Symbolic Solving of Extended Regular Expression Inequalities

Matthias Keil, Peter Thiemann
University of Freiburg,
Freiburg, Germany

Extended Regular Expressions

**Definition**

\[ r, s, t \ := \ \epsilon | A | r+s | r \cdot s | r^* | r & s | !r \]

- $\Sigma$ is a potentially infinite set of symbols
- $A, B, C \subseteq \Sigma$ range over sets of symbols
- $[r] \subseteq \Sigma^*$ is the language of a regular expression $r$, where $[A] = A$
Definition

Given two regular expressions $r$ and $s$,

$$ r \subseteq s \iff \llbracket r \rrbracket \subseteq \llbracket s \rrbracket $$

- $\llbracket r \rrbracket \subseteq \llbracket s \rrbracket \iff \llbracket r \rrbracket \cap \llbracket s \rrbracket = \emptyset$
- Decidable using standard techniques:
  Construct DFA for $r \& \neg s$ and check for emptiness
- Drawback is the expensive construction of the automaton
- PSPACE-complete
Antimirov’s Algorithm

- Deciding containment for *basic regular expressions*
- Based on derivatives and expression rewriting
- Avoid the construction of an automaton
- \( \partial_a(r) \) computes a regular expression for \( a^{-1}[r] \) (Brzozowski) with \( u \in [r] \) iff \( \epsilon \in [\partial_u(r)] \)

**Lemma**

For regular expressions \( r \) and \( s \),

\[
    r \sqsubseteq s \iff (\forall u \in \Sigma^*) \, \partial_u(r) \sqsubseteq \partial_u(s).
\]
Antimirov’s Algorithm (cont’d)

Lemma

\[ r \sqsubseteq s \iff (\nu(r) \Rightarrow \nu(s)) \land (\forall a \in \Sigma) \partial_a(r) \sqsubseteq \partial_a(s) \]

\[ \frac{\nu(r) \land \neg \nu(s)}{r \sqsubseteq s \vdash_{CC} false} \quad \text{CC-DISPROVE} \]

\[ \frac{\nu(r) \Rightarrow \nu(s)}{r \sqsubseteq s \vdash_{CC} \{ \partial_a(r) \sqsubseteq \partial_a(s) \mid a \in \Sigma \}} \quad \text{CC-UNFOLD} \]

- Choice of next step’s inequality is nondeterministic
- An infinite alphabet requires to compute for infinitely many \( a \in \Sigma \)
First Symbols

Lemma

\[ r \subseteq s \iff (\nu(r) \Rightarrow \nu(s)) \land (\forall a \in \text{first}(r)) \partial_a(r) \subseteq \partial_a(s) \]

- Let \( \text{first}(r) \) := \{ \( a \mid aw \in \llbracket r \rrbracket \) \} be the set of first symbols
- Restrict symbols to first symbols of the left hand side
- CC-Unfold does not have to consider the entire alphabet
- For extended regular expressions, \( \text{first}(r) \) may still be an infinite set of symbols
Problems

- Antimirov’s algorithm only works with basic regular expressions or requires a finite alphabet
- Extension of *partial derivatives* (Caron et al.) that computes an NFA from an extended regular expression
- Works on sets of sets of expressions
- Computing derivatives becomes more expensive
Goal

- Algorithm for deciding $[r] \subseteq [s]$ quickly
- Handle *extended regular expressions*
- Deal effectively with very large (or infinite) alphabets (e.g. Unicode character set)

Solution

- Require finitely many atoms, even if the alphabet is infinite
- Compute derivatives with respect to literals
Representing Sets of Symbols

A literal is a set of symbols $A \subseteq \Sigma$

**Definition**

$A$ is an element of an *effective* boolean algebra $(U, \sqcup, \sqcap, \neg, \bot, \top)$ where $U \subseteq \wp(\Sigma)$ is closed under the boolean operations.

- For finite (small) alphabets:
  $U = \wp(\Sigma), \ A \subseteq \Sigma$

- For infinite (or just too large) alphabets:
  $U = \{ A \in \wp(\Sigma) \mid A \text{ finite } \lor \overline{A} \text{ finite} \}$

- Second-level regular expressions:
  $\Sigma \subseteq \wp(\Gamma^*)$ with $U = \{ A \subseteq \wp(\Gamma^*) \mid A \text{ is regular} \}$

- Formulas drawn from a first-order theory over alphabets
  For example, $[a\mathchar`-z]$ represented by $x \geq 'a' \land x \leq 'z'$
Derivatives with respect to Literals

- Definition for $\partial_A(r)$?
- $\partial_a(r)$ computes a regular expression for $a^{-1}[r]$ (Brzozowski)

**Desired property**

$$\left[ \partial_A(r) \right] = A^{-1}[r] = \bigcup_{a \in A} a^{-1}[r] = \bigcup_{a \in A} \left[ \partial_a(r) \right]$$
Positive Derivatives on Literals

**Definition**

\[ \delta^+_A(B) := \begin{cases} 
\epsilon, & B \cap A \neq \perp \\
\emptyset, & \text{otherwise} 
\end{cases} \]

**Problem**

With \( A = \{a, b\} \) and \( r = (a \cdot c) \& (b \cdot c) \),

\[
\delta^+_A(r) = \delta^+_A(a \cdot c) \& \delta^+_A(b \cdot c)
= c \& c
\equiv \emptyset
\]
Negative Derivatives on Literals

Definition

\[ \delta_A^-(B) := \begin{cases} \epsilon, & \overline{B} \cap A = \perp \\ \emptyset, & \text{otherwise} \end{cases} \]

Problem

With \( A = \{a, b\} \) and \( r = (a \cdot c) + (b \cdot c) \),

\[ \delta_A^-(r) = \delta_A^-(a \cdot c) + \delta_A^-(b \cdot c) = \emptyset + \emptyset = \emptyset \subseteq c \]
Positive and Negative Derivatives

- Extends Brzozowski’s derivative operator to sets of symbols.
- Defined by induction and flip on the complement operator

**Definition**

From $\partial_a(!s) = !\partial_a(s)$, define:

$$\delta_A^+(!r) := !\delta_A^-(r)$$
$$\delta_A^-(!r) := !\delta_A^+(r)$$

**Lemma**

For any regular expression $r$ and literal $A$,

$$[[\delta_A^+(r)]] \supseteq \bigcup_{a \in A} [[\partial_a(r)]]$$
$$[[\delta_A^-(r)]] \subseteq \bigcap_{a \in A} [[\partial_a(r)]]$$
Literals of an Inequality

Lemma

\[ r \subseteq s \iff (\nu(r) \Rightarrow \nu(s)) \land (\forall a \in \text{first}(r)) \partial_a(r) \subseteq \partial_a(s) \]

- first\((r)\) may still be an infinite set of symbols
- Use \textit{first literals} as representatives of the \textit{first symbols}

Example

1. Let \( r = \{a, b, c, d\} \cdot d^* \), then \( \{a, b, c, d\} \) is a first literal
2. Let \( s = \{a, b, c\} \cdot c^* + \{b, c, d\} \cdot d^* \), then \( \{a, b, c\} \) and \( \{b, c, d\} \) are first literals
### Problem

Let \( r = \{a, b, c, d\} \cdot d^* \), \( s = \{a, b, c\} \cdot c^* + \{b, c, d\} \cdot d^* \), and \( A = \{a, b, c, d\} \), then

\[
\begin{align*}
\delta_+^A(r) & \subseteq \delta_+^A(s) \quad (1) \\
\delta_+^A(\{a, b, c, d\} \cdot d^*) & \subseteq \delta_+^A(\{a, b, c\} \cdot c^*) + \delta_+^A(\{b, c, d\} \cdot d^*) \quad (2) \\
d^* & \subseteq c^* + d^* \quad (3)
\end{align*}
\]

- Positive (negative) derivatives yield an upper (lower) approximation.
- To obtain the precise information, we need to restrict these literals suitably to *next literals*, e.g. \( \{\{a\}, \{b, c\}, \{d\}\} \)
Next Literals

\[
\begin{align*}
\text{next}(\epsilon) &= \{\emptyset\} \\
\text{next}(A) &= \{A\} \\
\text{next}(r+s) &= \text{next}(r) \boxtimes \text{next}(s) \\
\text{next}(r \cdot s) &= \begin{cases} \\
\text{next}(r) \boxtimes \text{next}(s), & \nu(r) \\
\text{next}(r), & \neg\nu(r) \\
\end{cases} \\
\text{next}(r^*) &= \text{next}(r) \\
\text{next}(r \& s) &= \text{next}(r) \sqcap \text{next}(s) \\
\text{next}(!r) &= \text{next}(r) \cup \{\bigcap\{\overline{A} \mid A \in \text{next}(r)\}\}
\end{align*}
\]

Definition

Let \( \mathcal{L}_1 \) and \( \mathcal{L}_2 \) be two sets of disjoint literals.

\[
\mathcal{L}_1 \boxtimes \mathcal{L}_2 := \\
\{(A_1 \sqcap A_2), (A_1 \sqcap \bigcup \mathcal{L}_2), (\bigcup \mathcal{L}_1 \sqcap A_2) \mid A_1 \in \mathcal{L}_1, A_2 \in \mathcal{L}_2\}
\]
Next Literals (cont’d)

Example

Let \( s = \{a, b, c\} \cdot c^* + \{b, c, d\} \cdot d^* \), then

\[
\text{next}(s) = \text{next}((a, b, c) \cdot c^*) \times \text{next}((b, c, d) \cdot d^*)
= \{\{a, b, c\}\} \times \{\{b, c, d\}\}
= \{\{a\}, \{b, c\}, \{d\}\}
\]

Lemma

For all \( r \),

- \( \bigcup \text{next}(r) \supseteq \text{first}(r) \)
- \( |\text{next}(r)| \) is finite
- \( (\forall A, B \in \text{next}(r)) \ A \cap B = \emptyset \)
Lemma

Let \( \mathcal{L} = \text{next}(r) \) and \( A \in \text{next}(r) \setminus \{\emptyset\} \).

1. \((\forall a, b \in A) \partial_a(r) = \partial_b(r) \land \delta^+_A(r) = \delta^-_A(r) = \partial_a(r)\)
2. \((\forall a \notin \bigcup \mathcal{L}) \partial_a(r) = \emptyset\)

Definition

Let \( A' \in \text{next}(r) \). For each \( \emptyset \neq A \subseteq A' \) define \( \partial_A(r) := \partial_a(r) \), where \( a \in A \).
Next Literals of an Inequality

- Next literal of \( \text{next}(r \sqsubseteq s) \)
- Sound to join literals of both sides \( \text{next}(r) \bowtie \text{next}(s) \)
- Contains also symbols from \( s \)
- First symbols of \( r \) are sufficient to prove containment

**Definition**

Let \( \mathcal{L}_1 \) and \( \mathcal{L}_2 \) be two sets of disjoint literals.

\[
\mathcal{L}_1 \times \mathcal{L}_2 := \{ (A_1 \cap A_2), (A_1 \cap \bigcup \mathcal{L}_2) \mid A_1 \in \mathcal{L}_1, A_2 \in \mathcal{L}_2 \}
\]

Left-based join corresponds to \( \text{next}(r \&(!s)) \).

**Definition**

Let \( r \sqsubseteq s \) be an inequality, define: \( \text{next}(r \sqsubseteq s) := \text{next}(r) \bowtie \text{next}(s) \)
Lemma

\[ r \sqsubseteq s \iff (\nu(r) \Rightarrow \nu(s)) \land (\forall a \in \text{first}(r)) \partial_a(r) \sqsubseteq \partial_a(s) \]

To determine a finite set of representatives

- select one symbol \( a \) from each equivalence class \( A \in \text{next}(r) \)
- calculate with \( \delta^+_A(r) \) or \( \delta^-_A(r) \) with \( A \in \text{next}(r) \)

Theorem (Containment)

\[ r \sqsubseteq s \iff (\nu(r) \Rightarrow \nu(s)) \land (\forall A \in \text{next}(r \sqsubseteq s)) \partial_A(r) \sqsubseteq \partial_A(s) \]
Conclusion

- Generalize Brzozowski’s derivative operator
- Extend Antimirov’s algorithm for proving containment
- Provides a symbolic decision procedure that works with extended regular expressions on infinite alphabets
- Literals drawn from an effective boolean algebra
- Main contribution is to identify a finite set that covers all possibilities
The language $[r] \subseteq \Sigma^*$ of a regular expression $r$ is defined inductively by:

\[
egin{align*}
  [\epsilon] & = \{ \epsilon \} \\
  [A] & = \{ a \mid a \in A \} \\
  [r+s] & = [r] \cup [s] \\
  [r \cdot s] & = [r] \cdot [s] \\
  [r^*] & = [r] \cdot [r^*] \\
  [r \& s] & = [r] \cap [s] \\
  [!r] & = [\overline{r}] 
\end{align*}
\]
The *nullable* predicate $\nu(r)$ indicates whether $[r]$ contains the empty word, that is, $\nu(r)$ iff $\epsilon \in [r]$.

\[
\begin{align*}
\nu(\epsilon) &= true \\
\nu(A) &= false \\
\nu(r+s) &= \nu(r) \lor \nu(s) \\
\nu(r\cdot s) &= \nu(r) \land \nu(s) \\
\nu(r^*) &= true \\
\nu(r&s) &= \nu(r) \land \nu(s) \\
\nu(!r) &= \neg \nu(r)
\end{align*}
\]
\( \partial_a(r) \) computes a regular expression for the left quotient \( a^{-1}[r] \).

\[
\begin{align*}
\partial_a(\epsilon) &= \emptyset \\
\partial_a(A) &= \begin{cases} 
\epsilon, & a \in A \\
\emptyset, & a \notin A
\end{cases} \\
\partial_a(r+s) &= \partial_a(r) + \partial_a(s) \\
\partial_a(r\cdot s) &= \begin{cases} 
\partial_a(r)\cdot s + \partial_a(s), & \nu(r) \\
\partial_a(r)\cdot s, & \neg \nu(r)
\end{cases} \\
\partial_a(r^*) &= \partial_a(r)\cdot r^* \\
\partial_a(r\& s) &= \partial_a(r)\& \partial_a(s) \\
\partial_a(!r) &= !\partial_a(r)
\end{align*}
\]
First Symbols

Let \( \text{first}(r) := \{ a \mid aw \in [r] \} \) be the set of first symbols derivable from regular expression \( r \).

\[
\begin{align*}
\text{first}(\epsilon) &= \emptyset \\
\text{first}(A) &= A \\
\text{first}(r+s) &= \text{first}(r) \cup \text{first}(s) \\
\text{first}(r \cdot s) &= \begin{cases} 
\text{first}(r) \cup \text{first}(s), & \nu(r) \\
\text{first}(r), & \neg \nu(r)
\end{cases} \\
\text{first}(r^*) &= \text{first}(r) \\
\text{first}(r \& s) &= \text{first}(r) \cap \text{first}(s) \\
\text{first}(!r) &= \Sigma \setminus \{ a \in \text{first}(r) \mid \partial_a(r) \neq \Sigma^* \}
\end{align*}
\]
Let $\text{first}(r) := \{a \mid aw \in [r]\}$ be the set of first symbols derivable from regular expression $r$.

$$
\begin{align*}
\text{literal}(\epsilon) &= \emptyset \\
\text{literal}(A) &= \{A\} \\
\text{literal}(r+s) &= \text{literal}(r) \cup \text{literal}(s) \\
\text{literal}(r \cdot s) &= \begin{cases} 
\text{literal}(r) \cup \text{literal}(s), & \nu(r) \\
\text{literal}(r), & \neg\nu(r)
\end{cases} \\
\text{literal}(r^*) &= \text{literal}(r) \\
\text{literal}(r\&s) &= \text{literal}(r) \cap \text{literal}(s) \\
\text{literal}(!r) &= \Sigma \cap \bigcup \{A \in \text{literal}(r) \mid \partial_A(r) = \Sigma^*\}
\end{align*}
$$
Lemma (Coverage)

For all $a$, $u$, and $r$ it holds that:

$$u \in \left[ \partial_a(r) \right] \iff \exists A \in \text{next}(r) : a \in A \land u \in \left[ \delta^+_A(r) \right] \land u \in \left[ \delta^-_A(r) \right]$$
Theorem (Finiteness)

Let $R$ be a finite set of regular inequalities. Define

$$F(R) = R \cup \{ \partial_A(r \sqsubseteq s) \mid r \sqsubseteq s \in R, A \in \text{next}(r \sqsubseteq s) \}$$

For each $r$ and $s$, the set $\bigcup_{i \in \mathbb{N}} F^{(i)}(\{ r \sqsubseteq s \})$ is finite.
(Disprove) \[ \nu(r) \quad \neg \nu(s) \] \[ \Gamma \vdash r \sqsubseteq s : false \]

(Cycle) \[ r \sqsubseteq s \in \Gamma \] \[ \Gamma \vdash r \sqsubseteq s : true \]

(Unfold-True) \[ r \sqsubseteq s \notin \Gamma \quad \nu(r) \Rightarrow \nu(s) \]
\[ \forall A \in \text{next}(r \sqsubseteq s) : \Gamma \cup \{r \sqsubseteq s\} \vdash \partial_A(r) \sqsubseteq \partial_A(s) : true \] \[ \Gamma \vdash r \sqsubseteq s : true \]

(Unfold-False) \[ r \sqsubseteq s \notin \Gamma \quad \nu(r) \Rightarrow \nu(s) \]
\[ \exists A \in \text{next}(r \sqsubseteq s) : \Gamma \cup \{r \sqsubseteq s\} \vdash \partial_A(r) \sqsubseteq \partial_A(s) : false \] \[ \Gamma \vdash r \sqsubseteq s : false \]
Prove and Disprove Axioms

(Prove-Identity)
\[ \Gamma \vdash r \sqsubseteq r : true \]

(Prove-Empty)
\[ \Gamma \vdash \emptyset \sqsubseteq s : true \]

(Prove-Nullable)
\[ \nu(s) \quad \vdash \epsilon \sqsubseteq s : true \]

(Disprove-Empty)
\[ \exists A \in \text{next}(r) : A \neq \emptyset \quad \vdash r \sqsubseteq \emptyset : false \]
Soundness

Theorem (Soundness)

For all regular expression $r$ and $s$:

$$\emptyset \vdash r \subseteq s : \top \iff r \subseteq s$$
Counterexample

Let \( r = \{a, b, c, d\} \cdot d^* \), \( s = \{a, b, c\} \cdot d^* + \{b, c, d\} \cdot d^* \), and \( A = \{a, b, c, d\} \), then

\[
\delta_A^-(r) \subseteq \delta_A^+(s) \tag{4}
\]

\[
\delta_A^-(\{a, b, c, d\} \cdot d^*) \subseteq \delta_A^-(\{a, b, c\} \cdot d^*) + \delta_A^-(\{b, c, d\} \cdot d^*) \tag{5}
\]

\[
d^* \subseteq \emptyset + \emptyset \tag{6}
\]
Next Literals of an Inequality

Example

Let $r = \{a, b, c, d\} \cdot d^*$, $s = \{a, b, c\} \cdot c^* + \{b, c, d\} \cdot d^*$ then

$\text{next}(r \sqsubseteq s) = \text{next}(\{a, b, c, d\} \cdot d^*) \times \text{next}(\{a, b, c\} \cdot d^* + \{b, c, d\} \cdot d^*)$

$= \{\{a\}, \{b, c\}, \{d\}\}$
Conjecture

\[ r \sqsubseteq s \iff (\nu(r) \Rightarrow \nu(s)) \land (\forall A \in \text{literal}(r)) \delta^+_A(r) \sqsubseteq \delta^-_A(s) \]