Translation of Semi-Extended Regular Expressions using Derivatives

Antoine Martin[1], Etienne Renault[2], and Alexandre Duret-Lutz[1]

[1] LRE, EPITA, Le Kremlin-Bicêtre, France
[2] SiPearl, France

Abstract. We generalize Antimirov's notion of *linear form* of a regular expression, to the Semi-Extended Regular Expressions typically used in the Property Specification Language or SystemVerilog Assertions. Doing so requires extending the construction to handle more operators, and dealing with expressions over alphabets $\Sigma = 2^{AP}$ of valuations of atomic propositions. Using linear forms to construct automata labeled by Boolean expressions suggests heuristics that we evaluate. Finally, we study a variant of this translation to automata with accepting transitions: this construction is more natural and provides smaller automata.

Keywords: Regular expressions · Automata · PSL · SVA · Derivative

1 Introduction

In this paper we discuss and compare techniques for translating (extended) regular expressions over alphabets $\Sigma = 2^{AP}$ where letters describe valuations of a set of atomic propositions AP. Such alphabets are typically used in formal methods such as *model checking* [3], *runtime verification* [4] or *synthesis* [14]. In our case, we are interested in supporting the regular expression operators of the PSL and SVA standards.

Property Specification Language (PSL) [13] and SystemVerilog Assertions (SVA) [1] are two industrial formal verification languages used in the field of hardware design and verification. These two languages include features for describing linear-time temporal properties or reasoning with clocks, but we restrict ourselves to the (semi-extended) regular expression properties. Both offer a nearly identical set of operators, albeit with different syntaxes.

For instance the expression, "`btn : ((red[=2] && opn[->]) || rst[->])`" is a PSL expression matching any sequence that starts with the `btn` signal on, and in which either the `red` signal is on exactly twice in the interval it takes for the signal `opn` turn on, or in which the `rst` signal turns on. In SVA this expression becomes "`btn ##0 ((red[=2] intersect opn[->1]) or rst[->1])`".

Other logics such as Linear Dynamic Logic (LDL) [16] or Propositional Dynamic Logic (PDL) [15] also use regular expressions over atomic propositions, so our work applies to them too, however, and unlike SVA or PSL these logics are usually defined only with classical regular operators plus a $\varphi?$ operator that is absent from PSL and SVA, and that we do not consider.

2 Definitions

In the entire paper, we assume an alphabet $\Sigma = 2^{AP}$ where letters describe valuations of a set of *atomic propositions* AP. For instance, if $AP = \{a, b\}$ then we denote the valuations as $\Sigma = \{\bar{a}\bar{b}, \bar{a}b, a\bar{b}, ab\}$. We use $\Sigma^\star$ to denote the set of finite sequences of valuations. For a sequence $\sigma \in \Sigma^\star$, we denote $|\sigma|$ its length, and we write $\sigma(i)$ for the letter at position $i \in \{0, 1, \ldots, |\sigma| - 1\}$ in σ. Using ";" as concatenation operator, we write $\sigma = \sigma(0)\,;\sigma(1)\,;\ldots;\sigma(|\sigma|-1)$. Finally, for two integers i, j such that $0 \leq i \leq j < |\sigma|$, we write $\sigma^{i..j}$ the (possibly empty if $i = j$) subsequence $\sigma(i)\,;\sigma(i+1)\,;\ldots;\sigma(j-1)$. For convenience, we write $\sigma^{i..}$ instead of $\sigma^{i..|\sigma|}$ and $\sigma^{..j}$ instead of $\sigma^{0..j}$.

Definition 1 (Boolean expression). *Any Boolean expression b is built using the following grammar, where $a \in AP$ can be any atomic proposition.*

$$b ::= \bot \mid \top \mid a \mid (b \vee b) \mid (b \wedge b) \mid \neg b$$

For convenience, we omit unnecessary parentheses, and use operators $\rightarrow$ and $\leftrightarrow$ as syntactic sugar with their obvious definitions.

Boolean expression are interpreted over a valuation $v \in \Sigma$ in the obvious way. We write $v \models b$ when the valuation v satisfies b, and $b \equiv b'$ when two Boolean expressions b and b' are satisfied by the same valuations.[3]

We use $\mathbb{B} = \{\bot, \top\}$ to denote the set of Boolean values, and $\mathbb{B}(AP)$ to denote the set of Boolean expressions over AP.

Definition 2 (SERE). *A Semi-Extended Regular Expression (SERE) r is built using the following grammar:*

$$r ::= b \mid \varepsilon \mid (r\,;r) \mid (r:r) \mid r^\star \mid (r \vee r) \mid (r \wedge r) \mid \mathsf{fm}(r)$$

The symbol "ε" is called the empty word*. Operators "$\vee$" (choice), ";" (concatenation) and "*" (Kleene star) are traditional regular operators. SERE extends those with "$\wedge$" (intersection) ":" (fusion), and "*fm*" (SVA's* first-match*). In practice, we omit parentheses when they are not necessary.*

The set of all SEREs is written SERE*.*

SEREs are interpreted over a finite sequence $\sigma \in \Sigma^\star$ of valuations defined inductively as follows:

$$\begin{aligned}
\sigma \models b \quad & \textit{iff} \quad |\sigma| = 1 \wedge \sigma(0) \models b \\
\sigma \models \varepsilon \quad & \textit{iff} \quad |\sigma| = 0 \\
\sigma \models (r_1\,;r_2) \quad & \textit{iff} \quad \exists i \geq 0,\ \sigma^{..i} \models r_1 \wedge \sigma^{i..} \models r_2 \\
\sigma \models (r_1 : r_2) \quad & \textit{iff} \quad \exists i \geq 0,\ \sigma^{..i+1} \models r_1 \wedge \sigma^{i..} \models r_2 \\
\sigma \models r^\star \quad & \textit{iff} \quad \textit{either } |\sigma| = 0 \textit{ or } \exists i > 0,\ \sigma^{..i} \models r \wedge \sigma^{i..} \models r^\star \\
\sigma \models (r_1 \vee r_2) \quad & \textit{iff} \quad \sigma \models r_1 \vee \sigma \models r_2 \\
\sigma \models (r_1 \wedge r_2) \quad & \textit{iff} \quad \sigma \models r_1 \wedge \sigma \models r_2 \\
\sigma \models \mathsf{fm}(r) \quad & \textit{iff} \quad (\sigma \models r) \wedge (\forall i < |\sigma|,\ \sigma^{..i} \not\models r)
\end{aligned}$$

[3] Testing $b \equiv b'$ is straightforward if b and b' are represented with BDDs [5].

The language *of a SERE r is the set $\mathscr{L}(r) = \{\sigma \in \Sigma^\star \mid \sigma \models r\}$ of all sequences satisfying r (or "matched" by r).*

In the above definition, regular expressions have been extended with three operators: the conjunction "$\wedge$" has obvious meaning, the fusion operator "$:$" ensures that the last letter matching the left operand is also the first letter of matching the right operand (this implies that a fusion can never match the empty word), and finally fm, the first-match operator of SVA, retains only the shortest possible match for a SERE r.

The PSL and SVA specifications defines other SERE operators (such as `[=`n`]` or `[->`n`]`) that can be seen as syntactic sugar on the above. In our syntax, the expression from the introduction becomes

$$\varphi = btn : \Big(\big(((\neg red)^\star ; red ; (\neg red)^\star ; red ; (\neg red)^\star) \wedge ((\neg opn)^\star ; opn)\big) \vee \big((\neg rst)^\star ; rst\big)\Big)$$

Definition 3 (Constant Term). *The* constant term *of an expression r, denoted $\lambda(r)$ is defined inductively as follows for any Boolean formula b and any SEREs r_1, r_2.*

$$\begin{aligned} \lambda(b) &= \bot & \lambda(r_1 \vee r_2) &= \lambda(r_1) \vee \lambda(r_2) \\ \lambda(\varepsilon) &= \varepsilon & \lambda(r_1 \wedge r_2) &= \lambda(r_1) \wedge \lambda(r_2) \\ \lambda(r_1 : r_2) &= \bot & \lambda(r_1 \,;\, r_2) &= \lambda(r_1) \,;\, \lambda(r_2) \\ \lambda(r_1^\star) &= \varepsilon & \lambda(\mathsf{fm}(r_1)) &= \lambda(r_1) \end{aligned}$$

Proposition 1. *With the above notation, $\lambda(r) = \varepsilon$ iff $\varepsilon \models r$.*

Definition 4 (Syntactic equivalence). *Given two SEREs r_1 and r_2, we say that they are syntactically equivalent, denoted $r_1 \mathrel{\mathring{=}} r_2$, if one can be rewritten into the other using the following so called ACI-rules (associativity, commutativity, and idempotence) and a few others:*

$$(r_1 \odot r_2) \odot r_3 \mathrel{\mathring{=}} r_1 \odot (r_2 \odot r_3) \mathrel{\mathring{=}} r_1 \odot r_2 \odot r_3 \qquad \text{for } \odot \in \{;, :, \vee, \wedge\} \qquad \text{(A)}$$

$$r_1 \odot r_2 \mathrel{\mathring{=}} r_2 \odot r_1 \qquad \text{for } \odot \in \{\vee, \wedge\} \qquad \text{(C)}$$

$$r_1 \odot r_1 \mathrel{\mathring{=}} r_1 \qquad \text{for } \odot \in \{\vee, \wedge\} \qquad \text{(I1)}$$

$$r^{\star\star} \mathrel{\mathring{=}} r^\star \qquad \mathsf{fm}(\mathsf{fm}(r)) \mathrel{\mathring{=}} \mathsf{fm}(r) \qquad \text{(I2)}$$

$$r \vee \bot \mathrel{\mathring{=}} r \qquad r \wedge \top \mathrel{\mathring{=}} r \qquad r \,;\, \varepsilon \mathrel{\mathring{=}} \varepsilon \,;\, r \mathrel{\mathring{=}} r \qquad \text{(Z)}$$

$$r \wedge \bot \mathrel{\mathring{=}} \bot \qquad r \,;\, \bot \mathrel{\mathring{=}} \bot \,;\, r \mathrel{\mathring{=}} \bot \qquad \text{(U1)}$$

$$r \vee \top \mathrel{\mathring{=}} \top \qquad r : \bot \mathrel{\mathring{=}} \bot : r \mathrel{\mathring{=}} \bot \qquad \text{(U2)}$$

$$r : \varepsilon \mathrel{\mathring{=}} \varepsilon : r \mathrel{\mathring{=}} \bot \qquad \text{(U3)}$$

$$\mathsf{fm}(r) \mathrel{\mathring{=}} \varepsilon \text{ if } \varepsilon \models r \qquad \text{(F)}$$

Proposition 2. *Two syntactically equivalent SEREs have the same language. I.e., $(r_1 \mathrel{\mathring{=}} r_2) \implies (\mathscr{L}(r_1) = \mathscr{L}(r_2))$.* See Appendix B.

From an implementation standpoint, it is straightforward to rewrite any SERE r into a unique representative of its equivalence class $[r]_{\triangleq} = \{s \in \mathsf{SERE} \mid r \triangleq s\}$. This can be achieved by applying the above rewriting rules during the construction of the syntax tree of r. In particular rule (A) can be implemented by considering these operators as n-ary (rather than binary), and then rules (C) and (I1) can be implemented by sorting operands and removing duplicates. Rule (F) can be implemented at construction as well, by deciding $\varepsilon \models r$ using Proposition 1.

Definition 5 (NFA). *A* nondeterministic finite automaton *is a tuple* $\mathcal{A} = \langle Q, \delta, \iota, F\rangle$ *where* Q *is a finite set of states,* $\delta \subseteq Q \times \mathbb{B}(AP) \times Q$ *is the transition relation,* $\iota \in Q$ *is the initial state, and* $F \subseteq Q$ *is the set of final states.*

We write $s \xrightarrow{f} d$ *when* $(s, f, d) \in \delta$.

A sequence of valuations $\sigma \in \Sigma^n$ *of size* n *is* accepted *by* $\mathcal{A}$ *if either* $n = 0$ *and* $\iota \in F$, *or* $n > 0$ *and there exists a sequence of transitions* $\rho = s_0 \xrightarrow{f_0} s_1 \xrightarrow{f_1} \cdots \xrightarrow{f_{n-1}} s_n$ *such that* $s_0 = \iota$, $s_n \in F$, *and for each* i, $\sigma(i) \models f_i$.

The language *of* $\mathcal{A}$, *denoted* $\mathscr{L}(\mathcal{A})$ *is the set of words accepted by* $\mathcal{A}$.

A Deterministic Finite Automaton *(DFA) is an NFA where transitions leaving each state have mutually exclusive Boolean expressions. Formally, an automaton is a DFA if for any two different transitions* $s \xrightarrow{f} d$ *and* $s \xrightarrow{f'} d'$ *with the same origin, we have* $f \wedge f' \equiv \bot$.

Such automata are sometimes called symbolic finite automata [9], however in our case the alphabet is always finite, so they can be handled in a usual way.

3 Building Automata using Linear Forms

In 1964, Brzozowsky [6] introduced the notion of *derivative* of a regular expression, allowing the construction of an equivalent deterministic finite automaton. This work was extended in 1995 by Antimirov [2], with a notion of *partial derivatives* allowing the construction of a non-deterministic finite automaton. More importantly, Antimirov introduced the concept of *linear form* of a regular expression as a more efficient way to compute the set of *partial derivatives*. An extension of *partial derivatives* was proposed by Caron et al. [7] to handle intersection and complement. Here, we adapt the concept of *linear form*, to SERE with an alphabet over 2^{AP} and their specific operators. In particular, the fact that our alphabet is exponential in the number of atomic propositions makes *linear forms* much more attractive than *partial derivatives*, because using the latter to build an automaton requires iterating over exponentially many letters.

3.1 Linear Forms

Definition 6 (Linear Form). *A* linear form *for a SERE* r *is a finite set of pairs* $\{(p_1, s_1), (p_2, s_2), \ldots\}$ *where* $p_i \in \mathbb{B}(AP)$, $p_i \not\equiv \bot$, $s_i \in \mathsf{SERE}$ *and* $s_i \not\triangleq \bot$, *such that* $\bigcup_i \mathscr{L}(p_i\,;s_i) = \mathscr{L}(r) \setminus \{\varepsilon\}$.

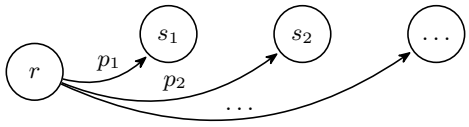

Fig. 1: Automaton view of a linear form $\{(p_1, s_1), (p_2, s_2), \ldots\}$ for a SERE r.

This definition differs from that of Antimirov [2] in that p_i is a satisfiable Boolean expression rather than a letter, and in that we explicitly forbid $s_i \stackrel{\circ}{=} \bot$. To simplify the upcoming notations we assume that any pair (p_i, s_i) in a linear form we construct is implicitly ignored when $p_i \equiv \bot$; for instance we shall write $\{\ldots, (p_i \wedge \neg p_j, s_i), \ldots\}$ with the implicit assumption that the pair $(p_i \wedge \neg p_j, s_i)$ must be omitted when $p_i \wedge \neg p_j$ is not satisfiable.

As we saw in Proposition 1, an empty sequence may only be matched by a SERE r if $\lambda(r) = \varepsilon$. If a non-empty sequence σ is matched by a SERE r, then Definition 6 implies that a linear form for r will have at least one pair (p_i, s_i) such that $\sigma(0) \models p_i$ and $\sigma^{1\cdot\cdot} \models s_i$. As a mental model for a linear form, it is useful to interpret it as the partial automaton shown in Figure 1: the p_is are Boolean formulas evaluated against $\sigma(0)$, and the p_is tell what SERE should be checked against the suffix $\sigma^{1\cdot\cdot}$. The constraints, in Definition 6, that $p_i \not\equiv \bot$, and $s_i \not\stackrel{\circ}{=} \bot$, prevent the creation of paths that will not recognize any word.

A SERE may have multiple linear forms; some will be called *deterministic*.

Definition 7 (Deterministic Linear Form). *A linear form* $\{(p_1, s_1), (p_2, s_2), \ldots\}$ *is* deterministic *if for any* $i \neq j$, $p_i \wedge p_j \equiv \bot$.

Here is a linear form for the formula φ of Section 2.

$$\begin{aligned} L_1 = \big\{ &\big(btn\wedge\neg red\wedge\neg opn, ((\neg red)^\star; red; (\neg red)^\star; red; (\neg red)^\star) \wedge ((\neg opn)^\star; opn)\big),\\ &\big(btn\wedge red\wedge\neg opn, \quad ((\neg red)^\star; red; (\neg red)^\star) \wedge ((\neg opn)^\star; opn)\big),\\ &\big(btn\wedge\neg rst, \qquad (\neg rst)^\star; rst\big),\\ &\big(btn\wedge rst, \qquad \varepsilon\big)\big\} \end{aligned}$$

L_1 is not deterministic because, for instance, $btn\wedge red\wedge\neg opn$ can hold together with $btn\wedge rst$. Here is a deterministic linear form for φ:

$$\begin{aligned} L_2 = \big\{ &\big(btn\wedge\neg red\wedge\neg opn\wedge\neg rst,\\ &\quad (((\neg red)^\star; red; (\neg red)^\star; red; (\neg red)^\star) \wedge ((\neg opn)^\star; opn)) \vee ((\neg rst)^\star; rst)\big),\\ &\big(btn\wedge\neg red\wedge\neg opn\wedge rst,\\ &\quad (((\neg red)^\star; red; (\neg red)^\star; red; (\neg red)^\star) \wedge ((\neg opn)^\star; opn)) \vee \varepsilon\big),\\ &\big(btn\wedge red\wedge\neg opn\wedge\neg rst,\\ &\quad (((\neg red)^\star; red; (\neg red)^\star) \wedge ((\neg opn)^\star; opn)) \vee ((\neg rst)^\star; rst)\big),\\ &\big(btn\wedge red\wedge\neg opn\wedge rst, \quad (((\neg red)^\star; red; (\neg red)^\star) \wedge ((\neg opn)^\star; opn)) \vee \varepsilon\big)\big\} \end{aligned}$$

Property 1 (determinization of a linear form). Any linear form $\{(p_1, s_1), (p_2, s_2), \ldots\}$ for an expression r can be converted into a deterministic linear form for r.

The idea is that if $p_1 \wedge p_2 \not\equiv \bot$, then $\{(p_1 \wedge \neg p_2, s_1), (\neg p_1 \wedge p_2, s_2), (p_1 \wedge p_2, s_1 \vee s_2), \ldots\}$ is also a linear form for r. This process can be repeated for any $i \neq j$ such that $p_i \wedge p_j \not\equiv \bot$. From now on, we assume the existence of a function det that determinizes a linear form.

See Appendix A.

3.2 Linearization of SEREs

We now discuss how to convert a SERE into a linear form. For now on, we assume that equations (A)–(F) are always applied, i.e., that we are only working with unique representatives of each equivalence classes of $\stackrel{\circ}{=}$, as discussed in Section 2.

To simplify the notations, we extend the concatenation and fusion operators to linear forms: given a linear form $L = \{(p_1, s_1), (p_2, s_2), \ldots, (p_n, s_n)\}$, an operator $\odot \in \{\,;, :\}$, and a SERE r, we write $L \odot r$ instead of $\{(p, s \odot r) \mid (p, s) \in L, s \odot r \neq \bot\}$. The notation works similarly for $r \odot L$.

We forgot the restriction $s \odot r \neq \bot$ in the CIAA'24 proceedings.

Definition 8 (Linearization of a SERE). *The following* LF *function turns a SERE into a linear form. It mostly extends the "lf" function of Antimirov [2, eq. (45)–(51)] to deal with SERE operators and Boolean formulas.*

$$
\begin{aligned}
\mathsf{LF}(\bot) &= \varnothing \\
\mathsf{LF}(\varepsilon) &= \varnothing \\
\mathsf{LF}(b) &= \{(b, \varepsilon)\} \\
\mathsf{LF}(r_1 \vee r_2) &= \mathsf{LF}(r_1) \cup \mathsf{LF}(r_2) \\
\mathsf{LF}(r^\star) &= \mathsf{LF}(r)\,;r^\star \\
\mathsf{LF}(r_1\,;r_2) &= (\mathsf{LF}(r_1)\,;r_2) \cup (\lambda(r_1)\,;\mathsf{LF}(r_2)) \\
\mathsf{LF}(r_1 : r_2) &= (\mathsf{LF}(r_1) : r_2) \cup \left\{ (p_i \wedge p_j, s_j) \;\middle|\; \begin{array}{l} (p_i, s_i) \in \mathsf{LF}(r_1),\ \lambda(s_i) = \varepsilon, \\ (p_j, s_j) \in \mathsf{LF}(r_2) \end{array} \right\} \\
\mathsf{LF}(r_1 \wedge r_2) &= \{(p_i \wedge p_j, s_i \wedge s_j) \mid (p_i, s_i) \in \mathsf{LF}(r_1),\ (p_j, s_j) \in \mathsf{LF}(r_2)\} \\
\mathsf{LF}(\mathsf{fm}(r)) &= \{(p_i, \mathsf{fm}(s_i)) \mid (p_i, s_i) \in \mathsf{det}(\mathsf{LF}(r))\}
\end{aligned}
$$

As noted below Definition 6, we assume that when one of these equations generates a pair (p_i, s_i) with $p_i \equiv \bot$, it is implicitly removed. Since we assume that rules (A)–(F) are always applied, these equations cannot produce a pair (p_i, s_i) such that $s_i \stackrel{\circ}{=} \bot$, but it could nonetheless be the case that $\mathscr{L}(s_i) = \varnothing$ (for instance if $s_i = a \wedge \neg a$).

To understand the definition for $\mathsf{LF}(\mathsf{fm}(r))$, it may be useful to give an intuition of how $\mathsf{fm}(r)$ works. The SERE $\mathsf{fm}(r)$ may match σ if only if σ is the shortest prefix of σ matching r. An easy way to construct an automaton for $\mathsf{fm}(r)$ is therefore to build a DFA for r, and then remove the outgoing edges of all accepting states of that DFA. This is actually what the above definition achieves.

The use of $\mathsf{det}(\mathsf{LF}(r))$ is making sure that the linear form is deterministic, and the use of $\mathsf{fm}(s_i)$ serves two purposes: (1) it ensures that upcoming choices will still be deterministic, and (2) more importantly, by applying rule (F), it cuts the successors of s_i when $\varepsilon \models s_i$. For instance, $\mathsf{LF}(\mathsf{fm}(a\,;a^\star)) = \{(a,\varepsilon)\}$ because $\mathsf{fm}(a^\star)$ gets reduced to ε by rule (F).

To use LF in an algorithm for building an automaton that recognizes $\mathscr{L}(r)$, we need two theorems. First, $\mathsf{LF}(r)$ should be a linear form, i.e., it should preserve the language of r, except for the empty word (Theorem 1 below). Then, the number of new expressions that can be created by applying LF recursively has to be finite (Theorem 2 below).

Theorem 1. $\mathsf{LF}(r) = \{(p_1, s_1), (p_2, s_2), \ldots\}$ *is a linear form for* $r \in \mathsf{SERE}$.

Proof (Sketch). The fact hat $\bigcup_i \mathscr{L}(p_i;s_i) = \mathscr{L}(r)\backslash\{\varepsilon\}$ can be shown by induction on the structure of r using Definitions 2 and 8. □ Details in App. C

Theorem 2 (Terms). *For* $r \in \mathsf{SERE}$, *let* $\mathrm{Terms}(r)$ *denote the smallest subset of* SERE *such that* $r \in \mathrm{Terms}(r)$ *and for each* $\phi \in \mathrm{Terms}(r)$ *and each* $(p_i, s_i) \in \mathsf{LF}(\phi)$ *we have* $s_i \in \mathrm{Terms}(r)$. *Then the set* $\mathrm{Terms}(r)$ *is finite.*

Proof (Sketch). Our proof, which we omit for brevity, is inspired by a similar theorem by Antimirov [2, Theorem 3.4], however the results differ because of the new operators we support. Specifically, Antimirov did not support operators $\wedge$, :, and fm. (Of these three, : is the least problematic.) Details in App. D.

We start by addapting Antimirov's notion of *partial derivative* [2, Definition 2.8] to our context. The partial derivative of r with respect to x is defined by:

$$\partial_x r = \{s \mid (p,s) \in \mathsf{LF}(r),\, x \models p\}$$

We extend the notation to support derivation by a nonempty word $w \in \Sigma^+$ with $\partial_w r = \partial_{w^{1..}} \partial_{w(0)} r$. Furthermore, we write $\partial_{\Sigma^+} r = \bigcup_{w \in \Sigma^+} \partial_w r$ for the set of all partial derivatives one can obtain using nonempty words of any length. With these conventions we have $\mathrm{Terms}(r) = \partial_{\Sigma^+} r \cup \{r\}$.

Working on the definition of LF and $\partial_x r$, we then establish the following inequalities for two SEREs r_1 and r_2:

$$\begin{aligned}
|\partial_{\Sigma^+}(r_1 \vee r_2)| &\leq |\partial_{\Sigma^+} r_1| + |\partial_{\Sigma^+} r_2| \\
|\partial_{\Sigma^+} r_1^\star| &\leq |\partial_{\Sigma^+} r_1| \\
|\partial_{\Sigma^+}(r_1\,; r_2)| &\leq |\partial_{\Sigma^+} r_1| + |\partial_{\Sigma^+} r_2| \\
|\partial_{\Sigma^+}(r_1 : r_2)| &\leq |\partial_{\Sigma^+} r_1| + |\partial_{\Sigma^+} r_2| \\
|\partial_{\Sigma^+}(r_1 \wedge r_2)| &\leq |\partial_{\Sigma^+} r_1| \times |\partial_{\Sigma^+} r_2| \\
|\partial_{\Sigma^+} \mathsf{fm}(r_1)| &\leq 2^{|\partial_{\Sigma^+} r_1|}
\end{aligned}$$

The finiteness of $\partial_{\Sigma^+} r$ and therefore of $\mathrm{Terms}(r)$ follows from the above. □

From the inequalities in the above sketch of proof, one can observe that the added operators do not have the same cost. In particular $\wedge$ incurs a quadratic cost, while fm leads to an exponential blow up.

```
input  : A SERE φ
output: An automaton A such that L(A) = L(φ)
Q, δ, F ← {φ}, ∅, ∅;
todo.push(φ);
while todo ≠ ∅ do
    f ← todo.pop();
    foreach (p, s) ∈ LF(f) do
        if s ∉ Q then
            Q ← Q ∪ {s};
            todo.push (s);
            if λ(s) = ε then
                F ← F ∪ {s};
        δ ← δ ∪ {f -p-> s};
return ⟨Q, δ, φ, F⟩;
```

Algorithm 1: Translation of a SERE ϕ to a NFA. Remember that rules (A)–(F) are always applied.

3.3 Automaton Construction

The traditional way to construct a finite automaton from such a linear form is to associate its states to regular expressions. For a state $r \in \mathrm{Terms}(r)$, we interpret the pairs (p_i, s_i) in $\mathsf{LF}(r)$ as a transition $r \xrightarrow{p_i} s_i$. Algorithm 1 shows a straightforward implementation of that construction. Final states are those that correspond to expressions that accept the empty word. At the end of this algorithm we naturally have $Q = \mathrm{Terms}(\phi)$ by construction, which ensures termination. The fact that $\mathscr{L}(\langle Q, \delta, \phi, F\rangle) = \mathscr{L}(\phi)$ follows from Theorem 1.

Although $\mathrm{Terms}(\phi)$ is defined as the smallest subset of SERE recursively produced by LF, the resulting automaton is not necessarily minimal in terms of number of states. Because we only use syntactic equivalence, the construction can produce two states labeled by SEREs $r_1 \neq r_2$ such that $\mathscr{L}(r_1) = \mathscr{L}(r_2)$.

This algorithm can be altered in several ways in attempt to simplify the resulting automata. In Section 4 we present ways to *simplify* the linear forms $\boxed{\mathsf{LF}(f)}$ before they get used in Algorithm 1. Then in Section 5 we propose larger modifications of Algorithm 1 meant to fuse states with identical linear forms.

4 Linear Form Simplifications

When constructing an automaton from a linear form, it is possible to alter the shape of the automaton constructed by transforming the linear forms it uses into other, equivalent linear forms. In this section we present a few transformations that aim at simplifying linear forms. By "*simplifying*" we mean to reduce the number of pairs in the hope that this results in a smaller automaton. Simplifying a finite automaton can of course be done after its construction using more traditional algorithms like bisimulation-based reductions [21], however it is always good to look for cheap opportunities to keep the intermediate automaton small.

Definition 9 (Unique Suffix and Unique Prefix simplifications). *Let L be a linear form, let* $\mathsf{MergePre}(L, s) = \bigvee_{(p,s)\in L} p$ *be the (Boolean) disjunction of all prefixes sharing a given suffix s, and let* $\mathsf{MergeSuf}(L, p) = \bigvee_{(p,s)\in L} s$ *be the (rational) disjunction of all suffixes sharing a given prefix p.*

We define US *(unique suffixes) and* UP *(unique prefixes) as follows:*

$$\mathsf{US}(L) = \{(\mathsf{MergePre}(L, s), s) \mid (p, s) \in L\}$$
$$\mathsf{UP}(L) = \{(p, \mathsf{MergeSuf}(L, p)) \mid (p, s) \in L\}$$

Replacing $\mathsf{LF}(f)$ by $\mathsf{US}(\mathsf{LF}(f))$ in Algorithm 1 is equivalent to merging the edges of the automaton that have the same source and same destination. For instance $\mathsf{US}(\{(a, r), (b, r)\}) = \{(a \vee b, r)\}$.

Replacing $\mathsf{LF}(f)$ by $\mathsf{UP}(\mathsf{LF}(f))$ in Algorithm 1 is merging outgoing edges that share the same label. In Antimorov's setup [2], where prefixes of linear forms are letters, using UP would create a deterministic automaton. However, because in our setup prefixes are Boolean formulas, this is not the case: UP can remove *some* non-determinism, but the result will not necessarily be deterministic. For instance the non-deterministic linear form $\{(a, q_1), (a \wedge b, q_2)\}$ is unchanged by UP. If we wish to construct a deterministic automaton, we can use $\mathsf{det}(\mathsf{LF}(f))$ (see Proposition 1). Our intent with UP is therefore not to produce a deterministic automaton, but to help reduce the size of a non-deterministic result.

We should point out that the equivalent of Theorem 2 still holds when $\mathsf{UP}(\mathsf{LF}(f))$ is used because the terms created by this new variant are disjunctions of terms created by the original construction.

Unfortunately, it is also possible that using UP will introduce new additional states in the automaton. For instance $\mathsf{UP}(\{(a, q_1), (a, q_2), (b, q_1), (\neg b, q_2)\}) = \{(a, q_1 \vee q_2), (b, q_1), (\neg b, q_2)\}$ would be introducing the state $q_1 \vee q_2$ that was not present initially.

Because US only merges edges, it sounds natural to use it together with UP. However, while it is possible to find cases where replacing $\mathsf{LF}(f)$ by $\mathsf{US}(\mathsf{UP}(\mathsf{LF}(f)))$ is better than replacing $\mathsf{LF}(f)$ by $\mathsf{UP}(\mathsf{US}(\mathsf{LF}(f)))$, the opposite also exists. See Appendix E

5 Signature and Transition-Based Variants

We now discuss variants of Algorithm 1, orthogonal to previous simplifications.

Consider the automaton of Figure 2, where the first two states are labeled by formulas that have the same linear form, and so are the last two states. If two expressions r_1 and r_2 have the same linear form $\mathsf{LF}(r_1) = \mathsf{LF}(r_2)$, it implies that $\mathscr{L}(r_1) \setminus \varepsilon = \mathscr{L}(r_2) \setminus \varepsilon$. Therefore, states that correspond to formulas with the same linear form (i.e., states that have identical sets of outgoing transitions) can be merged if they are both accepting, or both rejecting. Thus, the first two states of Figure 2 could be merged.

We can obtain such a merge automatically if we modify our translation as in Algorithm 2 to label each state by a pair $(\mathsf{LF}(\varphi), \lambda(\varphi))$ that we call the signature of φ. This gives the automaton of Figure 3.

The **return** line of Algorithms 2 and 3 had a typo in the CIAA'24 proceedings.

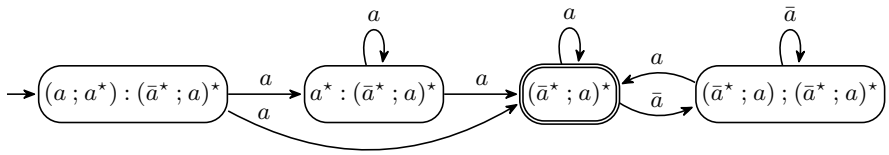

Fig. 2: Automaton for $\varphi = (a\,;a^\star):(\bar{a}^\star\,;a)^\star$. We have $\mathsf{LF}(\varphi) = \mathsf{LF}(a^\star:(\bar{a}^\star\,;a)^\star) = \{(a, a^\star:(\bar{a}^\star\,;a)^\star), (a, (\bar{a}^\star\,;a)^\star)\}$, and $\mathsf{LF}((\bar{a}^\star\,;a)^\star) = \mathsf{LF}((\bar{a}^\star\,;a)\,;(\bar{a}^\star\,;a)^\star) = \{(a, (\bar{a}^\star\,;a)^\star), (\bar{a}, (\bar{a}^\star\,;a)\,;(\bar{a}^\star\,;a)^\star)\}$.

input : A SERE ϕ
output: An NFA $\mathcal{A}$ such that $\mathscr{L}(\mathcal{A}) = \mathscr{L}(\phi)$

$L_\iota, b_\iota \leftarrow \mathsf{LF}(\phi), [\lambda(\phi) = \varepsilon]$;
$Q, \delta, F \leftarrow \{(L_\iota, b_\iota)\}, \varnothing, \varnothing$;
todo.push$((L_\iota, b_\iota))$;
while todo $\neq \varnothing$ **do**
  $(L, b) \leftarrow$ todo.pop();
  if b **then**
    $F \leftarrow F \cup \{(L, b)\}$;
  foreach $(p, s) \in L$ **do**
    $L', b' \leftarrow \mathsf{LF}(s), [\lambda(s) = \varepsilon]$;
    if $(L', b') \notin Q$ **then**
      $Q \leftarrow Q \cup \{(L', b')\}$;
      todo.push $((L', b'))$;
    $\delta \leftarrow \delta \cup \{(L, b) \xrightarrow{p} (L', b')\}$;
return $\langle Q, \delta, (L_\iota, b_\iota), F\rangle$;

Algorithm 2: Translation that identifies states with identical linear form and identical ε acceptance.

input : A SERE ϕ
output: A TFA $\mathcal{A}$ such that $\mathscr{L}(\mathcal{A}) = \mathscr{L}(\phi)$

$L_\iota, b_\iota \leftarrow \mathsf{LF}(\phi), [\lambda(\phi) = \varepsilon]$;
$Q, \delta \leftarrow \{L_\iota\}, \varnothing$;
todo.push(L_ι);
while todo $\neq \varnothing$ **do**
  $L \leftarrow$ todo.pop();
  foreach $(p, s) \in L$ **do**
    $L', b' \leftarrow \mathsf{LF}(s), [\lambda(s) = \varepsilon]$;
    if $L' \notin Q$ **then**
      $Q \leftarrow Q \cup \{L'\}$;
      todo.push (L');
    $\delta \leftarrow \delta \cup \{L \xrightarrow{p,b'} L'\}$;
return $\langle Q, \delta, L_\iota, b_\iota\rangle$;

Algorithm 3: Translation to transition-based automata, identifying states with identical linear form regardless of ε acceptance.

Currently, the last two states of Figure 3 may not be merged because one is accepting while the other is not. We could however merge them by changing our automaton formalism such that the notion of acceptance is carried by the transitions instead of the states. Although finite automata are seldom used with transition-based acceptance [24], ω-automata (i.e., automata over infinite words) with transition-based acceptance have been used for a long time as they often lead to simpler algorithms [22, 18, 19, 17, 23, to cite a few]. Let us define a transition-based finite automaton:

Definition 10 (TFA). *A* transition-based finite automaton *is a tuple* $\mathcal{A} = \langle Q, \delta, \iota, \beta\rangle$ *where* Q *is a finite set of states,* $\delta \subseteq Q \times \mathbb{B}(AP) \times \mathbb{B} \times Q$ *is the transition relation,* $\iota \in Q$ *is the initial state, and* $\beta \in \mathbb{B}$ *is a Boolean indicating whether* ε *should be accepted.*

We write $s \xrightarrow{f,b} d$ *when* $(s, f, b, d) \in \delta$.

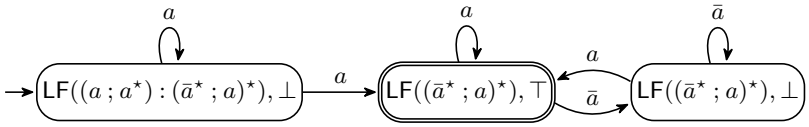

Fig. 3: Automaton obtained by merging states labeled by formulas that have the same linear form and the same acceptance of ε.

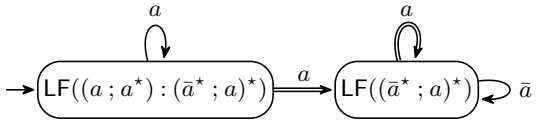

Fig. 4: Transition-based automaton obtained by merging states labeled by formulas that have the same linear form, regardless of the acceptance of ε, since the latter is decided on transitions.

A sequence of valuations $\sigma \in \Sigma^n$ of size n is accepted *by $\mathcal{A}$ if either $n = 0$ and $\beta = \top$, or $n > 0$ and there exists a sequence of transitions $\rho = s_0 \xrightarrow{f_0,b_0} s_1 \xrightarrow{f_1,b_1} \cdots \xrightarrow{f_{n-1},b_{n-1}} s_n$ such that $s_0 = \iota$, $b_{n-1} = \top$, and for all i, $\sigma(i) \models f_i$.*

The language *of $\mathcal{A}$, denoted $\mathscr{L}(\mathcal{A})$ is the set of words accepted by $\mathcal{A}$.*

In other words, transitions of a TFA carry an extra Boolean that is used to mark the transition as accepting, and a word is accepted if it is recognized by a run whose last transition is accepting. Graphically, we represent accepting transitions using arrows with double lines. The acceptance of the empty word is indicated by a special Boolean β in the definition, and can be represented graphically by using double lines on the arrow indicating the initial state.

TFA enjoy similar properties as traditional finite automata: they are as expressive as regular expressions, are closed under Boolean operations, etc. [24] However they can be slightly smaller, as we shall see in our evaluation.

Using the above definition, Algorithm 3 generates the automaton of Figure 4.

In our case, we have additional motivation for using TFAs. The reason we are working on translating SERE to automata is that SERE are part of the PSL and SVA standards. However, the PSL/SVA standards assume a SERE will always match a non-empty word. Therefore, the Boolean β that we added to the definition of a TFA to allow it to recognize ε can simply be ignored. Furthermore, as we translate a PSL formula into an ω-automaton, we build upon translation algorithms that naturally produce transition-based ω-automata [11]. In this context, it seems more natural to have SERE converted into TFA. The fact that TFA are more succinct comes as a bonus.

6 Experimental Evaluation

Our algorithms have been implemented in a development version of Spot [12]. A reproducibility package, archived at https://doi.org/10.5281/zenodo.10799850, See Appendix F.

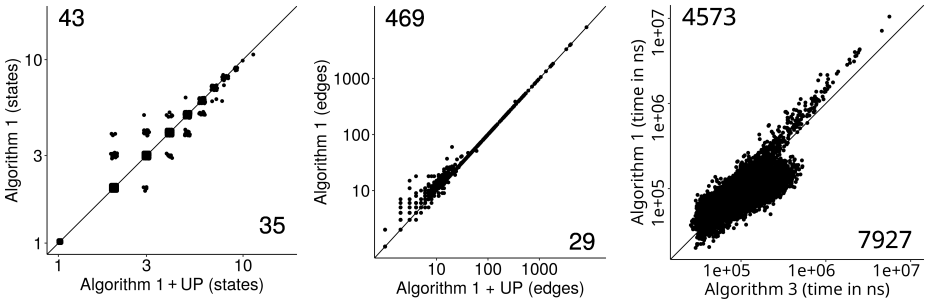

Fig. 5: Effect of UP on Algorithm 1.

Fig. 6: Time of Alg. 1–3.

contains our implementation, a Jupyter notebook to use it interactively, and scripts to reproduce our experiments.

We are not aware of any existing benchmark of SEREs. Therefore, to evaluate our work, we randomly generated a set of SEREs using Spot's `randltl` tool, with equal probability of occurrence for all SERE operators. We grouped the random SEREs into groups of expressions with equal size (number of nodes in their syntax tree) and equal number of unique atomic propositions, capping each group to 50 expressions. The resulting set has 12500 unique SEREs with sizes ranging from 1 to 35, and between 1 and 15 atomic propositions.

See Appendix G

Benchmarks were run on an AMD Ryzen 5 3600 CPU, with 16GB of RAM, and with core frequency capped at 2.2GHz to minimize the impact of throttling on timing measurements. For each SERE we evaluated variants of the translation by measuring the number of states of the produced automata, and the time needed to produce them. (We also measured the number of edges, but do not report it here.)

Scatter plots that show number of states use a jitter of ± 0.4 over their position to distinguish points. The numbers in the top left and bottom right corners of the plots count how many points are strictly above or below the diagonal.

We start by evaluating the impact of the simplification strategies of Section 4. Figure 5 presents the impact of UP on the number of states and edges of automata produced by Algorithm 1. As mentioned in Section 4, UP has mitigated results: it improves the number of states of the automaton almost as often as it worsens it. However the number of transitions is reduced in general.

Figure 7 shows that Algorithms 2–3 provide a more important reduction of automata sizes compared to Algorithm 1. Impact on translation time, as seen on Figure 6, is not significant (average speedup is -10%). Small automata have an overhead because of the labeling of states by linear forms rather than formulas, but the savings in size also yields savings in time for larger automata.

Figures 8 and 11 show that applying UP in Algorithm 3 has more effect than it had on Algorithm 1 (compare with Figure 5 where the impact of UP on the number of states was marginal). Figure 9 shows using $\mathsf{UP} \circ \mathsf{US}$ on Algorithm 3 does not yield significant changes in number of states, compared to only using

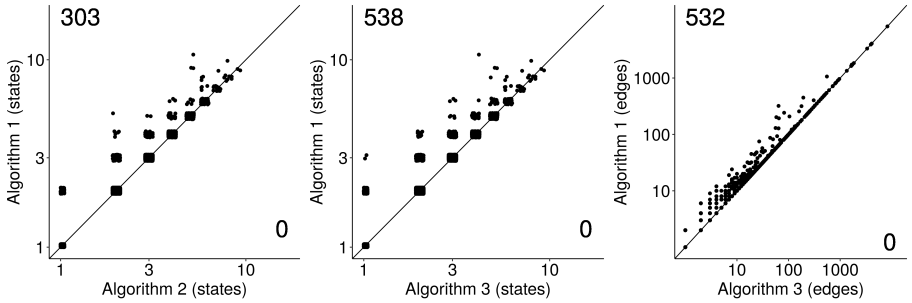

Fig. 7: Comparisons of Algorithms 1, 2, and 3.

UP. However Figure 12 shows an impressive reduction in the number of edges. Using US ∘ UP instead of UP would not change the number of states, as this is only merging edges, so we do not compare it to UP. Figures 10 and 13 shows that in practice, the impact of the order of application between UP and US discussed in Section 4 is rather limited, producing automata with a different number of states in only 21 cases out of our 12500 formulas.

7 Conclusion

We adapted Antimirov's non-deterministic automata construction based on linear forms, to the semi-extended regular expressions used by the PSL and SVA standards. As these SERE are defined on alphabet of the form 2^{AP}, we introduced some rewritings (UP, US) of these linear forms and evaluated their impact on a large benchmark. We also introduced alternative translation algorithms that use the linear form to simplify the automaton during its construction, or that build a transition-based automaton.

Our evaluation reveals that using transition-based automata, labeling them with linear forms, and simplifying those linear forms with UP are cheap and effective ways of keeping the output small. A compact output matters in applications where the automaton is constructed on-the-fly or only partially, and therefore cannot benefit from subsequent simplifications. (Satisfiability, which cannot be decided syntactically because of the intersection operator, is one such problem.)

Finally, we constructed a SERE benchmark dataset, which we hope can be reused in future work to compare different SERE, PSL or SVA translators.

References

1. *1800-2017 - IEEE Standard for SystemVerilog–Unified Hardware Design, Specification, and Verification Language*. IEEE, Feb. 2018. https://doi.org/10.1109/IEEESTD.2018.8299595.

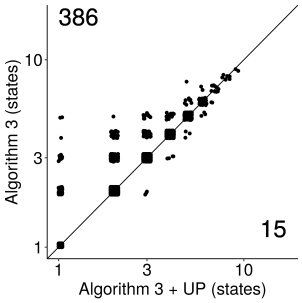

Fig. 8: Impact of UP on Algorithm 3

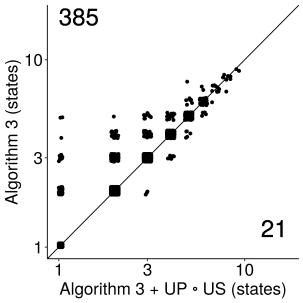

Fig. 9: Impact of UP ∘ US on Algorithm 3

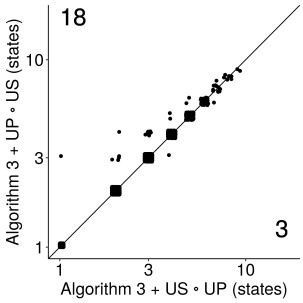

Fig. 10: Comparison of UP ∘ US and US ∘ UP

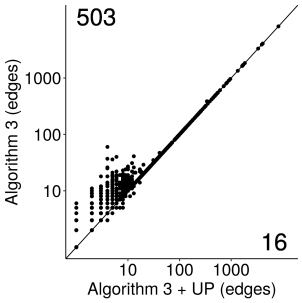

Fig. 11: Impact of UP on Algorithm 3

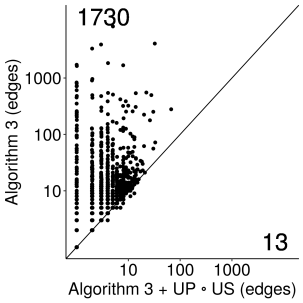

Fig. 12: Impact of UP ∘ US on Algorithm 3

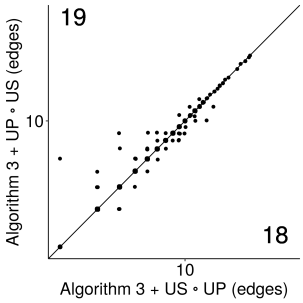

Fig. 13: Comparison of UP ∘ US and US ∘ UP

2. V. Antimirov. Partial derivatives of regular expressions and finite automaton constructions. *Theoretical Computer Science*, 155(2):291–319, Mar. 1996. https://doi.org/10.1016/0304-3975(95)00182-4.
3. C. Baier and J.-P. Katoen. *Principles of Model Checking*. MIT Press, 2008.
4. E. Bartocci, Y. Falcone, A. Francalanza, and G. Reger. *Introduction to Runtime Verification*, pages 1–33. Springer International Publishing, Cham, 2018. https://doi.org/10.1007/978-3-319-75632-5_1.
5. R. E. Bryant. Symbolic boolean manipulation with ordered binary-decision diagrams. *ACM Computing Surveys*, 24(3):293–318, Sept. 1992.
6. J. A. Brzozowski. Derivatives of regular expressions. *Journal of the ACM*, 11(4): 481–494, Oct. 1964. https://doi.org/10.1145/321239.321249.
7. P. Caron, J.-M. Champarnaud, and L. Mignot. Partial derivatives of an extended regular expression. In *Proceedings of the 5th International Conference on Language and Automata Theory and Applications (LATA'11)*, pages 179–191. Springer Berlin Heidelberg, 2011. https://doi.org/10.1007/978-3-642-21254-3_13.
8. J.-M. Couvreur. On-the-fly verification of temporal logic. In J. M. Wing, J. Woodcock, and J. Davies, editors, *Proceedings of the World Congress on Formal Methods in the Development of Computing Systems (FM'99)*, volume 1708 of *Lecture Notes in Computer Science*, pages 253–271, Toulouse, France, Sept. 1999. Springer-Verlag. ISBN 3-540-66587-0.

9. L. D'Antoni and M. Veanes. Minimization of symbolic automata. In S. Jagannathan and P. Sewell, editors, *Proceedings of the 41st Annual Symposium on Principles of Programming Languages (POPL'14)*, pages 541–554. ACM, Jan. 2014. https://doi.org/10.1145/2535838.2535849.
10. C. Dax, F. Klaedtke, and S. Leue. Specification languages for stutter-invariant regular properties. In Z. Liu and A. P. Ravn, editors, *Automated Technology for Verification and Analysis*, volume 5799 of *Lecture Notes in Computer Science*, pages 244–254. Springer, 2009. ISBN 978-3-642-04760-2. https://doi.org/10.1007/978-3-642-04761-9_19.
11. A. Duret-Lutz. LTL translation improvements in Spot 1.0. *International Journal on Critical Computer-Based Systems*, 5(1/2):31–54, Mar. 2014. https://doi.org/10.1504/IJCCBS.2014.059594.
12. A. Duret-Lutz, E. Renault, M. Colange, F. Renkin, A. G. Aisse, P. Schlehuber-Caissier, T. Medioni, A. Martin, J. Dubois, C. Gillard, and H. Lauko. From Spot 2.0 to Spot 2.10: What's new? In *Proceedings of the 34th International Conference on Computer Aided Verification (CAV'22)*, volume 13372 of *Lecture Notes in Computer Science*. Springer, Aug. 2022. https://doi.org/10.1007/978-3-031-13188-2_9.
13. C. Eisner and D. Fisman. *A Practical Introduction to PSL*. Series on Integrated Circuits and Systems. Springer, 2006. https://doi.org/10.1007/978-0-387-36123-9.
14. B. Finkbeiner. Synthesis of reactive systems. In J. Esparza, O. Grumberg, and S. Sickert, editors, *Dependable Software Systems Engineering*, volume 45 of *NATO Science for Peace and Security Series - D: Information and Communication Security*, pages 72–98. IOS Press Ebooks, 2016. https://doi.org/10.3233/978-1-61499-627-9-72.
15. M. J. Fischer and R. E. Ladner. Propositional dynamic logic of regular programs. *Journal of Computer and System Sciences*, 18(2):194–211, 1979. ISSN 0022-0000. https://doi.org/10.1016/0022-0000(79)90046-1.
16. G. D. Giacomo and M. Y. Vardi. Linear temporal logic and linear dynamic logic on finite traces. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI'13)*, page 854–860. AAAI Press, 2013. ISBN 9781577356332. https://doi.org/10.5555/2540128.2540252.
17. D. Giannakopoulou and F. Lerda. From states to transitions: Improving translation of LTL formulæ to Büchi automata. In D. Peled and M. Vardi, editors, *Proceedings of the 22nd IFIP WG 6.1 International Conference on Formal Techniques for Networked and Distributed Systems (FORTE'02)*, volume 2529 of *Lecture Notes in Computer Science*, Houston, Texas, Nov. 2002. Springer-Verlag. https://doi.org/10.5555/646220.682186.
18. R. P. Kurshan. Complementing deterministic Büchi automata in polynomial time. *J. Comput. Syst. Sci.*, 35(1):59–71, Aug. 1987. https://doi.org/10.1016/0022-0000(87)90036-5.
19. B. Le Saëc and I. Litovsky. On the minimization problem for ω-automata. In I. Prívara, B. Rovan, and P. Ruzicka, editors, *Proceedings of the 19th International Symposium on Mathematical Foundations of Computer Science (MFCS'94)*, volume 841 of *Lecture Notes in Computer Science*, pages 504–514, Kosice, Slovakia, Aug. 1994. Springer-Verlag. https://doi.org/10.5555/645723.666714.
20. J. Li, G. Pu, L. Zhang, Z. Wang, J. He, and K. G. Larsen. *On the Relationship between LTL Normal Forms and Büchi Automata*, pages 256–270. Springer, 2013. https://doi.org/10.1007/978-3-642-39698-4_16.
21. S. Lombardy and J. Sakarovitch. Two routes to automata minimization and the ways to reach it efficiently. In *Proceedings of the 23rd International Conference on*

Implementation and Application of Automata (CIAA'18), volume 10977 of *Lecture Notes in Computer Science*, pages 248–260. Springer, 2018. https://doi.org/10.1007/978-3-319-94812-6_21.

22. M. Michel. Algèbre de machines et logique temporelle. In M. Fontet and K. Mehlhorn, editors, *Proceedings of the Symposium on Theoretical Aspects of Computer Science (STACS'84)*, volume 166 of *Lecture Notes in Computer Science*, pages 287–298, Paris, Apr. 1984. https://doi.org/10.1007/3-540-12920-0_26.
23. T. Varghese. *Parity and Generalized Büchi Automata — determinisation and complementation*. PhD thesis, University of Liverpool, Nov. 2014.
24. S. Xiao, J. Li, S. Zhu, Y. Shi, G. Pu, and M. Vardi. On-the-fly synthesis for LTL over finite traces. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI'21)*, volume 35, pages 6530–6537, May 2021. https://doi.org/10.1609/aaai.v35i7.16809. Technical Tracks 7.

input : A linear form $\{(p_1, s_1), (p_2, s_2), \ldots, (p_n, s_n)\}$
output: An equivalent deterministic linear form

let $P \subseteq AP$ be the set of atomic propositions appearing in $p_1, p_2, \ldots, p_n$;
$L \leftarrow \varnothing$;
foreach $\ell \in 2^P$ **do**
  $S \leftarrow \bigvee_{\ell \models p_i} s_i$;
  if $S \neq \bot$ **then**
    $L \leftarrow L \cup \{(\ell, S)\}$;
return L;

Algorithm 4: Determinization of a linear form.

The following appendices are not part of the CIAA'24 proceedings because of size restrictions.

A Determinization of a Linear Form

We considered several implementations for the determinization $\mathsf{det}(L)$ of a linear form. The BDD representation of linear forms (see Section F) can be used to determinize the successor function [11], however since our linear forms are small in practice, we simply decided to work on the array representation using Algorithm 4.

B Proof of proposition 2

Proof. The proof follows from the fact that rewritings (A)–(F) are language-preserving. As most of these rules are either well known or obvious, we only prove those involving fm. For I2 we have:

$$\begin{aligned}
\sigma \models \mathsf{fm}(\mathsf{fm}(r)) &\iff (\sigma \models \mathsf{fm}(r)) \wedge (\forall i < |\sigma|,\, \sigma^{\cdot\cdot i} \not\models \mathsf{fm}(r)) \\
&\iff (\sigma \models r) \wedge (\forall i < |\sigma|,\, \sigma^{\cdot\cdot i} \not\models r) \wedge (\forall i < |\sigma|,\, \sigma^{\cdot\cdot i} \not\models \mathsf{fm}(r)) \\
&\iff (\sigma \models r) \wedge (\forall i < |\sigma|,\, \neg(\sigma^{\cdot\cdot i} \models r \vee \underbrace{\sigma^{\cdot\cdot i} \models \mathsf{fm}(r)}_{\text{implies } \sigma^{\cdot\cdot i} \models r \text{ by definition}})) \\
&\iff (\sigma \models r) \wedge (\forall i < |\sigma|,\, \sigma^{\cdot\cdot i} \not\models r) \\
&\iff \sigma \models \mathsf{fm}(r)
\end{aligned}$$

Therefore $\mathscr{L}(\mathsf{fm}(\mathsf{fm}(r))) = \mathscr{L}(\mathsf{fm}(r))$. Similarly, for rule (F), we have

$$\begin{aligned}
(\sigma \models \mathsf{fm}(r)) \wedge (\varepsilon \models r) &\iff (\sigma \models r) \wedge \underbrace{(\forall i < |\sigma|,\, \sigma^{\cdot\cdot i} \not\models r) \wedge (\varepsilon \models r)}_{\text{implies } |\sigma|=0 \text{ otherwise } \sigma^{\cdot\cdot 0}=\varepsilon \not\models r \text{ conflicts with } \varepsilon \models r} \\
&\iff |\sigma| = 0 \iff \sigma \models \varepsilon
\end{aligned}$$

Therefore when $\varepsilon \models r$, $\mathscr{L}(\mathsf{fm}(r)) = \mathscr{L}(\varepsilon)$. □

C Proof of Theorem 1: LF builds linear forms

Proof. Let $\mathscr{L}(\mathsf{LF}(r))$ designate $\bigcup_i \mathscr{L}(p_i\,;s_i)$. Using Definitions 2 and 8, we can show by induction on the structure of r that $\mathscr{L}(\mathsf{LF}(r)) = \mathscr{L}(r) \setminus \{\varepsilon\}$. For the regular operators, it was already established by Antomirov [2, Prop. 2.5].

We show here the case of $r_1 : r_2$. For $m \in \{1,2\}$ assume that $\mathsf{LF}(r_m) = \{(p_1^m, s_1^m), (p_2^m, s_2^m), \ldots\}$ is a linear form for r_m, i.e., that $\bigcup_i \mathscr{L}(p_i^m\,;s_i^m) = \mathscr{L}(r_m) \setminus \{\varepsilon\}$. We would like to prove that $\mathscr{L}(\mathsf{LF}(r_1 : r_2)) = \mathscr{L}(r_1 : r_2) \setminus \{\varepsilon\}$.

Since $\mathscr{L}(r_1 : r_2)$ may not contain ε by definition, we just have to prove that for any sequence $\sigma \in \Sigma^*$ we have $\sigma \models \mathscr{L}(\mathsf{LF}(r_1 : r_2)) \iff \sigma \models \mathscr{L}(r_1 : r_2)$.

($\Longleftarrow$) Consider σ in $\mathscr{L}(r_1 : r_2)$. By Definition 2 there exists $k \geq 0$ such that $\sigma^{..k+1} \models r_1$ and $\sigma^{k..} \models r_2$.

If $k = 0$, $\sigma^{..1} \models r_1$ implies that there exists a pair (p_i^1, s_i^1) in $\mathsf{LF}(r_1)$ such that $\sigma(0) \models p_i^1$ and $\lambda(s_i^1) = \emptyset$. Furthermore since $\sigma \models r_2$, there exists a pair (p_j^2, s_j^2) in $\mathsf{LF}(r_2)$ such that $\sigma(0) \models p_j^2$. We can see in Definition 8 that the pair $(p_i^1 \wedge p_j^2, s_j^2)$ exists in $\mathsf{LF}(r_1 : r_2)$, therefore $\sigma \in \mathscr{L}(\mathsf{LF}(r_1 : r_2))$.

If $k > 0$, $\sigma^{..k+1} \models r_1$ implies that there exists a pair (p_i^1, s_i^1) in $\mathsf{LF}(r_1)$ such that $\sigma(0) \models p_i^1$ and $\sigma^{1..k+1} \models s_i^1$. Since we know that $\sigma^{k..} \models r_2$, it follows that $\sigma^{1..} \models s_i^1 : r_2$. By definition $(p_i^1, s_i^1 : r_2)$ belongs to $\mathsf{LF}(p_i) : r_2$ which itself belongs to $\mathsf{LF}(r_1 : r_2)$. Therefore $\sigma \in \mathscr{L}(\mathsf{LF}(r_1 : r_2))$.

($\Longrightarrow$) Consider σ in $\mathscr{L}(\mathsf{LF}(r_1 : r_2))$. Looking at $\mathsf{LF}(r_1 : r_2)$ in Definition 8, either σ is matched by the left part of the union, or by the right part.

If it is matched by the left part, there exists a pair $(p_i^1, s_i^1 : r_2)$ such that $\sigma(0) \models p_i^1$ and $\sigma^{1..} \models s_i^1 : r_2$. The latter implies that there exists $k \geq 1$ such that $\sigma^{1..k+1} \models s_i^1$ and $\sigma^{k..} \models r_2$. Therefore we have $\sigma^{0..k+1} \models r_1$ and $\sigma^{k..} \models r_2$, which implies $\sigma \in \mathscr{L}(r_1 : r_2)$.

If it is matched by the right part, there exists a pair $(p_i^1 \wedge p_j^2, s_j^2) \in \mathsf{LF}(r_1 : r_2)$ such that $\sigma(0) \models p_i^1$, $\lambda(s_i^1) = \emptyset$, $\sigma(0) \models p_j^2$, and $\sigma^{1..} \models s_j^2$. The first two constraints imply that $\sigma^{0..1} \models r_1$, and the latter two that $\sigma^{1..} \models r_2$. It follows that $\sigma \models r_1 : r_2$.

Other operators can be proven similarly.

D Proof of Theorem 2: Terms(r) is finite

Our proof is inspired by a similar theorem by Antimirov [2, Theorem 3.4], however the results differ because of the new operators we support. Specifically, Antimirov did not support operators $\wedge$, :, and fm. (Of these three, : is the least problematic.)

Like Antimirov, we start by introducing a notion of *partial derivative* [2, Definition 2.8] which we adapt to our SEREs, where the alphabet is $\Sigma = 2^{AP}$.

Definition 11 (partial derivative). *Given an expression $r \in \mathsf{SERE}$ and a valuation $x \in \Sigma = 2^{AP}$, we denote $\partial_x r$ the partial derivative of r with respect to x defined by:*

$$\partial_x r = \{s \mid (p, s) \in \mathsf{LF}(r),\, x \models p\}$$

We extend partial derivatives to the case $\partial_x R$ where $R \subseteq \mathsf{SERE}$ is a set of expression, in a natural way: $\partial_x R = \cup_{r\in R}\partial_x r$. We also extend the notation to support derivation by a nonempty word $w \in \Sigma^+$ with $\partial_w r = \partial_{w^{1..}}\partial_{w(0)}r$.

Furthermore, we write $\partial_{\Sigma^+} r = \bigcup_{w\in\Sigma^+} \partial_w r$ for the set of all partial derivatives one can obtain using nonempty words of any length.

Using the above notation, we have $\mathrm{Terms}(r) = \partial_{\Sigma^+} r \cup \{r\}$.

Let us extend our Definition 3 of the constant term λ to cover a set $R \subseteq \mathsf{SERE}$ of expressions with $\lambda(R) = \bigvee_{r\in R} \lambda(r)$.

For $x \in \Sigma$, the following equalities follow from the Definitions 8 and 11:

$$\partial_x \bot = \varnothing \tag{1}$$
$$\partial_x \varepsilon = \varnothing \tag{2}$$
$$\partial_x b = \begin{cases} \{\varepsilon\} & \text{if } x \models b \\ \varnothing & \text{else} \end{cases} \tag{3}$$
$$\partial_x(r_1 \vee r_2) = \partial_x r_1 \cup \partial_x r_2 \tag{4}$$
$$\partial_x r^\star = (\partial_x r_1)\,;\, r^\star \tag{5}$$
$$\partial_x(r_1\,;\,r_2) = ((\partial_x r_1)\,;\,r_2) \cup (\lambda(r_1)\,;\,\partial_x r_2) \tag{6}$$
$$\partial_x(r_1 : r_2) = ((\partial_x r_1) : r_2) \cup (\lambda(\partial_x r_1)\,;\,\partial_x r_2) \tag{7}$$
$$\partial_x(r_1 \wedge r_2) = \{p_1 \wedge p_2 \mid p_1 \in \partial_x r_1,\, p_2 \in \partial_x r_2\} \tag{8}$$
$$\partial_x \mathsf{fm}(r) = \left\{ \bigvee_{s\in\partial_x r} s \right\} \tag{9}$$

Definition 12 (Suffix set). *Given a nonempty word $w \in \Sigma^+$, let $Sfx(w)$ denote the set $\{w^{i\cdot\cdot} \mid 0 \le i < |w|\}$ of nonempty suffixes of w.*

Lemma 1. *For $w \in \Sigma^+$ the following (in)equalities follow from (1)–(9):*

$$\partial_w(r_1 \vee r_2) = (\partial_w r_1) \cup (\partial_w r_2) \tag{10}$$
$$\partial_w r^\star \subseteq \bigcup_{v\in Sfx(w)} (\partial_v r)\,;\,r^\star \tag{11}$$
$$\partial_w(r_1\,;\,r_2) \subseteq ((\partial_w r_1)\,;\,r_2) \cup \bigcup_{v\in Sfx(w)} \partial_v r_2 \tag{12}$$
$$\partial_w(r_1 : r_2) \subseteq ((\partial_w r_1) : r_2) \cup \bigcup_{v\in Sfx(w)} \partial_v r_2 \tag{13}$$
$$\partial_w(r_1 \wedge r_2) = \{p_1 \wedge p_2 \mid p_1 \in \partial_w r_1,\, p_2 \in \partial_w r_2\} \tag{14}$$
$$\partial_w \mathsf{fm}(r_1) = \left\{ \bigvee_{s\in\partial_w r} s \right\} \tag{15}$$

Proof. Equations (10)–(12) are already known results [2, Lemma 3.3].

Let us prove (12) again, using our notations. Deriving $r_1 ; r_2$ using (6) and words of increasing lengths, we have the following:

$$\begin{aligned}
\partial_a(r_1 ; r_2) &= ((\partial_a r_1) ; r_2) \cup (\lambda(r_1) ; (\partial_a r_2)) && \text{for } a \in \Sigma \\
\partial_{ab}(r_1 ; r_2) &= ((\partial_{ab} r_1) ; r_2) \cup (\lambda(\partial_a r_1) ; (\partial_b r_2)) \cup (\lambda(r_1) ; (\partial_{ab} r_2)) && \text{for } ab \in \Sigma^2 \\
\partial_{abc}(r_1 ; r_2) &= ((\partial_{abc} r_1) ; r_2) \cup (\lambda(\partial_{ab} r_1) ; (\partial_c r_2)) && \text{for } abc \in \Sigma^3 \\
&\quad \cup (\lambda(\partial_a r_1) ; (\partial_{bc} r_2)) \\
&\quad \cup (\lambda(r_1) ; (\partial_{abc} r_2))
\end{aligned}$$

Since those $\lambda(\partial_a r_1)$, $\lambda(\partial_{ab} r_1)$, $\lambda(\partial_{abc} r_1)$... are used to conditionally enable the subsequent terms, it should be clear that for any word $w \in \Sigma^+$, the set $\partial_w(r_1;r_2)$ contains $((\partial_w r_1) ; r_2)$ and a subset of $\bigcup_{v \in Sfx(w)} \partial_v r_2$, justifying equation (12).

Equation (13) is proven similarly, the difference is only in the constant terms:

$$\begin{aligned}
\partial_a(r_1 : r_2) &= ((\partial_a r_1) : r_2) \cup (\lambda(\partial_a r_1) ; (\partial_a r_2)) && \text{for } a \in \Sigma \\
\partial_{ab}(r_1 : r_2) &= ((\partial_{ab} r_1) : r_2) \cup (\lambda(\partial_{ab} r_1) ; (\partial_b r_2)) && \text{for } ab \in \Sigma^2 \\
&\quad \cup (\lambda(\partial_a r_1) ; (\partial_{ab} r_2)) \\
\partial_{abc}(r_1 : r_2) &= ((\partial_{abc} r_1) : r_2) \cup (\lambda(\partial_{abc} r_1) ; (\partial_c r_2)) && \text{for } abc \in \Sigma^3 \\
&\quad \cup (\lambda(\partial_{ab} r_1) ; (\partial_{bc} r_2)) \\
&\quad \cup (\lambda(\partial_a r_1) ; (\partial_{abc} r_2))
\end{aligned}$$

Finally, equations (14)–(15) follow immediately from (8)–(9). □

Lemma 2. *For two SEREs r_1 and r_2 we have:*

$$\begin{aligned}
|\partial_{\Sigma^+}(r_1 \vee r_2)| &\leq |\partial_{\Sigma^+} r_1| + |\partial_{\Sigma^+} r_2| \\
|\partial_{\Sigma^+} r_1^\star| &\leq |\partial_{\Sigma^+} r_1| \\
|\partial_{\Sigma^+}(r_1 ; r_2)| &\leq |\partial_{\Sigma^+} r_1| + |\partial_{\Sigma^+} r_2| \\
|\partial_{\Sigma^+}(r_1 : r_2)| &\leq |\partial_{\Sigma^+} r_1| + |\partial_{\Sigma^+} r_2| \\
|\partial_{\Sigma^+}(r_1 \wedge r_2)| &\leq |\partial_{\Sigma^+} r_1| \times |\partial_{\Sigma^+} r_2| \\
|\partial_{\Sigma^+} \mathsf{fm}(r_1)| &\leq 2^{|\partial_{\Sigma^+} r_1|}
\end{aligned}$$

Proof. These inequalities are consequences of equations (10)–(15) and Definition 11.

$$\begin{aligned}
|\partial_{\Sigma^+}(r_1 \vee r_2)| &= \left| \bigcup_{w \in \Sigma^+} \partial_w(r_1 \vee r_2) \right| && \text{by Def. 11} \\
&= \left| \bigcup_{w \in \Sigma^+} (\partial_w r_1) \cup (\partial_w r_2) \right| && \text{by (10)} \\
&= |(\partial_{\Sigma^+} r_1) \cup (\partial_{\Sigma^+} r_2)| && \text{by Def. 11} \\
&\leq |\partial_{\Sigma^+} r_1| + |\partial_{\Sigma^+} r_2|
\end{aligned}$$

$$
\begin{aligned}
|\partial_{\Sigma^+} r_1^\star| &= \left| \bigcup_{w \in \Sigma^+} \partial_w r_1^\star \right| && \text{by Def. 11} \\
&\leq \left| \bigcup_{w \in \Sigma^+} \bigcup_{v \in Sfx(w)} (\partial_v r_1)\,;\, r_1^\star \right| && \text{by (11)} \\
&\leq \left| \bigcup_{w \in \Sigma^+} (\partial_w r_1)\,;\, r_1^\star \right| \overset{\text{Def. 11}}{=} |(\partial_{\Sigma^+} r_1)\,;\, r_1^\star| = |\partial_{\Sigma^+} r_1|
\end{aligned}
$$

$$
\begin{aligned}
|\partial_{\Sigma^+}(r_1\,;\, r_2)| &= \left| \bigcup_{w \in \Sigma^+} \partial_w (r_1\,;\, r_2) \right| && \text{by Def. 11} \\
&\leq \left| \bigcup_{w \in \Sigma^+} \left(((\partial_w r_1)\,;\, r_2) \cup \bigcup_{v \in Sfx(w)} \partial_v r_2 \right) \right| && \text{by (12)} \\
&= \left| \left(\bigcup_{w \in \Sigma^+} (\partial_w r_1)\,;\, r_2 \right) \cup \bigcup_{w \in \Sigma^+} \partial_w r_2 \right| \\
&= |((\partial_{\Sigma^+} r_1)\,;\, r_2) \cup (\partial_{\Sigma^+} r_2)| && \text{by Def. 11} \\
&\leq |\partial_{\Sigma^+} r_1| + |\partial_{\Sigma^+} r_2|
\end{aligned}
$$

The proof that $|\partial_{\Sigma^+}(r_1 : r_2)| \leq |\partial_{\Sigma^+} r_1| + |\partial_{\Sigma^+} r_2|$ is exactly the same as above, using (13) instead of (12).

$$
\begin{aligned}
|\partial_{\Sigma^+}(r_1 \wedge r_2)| &= \left| \bigcup_{w \in \Sigma^+} \partial_w (r_1 \wedge r_2) \right| && \text{by Def. 11} \\
&= \left| \bigcup_{w \in \Sigma^+} \{p_1 \wedge p_2 \mid p_1 \in \partial_w r_1,\, p_2 \in \partial_w r_2\} \right| && \text{by (14)} \\
&\leq |\{p_1 \wedge p_2 \mid p_1 \in \partial_{\Sigma^+} r_1,\, p_2 \in \partial_{\Sigma^+} r_2\}| \\
&\leq |\partial_{\Sigma^+} r_1| \times |\partial_{\Sigma^+} r_2|
\end{aligned}
$$

$$
\begin{aligned}
|\partial_{\Sigma^+} \mathsf{fm}(r_1)| &= |\bigcup_{w \in \Sigma^+} \partial_w \mathsf{fm}(r_1)| && \text{by Def. 11} \\
&= \left| \left\{ \mathsf{fm}\Big(\bigvee_{p \in \partial_w r_1} p \Big) \;\middle|\; w \in \Sigma^+ \right\} \right| && \text{by (15)} \\
&\leq \left| \left\{ \mathsf{fm}\Big(\bigvee_{p \in P} p \Big) \;\middle|\; P \in \partial_{\Sigma^+} r_1 \right\} \right| = 2^{|\partial_{\Sigma^+} r_1|}
\end{aligned}
$$

□

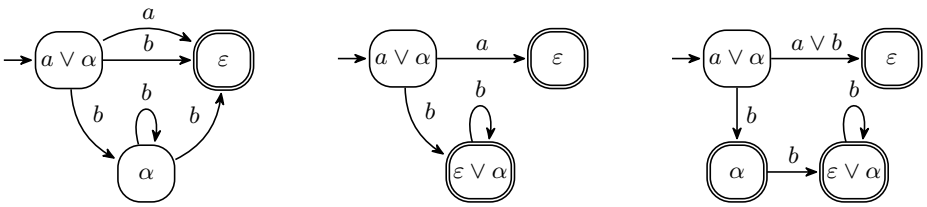

Fig. 14: Three equivalent automata built with variants of Algorithm 1 for the formula $a \vee \alpha$ where α abbreviates $b^\star : b$. (Note that α is equivalent to b^+ but the algorithm does not know that.) We have $\mathsf{LF}(\alpha) = \{(b, \varepsilon), (b, \alpha)\}$ so it follows that $\mathsf{LF}(a \vee \alpha) = \{(a, \varepsilon), (b, \varepsilon), (b, \alpha)\}$. Applying UP first on the initial formula does not leave anything for US to simplify: $\mathsf{US}(\mathsf{UP}(\mathsf{LF}(a \vee \alpha))) = \mathsf{UP}(\mathsf{LF}(a \vee \alpha)) = \{(a, \varepsilon), (b, \varepsilon \vee \alpha)\}$. Conversely, applying US first makes UP useless in this case: $\mathsf{UP}(\mathsf{US}(\mathsf{LF}(a \vee \alpha))) = \mathsf{US}(\mathsf{LF}(a \vee \alpha)) = \{(a \vee b, \varepsilon), (b, \alpha)\}$.

Corollary 1. *For any expression r, the set $\partial_{\Sigma^+} r$ is finite.*

Proof. Straightforward induction on the grammar of r, using Lemma 2. □

Since $\mathrm{Terms}(r) = \partial_{\Sigma^+} r \cup \{r\}$, this concludes the proof of Theorem 2.

Lemma 2 can actually be used to prove to some finer bounds on some subsets of SEREs:

Corollary 2. *For a SERE r, let $\|r\|$ denote the number of occurrences of maximal Boolean subformulas in r (i.e., the number of times rule "b" was used to produce r with the grammar given in Definition 2).*

1. *If r does not use operators $\wedge$ and fm, we have $|\partial_{\Sigma^+} r| \leq \|r\|$.*
   (In other words, adding the ":" operator to the set of classical regular operators preserves the bound established by Antimirov [2, Theorem 3.4].)
2. *If r does not use operator fm, we have $|\partial_{\Sigma^+} r| \leq 2^{\|r\|}$.*

Proof. Straightforward induction on the grammar of r, using Lemma 2. □

E US ∘ UP vs. UP ∘ US

As mentioned in Section 4 it is not clear if one should use US before or after UP. Figure 14 shows one case where replacing $\mathsf{LF}(f)$ by $\mathsf{US}(\mathsf{UP}(\mathsf{LF}(f)))$ in Algorithm 1 produces a smaller automaton than when using $\mathsf{UP}(\mathsf{US}(\mathsf{LF}(f)))$. Figure 15 shows an opposite case, where using $\mathsf{US}(\mathsf{UP}(\mathsf{LF}(f)))$ results in a larger automaton than with $\mathsf{UP}(\mathsf{US}(\mathsf{LF}(f)))$.

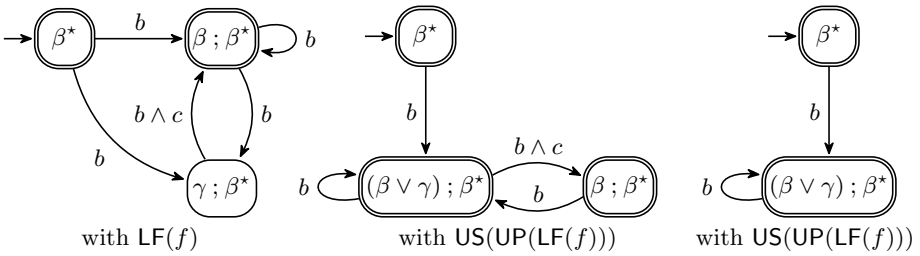

Fig. 15: Three equivalent automata built with variants of Algorithm 1 for the formula $\beta^\star$ where $\beta = (b^\star \wedge (1 \vee (1;c))^\star)$. We also use $\gamma = (b^\star \wedge (c;((1 \vee (1;c))^\star)))$. We have $\mathsf{LF}(\beta^\star) = \{(b, \beta\,;\beta^\star), (b, \gamma\,;\beta^\star)\}$, $\mathsf{LF}(\beta\,;\beta^\star) = \mathsf{LF}(\beta^\star)$, and $\mathsf{LF}(\gamma\,;\beta^\star) = \{(b\wedge c, \beta;\beta^\star)\}$. The order of US and UP does not change anything when processing the initial state; we have $\mathsf{US}(\mathsf{UP}(\mathsf{LF}(\beta^\star))) = \mathsf{UP}(\mathsf{US}(\mathsf{LF}(\beta^\star))) = \{(b, (\beta\vee\gamma);\beta^\star)\}$. However it does have an influence when processing $(\beta \vee \gamma)\,;\beta^\star$. Let us call $L = \mathsf{LF}((\beta \vee \gamma)\,;\beta^\star) = \{(b, \beta\,;\beta^\star), (b, \gamma\,;\beta^\star), (b \wedge c, \beta\,;\beta^\star)\}$. Then $\mathsf{US}(\mathsf{UP}(L)) = \mathsf{UP}(L) = \{(b, (\beta \vee \gamma)\,;\beta^\star), (b \wedge c, \beta\,;\beta^\star)\}$ but $\mathsf{US}(L) = \{(b, \beta\,;\beta^\star), (b, \gamma\,;\beta^\star)\}$ and therefore $\mathsf{US}(L) = \{(b, (\beta \vee \gamma)\,;\beta^\star)\}$.

F Implementation Considerations

We now discuss several topics related to our implementation of linear forms in Spot [12], and their use in Algorithm 1. We mention alternate implementations we attempted, in hope that that can give new ideas to the reader.

Since all SEREs handled during translation are combinations of sub-formulas of the original SERE, we represent all SEREs as a direct acyclic graph in which each sub-formula has a unique representation. A node in this DAG stores an operator and a list of pointers to the nodes of its operands. A global uniqueness table allows to map a pair (operator, operands) to the unique node that represents it if it exists in the graph. Additionally, we apply the rules (A)–(F) (from page 3) during the construction of the node. In particular, for rule (A), we handle associative operators as n-ary operators, and we inline their operands when they have the same top-level operator. Rule (C) is achieved by sorting all operands (for instance in their node creation order), and this then makes it easy to locate duplicate elements for rule (I1).

Linear forms, as introduced in Definition 6 are sets of pairs (p_i, s_i) where p_i is a Boolean formula which we represent as a BDD, and s_i is a SERE. Several data structures could be used to represent a set of such pairs, and the choice of the data structure could also depend on whether we later plan to use UP or US. For instance if one plans to use UP, it is tempting to represent linear forms as hash maps that map each p_i to its s_i, so that the latter can be updated whenever a new pair (p_i, s'_i) is introduced into the linear form. However, during development we noticed that linear forms were usually very small (average 3.12, median 1), making the construction of such mappings more expensive than simpler data structures. As a consequence, we simply represent them as arrays of pairs. If UP

or US needs to be applied, we first sort those pairs by prefixes or suffixes, and then merge the relevant pairs.

In classical algorithms to translate LTL formulas into Büchi automata, it is also common to rewrite any LTL formula φ into an equivalent form $\varphi = \bigvee_i p_i \wedge \mathsf{X}\, s_i$ where p_i is a Boolean formula that has to be verified now, while s_i is another LTL formula to be verified on the next step. This LTL rewriting, which has been called "disjunctive normal form" by some [20] is just a linear form in disguise. Couvreur's algorithm [8], which is implemented in Spot, has a convenient representation of these normal forms. It uses BDDs: since p_i is a Boolean formula, they are converted to BDDs in a straightforward way, and new BDD variable are introduced to represent each subformula of the form X_{s_i}. Once a BDD has been constructed for the full formula, its pairs (p_i, s_i) can be recovered by computing prime implicants. The whole process has a few advantages: firstly a BDD representation is unique for two equivalent formulas, so it can be used to label states in an algorithm similar to Algorithm 3, secondly extracting prime implicants will automatically remove some of the pairs (p_i, s_i) that were already covered by another one. As it was already used in Spot, we initially tried to use this representation for linear forms, however, and probably because of the average size of linear forms, we found that it was very slow compared to our array-based representation of linear forms.

G Distributions of SERE In Our Dataset

In this appendix we detail how the SEREs of our benchmark were generated. As mentioned in Section 6, we used Spot's `randltl --sere`[4] to generated random SERE.

We used option `--sere-prio=fstar=0,fstar_b=0,star_b=0,andnlm=0` to disable some additional operators not discussed in this article. `fstar` is a reference to the "fusion star" operator introduced by Dax et al. [10] which is to ":" what "*" is to ";". The `fstar_b` and `star_b` are bounded variant of these two star operators, where a minimum and maximum number of repetition is given. The `andnlm` is a "non length-matching" intersection, present in both SVA and PSL, but that can be defined as syntactic sugar. It should be pointed out that Spot's constructors for SEREs have more simplification rules than just (A)–(F). For instance $a\,;\,a\,;\,a^\star$ will automatically be turned into a "bounded star" $a^{\star 2..}$ that has to repeat at least twice. Although we did not mention those operators in the paper, our implementation support those in the obvious way. A complete list of reduction rules Spot systematically applies can be found in the "trivial identities" sections of https://spot.lre.epita.fr/tl.pdf.

`randltl` is also passed argument `--tree-size=20..40 25` to construct syntactic trees with between 20 and 40 nodes (operators or atomic propositions), using up to 25 distinct atomic propositions. Because nodes are counted during random generation, before actually constructing the SERE with rules (A)–(F),

[4] https://spot.lre.epita.fr/randltl.html

the resulting SERE may have a much smaller size. Eventually, we kept only formulas that had a size up to 35 and at most 15 atomic propositions.

We let `randltl` generate an infinite amount of unique SEREs, and grouped them according to their final size and count of unique atomic propositions, as shown in Table 1. Each group was capped to 50 SEREs, and we stopped the generation once we had 12500 formulas, as it became harder to stumble upon formulas that would fill incomplete groups.

As an example, the set of 5 formulas contained in the group for $|\varphi| = 2$ and $|AP| = 1$ is $\{\neg a, a^\star, a^{\star 1\cdot\cdot}, a^{\star 2\cdot\cdot}, a^{\star 2..2}\}$. These would correspond to the formulas $\{\neg a, a^\star, (a\,;a^\star),\ (a\,;a\,;a^\star), (a\,;a)\}$ if we restrict ourselves to the notations of Section 2.

Table 1: Distribution of SEREs based on their size and count of distinct atomic propositions

$\|\varphi\|$	$\|AP\|$														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1														
2	5														
3	10	6													
4	47	45	3												
5	50	50	50	3											
6	50	50	50	50	3										
7	50	50	50	50	36	2									
8	50	50	50	50	50	18									
9	50	50	50	50	50	50	8								
10	50	50	50	50	50	50	38	2							
11	33	50	50	50	50	50	50	8							
12	13	50	50	50	50	50	50	33	1						
13	5	50	50	50	50	50	50	50	3						
14	2	50	50	50	50	50	50	50	21						
15	1	40	50	50	50	50	50	50	50	6					
16		14	50	50	50	50	50	50	50	16	1				
17		8	50	50	50	50	50	50	50	50	2				
18		5	50	50	50	50	50	50	50	50	4				
19			50	50	50	50	50	50	50	50	25	1			
20			24	50	50	50	50	50	50	50	50	7			
21			15	50	50	50	50	50	50	50	50	15			
22			8	50	50	50	50	50	50	50	50	50	3		
23			4	50	50	50	50	50	50	50	50	50	14		
24			3	32	50	50	50	50	50	50	50	50	19		
25			1	15	50	50	50	50	50	50	50	50	49	2	
26				8	50	50	50	50	50	50	50	50	50	5	1
27				3	50	50	50	50	50	50	50	50	50	17	
28				1	27	50	50	50	50	50	50	50	50	50	1
29					13	50	50	50	50	50	50	50	50	50	11
30					18	50	50	50	50	50	50	50	50	50	14
31					7	50	50	50	50	50	50	50	50	50	29
32					2	37	50	50	50	50	50	50	50	50	49
33					2	23	50	50	50	50	50	50	50	50	50
34						7	50	50	50	50	50	50	50	50	50
35					1	1	43	50	50	50	50	50	50	50	50