KYBERNETIKA — VOLUME *56* (2020), NUMBER 4, PAGES 722–726

THE RANGE OF NON-LINEAR NATURAL POLYNOMIALS CANNOT BE CONTEXT-FREE

Dömötör Pálvölgyi

Suppose that some polynomial f with rational coefficients takes only natural values at natural numbers, i. e., $L = \{f(n) \mid n \in \mathbb{N}\} \subseteq \mathbb{N}$. We show that the base-$q$ representation of L is a context-free language if and only if f is linear, answering a question of Shallit. The proof is based on a new criterion for context-freeness, which is a combination of the Interchange lemma and a generalization of the Pumping lemma.

Keywords: contex-free languages, pumping lemma

Classification: 68Q45

Call a polynomial f over $\mathbb{Q}$ *natural* if $f(n) \in \mathbb{N}$ for every $n \in \mathbb{N}$. For example, $\frac{x^2-x}{2}$ is natural. Shallit [4, Reseach problem 3 in Section 4.11, page 138] proposed to study whether the base-q representation of the range, $L = \{f(n) \mid n \in \mathbb{N}\}$, of a natural polynomial is context-free or not. It is easy to see that if f is linear, i. e., its degree is at most one, then L is regular and, hence, context-free for any q. It was conjectured that L is not context-free for any other f. This conjecture was known to hold only in special cases, though Sándor Horváth had an unpublished manuscript that claimed a solution.[1] The goal of this note is to present a simple proof that uses a new lemma, which is a simple combination of two well-known necessary criteria for the context-freeness of a language.

A *context-free grammar* G is defined as a finite 4-tuple $G = (V, \Sigma, P, S)$, where V is the set of *non-terminal symbols*, Σ is the set of the *terminal symbols*, which we also call the letters of the *alphabet* (where $V \cap \Sigma = \emptyset$), P is the set of *production rules* and $S \in V$ is the *start symbol*. Each production rule is of the form $A \to \alpha$ where $A \in V$ and $\alpha \in (V \cup \Sigma)^*$ is a *string*. When such a rule is applied to an occurrence of A in some string β, that occurrence of the symbol A is replaced with α in β to obtain a new string. We say that a string $\gamma \in (V \cup \Sigma)^*$ can be *derived* from another string $\beta \in (V \cup \Sigma)^*$ if after applying some rules to certain occurrences of the appropriate non-terminals starting from β we can obtain γ. The *language* $L(G)$ of the grammar G is the set of words from Σ^* that can be derived from S. A derivation of a word $z \in L(G)$ from S can be described by a *derivation tree*; this is a rooted ordered tree whose non-leaf nodes are

DOI: 10.14736/kyb-2020-4-0722

[1] According to Shallit, see `https://cstheory.stackexchange.com/a/41864/419`..

labeled with non-terminal symbols such that the root is labeled with S, the labels of the children of any node labeled A are the right side of some rule $A \to \alpha$ in the given order, and leaves are labeled with terminal symbols that give z in the given order. A grammar is in *Chomsky normal form* if the right side of each production rule is either two non-terminal symbols, or one terminal symbol, or the empty string, but this latter is allowed only if the left side is S provided that S does not appear on the right side of any production; every context-free grammar has a Chomsky normal form. A language L is context-free if $L = L(G)$ for some context-free grammar G. For other basic definitions and statements about context-free grammars and languages, we direct the reader to [4].

Now we state two lemmas that we later combine.[2] The first is known as the Interchange Lemma.

Lemma 1. (Interchange Lemma Odgen et al. [3]) For every context-free language L there is a constant $p > 0$ such that for all $n \in \mathbb{N}$ for any collection of length n words $R \subset L$ there is a subset $Z = \{z_1, \ldots, z_k\} \subseteq R$ with $k \geq |R|/(pn^2)$, and decompositions $z_i = v_i w_i x_i$ such that each of $|v_i|$, $|w_i|$, and $|x_i|$ is independent of i, and the words $v_i w_j x_i$ are in L for every $1 \leq i, j \leq k$.

The second is the following generalization of the Pumping Lemma [1].

Lemma 2. (Dömösi and Kudlek [2]) For every context-free language L there is a constant p such that if in a word $z \in L$ we distinguish d positions and exclude e positions such that $d \geq p(e+1)$, then there is a decomposition $z = uvwxy$ such that vx contains at least one distinguished position, but no excluded positions and $uv^i wx^i y \in L$ for every $i \geq 0$.

A straight-forward combination of the proofs of Lemmas 1 and 2 gives the following.

Lemma 3. (Combined lemma) For every context-free language L there is a constant $p > 0$ such that for all n and for any collection of length n words $R \subseteq L$, if in each word of R we distinguish d positions and exclude e positions (not necessarily at the same place in different words) such that $d \geq p(e+1)$, then there is a $Z = \{z_1, \ldots, z_k\} \subset R$ with $k \geq |R|/(pn^4)$, and a decomposition $z_i = u_i v_i w_i x_i y_i$ for every $1 \leq i \leq k$ such that

- $|u_i|$, $|v_i|$, $|w_i|$, $|x_i|$, and $|y_i|$ are all independent of i,
- $v_i x_i$ contains at least one distinguished position, but no excluded positions,
- $u_{i_0} v_{i_1} \cdots v_{i_m} w_{i_{m+1}} x_{i_m} \cdots x_{i_1} y_{i_0} \in L$ for any sequence of indices $1 \leq i_0, \ldots, i_{m+1} \leq k$.

The proof of Lemma 3 can be found at the end of this note. Now we state an interesting corollary of Lemma 3 that we can apply to Shallit's problem.

Corollary 4. If in a context-free language L for infinitely many n there are $\omega(n^4)$ words of equal length in L whose prefixes of length $\omega(n)$ coincide and their suffixes of length n are pairwise different, then there is an integer B such that there are infinitely many pairs of words in L of equal length that differ only in their suffixes of length B.

[2]Note that we here we state them in a slightly weaker form as their original versions, as we do not use some parts of the original statements.

P r o o f . There is a p that satisfies the conditions of Lemma 3 for L. Take a large enough n for which there are pn^4+1 words of equal length in L whose prefixes of length $p(n+1)$ are the same, but their suffixes of length n are different; this will be R. Apply Lemma 3 to R, distinguishing the first $p(n+1)$ positions and excluding the last n positions to obtain some $Z = \{z_1 = u_1v_1w_1x_1y_1, z_2 = u_2v_2w_2x_2y_2\}$. It follows from the conditions that u_1 and u_2 must contain only distinguished positions, thus $u_1 = u_2$. Since v_i and x_i cannot contain excluded positions, either $y_1 \neq y_2$, or $|x_1| = |x_2| = 0$ and $w_1y_1 \neq w_2y_2$. In the former case the pairs of words $u_1v_1^jw_1x_1^jy_1$ and $u_2v_1^jw_1x_1^jy_2 = u_1v_1^jw_1x_1^jy_2$, in the latter case the pairs of words $u_1v_1^jw_1y_1$ and $u_1v_1^jw_2y_2$ satisfy the conclusion for $j > 0$. □

Now we are ready to prove our main result.

Theorem 5. The language $L = \{f(n) \mid n \in \mathbb{N}\}$ is not context-free for any non-linear natural polynomial f over any base-q.

P r o o f . First we show that the condition of Corollary 4 is satisfied for every natural polynomial f for infinitely many n for some words from $L = \{f(x) \mid x \in \mathbb{N}\}$. The plan is to take some numbers $x_1, \ldots, x_N$ (where $N = n^5$) for which $f(x_i) \neq f(x_j)$, and then add some large number s to each of them to obtain the desired words $f(x_i + s)$.

If the degree of f is d, then at most d numbers can take the same value, thus we can select $x_1, \ldots, x_N$ from the first dN natural numbers, which means that they have $O(\log n)$ digits (since d is a constant). In this case $f(x_i) = O((dN)^d)$, thus each $f(x_i)$ will also have $O(\log n)$ digits. If we pick s to be some natural number with n^2 digits, then $f(s) = \Theta(s^d)$ will have $D = dn^2 + \Theta(1)$ digits. We have $f(x_i + s) = f(s) + \Theta(s^{d-1}x_i)$, where $\Theta(s^{d-1}x_i)$ has $(d-1)n^2 + O(\log n)$ digits. Thus, each $f(x_i + s)$ will either have D digits, or (in case $f(s)$ starts with many 9's) D or $D+1$ digits, or (in case $f(s)$ starts with many 0's) D or $D-1$ digits. In either case, at least half, i. e., $N/2$ of them will have the same length; these will be the words we input to Corollary 4. We still need to show that for these $f(x_i + s)$ their first $\Omega(n^2)$ digits are the same and that their last $O(\log n)$ digits differ.

Let $M \in \mathbb{N}$ be such that $f(x) = \sum_{i=0}^{d} \frac{\alpha_i}{M} x^i$ for $\alpha_i \in \mathbb{Z}$. If s is a multiple of Mq^m, then the last m digits of $f(x)$ and $f(x + s)$ are the same for any x. This way it is easy to ensure that the last $O(\log n)$ digits in base-q stay different. Since $f(x + s) = \sum_{i=0}^{d} \frac{\alpha_i}{M}(x + s)^i = \frac{\alpha_d}{M} s^d + O(s^{d-1}(dN)^d)$, the first $n^2 - O(\log n)$ digits can take only two possible values (depending on whether there is a carry or not), thus one of these values is the same for $N/2$ of the $f(x_i + s)$. Thus we have shown that the condition of Corollary 4 is satisfied

If $L = \{f(x) \mid x \in \mathbb{N}\}$ was context-free, then from the conclusion of Corollary 4 we would obtain infinitely many pairs of numbers, $a_i, b_i \in L$, such that $|a_i - b_i| \leq 2^B$, but this is impossible for non-linear polynomials. □

We end with the omitted proof.

P r o o f. [Proof of Lemma 3] Fix a context-free grammar for L in Chomsky normal form, with t non-terminals. Fix a derivation tree for each word $z \in R$. We say that a node has a distinguished (resp. excluded) *descendant* if a distinguished (resp. excluded) position is derived from the given node in the tree, i. e., if there is a leaf among its descendants whose label is in a distinguished (resp. excluded) position of z.

Call a node of the derivation tree an *e-branch* node if both of its children have an excluded descendant. There are exactly $e - 1$ e-branch nodes in the derivation tree (if $e \geq 1$).

Call a node of the derivation tree a *d-branch* node if both of its children have a distinguished descendant. There are exactly $d - 1$ d-branch nodes in the derivation tree. Say that a d-branch node is the *d-parent* of its descendant d-branch node if there are no d-branch nodes between them. With this structure the d-branch nodes form a binary tree. The ith d-parent of a d-branch node is the i-times iteration of the d-parent operator.

Call a d-branch node *bad* if there is an e-branch node between it and its $(2t + 3)$th d-parent (excluding the node, but including its $(2t + 3)$th d-parent), or if it does not have a $(2t + 3)$th d-parent. Because of the binary structure of the d-branch nodes, each e-branch node can cause at most $2^{2t+4} - 1$ d-branch nodes to be bad, and a further $2^{2t+3} - 1$ d-branch nodes might not have a $(2t + 3)$th d-parent. Therefore in total there are at most $e2^{2t+4}$ bad d-branch nodes, so there is a d-branch node that is not bad if $p > 2^{2t+4}$. Consider the path from the $(2t + 3)$th d-parent of a non-bad d-branch node to the non-bad d-branch node. Note that, since there is no e-branch node, at most one node on this path can have an excluded descendant. Hence, we can conclude that there is a subpath with $t + 1$ d-branch nodes on it such that no sibling of any node along the subpath has an excluded descendant. (The worst case is when the excluded descendant(s) belong to the sibling of the $(t+2)^{\text{nd}}$ d-parent of a non-bad d-branch node.)

By the pigeonhole principle some non-terminal A appears twice on the left side of a rule along this subpath. While we reach one node from the other, some string $\alpha A\beta$ is derived from A. Apply the corresponding rules from the derivation tree to α and β to obtain the string vAx where $v, x \in \Sigma^*$. Thus, z can be written as $z = uvwxy$ such that vAx can be derived from A, w can be derived from A, the subwords v and x have no excluded position (since they are descendants of siblings of nodes along the path), but at least one of them has a distinguished position. For each $z \in R$ we fix such a decomposition $z = uvwxy$.

We partition R into at most $t\binom{n+4}{4}$ groups depending on which non-terminal A appeared on the left side of the rule, and the lengths of u, v, w, x and y. Let $c = t \max_n \binom{n+4}{4}/n^4 = 5t$ (if we only care about large n, then c would be close to $t/24$). By the pigeonhole principle one of the groups will have at least $|R|/(cn^4)$ words in it; this will be Z. Since we can arbitrarily apply the rules for A, the conclusion follows with $p \geq \max(c, 2^{2t+4} + 1)$. □

REMARKS AND ACKNOWLEDGMENT

This note started as a CSTheory.SE answer.[3]

I would like to thank the anonymous reviewers of Kybernetika for improving the presentation of the paper.

This work was supported by the Lendület program of the Hungarian Academy of Sciences (MTA), under grant number LP2017-19/2017

(Received October 19, 2019)

REFERENCES

[1] Y. Bar-Hillel, M. A. Perles, and E. Shamir: On formal properties of simple phrase-structure grammars. Zeitschrift für Phonetik, Sprachwissenschaft, und Kommunikationsforschung *14* (1961), 2, 143–172. DOI:10.1524/stuf.1961.14.14.143

[2] P. Dömösi and M. Kudlek: Strong iteration lemmata for regular, linear, context-free, and linear indexed languages. In: Fund. Comput. Theory 1999, pp. 226–233. DOI:10.1007/3-540-48321-7_18

[3] W. Ogden, R. J. Ross, and K. Winklmann: An "Interchange Lemma" for context-free languages. SIAM J. Comput. *14* (1982), 2, 410–415. DOI:10.1137/0214031

[4] J. Shallit: A Second Course in Formal Languages and Automata Theory. Cambridge University Press, New York 2008. DOI:10.1017/cbo9780511808876

Dömötör Pálvölgyi, MTA-ELTE Lendület Combinatorial Geometry Research Group, Institute of Mathematics, Eötvös Loránd University, Pázmány Péter sétány 1/C, Budapest, 1117. Hungary.

e-mail: dom@cs.elte.hu

[3] See my two answers for `https://cstheory.stackexchange.com/questions/41863/base-k-representations-of-the-co-domain-of-a-polynomial-is-it-context-free`; note that at that time I didn't know about Lemmas 1 and 2.