THEORIE-TAG 2017

H. Fernau (Herausgeber): Theorietag 2017, Bonn, 18. – 22.9.2017
Universität Trier, Technischer Bericht, S. 32–36.

Concise Description of Finite Languages, Revisited

Hermann Gruber[(A)] Markus Holzer[(B)] Simon Wolfsteiner[(C)]

[(A)]Knowledgepark GmbH, Leonrodstr. 68, 80636 München, Germany
hermann.gruber@kpark.de

[(B)]Institut für Informatik, Universität Giessen,
Arndtstr. 2, 35392 Giessen, Germany
holzer@informatik.uni-giessen.de

[(C)]Institut für Diskrete Mathematik und Geometrie, TU Wien,
Wiedner Hauptstr. 8–10, 1040 Wien, Austria
simon.wolfsteiner@tuwien.ac.at

Abstract

We investigate the grammatical complexity of finite languages w.r.t. context-free grammars and variants thereof. It is shown that the minimal number of productions necessary for a finite language encoded by a context-free grammar cannot be approximated within a ratio of $o(n^d)$, for all $d \geq 1$, unless $\mathsf{P} = \mathsf{NP}$. Here, n is the length of longest word in the finite language. Similar inapproximability results hold for linear context-free and right-linear (or regular) grammars.

1. Introduction

Questions regarding the economy of descriptions of formal languages by different formalisms such as automata, grammars, and formal systems have been studied quite extensively in the past, see, e.g., [13, 14]. The results in [4] mark the starting point of a theory of the grammatical complexity of finite languages where the chosen complexity measure is the number of productions. In particular, [4] gives a relative succinctness classification for various kinds of context-free grammars. Further results along these lines can be found in [1, 2, 3, 15] as well as some newer ones in, e.g., [6, 7, 9, 10]. It is worth mentioning that in [10] a method for proving lower bounds on the number of productions for context-free grammars was developed. For instance, it was shown that the set of all squares of a given length requires an exponential number of productions to be generated by a context-free grammar.

More recently, it was shown that there is a close relationship between a certain class of formal proofs in first-order logic and a certain class of (tree) grammars. In particular, the number of productions in such a grammar corresponds to the number of certain inference rules in

[(C)]This research was completed while the author was on leave at the Institut für Informatik, Universität Giessen, Arndtstr. 2, 35392 Giessen, Germany, in Summer 2017 and is partially supported by FWF project W1255-N23.

the proof [12, 8]. This correspondence sparked our interest in further investigating questions regarding the grammatical complexity of finite languages. The main result of this paper is that the minimal number of productions necessary for a finite language encoded by a context-free grammar cannot be approximated within a ratio of $o(n^d)$, for all $d \geq 1$, unless $\mathsf{P} = \mathsf{NP}$. Here, n is the length of a longest word in the finite language. This result nicely generalizes the inapproximability of the smallest grammar problem with approximation ratio less than $\frac{8569}{8568}$ unless $\mathsf{P} = \mathsf{NP}$ from [5]. Here, the smallest grammar problem asks for the smallest (in terms of the number of productions) context-free grammar that generates exactly *one* given word. As a byproduct of our inapproximability result, we show that the set of all cubes of a given length requires an exponential number of productions using elementary methods already developed in [4]. To be more precise, the language $T_n = \{\, w\$w\#w \mid w \in \{0,1\}^n \,\}$ requires $\Theta(2^n)$ context-free productions, for $n \geq 1$. This constitutes a drastic improvement of previous results obtained in [4] and, moreover, is more precise than using the lower bound method from [10] that results only in a lower bound of $\Omega(2^{n/8}/\sqrt{3n})$ many context-free productions.

2. Preliminaries

We assume the reader to be familiar with the basic notions on grammars and languages as contained in [11]. In particular, a *context-free grammar* (CFG) is a 4-tuple $G = (N, T, P, S)$, where N and T are disjoint alphabets of *nonterminals* and *terminals*, respectively, $S \in N$ is the *axiom*, and P is a finite set of *productions* of the form $A \to \alpha$, where $A \in N$ and $\alpha \in (N \cup T)^*$. As usual, the derivation relation of G is denoted by $\Rightarrow_G$, and the reflexive and transitive closure of $\Rightarrow_G$ is written as $\Rightarrow_G^*$. The *language generated* by G is defined as

$$L(G) = \{\, w \in T^* \mid S \Rightarrow_G^* w \,\}.$$

We also consider the following restrictions of context-free grammars: (i) a context-free grammar is said to be *linear context-free* (LIN) if the productions are of the form $A \to \alpha$, where $A \in N$ and $\alpha \in T^*(N \cup \{\varepsilon\})T^*$—here ε refers to the *empty word*, and (ii) a context-free grammar is said to be *right-linear* or *regular* (REG) if the productions are of the form $A \to \alpha$, where $A \in N$ and $\alpha \in T^*(N \cup \{\varepsilon\})$. Moreover, a grammar is said to have *weight* at most two, if every right-hand side α of each production $A \to \alpha$ in P is of length at most two, that is, $|\alpha| \leq 2$. Linear context-free and regular grammars of weight at most two are abbreviated by SLIN and SREG, respectively—the prefix S stands for *strict*—this naming was coined in [4]. Furthermore, Γ will denote the set of those abbreviations in the sequel, that is, $\Gamma = \{\mathsf{SREG}, \mathsf{REG}, \mathsf{SLIN}, \mathsf{LIN}, \mathsf{CFG}\}$.

We are interested in the complexity of finite languages w.r.t. different types of grammars. To be more precise: what is the smallest number of productions of a grammar required to generate the language L? Let $G = (N, T, P, S)$ be a context-free grammar. We define $|G|$ to be the number of productions if not stated otherwise, i.e., the number of elements in P. Then the *complexity* of a finite language L w.r.t. an X-grammar, for $X \in \Gamma$, also called the *X-complexity of L*, is defined as

$$X\mathsf{c}(L) = \min\{\, |G| \mid G \text{ is an } X\text{-grammar and } L = L(G) \,\}.$$

By definition, the following relations hold: $\mathsf{CFG} \leq \mathsf{LIN} \leq \mathsf{REG} \leq \mathsf{SREG}$ and moreover we have $\mathsf{CFG} \leq \mathsf{LIN} \leq \mathsf{SLIN} \leq \mathsf{SREG}$, where $X \leq Y$, for $X, Y \in \Gamma$, if and only if $X\mathsf{c}(L) \leq Y\mathsf{c}(L)$, for every finite language L. In the case that $X \leq Y$, we say that X is *more succinct than* Y.

3. Results

In the seminal paper [4] on concise description of finite languages by different types of grammars, certain languages were identified that can only be generated minimally by listing all words that belong to the language under consideration. For instance, the language

$$U_n = \{\, a^k b^k c a^\ell b^\ell d a^m b^m \mid 0 \leq k+\ell+m \leq n \,\}$$

contains a quadratic number of words and satisfies $\mathsf{CFGc}(U_n) = \Omega(n^2)$. The proof of this fact is based on [4, Lemma 2.1] which states some easy facts about *minimal* context-free grammars: let $G = (N, T, P, S)$ be a *minimal* context-free grammar for the finite language L. Then for every nonterminal $A \in N \setminus \{S\}$, there are words α_1 and α_2 with $\alpha_1 \neq \alpha_2$ such that $A \to \alpha_1$ and $A \to \alpha_2$ are in P. Moreover, for every $A \in N \setminus \{S\}$, the set $L_A(G) = \{\, w \in T^* \mid A \Rightarrow_G^* w \,\}$ contains at least two words, and there is no derivation of the form $A \Rightarrow_G^+ \alpha A \beta$ with $\alpha, \beta \in (N \cup T)^*$. Finally, for every $A \in N \setminus \{S\}$, there are $u_1, u_2, v_1, v_2 \in T^*$ such that $u_1 A u_2 \neq v_1 A v_2$ as well as $S \Rightarrow_G^* u_1 A u_2$ and $S \Rightarrow_G^* v_1 A v_2$. Using these facts we show that the set

$$T_n = \{\, w\$w\#w \mid w \in \{0,1\}^n \,\}$$

of all tripels of length n can be generated minimally by a context-free grammar only by listing all words. Thus, we have the following result—observe that this result is more precise than using the lower bound technique from [10] for the language under consideration:

Theorem 3.1 *Let $X \in \Gamma$ and $n \geq 1$. Then $X\mathsf{c}(T_n) = \Theta(2^n)$.*

The language T_n will be a basic building block for our main result, which states that the minimal number of context-free productions for a finite language cannot be approximated within a certain factor unless $\mathsf{P} = \mathsf{NP}$. The main result reads as follows:

Theorem 3.2 *Let $X \in \Gamma$. Given an X-grammar generating a finite language, it is impossible to approximate $X\mathsf{c}(L)$ within a factor of $o(n^d)$, for $n = \max\{\, |w| \mid w \in L \}$ and all $d \geq 1$, unless $\mathsf{P} = \mathsf{NP}$.*

The proof strategy is by a reduction from the coNP-complete unsatisfiability problem for 3SAT-formulae: given a formula F with m clauses and n variables, where each clause is the disjunction of at most 3 literals, it is coNP-complete to determine whether F is unsatisfiable—in other words whether the negation of F is a *tautology*. Then the core idea is to give a suitable presentation of non-satisfying assignments of F in $\{0,1\}^n$ for the n variables in form of a grammar G, such that F is unsatisfiable if and only if $L(G) = \{0,1\}^n$; by construction there is a one-to-one correspondence between assignments and words from the set $\{0,1\}^n$. In order to finish our reduction we embed G into a grammar that generates the language

$$L_F = L(G) \cdot \{0,1,\$,\#\}^{3c \cdot \log n + 2} \cup \{0,1\}^n \cdot T_{c \cdot \log n},$$

for some carefully chosen constant c. It is not hard to see that this reduction is polynomial, even if we force the grammar for L_F to be (strict) regular. Then we distinguish two cases: (i) clearly, if F is unsatisfiable then $L_F = \{0,1\}^n \cdot \{0,1,\$,\#\}^{3c \cdot \log n + 2}$ and there is a CFG-grammar with a constant number of productions that generates L_F. For the other types of X-grammars,

for $X \in \{\mathsf{REG}, \mathsf{SREG}, \mathsf{LIN}, \mathsf{SLIN}\}$, a linear number of productions suffices, i.e., the number is $O(n)$. (ii) On the other hand, if F is satisfiable, there is an assignment that evaluates F to *true*. Hence, there is a word $w \in \{0,1\}^n$ that corresponds to that assignment and is *not* a member of $L(G)$. But then the left-quotient of L_F w.r.t. the word w, that is, the language $w^{-1}L_F = \{\, v \in \{0,1,\$,\#\}^* \mid wv \in L_F \,\}$, is equal to the language of cubes $T_{c\cdot\log n}$. In order to estimate the number of productions for the set $w^{-1}L$, for some word $w \in T^*$ and a language $L \subseteq T^*$, we apply the following lemma.

Lemma 3.3 *Let $X \in \Gamma$ and $G = (N, T, P, S)$ be an X-grammar generating a finite language with $n = \max\{\, |w| \mid w \in L(G) \,\}$. Then one can effectively construct a grammar G' of the same type with $|G'| \leq |G|$, if $X \in \{\mathsf{REG}, \mathsf{SREG}\}$, and $|G'| = O(|G| \cdot n^4)$, if $X \in \{\mathsf{LIN}, \mathsf{SLIN}, \mathsf{CFG}\}$, satisfying $L(G') = w^{-1}L(G)$, for every $w \in T^*$.*

Before we continue with the outline of the proof strategy of the main theorem, we briefly explain the construction of the proof of the lemma. First, we transform the grammar into an equivalent grammar of the same type and *weight at most two*. This increases the number of productions at most by a factor of $O(n)$. Then we apply the triple construction of this grammar with the partial deterministic finite automaton that accepts wT^* in order to accept the intersection of both languages. Simultaneously during this construction, we take care of the triples that directly terminate to letters from the word w and replace them by the empty word ε. These triples can be easily identified, because the partial deterministic finite automaton for wT^* is actually a chain, where the word w is read, followed by the sole accepting state that has a trivial loop on all letters from T. The tedious details are left to the reader. The triple construction with the simultaneous modification increases the grammar at most by a factor of $O(n^3)$. Overall, this gives an increase by a factor of $O(n^4)$ from the original grammar. This proves the stated result for SLIN-, LIN-, and CFG-grammars. An alternative proof shows the linear bound for REG- and SREG-grammars for the quotient w.r.t a single word w.

Now let us come back to the proof outline for the main theorem. Assume that there is a polynomial time approximation algorithm for the minimal number of productions problem within $o(n^d)$, where n is the length of the longest word in the language under consideration and $d \geq 1$. Then this algorithm could be applied to decide whether $\mathsf{CFGc}(L_F) = O(1)$. To this end, set $c = d + 5$. If F is unsatisfiable, then $\mathsf{CFGc}(L_F) = O(1)$, as mentioned above. But if F is satisfiable, let w present a satisfying assignment for F. Then $w^{-1}L_F = T_{c\cdot\log n}$, and by Theorem 3.1, we have $\mathsf{CFGc}(T_{c\cdot n}) = \Theta(n^{d+5})$. By Lemma 3.3 we deduce that $\mathsf{CFGc}(L_F) = \Omega(n^{d+1})$ in this case. Thus, the putative approximation algorithm returns a grammar size of $o(n^{d+1})$ if and only if F is unsatisfiable. This solves a coNP-hard problem in deterministic polynomial time, which implies $\mathsf{P} = \mathsf{NP}$. A similar reasoning can be done with the other types of grammars from Γ. The details are left to the reader. This proves Theorem 3.2 and shows that the X-complexity, for $X \in \Gamma$, of a given finite language cannot be approximated within a factor of $o(n^d)$, for all $d \geq 1$, unless $\mathsf{P} = \mathsf{NP}$.

References

[1] B. ALSPACH, P. EADES, G. ROSE, A Lower-Bound For the Number of Productions Required For A Certain Class of Languages. *Discrete Appl. Math.* **6** (1983), 109–115.

[2] W. BUCHER, A Note on a Problem in the Theory of Grammatical Complexity. *Theoret. Comput. Sci.* **14** (1981) 3, 337–344.

[3] W. BUCHER, H. A. MAURER, K. CULIK II, Context-Free Complexity of Finite Languages. *Theoret. Comput. Sci.* **28** (1983) 3, 277–285.

[4] W. BUCHER, H. A. MAURER, K. CULIK II, D. WOTSCHKE, Concise Description of Finite Languages. *Theoret. Comput. Sci.* **14** (1981) 3, 227–246.

[5] M. CHARIKAR, E. LEHMAN, D. LIU, R. PANIGRAHY, M. PRABHAKARAN, A. SAHAI, S. SHELAT, The smallest grammar problem. *IEEE Trans. Inf. Theory.* **51** (2005) 7, 2554–2576.

[6] J. DASSOW, Descriptional Complexity and Operations—Two Non-classical Cases. In: G. PIGHIZZINI, C. CÂMPEANU (eds.), *Proceedings of the* 19*th Workshop on Descriptional Complexity of Formal Systems*. Number 10316 in LNCS, Springer, Milano, Italy, 2017, 33–44.

[7] J. DASSOW, R. HARBICH, Production Complexity of Some Operations on Context-Free Languages. In: M. KUTRIB, N. MOREIRA, R. REIS (eds.), *Proceedings of the* 14*th Workshop on Descriptional Complexity of Formal Systems*. Number 7386 in LNCS, Springer, Braga, Portugal, 2012, 141–154.

[8] S. EBERHARD, S. HETZL, Compressibility of Finite Languages by Grammars. In: J. SHALLIT, A. OKHOTIN (eds.), *Proceedings of the* 17*th Workshop on Deescriptional Complexity of Formal Systems*. Number 9118 in LNCS, Springer, Waterloo, Ontario, Canada, 2015, 93–104.

[9] K. ELLUL, B. KRAWETZ, J. SHALLIT, M.-W. WANG, Regular Expressions: New Results and Open Problems. *J. Autom., Lang. Comb.* **9** (2004) 2/3, 233–256.

[10] Y. FILMUS, Lower Bounds for Context-Free Grammars. *Inform. Process. Lett.* **111** (2011) 18, 895–898.

[11] M. A. HARRISON, *Introduction to Formal Language Theory*. Addison-Wesley, 1978.

[12] S. HETZL, Applying Tree Languages in Proof Theory. In: A. H. DEDIU, C. MARTÍN-VIDE (eds.), *Proceedings of the* 6*th International Conference Language and Automata Theory and Applications*. Number 7183 in LNCS, Springer, A Coruña, Spain, 2012, 301–312.

[13] M. HOLZER, M. KUTRIB, Descriptional Complexity—An Introductory Survey. In: C. MARTÍN-VIDE (ed.), *Scientific Applications of Language Methods*. World Scientific, 2010, 1–58.

[14] A. R. MEYER, M. J. FISCHER, Economy of description by automata, grammars, and formal systems. In: *Proceedings of the* 12*th Annual Symposium on Switching and Automata Theory*. IEEE Computer Society Press, 1971, 188–191.

[15] Z. TUZA, On the Context-Free Production Complexity of Finite Languages. *Discrete Appl. Math.* **18** (1987) 3, 293–304.