathrm{BD}\,(\mathsf{G}))_{\mathsf{p}}\right) \leqslant 2\,\ell(\mathsf{G}). \tag{38}$$

and, if $\ell(\mathsf{F}) \geqslant 1$,

$$\mathrm{card}\,(\mathrm{TBD}\,(\mathsf{F})) \leqslant 2\,\ell(\mathsf{F}) - 1. \tag{39}$$

The cases $\mathsf{E} = 0$, $\mathsf{E} = 1$ and $\mathsf{E} = a$ trivially satisfy (36) and (37). Moreover $\mathsf{E} = 0$ is not a base case of the induction since a non-zero expression does not contain any zero subexpression modulo the trivial identities.

Case $\mathsf{E} = \mathsf{F} + \mathsf{G}$: By (30), $(\mathrm{BD}\,(\mathsf{E}))_{\mathsf{p}} = (\mathrm{BD}\,(\mathsf{F}))_{\mathsf{p}} \cup (\mathrm{BD}\,(\mathsf{G}))_{\mathsf{p}}$ and (36) holds. By (27), $\mathrm{TBD}\,(\mathsf{E}) = \mathrm{TBD}\,(\mathsf{F}) \cup \mathrm{TBD}\,(\mathsf{G})$.

If $\ell(\mathsf{E}) \geqslant 1$, we then may assume that $\ell(\mathsf{F}) \geqslant 1$ and (39) holds.

Then, either $\ell(\mathsf{G}) = 0$, and $\mathrm{TBD}\,(\mathsf{G}) = \emptyset$, or $\ell(\mathsf{G}) \geqslant 1$ and then, by induction $\mathrm{card}\,(\mathrm{TBD}\,(\mathsf{G})) \leqslant 2\,\ell(\mathsf{G}) - 1$. In both cases, (37) holds.

Case $\mathsf{E} = \mathsf{F} \cdot \mathsf{G}$ (which implies F and G different from 1):

By (31), $\mathrm{BD}\,(\mathsf{E}) = (\mathrm{BD}\,(\mathsf{F}))_{\mathsf{p}} \cdot \mathsf{G} \cup \mathrm{TBD}\,(\mathsf{G}) \cup \delta_{\mathrm{BD}(\mathsf{F})}\mathrm{B}\,(\mathsf{G})$.

Since $\mathsf{G} \neq 1$, $\left((\mathrm{BD}\,(\mathsf{F}))_{\mathsf{p}} \cdot \mathsf{G}\right)_{\mathsf{p}} = (\mathrm{BD}\,(\mathsf{F}))_{\mathsf{p}} \cdot \mathsf{G}$ and thus:

$$\begin{aligned}(\mathrm{BD}\,(\mathsf{E}))_{\mathsf{p}} &= (\mathrm{BD}\,(\mathsf{F}))_{\mathsf{p}} \cdot \mathsf{G} \cup \left(\mathrm{TBD}\,(\mathsf{G}) \cup \delta_{\mathrm{BD}(\mathsf{F})}\mathrm{B}\,(\mathsf{G})\right)_{\mathsf{p}} \\ &\subseteq (\mathrm{BD}\,(\mathsf{F}))_{\mathsf{p}} \cdot \mathsf{G} \cup (\mathrm{BD}\,(\mathsf{G}))_{\mathsf{p}}\,.\end{aligned}$$

Then $\qquad \mathrm{card}\left((\mathrm{BD}\,(\mathsf{E}))_{\mathsf{p}}\right) \leqslant \mathrm{card}\left((\mathrm{BD}\,(\mathsf{F}))_{\mathsf{p}} \cdot \mathsf{G}\right) + \mathrm{card}\left((\mathrm{BD}\,(\mathsf{G}))_{\mathsf{p}}\right)$,

and (36) follows as above.

By (28), $\mathrm{TBD}\,(\mathsf{E}) = (\mathrm{TBD}\,(\mathsf{F}))_{\mathsf{p}} \cdot \mathsf{G} \cup \mathrm{TBD}\,(\mathsf{G}) \cup \delta_{\mathrm{TBD}(\mathsf{F})}\mathrm{B}\,(\mathsf{G}\,)$.

If $1 \in \mathrm{TBD}\,(\mathsf{F})$, then $\ell(\mathsf{F}) \geqslant 1$ and $\mathrm{TBD}\,(\mathsf{E}) = (\mathrm{TBD}\,(\mathsf{F}))_{\mathsf{p}} \cdot \mathsf{G} \cup \mathrm{BD}\,(\mathsf{G})$. The induction yields $\mathrm{card}\left((\mathrm{TBD}\,(\mathsf{F}))_{\mathsf{p}}\right) \leqslant 2\,\ell(\mathsf{F}) - 2$ and $\mathrm{card}\,(\mathrm{BD}\,(\mathsf{G})) \leqslant 2\,\ell(\mathsf{G}) + 1$ and thus (37) holds.

If $1 \notin \mathrm{TBD}\,(\mathsf{F})$,
then $\mathrm{TBD}\,(\mathsf{E}) = (\mathrm{TBD}\,(\mathsf{F}))_{\mathsf{p}} \cdot \mathsf{G} \cup \mathrm{TBD}\,(\mathsf{G}) \subseteq (\mathrm{BD}\,(\mathsf{F}))_{\mathsf{p}} \cdot \mathsf{G} \cup \mathrm{TBD}\,(\mathsf{G})$.
If $\ell(\mathsf{E}) \geqslant 1$ then,
either $\ell(\mathsf{G}) \geqslant 1$ and then $\mathrm{card}\,(\mathrm{TBD}\,(\mathsf{F} \cdot \mathsf{G})) \leqslant 2\,\ell(\mathsf{F}) + 2\,\ell(\mathsf{G}) - 1$,
or $\ell(\mathsf{G}) = 0$ and then both $\ell(\mathsf{F}) \geqslant 1$ and $\mathrm{TBD}\,(\mathsf{G}) = \emptyset$, then
$\mathrm{card}\,(\mathrm{TBD}\,(\mathsf{F} \cdot \mathsf{G})) \leqslant 2\,\ell(\mathsf{F}) - 1$.
In both cases, (37) holds.

Case $\mathsf{E} = \mathsf{F}^*$: The subexpression F is starred, thus not a constant expression and $\ell(\mathsf{F}) \geqslant 1$. By induction, $\mathrm{card}\,(\mathrm{TBD}\,(\mathsf{F})) \leqslant 2\,\ell(\mathsf{F}) - 1 = 2\,\ell(\mathsf{E}) - 1$.

By (32), $\mathrm{BD}\,(\mathsf{F}^*) = (\mathrm{TBD}\,(\mathsf{F}))_{\mathsf{p}} \cdot \mathsf{F}^* \,\cup\, \{\mathsf{F}^*\}$ and thus:

$$\mathrm{card}\left((\mathrm{BD}\,(\mathsf{E}))_{\mathsf{p}}\right) = \mathrm{card}\left((\mathrm{TBD}\,(\mathsf{F}))_{\mathsf{p}}\right) + 1 \leqslant 2\,\ell(\mathsf{E}).$$

By (29), $\mathrm{TBD}\,(\mathsf{F}^*) = \mathrm{TBD}\,(\mathsf{F}) \cdot \mathsf{F}^*$ and thus:

$$\mathrm{card}\,(\mathrm{TBD}\,(\mathsf{F}^*)) = \mathrm{card}\,(\mathrm{TBD}\,(\mathsf{F})) \leqslant 2\,\ell(\mathsf{E}) - 1.$$

□

3.2. A bound for expressions in star normal form

The definition of star normal form of an expression has been introduced by Brüggemann-Klein (in [2]) in order to design a quadratic algorithm for the construction of the position automaton. From any expression, it is possible to compute an equivalent expression which is in star normal form; this computation amounts to avoiding the star operation on subexpressions whose constant term is 1. In order to prove Theorem 2, we first introduce an inductive definition of the computation of the star normal form which is slightly different from the original one and better fitted to proofs by induction (on depth of expressions).

3.2.1. The star normal form revisited

Let us first recall the definition of an expression in star normal form.[5]

Definition 6 ([2]) *A rational expression* E *is in star normal form (SNF) if, and only if, for any* F *such that* F^* *is a subexpression of* E, $\mathsf{c}\,(\mathsf{F}) = 0$.

In order to compute expressions in SNF, two operators on expressions, • and ∘, were defined in [2]. Both are defined by induction on the depth of the expressions. The operator • computes an expression in SNF, and calls the operator ∘ when applied to a starred expression. The inductive definition goes as follows:

[5] This definition is not exactly the one given in [2] (*cf.* Definiton 2.6) where a somewhat technical supplementary condition is given. This condition is not necessary here; whether it is really necessary for the original purpose of SNF, that is, the quadratic algorithm, is the object of ongoing work of the authors.

$$\mathsf{0}^{\circ} = \mathsf{0}, \quad \mathsf{1}^{\circ} = \mathsf{0}, \quad \text{for all } a \in A \quad a^{\circ} = a, \tag{40}$$

$$(\mathsf{F} + \mathsf{G})^{\circ} = \mathsf{F}^{\circ} + \mathsf{G}^{\circ}, \tag{41}$$

$$(\mathsf{F} \cdot \mathsf{G})^{\circ} = \begin{cases} \mathsf{F} \cdot \mathsf{G} & \text{if } \mathsf{c}(\mathsf{F}) = \mathsf{c}(\mathsf{G}) = 0, \\ \mathsf{F}^{\circ} \cdot \mathsf{G} & \text{if } \mathsf{c}(\mathsf{F}) = 0 \text{ and } \mathsf{c}(\mathsf{G}) = 1, \\ \mathsf{F} \cdot \mathsf{G}^{\circ} & \text{if } \mathsf{c}(\mathsf{F}) = 1 \text{ and } \mathsf{c}(\mathsf{G}) = 0, \\ \mathsf{F}^{\circ} + \mathsf{G}^{\circ} & \text{if } \mathsf{c}(\mathsf{F}) = \mathsf{c}(\mathsf{G}) = 1, \end{cases} \tag{42}$$

$$(\mathsf{F}^*)^{\circ} = \mathsf{F}^{\circ}. \tag{43}$$

and:

$$\mathsf{0}^{\bullet} = \mathsf{0}, \quad \mathsf{1}^{\bullet} = \mathsf{1}, \quad \text{for all } a \in A \quad a^{\bullet} = a, \tag{44}$$

$$(\mathsf{F} + \mathsf{G})^{\bullet} = \mathsf{F}^{\bullet} + \mathsf{G}^{\bullet}, \tag{45}$$

$$(\mathsf{F} \cdot \mathsf{G})^{\bullet} = \mathsf{F}^{\bullet} \cdot \mathsf{G}^{\bullet}, \tag{46}$$

$$(\mathsf{F}^*)^{\bullet} = ((\mathsf{F}^{\bullet})^{\circ})^*. \tag{47}$$

The following is then established:

Proposition 10 ([2]) *For any rational expression* E*, the expression* $\mathsf{E}^{\bullet}$ *is in star normal form and is equivalent to* E*.*

The problem with the inductive definition (40)-(47) is that even though every subexpression of an expression in SNF is in SNF, the same heritage property is not true for expressions computed with the operator $\circ$. For instance:

$$((a^* b^*)^*)^{\circ} = a + b \qquad \text{whereas} \qquad (c\,(a^* b^*)^*)^{\circ} = c\,(a^* b^*)^*.$$

This difficulty can easily be overcome once it is noticed that in the course of the computation of $\mathsf{E}^{\bullet}$, the $\circ$ operator is only used in conjunction with the operator $\bullet$ (in (47)). We thus define a new operator $\square$ by the following:

$$\forall \mathsf{E} \in \mathsf{RatE}(A) \qquad \mathsf{E}^{\square} = (\mathsf{E}^{\bullet})^{\circ}. \tag{48}$$

We shall derive an inductive definition of $\square$ from the properties of $\circ$ that we first establish.

Proposition 11 *Let* E *be an expression. Then, the following holds:*

$$\mathsf{c}(\mathsf{E}^{\circ}) = 0, \tag{49}$$

$$\mathsf{c}(\mathsf{E}) = 0 \quad \Longrightarrow \quad \mathsf{E} = \mathsf{E}^{\circ} \qquad \textit{(which implies: } \mathsf{E} = \mathsf{E}^{\circ} \Leftrightarrow \mathsf{c}(\mathsf{E}) = 0\textit{)}, \tag{50}$$

$$(\mathsf{E}^{\circ})^{\circ} = \mathsf{E}^{\circ} \qquad \textit{(i.e., } \circ \textit{ is idempotent)}, \tag{51}$$

$$\mathsf{c}(\mathsf{E}^{\bullet}) = \mathsf{c}(\mathsf{E}). \tag{52}$$

Proof. All these properties are established by an easy induction. The most tedious one being (50), which we explicitly give. The base clauses obviously satisfy (50).

If $\mathsf{E} = \mathsf{F} + \mathsf{G}$, $\mathsf{c}\,(\mathsf{E}) = 0$ implies
$\mathsf{c}\,(\mathsf{F}) = \mathsf{c}\,(\mathsf{G}) = 0$ and thus, by induction, $(\mathsf{F} + \mathsf{G})^\circ = \mathsf{F}^\circ + \mathsf{G}^\circ = \mathsf{F} + \mathsf{G}$.

If $\mathsf{E} = \mathsf{F} \cdot \mathsf{G}$, $\mathsf{c}\,(\mathsf{E}) = 0$ implies $\mathsf{c}\,(\mathsf{F})\,\mathsf{c}\,(\mathsf{G}) = 0$.
If $\mathsf{c}\,(\mathsf{F}) = \mathsf{c}\,(\mathsf{G}) = 0$, then $(\mathsf{F} \cdot \mathsf{G})^\circ = \mathsf{F} \cdot \mathsf{G}$.
If $\mathsf{c}\,(\mathsf{F}) = 0$ and $\mathsf{c}\,(\mathsf{G}) = 1$, then $(\mathsf{F} \cdot \mathsf{G})^\circ = \mathsf{F}^\circ \cdot \mathsf{G} = \mathsf{F} \cdot \mathsf{G}$.
If $\mathsf{c}\,(\mathsf{F}) = 1$ and $\mathsf{c}\,(\mathsf{G}) = 0$, then $(\mathsf{F} \cdot \mathsf{G})^\circ = \mathsf{F} \cdot \mathsf{G}^\circ = \mathsf{F} \cdot \mathsf{G}$.

Finally, if $\mathsf{c}\,(\mathsf{E}) = 0$, then E cannot be of the form $\mathsf{E} = \mathsf{F}^*$. □

Proposition 12 *Let* F *and* G *be two rational expressions. Then the following holds:*

$$(\mathsf{F} + \mathsf{G})^\square = \mathsf{F}^\square + \mathsf{G}^\square, \tag{53}$$

$$(\mathsf{F} \cdot \mathsf{G})^\square = \begin{cases} \mathsf{F}^\square + \mathsf{G}^\square & \textit{if} \quad \mathsf{c}\,(\mathsf{F}) = \mathsf{c}\,(\mathsf{G}) = 1 \\ \mathsf{F}^\bullet \cdot \mathsf{G}^\bullet & \textit{otherwise} \end{cases}, \tag{54}$$

$$(\mathsf{F}^*)^\square = \mathsf{F}^\square. \tag{55}$$

Proof. From (41) and (45), immediately follows (53).

Let $(\mathsf{F} \cdot \mathsf{G})^\square = ((\mathsf{F} \cdot \mathsf{G})^\bullet)^\circ = (\mathsf{F}^\bullet \cdot \mathsf{G}^\bullet)^\circ$; the remainder of the proof of (54) requires a case examination (with the implicit use that $\mathsf{c}\,(\mathsf{H}^\bullet) = \mathsf{c}\,(\mathsf{H})$ for any H).
If $\mathsf{c}\,(\mathsf{F}) = 0$ or $\mathsf{c}\,(\mathsf{G}) = 0$ then $\mathsf{c}\,(\mathsf{E}) = 0$ and, by (50) $(\mathsf{F} \cdot \mathsf{G})^{\bullet\circ} = \mathsf{F}^\bullet \cdot \mathsf{G}^\bullet$.
If $\mathsf{c}\,(\mathsf{F}) = \mathsf{c}\,(\mathsf{G}) = 1$, then $(\mathsf{F}^\bullet \cdot \mathsf{G}^\bullet)^\circ = (\mathsf{F}^\bullet)^\circ + (\mathsf{G}^\bullet)^\circ = \mathsf{F}^\square + \mathsf{G}^\square$.

Finally, $(\mathsf{F}^*)^\square = ((\mathsf{F}^*)^\bullet)^\circ = (((\mathsf{F}^\bullet)^\circ)^*)^\circ = ((\mathsf{F}^\bullet)^\circ)^\circ = (\mathsf{F}^\bullet)^\circ$ by (51). □

Corollary 13 *The expression* $\mathsf{E}^\square$ *can be computed by induction on the depth of* E *by using Equations (53)–(55) and the base clauses:*

$$\mathsf{0}^\square = \mathsf{0}, \quad \mathsf{1}^\square = \mathsf{0}, \qquad \forall a \in A \quad a^\square = a. \tag{56}$$

Now, it is clear that (53)–(55), together with (44)–(46) and (47) rewritten as:

$$(\mathsf{F}^*)^\bullet = (\mathsf{F}^\square)^*, \tag{47'}$$

is a new inductive description of the computation of $\mathsf{E}^\bullet$, with the difference that any subexpression of $\mathsf{E}^\square$ is in SNF. For instance:

$$((a^* b^*)^*)^\square = a + b \qquad \text{and} \qquad (c\,(a^* b^*)^*)^\square = c\,(a + b)^*.$$

This new inductive definition of $\mathsf{E}^\bullet$ yields a linear time algorithm. As we did in Proposition 11 for the operator ○, we list a number of properties for the operators □ and ●, all established by obvious inductions.

Proposition 14 *Let* E *be an expression. Then, the following holds:*

$$\mathsf{c}\left(\mathsf{E}^{\square}\right) = 0, \tag{57}$$

$$\mathsf{c}\left(\mathsf{E}\right) = 0 \quad \Longleftrightarrow \quad \mathsf{E}^{\bullet} = \mathsf{E}^{\square}, \tag{58}$$

$$\ell(\mathsf{E}^{\square}) = \ell(\mathsf{E}^{\bullet}) = \ell(\mathsf{E}), \tag{59}$$

$$\mathsf{E}\ \textit{in SNF} \quad \Longrightarrow \quad \mathsf{E} = \mathsf{E}^{\bullet}. \tag{60}$$

Proof. *Ad* (58): True for the base cases.

$$\begin{aligned}
(\mathsf{F}+\mathsf{G})^{\square} = (\mathsf{F}+\mathsf{G})^{\bullet} &\Leftrightarrow \mathsf{F}^{\square} = \mathsf{F}^{\bullet} \text{ and } \mathsf{G}^{\square} = \mathsf{G}^{\bullet} \Leftrightarrow \\
&\Leftrightarrow \mathsf{c}\left(\mathsf{F}\right) = 0 \text{ and } \mathsf{c}\left(\mathsf{G}\right) = 0 \Leftrightarrow \mathsf{c}\left((\mathsf{F}+\mathsf{G})\right) = 0, \\
(\mathsf{F}\cdot\mathsf{G})^{\square} = (\mathsf{F}\cdot\mathsf{G})^{\bullet} = \mathsf{F}^{\bullet}\cdot\mathsf{G}^{\bullet} &\Leftrightarrow \mathsf{c}\left(\mathsf{F}\right)\mathsf{c}\left(\mathsf{G}\right) = \mathsf{c}\left((\mathsf{F}\cdot\mathsf{G})\right) = 0, \\
(\mathsf{F}^{*})^{\square} = (\mathsf{F}^{*})^{\bullet}&, \text{ that is } \mathsf{F}^{\square} = (\mathsf{F}^{\square})^{*} \text{ never happens, and } \mathsf{c}\left(\mathsf{F}^{*}\right) = 1.
\end{aligned}$$

Ad (60): True for the base cases. If $\mathsf{E} = \mathsf{F}+\mathsf{G}$ or $\mathsf{E} = \mathsf{F}\cdot\mathsf{G}$ and E is in SNF, then F and G are in SNF, thus $\mathsf{F} = \mathsf{F}^{\bullet}$ and $\mathsf{G} = \mathsf{G}^{\bullet}$, $(\mathsf{F}+\mathsf{G})^{\bullet} = \mathsf{F}+\mathsf{G}$ and $(\mathsf{F}\cdot\mathsf{G})^{\bullet} = \mathsf{F}\cdot\mathsf{G}$. If $\mathsf{E} = \mathsf{F}^{*}$ and E is in SNF, then $\mathsf{c}\left(\mathsf{F}\right) = 0$ and F in SNF, that is $\mathsf{F}^{\bullet} = \mathsf{F}^{\square} = \mathsf{F}$, thus $\mathsf{E}^{\bullet} = (\mathsf{F}^{*})^{\bullet} = (\mathsf{F}^{\square})^{*} = \mathsf{F}^{*} = \mathsf{E}$. □

For the sake of completeness, we give a proof of Proposition 10 with the new definitions of $\circ$ and $\bullet$; for the sake of fluent reading, we postpone it to the appendix.

3.2.2. Proof of Theorem 2

We are now able to prove the main result of this paper, already stated in the introduction.

Theorem 2 *If a rational expression* E *is in star normal form, then:*

$$\operatorname{card}\left(\mathrm{BD}\left(\mathsf{E}\right)\right) \leqslant \ell(\mathsf{E}) + 1.$$

Proof. The result trivially holds for $\mathsf{E} = \mathsf{0}$, and we suppose now that E, and thus all its subexpressions, are different from 0. We rather prove the theorem under the form:

$$\operatorname{card}\left(\left(\mathrm{BD}\left(\mathsf{E}\right)\right)_{\mathsf{p}}\right) \leqslant \ell(\mathsf{E}), \tag{61}$$

and together with two more precise statements that will be used in the induction:

$$\text{if} \quad \mathsf{c}\left(\mathsf{E}\right) = 0, \quad \operatorname{card}\left(\left(\mathrm{TBD}\left(\mathsf{E}\right)\right)_{\mathsf{p}}\right) \leqslant \ell(\mathsf{E}) - 1, \tag{62}$$

$$\text{if} \quad \mathsf{c}\left(\mathsf{E}\right) = 1 \quad \text{and} \quad 1 \in \mathrm{TBD}\left(\mathsf{E}\right), \quad \operatorname{card}\left(\left(\mathrm{TBD}\left(\mathsf{E}\right)\right)_{\mathsf{p}}\right) \leqslant \ell(\mathsf{E}) - 1. \tag{63}$$

We prove these three conditions by induction on the depth of E, using the inductive equations of Propositions 6 and 7. The base cases $\mathsf{E} = 1$ and $\mathsf{E} = a$ are obvious.

Case $\mathsf{E} = \mathsf{F}+\mathsf{G}$:

$$\mathrm{BD}\left(\mathsf{E}\right) = \mathrm{BD}\left(\mathsf{F}\right) \,\cup\, \mathrm{BD}\left(\mathsf{G}\right)$$

and (61) holds by induction. Moreover:

$$\mathrm{TBD}(\mathsf{E}) = \mathrm{TBD}(\mathsf{F}) \cup \mathrm{TBD}(\mathsf{G}) \subseteq \mathrm{BD}(\mathsf{F}) \cup \mathrm{TBD}(\mathsf{G}),$$

then:

$$(\mathrm{TBD}(\mathsf{E}))_{\mathsf{p}} = (\mathrm{TBD}(\mathsf{F}))_{\mathsf{p}} \cup (\mathrm{TBD}(\mathsf{G}))_{\mathsf{p}} \subseteq (\mathrm{BD}(\mathsf{F}))_{\mathsf{p}} \cup (\mathrm{TBD}(\mathsf{G}))_{\mathsf{p}}.$$

If $\mathsf{c}(\mathsf{E}) = 0$, then $\mathsf{c}(\mathsf{F}) = \mathsf{c}(\mathsf{G}) = 0$ and, by induction,

$$\mathrm{card}\left((\mathrm{TBD}(\mathsf{E}))_{\mathsf{p}}\right) \leqslant \ell(\mathsf{F}) - 1 + \ell(\mathsf{G}) - 1 = \ell(\mathsf{E}) - 2.$$

If $\mathsf{c}(\mathsf{E}) = 1$, one may assume that $\mathsf{c}(\mathsf{F}) = 1$. Then, either $\mathsf{c}(\mathsf{G}) = 0$ or $\mathsf{c}(\mathsf{G}) = 1$ and, in this latter case, if $1 \in \mathrm{TBD}(\mathsf{E})$ then one may assume that $1 \in \mathrm{TBD}(\mathsf{G})$ since F and G play the same role. In both cases, $\mathrm{card}\left((\mathrm{TBD}(\mathsf{G}))_{\mathsf{p}}\right) \leqslant \ell(\mathsf{G}) - 1$ holds by induction, and:

$$\mathrm{card}\left((\mathrm{TBD}(\mathsf{E}))_{\mathsf{p}}\right) \leqslant \ell(\mathsf{F}) + \ell(\mathsf{G}) - 1 = \ell(\mathsf{E}) - 1.$$

Case $\mathsf{E} = \mathsf{F} \cdot \mathsf{G}$:

$$\mathrm{BD}(\mathsf{E}) = (\mathrm{BD}(\mathsf{F}))_{\mathsf{p}} \cdot \mathsf{G} \cup \mathrm{TBD}(\mathsf{G}) \cup \delta_{\mathrm{BD}(\mathsf{F})} \subseteq (\mathrm{BD}(\mathsf{F}))_{\mathsf{p}} \cdot \mathsf{G} \cup \mathrm{BD}(\mathsf{G}).$$

As $\mathsf{G} \neq 1$, (61) holds by induction.

$$\mathrm{TBD}(\mathsf{E}) = (\mathrm{TBD}(\mathsf{F}))_{\mathsf{p}} \cdot \mathsf{G} \cup \mathrm{TBD}(\mathsf{G}) \cup \delta_{\mathrm{TBD}(\mathsf{F})} \subseteq (\mathrm{BD}(\mathsf{F}))_{\mathsf{p}} \cdot \mathsf{G} \cup \mathrm{BD}(\mathsf{G}).$$

If $\mathsf{c}(\mathsf{E}) = 0$ then either $\mathsf{c}(\mathsf{F}) = 0$ (case 1) or $\mathsf{c}(\mathsf{F}) = 1$ in which case, either $1 \in \mathrm{TD}(\mathsf{F})$ (case 2) or not (case 3). In cases 1 and 2, by induction:

$$\mathrm{card}\left((\mathrm{TBD}(\mathsf{E}))_{\mathsf{p}}\right) \leqslant \mathrm{card}\left((\mathrm{TBD}(\mathsf{F}))_{\mathsf{p}}\right) + \mathrm{card}\left((\mathrm{BD}(\mathsf{G}))_{\mathsf{p}}\right) \leqslant \ell(\mathsf{F}) - 1 + \ell(\mathsf{G}) = \ell(\mathsf{E}) - 1.$$

In case 3, we have both $\mathsf{c}(\mathsf{G}) = 0$ and $\delta_{\mathrm{TBD}(\mathsf{F})} = 0$ and thus, by induction:

$$\mathrm{card}\left((\mathrm{TBD}(\mathsf{E}))_{\mathsf{p}}\right) \leqslant \mathrm{card}\left((\mathrm{BD}(\mathsf{F}))_{\mathsf{p}}\right) + \mathrm{card}\left((\mathrm{TBD}(\mathsf{G}))_{\mathsf{p}}\right) \leqslant \ell(\mathsf{F}) + \ell(\mathsf{G}) - 1 = \ell(\mathsf{E}) - 1.$$

Case $\mathsf{E} = \mathsf{F}^*$:

$$\mathrm{BD}(\mathsf{E}) = (\mathrm{TBD}(\mathsf{F}))_{\mathsf{p}} \cdot \mathsf{F}^* \cup \{\mathsf{F}^*\} \text{ and } \mathrm{TBD}(\mathsf{E}) = (\mathrm{TBD}(\mathsf{F}))_{\mathsf{p}} \cdot \mathsf{F}^*.$$

Since E is in SNF, $\mathsf{c}(\mathsf{F}) = 0$, thus $\mathrm{card}\left((\mathrm{BD}(\mathsf{E}))_{\mathsf{p}}\right) \leqslant (\ell(\mathsf{F}) - 1) + 1$ and (61) holds. Since $\mathsf{c}(\mathsf{E}) = 1$ and $1 \notin \mathrm{TBD}(\mathsf{E})$ (as $\mathsf{F} \neq 1$) both (62) and (63) hold. □

Remark 1 There is no general relation between $\mathrm{card}(\mathrm{D}(\mathsf{E}))$ and $\mathrm{card}(\mathrm{BD}(\mathsf{E}))$ for a rational expression E even if E is in star normal form.

For instance, $\mathsf{H}_k = a \cdot (b_1 + b_2 + \ldots + b_k)$ has 3 derived terms and $k+2$ broken derived terms (Figure 3 shows the derived term and the broken derived term automata for $\mathsf{H}_3 = a \cdot (b_1 + b_2 + b_3)$). On the contrary, the expression $\mathsf{E}_k = (a^* + b^*) \cdot (a \cdot (a^* + b^*))^k$ has $3k+3$ derived terms and $2k+2$ broken derived terms. In the previous examples, Figure 1 and 2 show the derived term and the broken derived term automata of E_1 respectively.

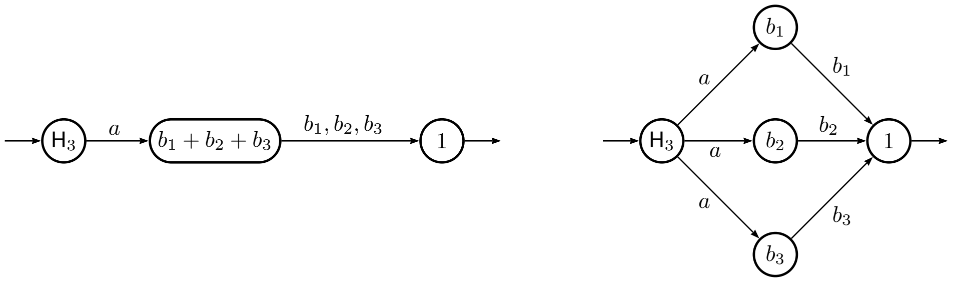

Figure 3: The derived term and the broken derived term automata of H_3.

4. Derivation and bracketing

We now consider the influence of the bracketing of expressions on the derivation and on the number of derived terms (broken or not). As the product of languages is associative, rational expressions are most often written as if the product operator be associative. This is indeed not the case and $(\mathsf{E} \cdot \mathsf{F}) \cdot \mathsf{G}$ and $\mathsf{E} \cdot (\mathsf{F} \cdot \mathsf{G})$ are two distinct expressions that yield different syntactic trees. We prove here that the derived terms, as well as the broken derived terms, of such expressions may also be different.

Example 8 Let us consider the not completely bracketed expression $a\,b\,(c\,(a\,b))^*$ and observe the result of the derivation of the two expressions $a\,(b\,(c\,(a\,b))^*)$ and $(a\,b)\,(c\,(a\,b))^*$ obtained by different bracketings.

$$\mathrm{D}\,(a\,(b\,(c\,(a\,b))^*)) = \{a\,(b\,(c\,(a\,b))^*)\,,\ b\,(c\,(a\,b))^*, (c\,(a\,b))^*\,,\ (a\,b)\,(c\,(a\,b))^*\}.$$
$$\mathrm{D}\,((a\,b)\,(c\,(a\,b))^*) = \{(a\,b)\,(c\,(a\,b))^*\,,\ b\,(c\,(a\,b))^*, (c\,(a\,b))^*\}.$$

In both cases, the breaking derivation yields the same result as the derivation: $\mathrm{D}\,(a\,(b\,(c\,(a\,b))^*)) = \mathrm{BD}\,(a\,(b\,(c\,(a\,b))^*))$ and $\mathrm{D}\,((a\,b)\,(c\,(a\,b))^*) = \mathrm{BD}\,((a\,b)\,(c\,(a\,b))^*)$.

Theorem 15 *Let* E*,* F *and* G *be three rational expressions. Then, the following holds:*

$$\operatorname{card}\,(\mathrm{D}\,((\mathsf{E} \cdot \mathsf{F}) \cdot \mathsf{G})) \leqslant \operatorname{card}\,(\mathrm{D}\,(\mathsf{E} \cdot (\mathsf{F} \cdot \mathsf{G})))\,, \tag{64}$$
$$\operatorname{card}\,(\mathrm{BD}\,((\mathsf{E} \cdot \mathsf{F}) \cdot \mathsf{G})) \leqslant \operatorname{card}\,(\mathrm{BD}\,(\mathsf{E} \cdot (\mathsf{F} \cdot \mathsf{G})))\,. \tag{65}$$

Proof. We only prove (65) since the same proof goes for (64). The idea of this proof is to compute $\mathrm{BD}\,((\mathsf{E} \cdot \mathsf{F}) \cdot \mathsf{G})$ and $\mathrm{BD}\,(\mathsf{E} \cdot (\mathsf{F} \cdot \mathsf{G}))$ as union of sets and to study the intersection of these two unions.

The trivial cases are not relevant for the proof and we assume that $\mathsf{E}, \mathsf{F}, \mathsf{G} \neq 1$. By

(28) and (31) the following holds:

$$\begin{aligned}
\mathrm{BD}\,(\mathsf{E}\cdot(\mathsf{F}\cdot\mathsf{G})) &= (\mathrm{BD}\,(\mathsf{E}))_{\mathsf{p}}\cdot(\mathsf{F}\cdot\mathsf{G})\;\cup\;\mathrm{TBD}\,(\mathsf{F}\cdot\mathsf{G})\;\cup\;\delta_{\mathrm{BD}(\mathsf{E})}\mathrm{B}\,(\mathsf{F}\cdot\mathsf{G})\\
&= (\mathrm{BD}\,(\mathsf{E}))_{\mathsf{p}}\cdot(\mathsf{F}\cdot\mathsf{G})\;\cup\;(\mathrm{TBD}\,(\mathsf{F}))_{\mathsf{p}}\cdot\mathsf{G}\;\cup\;\delta_{\mathrm{TBD}(\mathsf{F})}\mathrm{B}\,(\mathsf{G})\\
&\quad\cup\;\delta_{\mathrm{BD}(\mathsf{E})}\,(\mathrm{B}\,(\mathsf{F}))_{\mathsf{p}}\cdot\mathsf{G}\;\cup\;\mathrm{TBD}\,(\mathsf{G})\;\cup\;\delta_{\mathrm{BD}(\mathsf{E})}\delta_{\mathrm{B}(\mathsf{F})}\mathrm{B}\,(\mathsf{G})\,.\\
\mathrm{BD}\,((\mathsf{E}\cdot\mathsf{F})\cdot\mathsf{G}) &= (\mathrm{BD}\,(\mathsf{E}\cdot\mathsf{F}))_{\mathsf{p}}\cdot\mathsf{G}\;\cup\;\mathrm{TBD}\,(\mathsf{G})\;\cup\;\delta_{\mathrm{BD}(\mathsf{E}\cdot\mathsf{F})}\mathrm{B}\,(\mathsf{G})\\
&= \Big((\mathrm{BD}\,(\mathsf{E}))_{\mathsf{p}}\cdot\mathsf{F}\Big)\cdot\mathsf{G}\;\cup\;(\mathrm{TBD}\,(\mathsf{F}))_{\mathsf{p}}\cdot\mathsf{G}\;\cup\;\delta_{\mathrm{BD}(\mathsf{E})}\,(\mathrm{B}\,(\mathsf{F}))_{\mathsf{p}}\cdot\mathsf{G}\\
&\quad\cup\;\mathrm{TBD}\,(\mathsf{G})\;\cup\;\delta_{\mathrm{BD}(\mathsf{E}\cdot\mathsf{F})}\mathrm{B}\,(\mathsf{G})\,.
\end{aligned}$$

Let us first rewrite the last term of the last equation. Since $\delta_{(\mathrm{TBD}(\mathsf{E}))_{\mathsf{p}}} = 0$, and by (31), the following holds:

$$\delta_{\mathrm{BD}(\mathsf{E}\cdot\mathsf{F})} = \delta_{\mathrm{TBD}(\mathsf{F})} + \delta_{\mathrm{BD}(\mathsf{E})}\delta_{\mathrm{B}(\mathsf{F})}\cdot$$

and hence:

$$\delta_{\mathrm{BD}(\mathsf{E}\cdot\mathsf{F})}\mathrm{B}\,(\mathsf{G}) = \delta_{\mathrm{TBD}(\mathsf{F})}\mathrm{B}\,(\mathsf{G})\;\cup\;\delta_{\mathrm{BD}(\mathsf{E})}\delta_{\mathrm{B}(\mathsf{F})}\mathrm{B}\,(\mathsf{G})\,.$$

Let us denote:

$$X = \Big(\mathrm{TBD}\,(\mathsf{F})_{\mathsf{p}}\Big)\cdot\mathsf{G}\;\cup\;\delta_{\mathrm{BD}(\mathsf{E})}\mathrm{B}\,(\mathsf{F})_{\mathsf{p}}\cdot\mathsf{G},\qquad\text{and}\qquad Y = \mathrm{TBD}\,(\mathsf{G})\;\cup\;\delta_{\mathrm{BD}(\mathsf{E}\cdot\mathsf{F})}\mathrm{B}\,(\mathsf{G})\,.$$

The following holds:

$$\begin{aligned}
\mathrm{BD}\,((\mathsf{E}\cdot\mathsf{F})\cdot\mathsf{G}) &= \Big((\mathrm{BD}\,(\mathsf{E}))_{\mathsf{p}}\cdot\mathsf{F}\Big)\cdot\mathsf{G}\;\cup\;X\;\cup\;Y,\\
\mathrm{BD}\,(\mathsf{E}\cdot(\mathsf{F}\cdot\mathsf{G})) &= (\mathrm{BD}\,(\mathsf{E}))_{\mathsf{p}}\cdot(\mathsf{F}\cdot\mathsf{G})\;\cup\;X\;\cup\;Y.
\end{aligned}$$

There is a 1-1 correspondence between the sets $((\mathrm{BD}\,(\mathsf{E}))_{\mathsf{p}}\cdot\mathsf{F})\cdot\mathsf{G}$ and $(\mathrm{BD}\,(\mathsf{E}))_{\mathsf{p}}\cdot(\mathsf{F}\cdot\mathsf{G})$ which have then the same cardinality.

It is then sufficient to establish that $(\mathrm{BD}\,(\mathsf{E}))_{\mathsf{p}}\cdot(\mathsf{F}\cdot\mathsf{G})\;\cap\;(X\;\cup\;Y) = \emptyset$ since $\mathrm{BD}\,((\mathsf{E}\cdot\mathsf{F})\cdot\mathsf{G})$ will then have at most as many elements as $\mathrm{BD}\,(\mathsf{E}\cdot(\mathsf{F}\cdot\mathsf{G}))$. More precisely, we study the intersection of $(\mathrm{BD}\,(\mathsf{E}))_{\mathsf{p}}\cdot(\mathsf{F}\cdot\mathsf{G})$ with X and Y.

Every expression in X consists of a concatenation whose right term is G while every expression in $(\mathrm{BD}\,(\mathsf{E}))_{\mathsf{p}}\cdot(\mathsf{F}\cdot\mathsf{G})$ consists of a concatenation whose right term is $\mathsf{F}\cdot\mathsf{G}$. Since $\mathsf{F}\cdot\mathsf{G}\neq\mathsf{G}$, we have $(\mathrm{BD}\,(\mathsf{E}))_{\mathsf{p}}\cdot(\mathsf{F}\cdot\mathsf{G})\;\cap\;X = \emptyset$.

By definition of the broken derived terms, if H is a broken derived term of G and H is a concatenation $\mathsf{H} = \mathsf{H}_1\cdot\mathsf{H}_2$ then H_2 is a subexpression of G. Since $\mathsf{F}\cdot\mathsf{G}$ cannot be a subexpression of G, we have: $(\mathrm{BD}\,(\mathsf{E}))_{\mathsf{p}}\cdot(\mathsf{F}\cdot\mathsf{G})\;\cap\;Y = \emptyset$. □

Remark 2 The derivatives as defined by Brzozowski in [3] are also sensitive to the bracketing of the product but there is no clear advantage for left or right bracketing. For instance, the left bracketing of $ab(cab)^*$ yields a smaller set of derivatives than the right one (cf. Example 8) whereas the converse holds for $(a+b)^*ab(a+b)^*$ (cf. [10, Exercise I.4.10] for instance).

References

[1] V. Antimirov. Partial derivatives of regular expressions and finite automaton constructions. *Theoret. Comput. Sci.* **155** (1996), 291–319.

[2] A. Brüggemann-Klein. Regular expressions into finite automata. *Theoret. Comput. Sci.* **120** (1993), 197–213.

[3] J. A. Brzozowski. Derivatives of regular expressions. *J. Assoc. Comput. Mach.* **11** (1964), 481–494.

[4] J.-M. Champarnaud, D. Ziadi. Canonical derivatives, partial derivatives and finite automaton constructions. *Theoret. Comput. Sci.* **289** (2002), 137–163.

[5] J. H. Conway. *Regular Algebra and Finite Machines*, Chapman and Hall, 1971.

[6] V. Glushkov. The abstract theory of automata. *Russian Mathematical Surveys* **16** (1961), 1–53.

[7] S. Lombardy, J. Sakarovitch. Derivatives of rational expressions with multiplicity, *Theoret. Comput. Sci.* **332** (2005), 141–177.

[8] S. Lombardy, J. Sakarovitch. How expressions can code for automata, *RAIRO – Theoret. Inform. and Appl.* **39** (2005), 217–237.

[9] S. Lombardy, J. Sakarovitch. Corrigendum to our paper 'How expressions can code for automata', *RAIRO – Theoret. Inform. and Appl.* to appear.
`http://www.enst.fr/~jsaka/PUB/Files/HCE.pdf`

[10] J. Sakarovitch. *Eléments de théorie des automates*, Vuibert, 2003. Corrected English edition: *Elements of Automata Theory*, Cambridge University Press, 2009.

APPENDIX

A. Proof of Proposition 4

Proposition 4 ([7]) *Let* F *and* G *be two rational expressions. Then the following holds.*

$$\begin{aligned} \mathrm{TD}\left((\mathsf{F}+\mathsf{G})\right) &= \mathrm{TD}\left(\mathsf{F}\right) \cup \mathrm{TD}\left(\mathsf{G}\right), && (12)\\ \mathrm{TD}\left((\mathsf{F}\cdot\mathsf{G})\right) &= \left(\mathrm{TD}\left(\mathsf{F}\right)\right)\cdot\mathsf{G} \cup \mathrm{TD}\left(\mathsf{G}\right), && (13)\\ \mathrm{TD}\left((\mathsf{F}^*)\right) &= \left(\mathrm{TD}\left(\mathsf{F}\right)\right)\cdot\mathsf{F}^*. && (14)\end{aligned}$$

Proof. These equations follow from the application of the inductive definition of derivation by words (Equation (8)) to the derivation of a sum, a product, and a star, of an expression.

For the sum, let us prove the following:

$$\forall f \in A^+ \qquad \frac{\partial}{\partial f}(\mathsf{F}+\mathsf{G}) = \frac{\partial}{\partial f}\mathsf{F} \cup \frac{\partial}{\partial f}\mathsf{G}. \tag{66}$$

Equation (66) is (3) for words f of length 1; the following sequence of equalities establishes it by induction for words of greater length.

$$\begin{aligned} &\forall f \in A^+,\ \forall a \in \mathsf{RatE}(A)\\ &\frac{\partial}{\partial fa}(\mathsf{F}+\mathsf{G}) = \frac{\partial}{\partial a}\left(\frac{\partial}{\partial f}(\mathsf{F}+\mathsf{G})\right) = \frac{\partial}{\partial a}\left(\frac{\partial}{\partial f}\mathsf{F} \cup \frac{\partial}{\partial f}\mathsf{G}\right)\\ &\qquad = \frac{\partial}{\partial a}\left(\frac{\partial}{\partial f}\mathsf{F}\right) \cup \frac{\partial}{\partial a}\left(\frac{\partial}{\partial f}\mathsf{G}\right) = \frac{\partial}{\partial fa}\mathsf{F} \cup \frac{\partial}{\partial fa}\mathsf{G}.\end{aligned}$$

The definition (9) of $\mathrm{TD}\left(\mathsf{E}\right)$ applied to (66) yields (12).

For the product, let us prove the following:

$$\forall f \in A^+ \quad \frac{\partial}{\partial f}(\mathsf{F}\cdot\mathsf{G}) = \left(\frac{\partial}{\partial f}\mathsf{F}\right)\cdot\mathsf{G} \cup \left(\bigcup_{\substack{g,h\in A^+\\ gh=f}} \mathsf{c}\left(\frac{\partial}{\partial g}\mathsf{F}\right)\frac{\partial}{\partial h}\mathsf{G}\right) \cup \mathsf{c}\left(\mathsf{F}\right)\frac{\partial}{\partial f}\mathsf{G}. \tag{67}$$

Equation (67) is (4) for words f of length 1; the following sequence of equalities establishes it by induction for words of greater length.

$$\begin{aligned} &\forall f \in A^+,\ \forall a \in \mathsf{RatE}(A)\\ &\frac{\partial}{\partial fa}(\mathsf{F}\cdot\mathsf{G}) = \frac{\partial}{\partial a}\left(\frac{\partial}{\partial f}(\mathsf{F}\cdot\mathsf{G})\right)\\ &\qquad = \frac{\partial}{\partial a}\left(\left(\frac{\partial}{\partial f}\mathsf{F}\right)\cdot\mathsf{G} \cup \left(\bigcup_{\substack{g,h\in A^+\\ gh=f}} \mathsf{c}\left(\frac{\partial}{\partial g}\mathsf{F}\right)\frac{\partial}{\partial h}\mathsf{G}\right) \cup \mathsf{c}\left(\mathsf{F}\right)\frac{\partial}{\partial f}\mathsf{G}\right)\end{aligned}$$

$$\begin{aligned}
&= \left(\frac{\partial}{\partial a}\left(\frac{\partial}{\partial f}\,\mathsf{F}\right)\right)\cdot\mathsf{G} \cup \mathsf{c}\left(\frac{\partial}{\partial f}\,\mathsf{F}\right)\frac{\partial}{\partial a}\,\mathsf{G}\\
&\qquad\qquad \cup \left(\bigcup_{\substack{g,h\in A^+\\ gh=f}} \mathsf{c}\left(\frac{\partial}{\partial g}\,\mathsf{F}\right)\frac{\partial}{\partial a}\left(\frac{\partial}{\partial h}\,\mathsf{G}\right)\right)\\
&\qquad\qquad \cup \mathsf{c}\,(\mathsf{F})\,\frac{\partial}{\partial a}\left(\frac{\partial}{\partial f}\,\mathsf{G}\right)\\
&= \left(\frac{\partial}{\partial fa}\,\mathsf{F}\right)\cdot\mathsf{G}\cup\left(\bigcup_{\substack{g,h\in A^+\\ gh=fa}}\mathsf{c}\left(\frac{\partial}{\partial g}\,\mathsf{F}\right)\frac{\partial}{\partial h}\,\mathsf{G}\right)\cup\mathsf{c}\,(\mathsf{F})\,\frac{\partial}{\partial fa}\,\mathsf{G}.
\end{aligned}$$

By Property 2 (and as F is not a constant), there exists a g in A^+ such that $\mathsf{c}\left(\frac{\partial}{\partial g}\,\mathsf{F}\right) = 1$ and thus the definition (9) of $\mathrm{TD}\,(\mathsf{E})$ applied to (67) yields (13).

For the star, the exact expression of $\frac{\partial}{\partial f}\,(\mathsf{F}^*)$ is long, unnecessarily complicated for establishing (14) which is our aim. Let us prove instead the following double inclusion:

$$\forall f\in A^+ \qquad \left(\frac{\partial}{\partial f}\,\mathsf{F}\right)\cdot\mathsf{F}^*\subseteq\frac{\partial}{\partial f}\,(\mathsf{F}^*)\subseteq\bigcup_{\substack{h\in A^+\\ gh=f}}\left(\frac{\partial}{\partial h}\,\mathsf{F}\right)\cdot\mathsf{F}^*. \tag{68}$$

Equation (5) gives the equality among the three sets in (68) for words f of length 1. For the words of greater length, both inclusions are established by an easy induction.

$$\begin{aligned}
&\forall f\in A^+\,,\ \forall a\in\mathsf{RatE}(A)\\
&\quad\frac{\partial}{\partial fa}\,(\mathsf{F}^*) = \frac{\partial}{\partial a}\left(\frac{\partial}{\partial f}\,(\mathsf{F}^*)\right)\\
&\qquad\qquad\supseteq\frac{\partial}{\partial a}\left(\left(\frac{\partial}{\partial f}\,\mathsf{F}\right)\cdot\mathsf{F}^*\right) = \left(\frac{\partial}{\partial a}\left(\frac{\partial}{\partial f}\,\mathsf{F}\right)\cup\mathsf{c}\left(\frac{\partial}{\partial f}\,\mathsf{F}\right)\frac{\partial}{\partial a}\,\mathsf{F}\right)\cdot\mathsf{F}^*\\
&\qquad\qquad\supseteq\left(\frac{\partial}{\partial fa}\,\mathsf{F}\right)\cdot\mathsf{F}^*.
\end{aligned}$$

$$\begin{aligned}
&\forall f\in A^+\,,\forall a\in\mathsf{RatE}(A)\\
&\frac{\partial}{\partial fa}\,(\mathsf{F}^*)\subseteq\frac{\partial}{\partial a}\left(\bigcup_{\substack{h\in A^+\\ gh=f}}\left(\frac{\partial}{\partial h}\,\mathsf{F}\right)\cdot\mathsf{F}^*\right) = \bigcup_{\substack{h\in A^+\\ gh=f}}\left(\left(\frac{\partial}{\partial ha}\,\mathsf{F}\right)\cdot\mathsf{F}^*\cup\mathsf{c}\left(\frac{\partial}{\partial h}\,\mathsf{F}\right)\frac{\partial}{\partial a}\,\mathsf{F}\cdot\mathsf{F}^*\right)\\
&\qquad\subseteq\bigcup_{\substack{h\in A^+\\ gh=fa}}\left(\frac{\partial}{\partial h}\,\mathsf{F}\right)\cdot\mathsf{F}^*.
\end{aligned}$$

If one takes the union of the double inclusion (68) for all f in A^+, the two extreme terms are equal, and thus equal to the middle one, which yields (14). □

B. Proof of Proposition 10

Proposition 10 relies first on an easy lemma on languages.

Lemma 16 *Let L and K be two languages over A^*. If $\mathsf{c}(L) = \mathsf{c}(K) = 1$, then $(L \cup K)^* = (LK)^*$.*

Proof. For a language L, let us denote by L_p the proper part of L : $L_p = L \setminus 1_{A^*}$. If $\mathsf{c}(L) = 1$, then $L = \{1_{A^*}\} \cup L_p$. As $*$ is an isotone operator, that is, $L \subseteq H$ implies $L^* \subseteq H^*$, the following inclusion holds:

$$\begin{aligned}(LK)^* &= \big((1_{A^*} \cup L_p)(1_{A^*} \cup K_p)\big)^* \\ &= (1_{A^*} \cup L_p \cup K_p \cup L_pK_p)^* \supseteq (1_{A^*} \cup L_p \cup K_p)^* = (L \cup K)^*.\end{aligned}$$

For the converse inclusion, we first note that, for any language L, $(L \cup L^2)^* = L^*$. Indeed, for all n:

$$\bigcup_{0 \leqslant i \leqslant n} L^i \subseteq \bigcup_{0 \leqslant i \leqslant n} (L \cup L^2)^i \subseteq \bigcup_{0 \leqslant i \leqslant 2n} L^i.$$

and the former equality is obtained when n tends to infinity. We then have:

$$(L \cup K)^* = ((L \cup K) \cup (L \cup K)^2)^* = (L \cup K \cup L^2 \cup K^2 \cup LK \cup KL)^* \supseteq (LK)^*.$$

□

Proposition 10 ([2]) *For any rational expression E, the expression $\mathsf{E}^\bullet$ is in star normal form and is equivalent to E.*

Proof. In order to lighten the writing, we denote by $\mathsf{F} \simeq \mathsf{G}$ the fact that F and G are equivalent rational expressions, that is, if $L(\mathsf{F}) = L(\mathsf{G})$.

We establish by a simultaneous induction the following two equations:

$$\mathsf{E}^\bullet \text{ is in SNF} \quad \text{and} \quad \mathsf{E}^\bullet \simeq \mathsf{E} \tag{69}$$

$$\mathsf{E}^\square \text{ is in SNF} \quad \text{and} \quad (\mathsf{E}^\square)^* \simeq \mathsf{E}^*. \tag{70}$$

Both (69) and (70) clearly hold for the base clauses.

Let $\mathsf{E} = \mathsf{F} + \mathsf{G}$: (69) holds trivially by induction.

The expression $(\mathsf{F} + \mathsf{G})^\square = \mathsf{F}^\square + \mathsf{G}^\square$ is in SNF by induction. A double application of the well-known 'sum-star' identity (*cf.* [5] for instance) then yields:

$$\begin{aligned}((\mathsf{F} + \mathsf{G})^\square)^* &\simeq (\mathsf{F}^\square + \mathsf{G}^\square)^* \simeq ((\mathsf{F}^\square)^*\mathsf{G}^\square)^*(\mathsf{F}^\square)^* \simeq (\mathsf{F}^*\mathsf{G}^\square)^*\mathsf{F}^* \simeq (\mathsf{F} + \mathsf{G}^\square)^* \\ &\simeq ((\mathsf{G}^\square)^*\mathsf{F})^*(\mathsf{G}^\square)^* \simeq (\mathsf{G}^*\mathsf{F})^*\mathsf{G}^* \simeq (\mathsf{F} + \mathsf{G})^*.\end{aligned}$$

Let $\mathsf{E} = \mathsf{F} \cdot \mathsf{G}$: $\mathsf{E}^\bullet = \mathsf{F}^\bullet \cdot \mathsf{G}^\bullet$ and (69) holds trivially by induction.

The expression $(\mathsf{F}\cdot\mathsf{G})^{\square}$, $(\mathsf{F}\cdot\mathsf{G})^{\square}=\mathsf{F}^{\square}+\mathsf{G}^{\square}$ or $(\mathsf{F}\cdot\mathsf{G})^{\square}=\mathsf{F}^{\bullet}\cdot\mathsf{G}^{\bullet}$, is in SNF by induction.

If $\mathsf{c}\,(\mathsf{F})=\mathsf{c}\,(\mathsf{G})=1$, then:

$$\begin{aligned}((\mathsf{F}\cdot\mathsf{G})^{\square})^{*}=((\mathsf{F}+\mathsf{G})^{\square})^{*}&\simeq(\mathsf{F}+\mathsf{G})^{*}\text{ as above}\\&\simeq(\mathsf{F}\cdot\mathsf{G})^{*}\text{ by lemma}\end{aligned}$$

If $\mathsf{c}\,(\mathsf{F})\,\mathsf{c}\,(\mathsf{G})=0$, then $(\mathsf{F}\cdot\mathsf{G})^{\square}=\mathsf{F}^{\bullet}\cdot\mathsf{G}^{\bullet}\simeq\mathsf{F}\cdot\mathsf{G}$ by induction. Thus $((\mathsf{F}\cdot\mathsf{G})^{\square})^{*}\simeq(\mathsf{F}\cdot\mathsf{G})^{*}$.

Let $\mathsf{E}=\mathsf{F}^{*}$: $\mathsf{E}^{\bullet}=(\mathsf{F}^{\square})^{*}$ is in SNF, by induction and since $\mathsf{c}\,(\mathsf{F}^{\square})=0$. Moreover $(\mathsf{F}^{\square})^{*}\simeq\mathsf{F}^{*}$ by induction.

The expression $(\mathsf{F}^{*})^{\square}=(\mathsf{F}^{\square})^{*}$ is in SNF by induction and $(\mathsf{F}^{\square})^{*}\simeq\mathsf{F}^{*}\simeq(\mathsf{F}^{*})^{*}$. □