Decision Procedures for Sequence Theories

Artur Jeż[1], Anthony W. Lin[2,3], Oliver Markgraf[2](✉), and Philipp Rümmer[4,5]

[1] University of Wrocław, Wrocław, Poland
[2] TU Kaiserslautern, Kaiserslautern, Germany
markgraf@cs.uni-kl.de
[3] Max Planck Institute for Software Systems, Kaiserslautern, Germany
[4] University of Regensburg, Regensburg, Germany
[5] Uppsala University, Uppsala, Sweden

Abstract. Sequence theories are an extension of theories of strings with an infinite alphabet of letters, together with a corresponding alphabet theory (e.g. linear integer arithmetic). Sequences are natural abstractions of extendable arrays, which permit a wealth of operations including append, map, split, and concatenation. In spite of the growing amount of tool support for theories of sequences by leading SMT-solvers, little is known about the decidability of sequence theories, which is in stark contrast to the state of the theories of strings. We show that the decidable theory of strings with concatenation and regular constraints can be extended to the world of sequences over an alphabet theory that forms a Boolean algebra, while preserving decidability. In particular, decidability holds when regular constraints are interpreted as parametric automata (which extend both symbolic automata and variable automata), but fails when interpreted as register automata (even over the alphabet theory of equality). When length constraints are added, the problem is Turing-equivalent to word equations with length (and regular) constraints. Similar investigations are conducted in the presence of symbolic transducers, which naturally model sequence functions like map, split, filter, *etc.* We have developed a new sequence solver, SeCo, based on parametric automata, and show its efficacy on two classes of benchmarks: (i) invariant checking on array-manipulating programs and parameterized systems, and (ii) benchmarks on symbolic register automata.

1 Introduction

Sequences are an extension of strings, wherein elements might range over an infinite domain (e.g., integers, strings, and even sequences themselves). Sequences

A. Jeż was supported under National Science Centre, Poland project number 2017/26/E/ST6/00191. A. Lin and O. Markgraf were supported by the ERC Consolidator Grant 101089343 (LASD). P. Rümmer was supported by the Swedish Research Council (VR) under grant 2018-04727, the Swedish Foundation for Strategic Research (SSF) under the project WebSec (Ref. RIT17-0011), and the Wallenberg project UPDATE.

© The Author(s) 2023
C. Enea and A. Lal (Eds.): CAV 2023, LNCS 13965, pp. 18–40, 2023.
https://doi.org/10.1007/978-3-031-37703-7_2

are ubiquitous and commonly used data types in modern programming languages. They come under different names, e.g., Python/Haskell/Prolog lists, Java ArrayList (and to some extent Streams) and JavaScript arrays. Crucially, sequences are *extendable*, and a plethora of operations (including append, map, split, filter, concatenation, etc.) can naturally be defined and are supported by built-in library functions in most modern programming languages.

Various techniques in software model checking [30] — including symbolic execution, invariant generation — require an appropriate SMT theory, to which verification conditions could be discharged. In the case of programs operating on sequences, we would consequently require an SMT theory of sequences, for which leading SMT solvers like Z3 [6,38] and cvc5 [4] already provide some basic support for over a decade. The basic design of sequence theories, as done in Z3 and cvc5, as well as in other formalisms like symbolic automata [15], is in fact quite natural. That is, sequence theories can be thought of as extensions of theories of strings with an infinite alphabet of letters, together with a corresponding alphabet theory, e.g. Linear Integer Arithmetic (LIA) for reasoning about sequences of integers. Despite this, very little is known about what is decidable over theories of sequences.

In the case of finite alphabets, sequence theories become theories over strings, in which a lot of progress has been made in the last few decades, barring the long-standing open problem of string equations with length constraints (e.g. see [26]). For example, it is known that the existential theory of concatenation over strings with regular constraints is decidable (in fact, PSPACE-complete), e.g., see [17,29,36,40,43]. Here, a *regular constraint* takes the form $x \in L(E)$, where E is a regular expression, mandating that the expression E matches the string represented by x. In addition, several natural syntactic restrictions — including straight-line, acylicity, and chain-free (e.g. [1,2,5,11,12,26,35]) — have been identified, with which string constraints remain decidable in the presence of more complex string functions (e.g. transducers, replace-all, reverse, etc.). In the case of infinite alphabets, only a handful of results are available. Furia [25] showed that the existential theory of sequence equations over the alphabet theory of LIA is decidable by a reduction to the existential theory of concatenation over strings (over a finite alphabet) *without regular constraints*. Loosely speaking, a number (e.g. 4) can be represented as a string in unary (e.g. 1111), and addition is then simulated by concatenation. Therefore, his decidability result does not extend to other data domains and alphabet theories. Wang et al. [45] define an extension of the array property fragment [9] with concatenation. This fragment imposes strong restrictions, however, on the equations between sequences (here called finite arrays) that can be considered.

"Regular Constraints" Over Sequences. One answer of what a regular constraint is over sequences is provided by *automata modulo theories*. Automata modulo theories [15,16] are an elegant framework that can be used to capture the notion of regular constraints over sequences: Fix an alphabet theory T that forms a Boolean algebra; this is satisfied by virtually all existing SMT theories. In this framework, one uses formulas in T to capture multiple (possibly infinitely many)

transitions of an automaton. More precisely, between two states in a *symbolic automaton* one associates a unary[1] formula $\varphi(x) \in T$. For example, $q \to_\varphi q'$ with $\varphi := x \equiv 0 \pmod 2$ over LIA corresponds to all transitions $q \to_i q'$ with any even number i. Despite their nice properties, it is known that many simple languages cannot be captured using symbolic automata; e.g., one cannot express the language consisting of sequences containing the same even number i *throughout* the sequence.

There are essentially two (expressively incomparable) extensions of symbolic automata that address the aforementioned problem: (i) Symbolic Register Automata (SRA) [14] and (ii) Parametric Automata (PA) [21,23,24]. The model SRA was obtained by combining register automata [31] and symbolic automata. The model PA extends symbolic automata by allowing *free variables* (a.k.a. *parameters*) in the transition guards, i.e., the guard will be of the form $\varphi(x, \bar{p})$, for parameters $\bar{p}$. In an accepting path of PA, a parameter p used in multiple transitions has to be instantiated with the same value, which enables comparisons of different positions in an input sequence. For example, we can assert that only sequences of the form i^*, for an even number i, are accepted by the PA with a single transition $q \to_\varphi q$ with $\varphi(x, p) := x = p \wedge x \equiv 0 \pmod 2$ and q being the start and final state. PA can also be construed as an extension of both variable automata [27] and symbolic automata. SRA and PA are not comparable: while parameters can be construed as read-only registers, SRA can only compare two different positions using equality, while PA may use a general formula in the theory in such a comparison (e.g., order).

Contributions. The main contribution of this paper is to provide *the first decidable fragments of a theory of sequences parameterized in the element theory.* In particular, we show how to leverage string solvers to solve theories over sequences. We believe this is especially interesting, in view of the plethora of existing string solvers developed in the last 10 years (e.g. see the survey [3]). This opens up new possibilities for verification tasks to be automated; in particular, we show how verification conditions for Quicksort, as well as Bakery and Dijkstra protocols, can be captured in our sequence theory. This formalization was done in the style of *regular model checking* [8,34], whose extension to infinite alphabets has been a longstanding challenge in the field. We also provide a new (dedicated) sequence solver SeCo We detail our results below.

We first show that the quantifier-free theory of sequences with concatenation and PA as regular constraints is decidable. Assuming that the theory is solvable in PSpace (which is reasonable for most SMT theories), we show that our algorithm runs in ExpSpace (i.e., double-exponential time and exponential space). We also identify conditions on the SMT theory T under which PSpace can be achieved and as an example show that Linear Real Arithmetic (LRA) satisfies those conditions. This matches the PSpace-completeness of the theory of strings with concatenation and regular constraints [18].

We consider three different variants/extensions:

[1] This can be generalized to any arity, which has to be set uniformly for the automaton.

(i) *Add length constraints.* Length constraints (e.g., $|\mathbf{x}| = |\mathbf{y}|$ for two sequence variables $\mathbf{x}, \mathbf{y}$) are often considered in the context of string theories, but the decidability of the resulting theory (i.e., strings with concatenation and length constraints) is still a long-standing open problem [26]. We show that the case for sequences is Turing-equivalent to the string case.

(ii) *Use SRA instead of PA.* We show that the resulting theory of sequences is undecidable, even over the alphabet theory T of equality.

(iii) *Add symbolic transducers.* Symbolic transducers [15,16] extend finite-state input/output transducers in the same way that symbolic automata extend finite-state automata. To obtain decidability, we consider formulas satisfying the straight-line restriction that was defined over strings theories [35]. We show that the resulting theory is decidable in 2-ExpTime and is ExpSpace-hard, if T is solvable in PSpace.

We have implemented the solver SeCo based on our algorithms, and demonstrated its efficacy on two classes of benchmarks: (i) invariant checking on array-manipulating programs and parameterized systems, and (ii) benchmarks on Symbolic Register Automata (SRA) from [14]. For the first benchmarks, we model as sequence constraints invariants for QuickSort, Dijkstra's Self-Stabilizing Protocol [20] and Lamport's Bakery Algorithm [33]. For (ii), we solve decision problems for SRA on benchmarks of [14] such as emptiness, equivalence and inclusion on regular expressions with back-references. We report promising experimental results: our solver SeCo is up to three orders of magnitude faster than the SRA solver in [14].

Organization. We provide a motivating example of sequence theories in Sect. 2. Section 3 contains the syntax and semantics of the sequence constraint language, as well as some basic algorithmic results. We deal with equational and regular constraints in Sect. 4. In Sect. 5, we deal with the decidable fragments with equational constraints, regular constraints, and transducers. We deal with extensions of these languages with length and SRA constraints in Sect. 6. In Sect. 7 we report our implementation and experimental results. We conclude in Sect. 8. Missing details and proofs can be found in the full version.

2 Motivating Example

We illustrate the use of sequence theories in verification using a implementation of QuickSort [28], shown in Listing 1. The example uses the Java Streams API and resembles typical implementations of QuickSort in functional languages; the program uses high-level operations on streams and lists like *filter* and *concatenation.* As we show, the data types and operations can naturally be modelled using a theory of sequences over integer arithmetic, and our results imply decidability of checks that would be done by a verification system.

The function `quickSort` processes a given list `l` by picking the first element as the pivot `p`, then creating two sub-lists `left`, `right` in which all numbers

```
/*@
 * ensures \forall int i; \result.contains(i) == l.contains(i);
 */
public static List<Integer> quickSort(List<Integer> l) {
  if (l.size() < 1) return l;
  Integer p = l.get(0);
  List<Integer> left   = l.stream().filter(i -> i < p)
                          .collect(Collectors.toList());
  List<Integer> right  = l.stream().skip(1).filter(i -> i >= p)
                          .collect(Collectors.toList());
  List<Integer> result = quickSort(left);
  result.add(p); result.addAll(quickSort(right));
  return result;
}
```

Listing 1. Implementation of QuickSort with Java Streams.

$\geq$`p` (resp., $<$`p`) have been eliminated. The function `quickSort` is then recursively invoked on the two sub-lists, and the results are finally concatenated and returned.

We focus on the verification of the post-condition shown in the beginning of Listing 1: sorting does not change the set of elements contained in the input list. This is a weaker form of the permutation property of sorting algorithms, and as such known to be challenging for verification methods (e.g., [42]). Sortedness of the result list can be stated and verified in a similar way, but is not considered here. Following the classical design-by-contract approach [37], to verify the partial correctness of the function it is enough to show that the post-condition is established in any top-level call of the function, assuming that the post-condition holds for all recursive calls. For the case of non-empty lists, the verification condition, expressed in our logic, is:

$$\begin{pmatrix} \mathbf{left} = T_{<\mathbf{l}_0}(\mathbf{l}) \wedge \mathbf{right} = T_{\geq \mathbf{l}_0}(\mathit{skip}_1(\mathbf{l})) \wedge \\ \forall i.\,(i \in \mathbf{left} \leftrightarrow i \in \mathbf{left}') \wedge \forall i.\,(i \in \mathbf{right} \leftrightarrow i \in \mathbf{right}') \wedge \\ \mathbf{res} = \mathbf{left}' \,.\, [\mathbf{l}_0] \,.\, \mathbf{right}' \end{pmatrix} \rightarrow \forall i.\,(i \in \mathbf{l} \leftrightarrow i \in \mathbf{res})$$

The variables $\mathbf{l}, \mathbf{res}, \mathbf{left}, \mathbf{right}, \mathbf{left}', \mathbf{right}'$ range over sequences of integers, while i is a bound integer variable. The formula uses several operators that a useful sequence theory has to provide: (i) $\mathbf{l}_0$: the first element of input list $\mathbf{l}$; (ii) $\in$ and $\notin$: membership and non-membership of an integer in a list, which can be expressed using symbolic parametric automata; (iii) skip_1, $T_{<\mathbf{l}_0}$, $T_{\geq \mathbf{l}_0}$: sequence-to-sequence functions, which can be represented using symbolic parametric transducers; (iv) $\cdot\,.\,\cdot$: concatenation of several sequences. The formula otherwise is a direct model of the method in Listing 1; the variables $\mathbf{left}', \mathbf{right}'$ are the results of the recursive calls, and concatenated to obtain the result sequence.

In addition, the formula contains quantifiers. To demonstrate validity of the formula, it is enough to eliminate the last quantifier $\forall i$ by instantiating with a Skolem symbol k, and then instantiate the other quantifiers (left of the implication) with the same k:

$$\begin{pmatrix} \mathbf{left} = T_{<\mathbf{l}_0}(\mathbf{l}) \wedge \mathbf{right} = T_{\geq \mathbf{l}_0}(\mathit{skip}_1(\mathbf{l})) \wedge \\ (k \in \mathbf{left} \leftrightarrow k \in \mathbf{left}') \wedge (k \in \mathbf{right} \leftrightarrow k \in \mathbf{right}') \wedge \\ \mathbf{res} = \mathbf{left}' \,.\, [\mathbf{l}_0] \,.\, \mathbf{right}' \end{pmatrix} \rightarrow (k \in \mathbf{l} \leftrightarrow k \in \mathbf{res})$$

As one of the results of this paper, we prove that this final formula is in a decidable logic. The formula can be rewritten to a disjunction of straight-line formulas, and shown to be valid using the decision procedure presented in Sect. 5.

3 Models

In this section, we will define our sequence constraint language, and prove some basic results regarding various constraints in the language. The definition is a natural generalization of string constraints (e.g. see [12,17,26,29,35]) by employing an alphabet theory (a.k.a. element theory), as is done in symbolic automata and automata modulo theories [15,16,44].

For simplicity, our definitions will follow a model-theoretic approach. Let σ be a vocabulary. We fix a σ-structure $\mathfrak{S} = (D; I)$, where D can be a finite or an infinite set (i.e., the universe) and I maps each function/relation symbol in σ to a function/relation over D. The elements of our sequences will range over D. We assume that the quantifier-free theory $T_{\mathfrak{S}}$ over $\mathfrak{S}$ (including equality) is decidable. Examples of such $T_{\mathfrak{S}}$ are abound from SMT, e.g., LRA and LIA. We write T instead of $T_{\mathfrak{S}}$, when $\mathfrak{S}$ is clear. Our quantifier-free formula will use *uninterpreted T-constants* $a, b, c, \ldots$, and may also use variables $x, y, z, \ldots$. (The distinction between uninterpreted constants and variables is made only for the purpose of presentation of sequence constraints, as will be clear shortly.) We use $\mathcal{C}$ to denote the set of all uninterpreted T-constants. A formula φ is satisfiable if there is an assignment that maps the uninterpreted constants and variables to concrete values in D such that the formula becomes true in $\mathfrak{S}$.

Next, we define how we lift T to sequence constraints, using T as the *alphabet theory* (a.k.a. *element theory*). As in the case of strings (over a finite alphabet), we use standard notation like D^* to refer to the set of all sequences over D. By default, elements of D^* are written as standard in mathematics, e.g., $7, 8, 100$, when $D = \mathbb{Z}$. Sometimes we will disambiguate them by using brackets, e.g., $(7, 8, 100)$ or $[7, 8, 100]$. We will use the symbol s (with/without subscript) to refer to concrete sequences (i.e., a member of D^*). We will use $\mathbf{x}, \mathbf{y}, \mathbf{z}$ to refer to T-sequence variables. Let $\mathcal{V}$ denote the set of all T-sequence variables, and $\Gamma := \mathcal{C} \cup D$. We will define constraint languages syntactically at the beginning, and will instantiate them to specific sequence operations. The theory T^* of T-sequences consists of the following constraints:

$$\varphi ::= R(\mathbf{x}_1, \ldots, \mathbf{x}_r) \mid \varphi \wedge \varphi$$

where R is an r-ary relation symbol. In our definition of each atom R below, we will specify if an assignment μ, which maps each $\mathbf{x}_i$ to a T-sequence and each uninterpreted constant to a T-element, satisfies R. If μ satisfies all atoms, we say that μ is a *solution* and the *satisfiability problem* is to decide whether there is a solution for a given φ.

A few remarks about the missing boolean operators in the constraint language above are in order. Disjunctions can be handled easily using the DPLL(T) framework (e.g. see [32]), so we have kept our theory conjunctive. As in the case of strings, negations are usually handled separately because they can sometimes (but not in all cases) be eliminated while preserving decidability.

Equational Constraints. A *T-sequence equation* is of the form

$$L = R$$

where each of L and R is a concatenation of concrete T-elements, uninterpreted constants, and T-sequence variables. That is, if $\Theta := \Gamma \cup \mathcal{V}$, then $L, R \in \Theta^*$.

For example, in the equation

$$0.1.\mathbf{x} = \mathbf{x}.0.1$$

the set of all solutions is of the form $\mathbf{x} \mapsto (01)^*$. To make this more formal, we extend each assignment μ to a homomorphism on Θ^*. We write $\mu \models L = R$ if $\mu(L) = \mu(R)$. Notice that this definition is just direct extension of that of *word equations* (e.g. see [17]), i.e., when the domain D is finite.

In most cases the inequality constraints $L \neq R$ can be reduced to equality in our case this requires also element constraints, described below.

Element Constraints. We allow T-formulas to constrain the uninterpreted constants. More precisely, given a T-sentence (i.e., no free variables) φ that uses $\mathcal{C}$ as uninterpreted constants, we obtain a proposition P (i.e., 0-ary relation) that $\mu \models P$ iff $T \models_\mu \varphi$.

Negations in the equational constraints can be removed just like in the case of strings, i.e., by means of additional variables/constants and element constraints. For example, $\mathbf{x} \neq \mathbf{y}$ can be replaced by $(\mathbf{x} = \mathbf{z}a\mathbf{x}' \wedge \mathbf{y} = \mathbf{z}b\mathbf{y}' \wedge a \neq b) \vee \mathbf{x} = \mathbf{y}a\mathbf{z} \vee \mathbf{x}a\mathbf{z} = \mathbf{y}$. Notice that $a \neq b$ is a T-formula because we assume the equality symbol in T.

Regular Constraints. Over strings, regular constraints are simply unary constraints $U(\mathbf{x})$, where U is an automaton. The interpretation is $\mathbf{x}$ is in the language of U. We define an analogue of regular constraints over sequences using *parametric automata* [21,23,24], which generalize both symbolic automata [15,16] and variable automata [27].

A *parametric automaton* (PA) over T is of the form $\mathcal{A} = (\mathcal{X}, Q, \Delta, q_0, F)$, where $\mathcal{X}$ is a finite set of parameters, Q is a finite set of control states, $q_0 \in Q$ is the initial state, $F \subseteq Q$ is the set of final states, and $\Delta \subseteq_{\text{fin}} Q \times T(\mathit{curr}, \mathcal{X}) \times Q$. Here, *parameters* are simply uninterpreted T-constants, i.e., $\mathcal{X} \subseteq \mathcal{C}$. Formulas

that appear in transitions in Δ will be referred to as *guards*, since they restrict which transitions are enabled at a given state. Note that *curr* is an uninterpreted constant that refers to the "current" position in the sequence. The semantics is quite simply defined: a sequence $(d_1, d_2, \ldots, d_n)$ is in the language of $\mathcal{A}$ under the assignment of parameters μ, written as $(d_1, \ldots, d_n) \in L_\mu(\mathcal{A})$, when there is a sequence of Δ-transitions

$$(q_0, \varphi_1(curr, \mathcal{X}), q_1), (q_1, \varphi_2(curr, \mathcal{X}), q_2), \ldots, (q_{n-1}, \varphi_n(curr, \mathcal{X}), q_n),$$

such that $q_n \in F$ and $T \models \varphi_i(d_i, \mu(\mathcal{X}))$. Finally, for a regular constraint $\mathcal{A}(\mathbf{x})$ is satisfied by μ, when $\mu(\mathbf{x}) \in L_\mu(\mathcal{A})$.

Note, that it is possible to complement a PA $\mathcal{A}$, one has to be careful with the semantics: we treat $\mathcal{A}$ as a symbolic automaton, which are closed under boolean operations [15]. So we are looking for μ such that $\mu(\mathbf{x}) \in \overline{L_\mu(\mathbf{x})}$. What we cannot do using the complementation, is a universal quantification over the parameters; note that already theory of strings with universal and existential quantifiers is undecidable.

We state next a lemma showing that PAs using only "local" parameters, together with equational constraints, can encode the constraint language that we have defined so far.

Lemma 1. *Satisfiability of sequence constraints with equation, element, and regular constraints can be reduced in polynomial-time to satisfiability of sequence constraints with equation and regular constraints (i.e., without element constraints). Furthermore, it can be assumed that no two regular constraints share any parameter.*

Proposition 1. *Assume that T is solvable in* NP *(resp.* PSPACE*). Then, deciding nonemptiness of a parametric automaton over T is in* NP *(resp.* PSPACE*).*

The proof is standard (e.g. see [21,23,24]), and only sketched here. The algorithm first nondeterministically guesses a simple path in the automaton $\mathcal{A}$ from an initial state q_0 to some final state q_F. Let us say that the guards appearing in this path are $\psi_1(curr, \mathcal{X}), \ldots, \psi_k(curr, \mathcal{X})$. We need to check if this path is realizable by checking T-satisfiability of

$$\exists \mathcal{X}. \bigwedge_{i=1}^{k} \exists curr. (\psi_i(curr, \mathcal{X})).$$

It is easy to see that this is an NP (resp. NPSPACE = PSPACE) procedure.

Parametric Transducers. We define a suitable extension of symbolic transducers over parameters following the definition from Veanes et al. [44]. A *transducer constraint* is of the form $\mathbf{y} = \mathcal{T}(\mathbf{x})$, for a parametric transducer $\mathcal{T}$. A *parametric transducer* over T is of the form $\mathcal{T} = (\mathcal{X}, Q, \Delta, q_0, F)$, where $\mathcal{X}$, Q, q_0, F are just like in parametric automata. Unlike parametric automata, Δ is a finite set of tuples of the form $(p, (\varphi, \mathbf{w}), q)$, where (p, φ, q) is a standard transition in

parametric automaton, and $\mathbf{w}$ is a (possibly empty) sequence of T-terms over variable *curr* and constants $\mathcal{X}$, e.g., $\mathbf{w} = (curr+7, curr+2)$. One can think of $\mathbf{w}$ as the output produced by the transition. Given an assignment μ of parameters and the sequence variables, the constraint $\mathbf{y} = \mathcal{T}(\mathbf{x})$ is satisfied when there is a sequence of Δ-transitions

$$(q_0, \varphi_1(curr, \mathcal{X}), \mathbf{w}_1, q_1), (q_1, \varphi_2(curr, \mathcal{X}), \mathbf{w}_2, q_2), \ldots (q_{n-1}, \varphi_n(curr, \mathcal{X}), \mathbf{w}_n, q_n),$$

such that $q_n \in F$ and $T \models \varphi_i(d_i, \mu(\mathcal{X}))$, where $\mu(\mathbf{x}) = (d_1, \ldots, d_n)$, and finally

$$\mu(\mathbf{y}) = \mu_1(\mathbf{w}_1) \cdots \mu_n(\mathbf{w}_n)$$

where μ_i is μ but maps *curr* to d_i. The definition assumes that μ_i is extended to terms and concatenation thereof by homomorphism, e.g., in LRA, if $\mathbf{w}_1 = (curr + 7, curr + 2)$ and μ_1 maps *curr* to 10, then $\mathbf{w}_1$ will get mapped to $17, 12$. Given a set $S \subseteq D^*$ and an assignment μ (mapping the constants to D), we define the *pre-image* $\mathcal{T}_\mu^{-1}(S)$ of S under $\mathcal{T}$ with respect to μ as the set of sequences $\mathbf{w} \in D^*$ such that $\mathbf{w}' = \mathcal{T}(\mathbf{w})$ holds with respect to μ.

4 Solving Equational and Regular Constraints

Here we present results on solving equational constraints, together with regular constraints, by a reduction to the string case, for which a wealth of results are already available. In general, this reduction causes an exponential blow-up in the resulting string constraint, which we show to be unavoidable in general. That said, we also provide a more refined analysis in the case when the underlying theory is LRA, where we can avoid this exponential blow-up.

Prelude: The Case of Strings. We start with some known results about the case of strings. The satisfiability of word equations with regular constraints is PSpace-complete [18,19]. This upper bound can be extended to full quantifier-free theory [10]. When no regular constraints are given, the problem is only known to be NP-hard, and it is widely believed to be in NP. In the absence of regular constraints, without loss of generality Γ can be assumed to contain only letters from the equations; this is not the case in presence of regular constraints. The algorithm solving word equations [19] does not need an explicit access to Γ: it is enough to know whether there is a letter which labels a given set of transitions in the NFAs used in the regular constraints. In principle, there could be exponentially many different (i.e., inducing different transitions in the NFAs) letters. When oracle access to such alphabet is provided, the satisfiability can still be decided in PSpace: while not explicitly claimed, this is exactly the scenario in [19, Sect. 5.2]

Other constraints are also considered for word equations; perhaps the most widely known are the length constraints, which are of the form: $\sum_{x \in \mathcal{V}} a_x \cdot |x| \leq c$, where $\{a_x\}_{x \in \mathcal{V}}, c$ are integer constants and $|x|$ denotes the length $|\mu(x)|$, with an obvious semantics. It is an open problem, whether word equations with length constraints are decidable, see [26].

Reduction to Word Equations. We assume Lemma 1, i.e. that the parameters used for different automata-based constraints are pairwise different. In particular, when looking for a satisfying assignment μ we can first fix assignment for $\mathcal{X}$ and then try to extend it to $\mathcal{V}$. To avoid confusion, we call this partial assignment $\pi : \mathcal{X} \to D$.

Consider a set Φ of all atoms in all guards in the regular constraints together with the set of formulas $\{x = c\}$ over all constants $c \in D$ that appear in all equational constraints and the negations of both types of formulas. Fix an assignment $\pi : \mathcal{X} \to D$. The type $\text{type}_\pi(a)$ of a (under assignment π) is the set of formulas in Φ satisfied by a, i.e. $\{\varphi \in \Phi \ : \ \varphi(\pi(\mathcal{X}), a) \text{ holds}\}$. Clearly there are at most exponentially many different types (for a fixed π). A type t is realizable (for π) when $t = \text{type}_\pi(a)$ and it is realized by a.

If the constraints are satisfiable (for some parameters assignment π) then they are satisfiable over a subset $D_\pi \subseteq_{\text{fin}} D$, in the sense that we assign uniterpreted constants elements from D_π and T-sequence variables elements of D_π^*, where D_π is created by taking (arbitrarily) one element of a realizable type. Note that for each constant c in the equational constraints there is a formula "$x = c$" in Φ, in particular $\text{type}_\pi(c)$ is realizable (only by c) and so $c \in D_\pi$.

Lemma 2. *Given a system of constraints and a parameter assignment π let $D_\pi \subseteq D$ be obtained by choosing (arbitrarily) for each realizable type a single element of this type. Then the set of constraints is satisfiable (for π) over D if and only if they are satisfiable (for π) over D_π. To be more precise, there is a letter-to-letter homomorphism $\psi : D^* \to D_\pi^*$ such that if μ is a solution of a system of constraints then $\psi \circ \mu$ is also a solution.*

The proof can be found in the full version, its intuition is clear: we map each letter $a \in D$ to the unique letter in D_π of the same type.

Once the assignment is fixed (to π) and domain restricted to a finite set (D_π), the equational and regular constraints reduce to word equations with regular constraints: treat D_π as a finite alphabet, for a parametric automaton $\mathcal{A} = (\mathcal{X}, Q, \Delta, q_0, F)$ create an NFA $\mathcal{A}' = (D_\pi, Q, \Delta', q_0, F)$, i.e. over the alphabet D_π, with the same set of states Q, same starting state q_0 and accepting states F and the relation defined as $(q, a, q') \in \Delta'$ if and only if there is $(q, \varphi(curr, \mathcal{X}), q') \in \Delta$ such that $\varphi(a, \pi(\mathcal{X}))$ holds, i.e. we can move from q to q' by a in $\mathcal{A}'$ if and only if we can make this move in $\mathcal{A}$ under assignment π. Clearly, from the construction

Lemma 3. *Given an assignment of parameters π let D_μ be a set from Lemma 2, $\mathcal{A}$ be a parametric automaton and $\mathcal{A}'$ the automaton as constructed above. Then*

$$L_\pi(\mathcal{A}) \cap D_\pi^* = L(\mathcal{A}') \ .$$

We can rewrite the parametric automata-constraints with regular constraints and treat equational constraints as word equations (over the finite alphabet D_π). From Lemma 2 and Lemma 3 it follows that the original constraints have a solution for assignment π if and only if the constructed system of constraints has a solution. Therefore once the appropriate assignment π is fixed, the validity

of constraints can be verified [19]. It turns out that we do not need the actual π, it is enough to know which types are realisable for it, which translates to an exponential-size formula. We will use letter τ to denote subset of Φ; the idea is that $\tau = \{\text{type}_\pi(a) \ : \ a \in D\} \subseteq 2^\Phi$ and if different π, π' give the same sets of realizable types, then they both yield a satisfying assignment or both not. Hence it is enough to focus on τ and not on actual π.

Lemma 4. *Given a system of equational and regular constraints we can non-deterministically reduce them to a formula of a form*

$$\exists_{t \in \tau} a_t \in D. \exists \mathcal{X} \in D^+. \bigwedge_{t \in \tau} \bigwedge_{\varphi \in t} \varphi(\mathcal{X}, a_t) \ , \tag{1}$$

where $\tau \subseteq 2^\Phi$ is of at most exponential size, and a system of word equations with regular constraints of linear size and over an $|\tau|$-size alphabet, using auxiliary $\mathcal{O}(n|\tau|)$ space. The solution of the latter word equations (for which also (1) holds) are solutions of the original system, by appropriate identifications of symbols.

Proof. We guess the set τ of types of the assignment of parameters π, i.e. $\tau = \{\text{type}_\pi(a) \ : \ a \in D\}$ such that there is an assignment μ extending π; note that as Φ has linearly many atoms and $\tau \subseteq 2^\Phi$, then $|\tau|$ may be of exponential size, in general. The (1) verifies the guess: we validate whether there are values of $\mathcal{X}$ such that for each type $t \in \tau$ there is a value a such that $\text{type}_\pi(a) = t$.

Let D_π be a set having one symbol per every type in τ, as in Lemma 2; note that this includes all constants in the equational constraints. The algorithm will not have access to particular values, instead we store each $t \in \tau$, say as a bitvector describing which atoms in Φ this letter satisfies. In particular, $|D_\pi| = |\tau|$ and it is at most exponential. In the following we will consider only solutions over D_π.

For each $a \in D_\pi$ we can validate, which transitions in $\mathcal{A}$ it can take: the transition is labelled by a guard which is a conjunction of atoms from Φ and either each such atom is in $\text{type}_\pi(a)$ or not. Hence we can treat $\mathcal{A}$ as an NFA for D_π. We do not need to construct nor store it, we can use $\mathcal{A}$: when we want to make a transition by $\varphi(\mathcal{X}, a)$ we look up, whether each atom of φ is in $\text{type}_\pi(a)$ or not. Similarly, the constraint $\mathcal{A}(\mathbf{x})$ is restricted to $\mathbf{x} \in L_\pi(\mathcal{A})$ and for $\mathbf{x} \in D_\pi^*$ this is a usual regular constraint.

We treat equational constraints as word equations over alphabet D_π.

Concerning the correctness of the reduction: if the system of word equations (with regular constraints) is satisfiable and the formula (1) is also satisfiable, then there is a satisfying assignment μ over D_π and D_π^* in particular, there is an assignment of parameters for which there are letters of the given types (note that in principle it could be that μ induces more types, i.e. there is a value a such that $\text{type}_\mu(a) \notin \tau$ and so it is not represented in D_π, but this is fine: enlarging the alphabet cannot invalidate a solution), i.e. the transitions for a_t in the automata after the reduction are the same as in the corresponding parametric automata for the assignment π, this is guaranteed by the satisfiability of (1) and the way we construct the instance, see Lemma 3.

On the other hand, when there is a solution of the input constraints, there is one for some assignment of parameters π. Hence, by Lemma 2, there is a solution over D_π. The algorithm guesses $\tau = \{\text{type}_\pi(a) : a \in D\}$ and (1) is true for it. Then by Lemma 2 there is a solution over D_π as constructed in the reduction and by Lemma 3 the regular constraints define the same subsets of D_π^* both when interpreted as parametric automata and NFAs. □

Theorem 1. *If theory T is in* PSpace *then sequence constraints are in* ExpSpace.

If τ is polynomial size and the formula (1) *can be verified in* PSpace, *then sequence constraints can be verified in* PSpace.

One of the difficulties in deciding sequence constraints using the word equations approach is the size of set of realizable types τ, which could be exponential. For some concrete theories it is known to be smaller and thus a lower upper bound on complexity follows. For instance, it is easy to show that for LRA there are linearly many realizable types, which implies a PSpace upper bound.

Corollary 1. *Sequence constraints for Linear Real Arithmetic are in* PSpace.

In general, the ExpSpace upper bound from Theorem 1 cannot be improved, as even non-emptiness of intersection of parametric automata is ExpSpace-complete for some theories decidable in PSpace. This is in contrast to the case of symbolic automata, for which the non-emptiness of intersection (for a theory T decidable in PSpace) is in PSpace. This shows the importance of parameters in our lower bound proof.

Theorem 2. *There are theories with existential fragment decidable in* PSpace *and whose non-emptiness of intersection of parametric automata is* ExpSpace-*complete.*

When no regular constraints are allowed, we can solve the equational and element constraints in PSpace (note that we do not use Lemma 1).

Theorem 3. *For a theory T decidable in* PSpace, *the element and equational constraints (so no regular constraints) can be decided in* PSpace.

5 Algorithm for Straight-Line Formulas

It is known that adding finite transducers into word equations results in an undecidable model (e.g. see [35]). Therefore, we extend the *straight-line restriction* [12,35] to sequences, and show that it suffices to recover decidability for equational constraints, together with regular and transducer constraints. In fact, we will show that deciding problems in the straight-line fragment is solvable in doubly exponential time and is ExpSpace-hard, if T is solvable in PSpace. It has been observed that the straight-line fragment for the theory of strings already covers many interesting benchmarks [12,35], and similarly many properties of sequence-manipulating programs can be proven using the fragment, including the QuickSort example from Sect. 2 and other benchmarks shown in Sect. 7.

The Straight-Line Fragment SL. We start by defining recognizable formulas over sequences, followed by the syntactic and semantic restrictions on our constraint language. This definition follows closely the definition of recognizable relations over finite alphabets, except that we replace finite automata with parametric automata.

Definition 1 (Recognizable formula). *A formula* $R(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_r)$ *is* recognizable *if it is equivalent to a positive Boolean combination of regular constraints.*

Note that this is simply a generalization of regular constraints to multiple variables, i.e., 1-ary recognizable formula can be turned into a regular constraint, which is closed under intersection and union.

To define the straight-line fragment, we use the approach of [12]; that is, the fragment is defined in terms of "feasibility of a symbolic execution". Here, a *symbolic execution* is just a sequence of assignments and assertions, whereas the *feasibility* problem amounts to deciding whether there are concrete values of the variables so that the symbolic execution can be run and none of the assertions are violated. We now make this intuition formal. A symbolic execution is syntactically generated by the following grammar:

$$S \quad ::= \quad \mathbf{y} := f(\mathbf{x}_1, \ldots, \mathbf{x}_k, \mathcal{X}) \mid \mathbf{assert}(R(\mathbf{x}_1, \ldots, \mathbf{x}_r)) \mid \mathbf{assert}(\varphi) \mid S; S \qquad (2)$$

where $f : (D^*)^k \times D^{|\mathcal{X}|} \to D$ is a function, R are recognizable formulas, and φ are element constraints.

The symbolic execution S can be turned into a sequence constraint as follows. Firstly, we can turn S into the standard *Static Single Assignment (SSA)* form by means of introducing new variables on the left-hand-side of an assignment. For example, $\mathbf{y} := f(\mathbf{x}); \mathbf{y} := g(\mathbf{z})$ becomes $\mathbf{y} := f(\mathbf{x}_1); \mathbf{y}' := g(\mathbf{z})$. Then, in the resulting constraint, each variable appears *at most once* on the left-hand-side of an assignment. That way, we can simply replace each assignment symbol $:=$ with an equality symbol $=$. We then treat each sequential composition as the conjunction operator $\wedge$ and assertion as a conjunct. Note that individual assertions are already sequence constraints. Next, we define how an interpretation μ satisfies the constraint $\mathbf{y} = f(\mathbf{x}_1, \ldots, \mathbf{x}_r, \mathcal{X})$:

$$\mu \models \mathbf{y} = f(\mathbf{x}_1, \ldots, \mathbf{x}_r, \mathcal{X}) \quad \text{iff} \quad \mu(\mathbf{y}) = f(\mu(\mathbf{x}_1), \ldots, \mu(\mathbf{x}_r), \mu(\mathcal{X})).$$

Note that '=' on the l.h.s. is syntactic, while the '=' on the r.h.s. is in the metalanguage. The definition of the semantics of the language is now inherited from Sect. 3.

In addition to the syntactic restrictions, we also need a semantic condition: in our language, we only permit functions f such that the pre-image of each regular constraint under f is effectively a recognizable formula:

(RegInvRel) A function f is permitted if for each regular constraint $\mathcal{A}(\mathbf{y})$, it is possible to compute a recognizable formula that is equivalent to the formula $\exists \mathbf{y} : \mathcal{A}(\mathbf{y}) \wedge \mathbf{y} = f(\mathbf{x}_1, \ldots, \mathbf{x}_r, \mathcal{X})$.

Two functions satisfying (RegInvRel) are the concatenation function $\mathbf{x} := \mathbf{y}.\mathbf{z}$ (here $\mathbf{y}$ could be the same as $\mathbf{z}$) and parametric transducers $\mathbf{y} := \mathcal{T}(\mathbf{x})$. We will only use these two functions in the paper, but the result is generalizable to other functions.

Proposition 2. *Given a regular constraint $\mathcal{A}(\boldsymbol{y})$ and a constraint $\boldsymbol{y} = \boldsymbol{x}.\boldsymbol{z}$, we can compute a recognizable formula $\psi(\boldsymbol{x}, \boldsymbol{z})$ equivalent to $\exists \boldsymbol{y} : \mathcal{A}(\boldsymbol{y}) \wedge \boldsymbol{y} = \boldsymbol{x}.\boldsymbol{z}$. Furthermore, this can be achieved in polynomial time.*

The proof of this proposition is exactly the same as in the case of strings, e.g., see [12,35].

Proposition 3. *Given a regular constraint $\mathcal{A}(\boldsymbol{y})$ and a parametric transducer constraint $\boldsymbol{y} = \mathcal{T}(\boldsymbol{x})$, we can compute a regular constraint $\mathcal{A}'(\boldsymbol{x})$ that is equivalent to $\exists \boldsymbol{y} : \mathcal{A}(\boldsymbol{y}) \wedge \boldsymbol{y} = \mathcal{T}(\boldsymbol{x})$. This can be achieved in exponential time.*

The construction in Proposition 3 is essentially the same as the pre-image computation of a symbolic automaton under a symbolic transducer [44]. The complexity is exponential in the maximum number of output symbols of a single transition (i.e. the maximum length of $\mathbf{w}$ in the transducer), which is in practice a small natural number.

The following is our main theorem on the SL fragment with equational constraints, regular constraints, and transducers.

Theorem 4. *If T is solvable in* PSpace*, then the SL fragment with concatenation and parametric transducers over T is in 2*-ExpTime *and is* ExpSpace*-hard.*

Proof. We give a decision procedure. We assume that S is already in SSA (i.e. each variable appears at most once on the left-hand side). Let us assume that S is of the form $S'; \mathbf{y} := f(\mathbf{x}_1, ...\mathbf{x}_r)$, for some symbolic execution S'. Without loss of generality, we may assume that each recognizable constraint is of the form $\mathcal{A}(\mathbf{x})$. This is no limitation: (1) since each R in the assertion is a recognizable formula, we simply have to "guess" one of the implicants for each R, and (2) $\mathbf{assert}(\psi_1 \wedge \psi_2)$ is equivalent to $\mathbf{assert}(\psi_1); \mathbf{assert}(\psi_2)$.

Assume now that $\{\mathcal{A}_1(\mathbf{y}), \ldots, \mathcal{A}_m(\mathbf{y})\}$ are all the regular constraints on $\mathbf{y}$ in S. By our assumption, it is possible to compute a recognizable formula equivalent to

$$\psi(\mathbf{x}_1, \ldots, \mathbf{x}_r) := \exists \mathbf{y} : \bigwedge_{i=1}^{m} \mathcal{A}_i(\mathbf{y}) \wedge \mathbf{y} = f(\mathbf{x}_1, \ldots, \mathbf{x}_r).$$

There are two ways to see this. The first way is that regular constraints are closed under intersection. This is in general computationally quite expensive because of a product automata construction before applying the pre-image computation. A better way to do this is to observe that ψ is equivalent to the conjunction of ψ_i's over $i = 1, \ldots, m$, where

$$\psi_i := \exists \mathbf{y} : \mathcal{A}_i(\mathbf{y}) \wedge \mathbf{y} = f(\mathbf{x}_1, \ldots, \mathbf{x}_r).$$

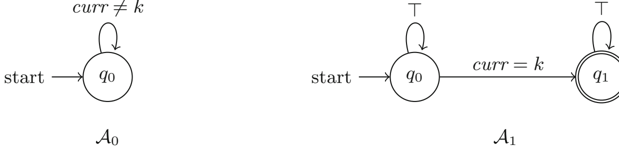

Fig. 1. $\mathcal{A}_0$ accepts all words not containing k and $\mathcal{A}_1$ accepts all words containing k.

By our semantic condition, we can compute recognizable formulas $\psi'_i, \ldots, \psi'_m$ equivalent to $\psi_1, \ldots, \psi_m$ respectively. Therefore, we simply replace S by

$$S'; \mathbf{assert}(\psi'_1); \cdots ; \mathbf{assert}(\psi'_m),$$

in which every occurrence of $\mathbf{y}$ has been completely eliminated. Applying the above variable elimination iteratively, we obtain a conjunction of regular constraints. We now end up with a conjunction of regular constraints and element constraints, which as we saw from Sect. 4 is decidable. □

Example 1. We consider the example from Sect. 2 where a weaker form of the permutation property is shown for QuickSort. The formula that has to be proven is a disjunction of straight-line formulas and in the following we execute our procedure only on one disjunct without redundant formulas:

$$\mathbf{assert}(\mathcal{A}_0(\mathbf{left'})); \mathbf{assert}(\mathcal{A}_0(\mathbf{right'})); \mathbf{res} = \mathbf{left'} \,.\, [\mathbf{l}_0] \,.\, \mathbf{right'}; \mathbf{assert}(\mathcal{A}_1(\mathbf{res}))$$

We model $L(\mathcal{A}_1)$ as the language which accepts all words which contain one letter equal to k and $L(\mathcal{A}_0)$ as the language which accepts only words not containing k, where k is an uninterpreted constant, so a single element. See Fig. 1. We begin by removing the operation $\mathbf{res} = \mathbf{left'} \,.\, [\mathbf{l}_0] \,.\, \mathbf{right'}$. The product automaton for all assertions that contain $\mathbf{res}$ is just $\mathcal{A}_1$. Hence, we can remove the assertion $\mathbf{assert}(\mathcal{A}_1(\mathbf{res}))$. The concatenation function . satisfies **RegInvRel** and the pre-image g can be represented by

$$\bigvee_{0 \le i,j \le 1} \mathcal{A}_1^{q_0,\{q_i\}}(\mathbf{left'}) \wedge \mathcal{A}_1^{q_i,\{q_j\}}([\mathbf{l}_0]) \wedge \mathcal{A}_1^{q_j,\{q_1\}}(\mathbf{right'}),$$

where $\mathcal{A}_i^{p,F'}$ is $\mathcal{A}_i$ with start state set to p and finals to F'.

In the next step, the assertion g is added to the program and all assertions containing $\mathbf{res}$ and the concatenation function are removed.

$$\mathbf{assert}(\mathcal{A}_0(\mathbf{left'})); \mathbf{assert}(\mathcal{A}(\mathbf{right'})); \mathbf{assert}(g(\mathbf{left'}, [\mathbf{l}_0], \mathbf{right'}))$$

From here, we pick a tuple from g, lets say $i = j = 1$, and obtain

$$\mathbf{assert}(\mathcal{A}_0(\mathbf{left'})); \mathbf{assert}(\mathcal{A}_0(\mathbf{right'})); \mathbf{assert}(\mathbf{left'} \in \mathcal{A}_1^{q_0,\{q_1\}});$$
$$\mathbf{assert}([\mathbf{l}_0] \in \mathcal{A}_1^{q_1,\{q_1\}}); \mathbf{assert}(\mathbf{right'} \in \mathcal{A}_1^{q_1,\{q_1\}})$$

Finally, the product automata $\mathcal{A}_0 \times \mathcal{A}_1^{q_0,\{q_1\}}$ and $\mathcal{A}_0 \times \mathcal{A}_1^{q_0,\{q_1\}}$ are computed for the variables $\mathbf{left}', \mathbf{right}'$ and a non-emptiness check over the product automata and the automaton for $[\mathbf{l}_0]$ is done. The procedure will find no combination of paths for each automaton which can be satisfied, since $\mathbf{left}'$ is forced to accept no words containing k by $\mathcal{A}_0$ and only accepts by reading a k from $\mathcal{A}_1^{q_0,\{q_1\}}$. Next, the procedure needs to exhaust all tuples from $(\mathcal{A}_1^{q_0,\{q_i\}}, \mathcal{A}_1^{q_i,\{q_j\}}, \mathcal{A}_1^{q_j,\{q_1\}})_{0 \le i,j \le 1}$ before it is proven that this disjunct is unsatisfiable.

6 Extensions and Undecidability

Length Constraints. We consider the extension of our model by allowing *length-constraints* on the sequence variables: for each sequence variable $\mathbf{x}$ we consider the associated length variable $\ell_{\mathbf{x}}$, let the set of length variables be $\mathcal{L} = \{\ell_{\mathbf{x}} : \mathbf{x} \in \mathcal{V}\}$, we extend μ to $\mathcal{L}$, it assigns natural numbers to them. The length constraints are of the form $\sum_{\mathbf{x}} a_{\mathbf{x}} \ell_{\mathbf{x}} ?0$, where $? \in \{<, \le, =, \ne, \ge, >\}$ and each $a_{\mathbf{x}}$ is an integer constant, i.e., linear arithmetic formulas on the length-variables. The semantics is natural: we require that $|\mu(\mathbf{x})| = \mu(\ell_{\mathbf{x}})$ (the assigned values are the true lengths of sequences) and that $\mu(\mathcal{L})$ satisfies each length constraint.

There is, however, another possible extensions: if we the theory $T_{\mathfrak{S}}$ is the Presburger arithmetic, then the parameter automata could use the values $\ell_{\mathbf{x}}$. We first deal with a more generic, though restricted case, when this is not allowed: then all reductions from Sect. 4 generalize and we can reduce to the word equations with regular and length constraints. However, the decidability status of this problem is unknown. When we consider Presburger arithmetic and allow the automata to employ the length variables, then it turns out that we can interpret the formula (1) as a collection of length constraints, and again we reduce to word equations with regular and length constraints.

Automata Oblivious of Lengths. We first consider the setting, in which the length variables $\mathcal{L}$ can only be used in length constraints. It is routine to verify that the reduction from Sect. 4 generalize to the case of length constraints: it is possible to first fix μ for parameters, calling it again π. Then Lemma 2 shows that each solution μ can be mapped by a letter-to-letter homomorphism to a finite alphabet D_π, and this mapping preserves the satisfiability/unsatisfiability of length constraints, so Lemma 2 still holds when also length constraints are allowed. Similarly, Lemma 3 is also not affected by the length constraints and finally Lemma 4 deals with regular and equational constraints, ignoring the other possible constraints and the length of substitutions for variables are the same. Hence it holds also when the length constraints are allowed then the resulting word equations use regular and length constraints.

Unfortunately, the decidability of word equations with linear length constraints (even without regular constraints) is a notorious open problem. Thus instead of decidability, we get Turing-equivalent problems.

Theorem 5. *Deciding regular, equational and length constraints for T-sequences of a decidable theory T is Turing-equivalent to word equations with regular and length constraints.*

Automata Aware of the Sequence Lengths. We now consider the case when the underlying theory $T_{\mathfrak{S}}$ is the Presburger arithmetic, i.e. $\mathfrak{S}$ is the natural numbers and we can use addition, constants $0, 1$ and comparisons (and variables). The additional functionality of the parametric automaton $\mathcal{A}$ is that $\Delta \subseteq_{\text{fin}} Q \times T(curr, \mathcal{X}, \mathcal{L}) \times Q$, i.e. the guards can also use the length variables; the semantics is extended in the natural way.

Then the type $\text{type}_\pi(a)$ of $a \in \mathbb{N}$ now depends on μ values on $\mathcal{X}$ and $\mathcal{L}$, hence we denote by π the restriction of μ to $\mathcal{X} \cup \mathcal{L}$. Then Lemma 2, 3 still hold, when we fix π. Similarly, Lemma 4 holds, but the analogue of (1) now uses also the length variables, which are also used in the length constraints. Such a formula can be seen as a collection of length constraints for original length variables $\mathcal{L}$ as well as length variables $\mathcal{X} \cup \{a_t \,:\, t \in \tau\}$. Hence we validate this formula as part of the word equations with length constraints. Note that a_t has two roles: as a letter in D_π and as a length variable. However, the connection is encoded in the formula from the reduction (analogue of (1)) and we can use two different sets of symbols.

Theorem 6. *Deciding conjunction of regular, equational and length constraints for sequences of natural numbers with Presburger arithmetic, where the regular constraints can use length variables, is Turing-equivalent to word equations with regular and (up to exponentially many) length constraints.*

Undecidability of Register Automata Constraints. One could use more powerful automata for regular constraints; one such popular model are register automata; informally, such automaton has k registers $r_1, \ldots, r_k$ and its transition depends on state and a value of formula using the registers and *curr*: the read value [23]; note that the registers can be updated: to *curr* or to one of register's values; this is specified in the transition. In "classic" register automata guards can only use equality and inequality between registers and *curr*; in SRA model more powerful atoms are allowed. We show that sequence constraints and register automata constraints (which use quantifier-free formulas with equality and inequality as only atoms, i.e. do not employ the SRA extension) lead to undecidability (over infinite domain D).

Theorem 7. *Satisfiability of equational constraints and register automata constraints, which use equality and inequality only, over infinite domain, is undecidable.*

7 Implementations, Optimizations and Benchmarks

Implementation. We have implemented our decision procedure for problems in the constraint language SL for the theory of sequences in a new tool SeCo

(Sequence Constraint Solver) on top of the SMT solver Princess [41]. We extend a publicly available library for symbolic automata and transducers [13] to parametric automata and transducers by connecting them to the uninterpreted constants in our theory of sequences. Our tool supports symbolic transducers, concatenation of sequences and reversing of sequences. Any additional function which satisfies **RegInvRel** such as a replace function which replaces only the first and leftmost longest match can be added in the future.

Our algorithm is an adaption of the tool OSTRICH [12] and closely follows the proof of Theorem 4. To summarize the procedure, a depth-first search is employed to remove all functions in the given input and splitting on the pre-images of those functions. When removing a function, new assertions are added to the pre-image constraints. After all functions have been removed and only assertions are left a nonemptiness check is called over all parametric automata which encoded the assertions. If the check is successful a corresponding model can be constructed, otherwise the procedure computes a conflict set and back-jumps to the last split in the depth search.[2]

Benchmarks. We have performed experiments on two benchmark suites. The first one concerns itself with the verification of properties for programs manipulating sequences. The second benchmark suite compares our tool against an algorithm using symbolic register automata [13] on decision procedures of regular expressions with back-references such as emptiness, equivalence and inclusion.

Both benchmark suites require universal quantification over the parameters; there are existing methods for eliminating these universal quantifiers, one such class are the *semantically deterministic* (SD) [22] PAs; despite its name, being SD is algorithmically checkable. Most of considered the PAs are SD, in particular all in benchmark suite 2.

Experiments were conducted on an AMD Ryzen 5 1600 Six-Core CPU with 16 GB of RAM running on Windows 10. The results for second benchmark suite is shown Table 1. The timeout for all benchmarks is 300 s.

In the first benchmarks suite we are looking to verify a weaker form of the permutation property of sorting as shown in Sect. 2. Furthermore, we verify properties of two self-stabilizing algorithms for mutual exclusion on parameterized systems. The first one is Lamport's bakery algorithm [33], for which we proved that the algorithm ensures mutual exclusion. The system is modelled in the style of regular model checking [8], with system states represented as words, here over an infinite alphabet: the character representing a thread stores the thread control state, a Boolean flag, and an integer as the number drawn by the thread. The system transitions are modelled as parametric transducers, and invariants as parametric automata. The second algorithm is known as Dijkstra's Self-Stabilizing Protocol [20], in which system states are encoded as sequences of integers, and in which we verify that the set of states in which exactly one processor is privileged forms an invariant. The mentioned benchmarks require

[2] For a more detailed write-up of the depth-first search algorithm see OSTRICH [12] Algorithm 1.

Table 1. Benchmark suite 2. *SRA* is used for the algorithm for symbolic register automata and *SEQ* for our tool. The symbol $\emptyset$ indicates the column where emptiness was checked, $\equiv$ indicates self equivalence and $\subseteq$ inclusion of languages.

$\mathcal{L}_1$	$\mathcal{L}_2$	$SRA_\emptyset(\mathcal{L}_1)$	SeCo$_\emptyset(\mathcal{L}_1)$	$SRA_\equiv(\mathcal{L}_1)$	SeCo$_\equiv(\mathcal{L}_1)$	$SRA_\subseteq(\mathcal{L}_2, \mathcal{L}_1)$	SeCo$_\subseteq(\mathcal{L}_2, \mathcal{L}_1)$
Pr-C2	Pr-CL2	0.03 s	0.65 s	0.43 s	0.10 s	4.7 s	0.10 s
Pr-C3	Pr-CL3	0.58 s	0.70 s	10.73 s	0.12 s	36.90 s	0.10 s
Pr-C4	Pr-CL4	18.40 s	0.77 s	98.38 s	0.14 s	–	0.10 s
Pr-C6	Pr-CL6	–	1.00 s	–	0.12 s	–	0.10 s
Pr-CL2	Pr-C2	0.33 s	0.30 s	1.03 s	0.13 s	0.52 s	0.76 s
Pr-CL3	Pr-C3	14.04 s	0.38 s	20.44 s	0.13 s	10.52 s	0.76 s
Pr-CL4	Pr-C4	–	0.41 s	0.43 s	0.12 s	–	0.82 s
Pr-CL6	Pr-C6	–	0.62 s	0.43 s	0.12 s	–	1.27 s
IP-2	IP-3	0.11 s	1.53 s	0.63 s	0.14 s	2.43 s	0.15 s
IP-3	IP-4	1.83 s	1.45 s	4.66 s	0.14 s	28.60 s	0.17 s
IP-4	IP-6	30.33 s	1.75 s	80.03 s	0.14 s	–	0.17 s
IP-6	IP-9	–	1.60 s	0.43 s	0.13 s	–	0.17 s

universal quantification, but similar to the motivating example from Sect. 2 one can eliminate quantifiers by Skolemization and instantiation which was done by hand.

The second benchmark suite consists of three different types of benchmarks, summarized in Table 1. The benchmark PR-Cn describes a regular expression for matching products which have the same code number of length n, and PR-CLn matches not only the code number but also the lot number. The last type of benchmark is IP-n, which matches n positions of 2 IP addresses. The benchmarks are taken from the regular-expression crowd-sourcing website RegExLib [39] and are also used in experiments for symbolic register automata [14] which we also compare our results against. To apply our decision procedure to the benchmarks, we encode each of the benchmarks as a parametric automaton, using parameters for the (bounded-size) back-references. The task in the experiments is to check emptiness, language equivalence, and language inclusion for the same combinations of the benchmarks as considered in [14].

Results of the Experiments. All properties can be encoded by parametric automata with very few states and parameters. As a result the properties for each program can be verified in < 2.6 s, in detail the property for Dijkstra's algorithm was proven in 0.6 s, QuickSort in 1.1 s and Lamport's bakery algorithm in 2.5 s.

The results for the second benchmark suite are shown in Table 1. The algorithm for symbolic register automata times out on 11 of the 36 benchmarks and our tool solves most benchmarks in $<$1 s. One thing to observe that the symbolic register automata scales poorly when more registers are needed to capture the back-references while the performance of our approach does not change noticeably when more parameters are introduced.

8 Conclusion and Future Work

In this paper, we have performed a systematic investigation of decidability and complexity of constraints on sequences. Our starting point is the subcase of string constraints (i.e. over a finite set of sequence elements), which include equational constraints with concatenation, regular constraints, length constraints, and transducers. We have identified parametric automata (extending symbolic automata and variable automata) as suitable notion of "regular constraints" over sequences, and parametric transducers (extending symbolic transducers) as suitable notion of transducers over sequences. We showed that decidability results in the case of strings carry over to sequences, although the complexity is in general higher than in the case of strings (sometimes exponentially higher). For certain element theory (e.g. Linear Real Arithmetic), it is possible to retain the same complexity as in the string case. We also delineate the boundary of the suitable notion of "regular constraints" by showing that the equational constraints with symbolic register automata [14] yields undecidable satisfiability. Finally, our new sequence solver SeCo shows promising experimental results.

There are several future research avenues. Firstly, the complexity of sequence constraints over other specific element theories (e.g. Linear Integer Arithmetic) should be precisely determined. Secondly, is it possible to recover decidability with other fragments of register automata (e.g., single-use automata [7])? On the implementation side, there are some algorithmic improvements, e.g., better nonemptiness checks for parametric automata in the case of a single automaton, as well as product of multiple automata.

Acknowledgment. We thank anonymous reviewers for their thorough and helpful feedback. We are grateful to Nikolaj Bjørner, Rupak Majumdar and Margus Veanes for the inspiring discussion.

References

1. Abdulla, P.A., et al.: String constraints for verification. In: Biere, A., Bloem, R. (eds.) CAV 2014. LNCS, vol. 8559, pp. 150–166. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-08867-9_10
2. Abdulla, P.A., Atig, M.F., Diep, B.P., Holík, L., Janků, P.: Chain-free string constraints. In: Chen, Y.-F., Cheng, C.-H., Esparza, J. (eds.) ATVA 2019. LNCS, vol. 11781, pp. 277–293. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-31784-3_16
3. Amadini, R.: A survey on string constraint solving. ACM Comput. Surv. **55**(2), 16:1-16:38 (2023). https://doi.org/10.1145/3484198
4. Barbosa, H., et al.: cvc5: a versatile and industrial-strength SMT solver. In: TACAS 2022. LNCS, vol. 13243, pp. 415–442. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-99524-9_24
5. Barceló, P., Figueira, D., Libkin, L.: Graph logics with rational relations. Log. Methods Comput. Sci. **9**(3) (2013). https://doi.org/10.2168/LMCS-9(3:1)2013
6. Bjørner, N., de Moura, L., Nachmanson, L., Wintersteiger, C.M.: Programming Z3. In: Bowen, J.P., Liu, Z., Zhang, Z. (eds.) SETSS 2018. LNCS, vol. 11430, pp. 148–201. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-17601-3_4

7. Bojanczyk, M., Stefanski, R.: Single-use automata and transducers for infinite alphabets. In: Czumaj, A., Dawar, A., Merelli, E. (eds.) 47th International Colloquium on Automata, Languages, and Programming, ICALP 2020, July 8–11, 2020, Saarbrücken, Germany (Virtual Conference). LIPIcs, vol. 168, pp. 113:1–113:14. Schloss Dagstuhl - Leibniz-Zentrum für Informatik (2020). https://doi.org/10.4230/LIPIcs.ICALP.2020.113
8. Bouajjani, A., Jonsson, B., Nilsson, M., Touili, T.: Regular Model Checking. In: Emerson, E.A., Sistla, A.P. (eds.) CAV 2000. LNCS, vol. 1855, pp. 403–418. Springer, Heidelberg (2000). https://doi.org/10.1007/10722167_31
9. Bradley, A.R., Manna, Z., Sipma, H.B.: What's decidable about arrays? In: Emerson, E.A., Namjoshi, K.S. (eds.) VMCAI 2006. LNCS, vol. 3855, pp. 427–442. Springer, Heidelberg (2005). https://doi.org/10.1007/11609773_28
10. Büchi, J.R., Senger, S.: Definability in the existential theory of concatenation and undecidable extensions of this theory. In: The Collected Works of J. Richard Büchi, pp. 671–683. Springer, New York (1990). https://doi.org/10.1007/978-1-4613-8928-6_37
11. Chen, T., et al.: Solving string constraints with regex-dependent functions through transducers with priorities and variables. Proc. ACM Program. Lang. **6**(POPL), 1–31 (2022). https://doi.org/10.1145/3498707
12. Chen, T., Hague, M., Lin, A.W., Rümmer, P., Wu, Z.: Decision procedures for path feasibility of string-manipulating programs with complex operations. Proc. ACM Program. Lang. **3**(POPL), 49:1–49:30 (2019). https://doi.org/10.1145/3290362
13. D'Antoni, L.: SVPAlib. Symbolic Automata Library (2018). https://github.com/lorisdanto/symbolicautomata. Accessed 2 Feb 2023
14. D'Antoni, L., Ferreira, T., Sammartino, M., Silva, A.: Symbolic register automata. In: Dillig, I., Tasiran, S. (eds.) CAV. vol. 11561, pp. 3–21. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-25540-4_1
15. D'Antoni, L., Veanes, M.: The power of symbolic automata and transducers. In: Majumdar, R., Kunčak, V. (eds.) CAV 2017. LNCS, vol. 10426, pp. 47–67. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-63387-9_3
16. D'Antoni, L., Veanes, M.: Automata modulo theories. Commun. ACM **64**(5), 86–95 (2021). https://doi.org/10.1145/3419404
17. Diekert, V.: Makanin's algorithm. In: Lothaire, M. (ed.) Algebraic Combinatorics on Words, Encyclopedia of Mathematics and its Applications, vol. 90, chap. 12, pp. 387–442. Cambridge University Press (2002)
18. Diekert, V., Gutiérrez, C., Hagenah, C.: The existential theory of equations with rational constraints in free groups is PSPACE-complete. Inf. Comput. **202**(2), 105–140 (2005). https://doi.org/10.1016/j.ic.2005.04.002
19. Diekert, V., Jeż, A., Plandowski, W.: Finding all solutions of equations in free groups and monoids with involution. Inf. Comput. **251**, 263–286 (2016). https://doi.org/10.1016/j.ic.2016.09.009
20. Dijkstra, E.W.: Self-stabilizing systems in spite of distributed control. Commun. ACM **17**(11), 643–644 (1974). https://doi.org/10.1145/361179.361202
21. Faran, R., Kupferman, O.: On synthesis of specifications with arithmetic. In: Chatzigeorgiou, A., et al. (eds.) SOFSEM 2020. LNCS, vol. 12011, pp. 161–173. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-38919-2_14
22. Faran, R., Kupferman, O.: On synthesis of specifications with arithmetic. In: Chatzigeorgiou, A., et al. (eds.) SOFSEM 2020. LNCS, vol. 12011, pp. 161–173. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-38919-2_14

23. Figueira, D., Jeż, A., Lin, A.W.: Data path queries over embedded graph databases. In: PODS '22: International Conference on Management of Data, Philadelphia, 12–17 June, 2022. pp. 189–201 (2022). https://doi.org/10.1145/3517804.3524159
24. Figueira, D., Lin, A.W.: Reasoning on data words over numeric domains. In: LICS '22: 37th Annual ACM/IEEE Symposium on Logic in Computer Science, Haifa, Israel, 2–5 August 2022, pp. 37:1–37:13 (2022). https://doi.org/10.1145/3531130.3533354
25. Furia, C.A.: What's decidable about sequences? In: Bouajjani, A., Chin, W.-N. (eds.) ATVA 2010. LNCS, vol. 6252, pp. 128–142. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15643-4_11
26. Ganesh, V., Minnes, M., Solar-Lezama, A., Rinard, M.: Word equations with length constraints: what's decidable? In: Biere, A., Nahir, A., Vos, T. (eds.) HVC 2012. LNCS, vol. 7857, pp. 209–226. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39611-3_21
27. Grumberg, O., Kupferman, O., Sheinvald, S.: Variable automata over infinite alphabets. In: Dediu, A.-H., Fernau, H., Martín-Vide, C. (eds.) LATA 2010. LNCS, vol. 6031, pp. 561–572. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-13089-2_47
28. Hoare, C.A.R.: Quicksort. Comput. J. **5**(1), 10–15 (1962). https://doi.org/10.1093/comjnl/5.1.10
29. Jeż, A.: Recompression: a simple and powerful technique for word equations. J. ACM **63**(1), 4:1–4:51 (2016). https://doi.org/10.1145/2743014
30. Jhala, R., Majumdar, R.: Software model checking. ACM Comput. Surv. **41**(4), 21:1–21:54 (2009). https://doi.org/10.1145/1592434.1592438
31. Kaminski, M., Francez, N.: Finite-memory automata. Theor. Comput. Sci. **134**(2), 329–363 (1994). https://doi.org/10.1016/0304-3975(94)90242-9
32. Kroening, D., Strichman, O.: Decision Procedures. Springer (2008)
33. Lamport, L.: A new solution of Dijkstra's concurrent programming problem. Commun. ACM **17**(8), 453–455 (1974). https://doi.org/10.1145/361082.361093
34. Lin, A.W., Rümmer, P.: Regular model checking revisited. In: Olderog, E.-R., Steffen, B., Yi, W. (eds.) Model Checking, Synthesis, and Learning. LNCS, vol. 13030, pp. 97–114. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-91384-7_6
35. Lin, A.W., Barceló, P.: String solving with word equations and transducers: towards a logic for analysing mutation XSS. In: Proceedings of the 43rd Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages, POPL 2016, St. Petersburg, 20–22 January 2016, pp. 123–136 (2016). https://doi.org/10.1145/2837614.2837641
36. Makanin, G.S.: The problem of solvability of equations in a free semigroup. Sbornik: Mathematics **32**(2), 129–198 (1977)
37. Meyer, B.: Applying "Design by contract." IEEE Comput. **25**(10), 40–51 (1992). https://doi.org/10.1109/2.161279
38. de Moura, L.M., Bjørner, N.: Z3: An efficient SMT solver. In: TACAS (2008)
39. None: RegExLib (2017). https://regexlib.com/. Accessed 2 Feb 2023
40. Plandowski, W.: On PSPACE generation of a solution set of a word equation and its applications. Theor. Comput. Sci. **792**, 20–61 (2019). https://doi.org/10.1016/j.tcs.2018.10.023
41. Rümmer, P.: A constraint sequent calculus for first-order logic with linear integer arithmetic. In: Cervesato, I., Veith, H., Voronkov, A. (eds.) LPAR 2008. LNCS (LNAI), vol. 5330, pp. 274–289. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-89439-1_20

42. Safari, M., Huisman, M.: A generic approach to the verification of the permutation property of sequential and parallel swap-based sorting algorithms. In: Dongol, B., Troubitsyna, E. (eds.) IFM 2020. LNCS, vol. 12546, pp. 257–275. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-63461-2_14
43. Schulz, K.U.: Makanin's algorithm for word equations–two improvements and a generalization. In: Schulz, K.U. (ed.) IWWERT. Lecture Notes in Computer Science, vol. 572, pp. 85–150. Springer, Cham (1990). https://doi.org/10.1007/3-540-55124-7_4
44. Veanes, M., Hooimeijer, P., Livshits, B., Molnar, D., Bjorner, N.: Symbolic finite state transducers: algorithms and applications. SIGPLAN Not. **47**(1), 137–150 (2012). https://doi.org/10.1145/2103621.2103674
45. Wang, Q., Appel, A.W.: A solver for arrays with concatenation. J. Autom. Reason. **67**(1), 4 (2023). https://doi.org/10.1007/s10817-022-09654-y

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.