On Some Succinct Representations of Regular Languages
Extended Abstract

Bruno Guillon, Giovanni Pighizzini, and Luca Prigioniero

Dipartimento di Informatica, Università degli Studi di Milano, Italy
{guillonb, pighizzini, prigioniero}@di.unimi.it

Abstract. Non-self-embedding grammars, constant-height pushdown automata and 1-limited automata are restrictions of context-free grammars, pushdown automata and Turing machines, respectively. All of them characterize the class of regular languages. There is a double size exponential gap from each of these models to deterministic finite automata. Non-self-embedding grammars and constant-height pushdown automata are polynomially related in size. Moreover, there exists a polynomial size simulation by 1-limited automata. In contrast, the converse transformation costs exponential.

1 Introduction

Regular languages are usually represented using regular expression or finite automata. It is well known that, in the worst case, deterministic automata can require exponentially many states with respect to equivalent nondeterministic automata. Hence, there is an exponential size gap from nondeterministic to deterministic automata. Further representations which can be more succinct than nondeterministic automata have been discovered and investigated. Three of them are considered in this work: non-self-embedding grammars, constant-height pushdown automata, and 1-limited automata. The size gap from each one of these representations to equivalent deterministic automata is double exponential.

To describe *non-self-embedding grammars*, we first recall that the extra capability of context-free grammars with respect to regular ones is that of describing recursive structures as, for instance, nested parentheses, arithmetic expressions, typical programming language constructs. In terms of recognizing devices, this capability is implemented through the pushdown store, which is used to extend finite automata in order to make the resulting model, namely pushdown automata, equivalent to context-free grammars.

To emphasize this capability, in one of his pioneering papers, Chomsky investigated the *self-embedding* property [4]: a context-free grammar is self-embedding if it contains a variable A which, in some sentential form, is able to reproduce itself surrounded by two nonempty strings α and β, in symbols $A \overset{\star}{\Longrightarrow} \alpha A \beta$. Roughly speaking, this means that such *self-embedded* variable A is "truly" recursive. He proved that, among all context-free grammars, only self-embedding ones can generate nonregular languages. Hence, *non-self-embedding grammars* are no more powerful than finite automata.

The proof given by Chomsky of this result is constructive, namely it provides a method for obtaining a finite automaton equivalent to a given non-self-embedding grammar [3,4]. A different constructive proof of the same result was given by Anselmo, Giammarresi, and Varricchio [1], by showing a decomposition of non-self-embedding grammars in regular grammars and then iteratively applying regular substitutions to obtain equivalent finite automata. In the same paper, the authors also proved that the size gap from non-self-embedding grammars to equivalent automata is at least exponential.

It is worthwhile to mention that, in 1971, Meyer and Fischer proved that for any recursive function f and arbitrarily large integer n, there exists a context-free grammar whose description has size n and which generates a regular language, such that any equivalent finite automaton requires at least $f(n)$ states [9]. This means that it is not possible to obtain a recursive bound relating the size of context-free grammars generating regular languages with the number of states of equivalent deterministic finite automata. It is important to notice that the result of Meyer and Fischer was obtained by considering grammars with a two-letter terminal alphabet. The unary, i.e., one-letter, case was studied in 2002 by Pighizzini, Shallit, and Wang, who obtained optimal recursive bounds [13].

We recently proved that also in the case of non-self-embedding grammars, the bounds are recursive, independently on the alphabet size [12]. In particular, by inspecting and refining the construction presented in [1], we showed that each non-self-embedding grammar of size s can be converted into equivalent nondeterministic and deterministic automata with $2^{O(s)}$ and $2^{2^{O(s)}}$ states, respectively. We also obtained a family of languages that witness that these gaps cannot be reduced. Furthermore, these gaps do not change if we allow the variables which generate only unary strings (i.e., strings consisting of occurrences of only one terminal) to be self-embedded. Such grammars, which are also equivalent to finite automata, are called *quasi-non-self-embedding grammars.*

Constant-height pushdown automata are standard nondeterministic pushdown automata where the amount of available pushdown store is fixed. Hence, the number of their possible configurations is finite, thus implying that they are no more powerful than finite automata. Exponential and double exponential gaps from constant-height pushdown automata to nondeterministic and deterministic automata, respectively, have been proved in [5]. Furthermore, in [2] the authors showed the interesting result that also the gap from nondeterministic to deterministic constant-height pushdown automata is double exponential. As non-self-embedding grammars, constant-height pushdown automata are restrictions of the corresponding general model, where true recursions are not possible. By comparing these two models, we proved that they are polynomially related in size [6].

For each integer $d > 0$, a *d-limited automaton* is a one-tape nondeterminstic Turing machine which is allowed to rewrite the content of each tape cell only in the first d visits. These models have been introduced by Hibbard in 1967, who proved that for each $d \geq 2$ they characterize context-free languages [7]. This yields a hierarchy of acceptors, merely obtained by restricting one-tape Turing machines, corresponding to Chomsky's classification. Furthermore, as shown in [14, Thm. 12.1], 1-limited automata are equivalent to finite automata. This equivalence has been investigated from the descriptional complexity point of view in [11], by proving exponential and double exponential gaps from 1-limited automata to nondeterministic and deterministic finite automata, respectively. Our main result is a construction transforming each non-self-embedding grammar into a 1-limited automaton of polynomial size. For the converse transformation, we show that an exponential size is necessary. Indeed, we prove a stronger result by exhibiting, for each $n > 0$, a language L_n accepted by a two-way deterministic finite automaton with $O(n)$ states, which requires exponentially many states to be accepted even by an unrestricted pushdown automaton. From the cost of the conversion of 1-limited automata into nondeterministic automata, it turns out that for the conversion of 1-limited automata into non-self-embedding grammars an exponential size is also sufficient.

We use standard abbreviations as CFG, PDA, etc. The prefix 2 before NFA or DFA is used to indicate *two-way automata*. PDAs with pushdown store height bounded by h are indicated as h-PDAs. NSE is an abbreviation for non-self-embedding. Figure 1 summarizes some of the results discussed in this extended abstract. More details can be found in [12] and [6].

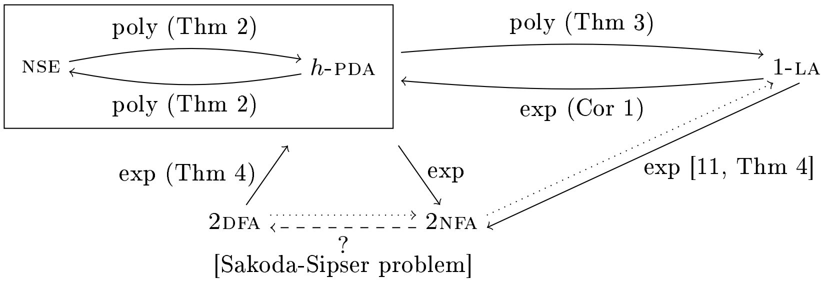

Fig. 1. Some bounds discussed in the paper. Dotted arrows denote trivial relationships, while the dashed arrow indicates the famous Sakoda and Sipser's question. The exponential cost of the simulation of h-PDAs by 2NFAs is discussed at the end of Section 2.

2 A Summary of the Results

For each model under consideration, we evaluate its size as the total number of symbols used to write down its description. For instance, the *size of a grammar* is linear in the sum of the lengths of its productions.

Theorem 1 ([12]). *Given an* NSE *grammar of size s, there exist an equivalent* NFA *and an equivalent* DFA *with a number of states exponential and double exponential in s, respectively. In the worst case these sizes cannot be reduced.*

To show that the bounds in Theorem 1 cannot be reduced, we can consider, for any integer $h > 0$, the language $L_h \subseteq \{a, b\}^*$ defined as the set of strings composed of k blocks $w_1 w_2 \cdots w_k$ each of length h, for some $k > 1$, such that the last block w_k is the reverse of one of the first $k - 1$ blocks, i.e.,

$$L_h = \{w_1 w_2 \cdots w_{k-1} w_k \mid k > 1, w_i \in \{a, b\}^h, i = 1, \ldots, k, \\ \text{and } \exists j, 1 \leq j < k, \text{ s.t. } w_j = w_k^R\}.$$

It can be proved that L_h is generated by a NSE of size $O(n)$, while, by a standard distinguish ability argument, each DFA accepting it requires 2^{2^h} many states.

The statement of Theorem 1 remains true if the grammar is *quasi*-NSE, i.e., it is allowed to contain self-embedded variables, provided that each terminal string generated by them is unary, namely it consists only of occurrences of a same symbol. In contrast, for *quasi*-NSE grammars generating letter bounded languages, the cost of the conversion into DFAs reduces to a simple exponential in the size cost.

Theorem 2 ([6]). *h-PDAs and NSE grammars are polynomially related in size.*

In the proof of Theorem 2, the transformation from h-PDAs to NSE grammars is an adaption of a standard transformation from PDAs to CFGs. For the converse, a modification of a decomposition of NSE grammars presented in [1] is used.

Now, we consider the size relationships of NSE grammars and h-PDAs with 1-limited automata.

Theorem 3 ([6]). *For every NSE grammar G, there exist a 1-state letter-to-letter nondeterministic transducer $\mathcal{T}$ and a 2NFA $\mathcal{A}$ of polynomial size, such that a word w is generated by G if and only if $\mathcal{A}$ accepts an image u of w by $\mathcal{T}$. As a consequence, G can be transformed into a 1-LA of polynomial size.*

Given an input w, the transducer $\mathcal{T}$ nondeterministically generates a *compression* of a derivation tree of w. The 2NFA $\mathcal{A}$ verifies the validity of such a guess. The resulting 1-LA is a composition of $\mathcal{T}$ and $\mathcal{A}$. The converse transformation is exponential. Actually we have a stronger result:

Theorem 4 ([6]). *For each $n > 0$, let L_n be the language of the powers of any string of length n over $\{0, 1\}$,* i.e., *$L_n = \{u^k \mid u \in \{0, 1\}^n,\ k \geq 0\}$. Then:*

- *L_n is accepted by a 2DFA of size $O(n)$;*
- *each context-free grammar in Chomsky normal form needs exponentially many variables in n to generate L_n;*
- *the size of any PDA accepting L_n is at least exponential in n.*

Corollary 1. *The size cost of the conversion of 1-LAs into NSE grammars and h-PDAs is exponential.*

Proof. The lower bound derives from Theorem 4. For the upper bound, in [11] it was proved that each 1-LA can be transformed into a 1NFA of exponential size from which, by standard construction, we can obtain a regular (and, so, NSE) grammar, without increasing the size asymptotically. □

In [2], the question of the cost of the conversion of deterministic h-PDAs into 1NFAs was raised. To this regard, we observe that the language $(a^{2^n})^*$ is accepted by a deterministic h-PDA of size polynomial in n for large enough h (see, *e.g.*, [10]) but, by a standard pumping argument, it requires at least 2^n states to be accepted by 1NFAs. Actually, as a consequence of state lower bound presented in [8], 2^n states are also necessary to accept it on each 2NFA. Considering Theorem 4, we can conclude that both simulations from two-way automata to h-PDAs and from h-PDAs to two-way automata cost at least exponential.

References

1. Anselmo, M., Giammarresi, D., Varricchio, S.: Finite automata and non-self-embedding grammars. In: CIAA 2002. LNCS, vol. 2608, pp. 47–56 (2002)
2. Bednárová, Z., Geffert, V., Mereghetti, C., Palano, B.: Removing nondeterminism in constant height pushdown automata. Inf. Comput. **237**, 257–267 (2014)
3. Chomsky, N.: On certain formal properties of grammars. Information and Control **2**(2), 137–167 (1959)
4. Chomsky, N.: A note on phrase structure grammars. Information and Control **2**(4), 393–395 (1959)
5. Geffert, V., Mereghetti, C., Palano, B.: More concise representation of regular languages by automata and regular expressions. Inf. Comput. **208**(4), 385–394 (2010)
6. Guillon, B., Pighizzini, G., Prigioniero, L.: Non-self-embedding grammars, constant-height pushdown automata, and limited automata. In: CIAA 2018. LNCS (2018), to appear
7. Hibbard, T.N.: A generalization of context-free determinism. Information and Control **11**(1/2), 196–238 (1967)
8. Mereghetti, C., Pighizzini, G.: Two-way automata simulations and unary languages. Journal of Automata, Languages and Combinatorics **5**(3), 287–300 (2000)
9. Meyer, A.R., Fischer, M.J.: Economy of description by automata, grammars, and formal systems. In: 12th Annual Symposium on Switching and Automata Theory, East Lansing, Michigan, USA, October 13-15, 1971. pp. 188–191. IEEE Computer Society (1971)
10. Pighizzini, G.: Deterministic pushdown automata and unary languages. Int. J. Found. Comput. Sci. **20**(4), 629–645 (2009)
11. Pighizzini, G., Pisoni, A.: Limited automata and regular languages. Int. J. Found. Comput. Sci. **25**(7), 897–916 (2014)
12. Pighizzini, G., Prigioniero, L.: Non-self-embedding grammars and descriptional complexity. In: NCMA 2017. pp. 197–209 (2017)
13. Pighizzini, G., Shallit, J., Wang, M.: Unary context-free grammars and pushdown automata, descriptional complexity and auxiliary space lower bounds. J. Comput. Syst. Sci. **65**(2), 393–414 (2002)
14. Wagner, K.W., Wechsung, G.: Computational Complexity. D. Reidel Publishing Company, Dordrecht (1986)