The range of non-linear natural polynomials cannot be context-free

Dömötör Pálvölgyi

MTA-ELTE Lendület Combinatorial Geometry Research Group,
Institute of Mathematics, Eötvös Loránd University, Budapest, Hungary

April 10, 2019

Abstract

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-\( k \)-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.

Call a polynomial \( f \) over \( \mathbb{Q} \) natural if \( f(n) \in \mathbb{N} \) for every \( n \in \mathbb{N} \). For example, \( x^2 + x^2 \) is natural.

Shallit [4, Research problem 3 in Section 4.11, page 138] proposed to study whether the base-\( k \)-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 context-free for any \( k \). 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 This conjecture was known to hold only in special cases, though Sándor Horváth had an unpublished manuscript that claimed a solution. 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 context-freeness.

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, \( \Sigma \) 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 \( A \rightarrow \alpha \) where \( A \in V \) and \( \alpha \in (V \cup \Sigma)^* \) is an expression. When such a rule is applied to an occurrence of \( A \) in some expression \( \beta \), that occurrence of the symbol \( A \) is replaced with \( \alpha \) in \( \beta \) to obtain a new expression. We say that an expression \( \gamma \in (V \cup \Sigma)^* \) can be derived from another expression \( \beta \in (V \cup \Sigma)^* \) if after applying some rules to some appropriate non-terminals starting from \( \beta \) we can obtain \( \gamma \). The language \( L(G) \) of the grammar \( G \) is the set of words from \( \Sigma^* \) 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 nodes are 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 \( A \rightarrow \alpha \) in the given order, and the labels of the leaves 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; every context-free grammar has a Chomsky normal form. A language \( L \) is context-free if \( L = L(G) \) for some context-free grammar \( G \).

\[1\text{According to Shallit, see } \url{https://cstheory.stackexchange.com/a/41864/419}.\]
For other basic definitions and statements about context-free grammars and languages, we direct the reader to [4].

Now we state a slightly weaker form of the two lemmas we later combine. The first is known as the Interchange lemma.

**Lemma 1 (Interchange lemma [3])**. For every context-free language \( L \) there is a constant \( p > 0 \) such that for all \( n \) for any collection of length \( n \) words \( R \subseteq L \) there is a \( Z = \{z_1, \ldots, z_k\} \subseteq R \) with \( k \geq |R|/(pn^2) \), and decompositions \( z_i = u_i v_i x_i \) 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 (Dömősi-Kudlek [2])**. For every context-free language \( L \) there is a constant \( p > 0 \) 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 \( z = uvwxy \in L \) 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 \( n \) for any collection of length \( n \) words \( R \subseteq L \), if we distinguish \( d \) positions and exclude \( e \) positions such that \( d \geq p(e + 1) \), then there is a \( Z = \{z_1, \ldots, z_k\} \subseteq R \) with \( k \geq |R|/(pn^4) \), and a decomposition \( z_i = u_i v_i w_i x_i y_i \) such that

- \( |u_i|, |v_i|, |w_i|, |x_i|, \) and \( |y_i| \) are all independent of \( i \),
- \( v_i x_i \) has a distinguished position, but no excluded positions,
- \( u_{i_0} v_{i_1} \ldots v_{i_m} w_{i_{m+1}} x_{i_{m+2}} \ldots x_{i_{m+i_0}} y_{i_0} \in L \) for every sequence of indices \( 1 \leq i_0, i_1, \ldots, i_{m+1} \leq k \).

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 first \( \omega(n) \) letters are the same and their last \( n \) letters are different (pairwise), 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 last \( B \) letters.

**Proof.** 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 first \( p(n+1) \) letters are the same, but their last \( n \) letters 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 \( Z = \{z_1 = u_1 v_1 w_1 x_1 y_1, z_2 = u_2 v_2 w_2 x_2 y_2\} \). It follows from the conditions that \( u_1 \) and \( u_2 \) must contain only distinguished positions, thus \( u_1 = u_2 \). Since \( v_1 \) and \( x_1 \) cannot contain excluded positions, either \( y_1 = y_2 \), or \( x_1 = x_2 = \emptyset \) and \( w_1 y_1 = w_2 y_2 \). In the former case the pairs of words \( u_1 v_1^j w_1 x_1^j y_1 \) and \( u_2 v_2^j w_1 x_2^j y_2 = u_1 v_1^j w_1 x_1^j y_2 \), in the latter case the pairs of words \( u_1 v_1^j w_1 y_1 \) and \( u_1 v_1^j w_2 y_2 \) satisfy the conclusion. \( \square \)

Now we are ready to prove our main result.

**Theorem 5.** \( L \) is not context-free for non-linear natural polynomials over any base-\( k \).
Proof. 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 \( dN \) 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 number with \( n^2 \) digits, then \( f(s) \) will have \( D = dn^2 + \Theta(1) \) digits, and each \( f(x_i + s) \) will have \( D \) or \( D + 1 \) digits, thus at least half, i.e., \( N/2 \) of them 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 \( q \in \mathbb{N} \) be such that \( f(x) = \sum_{i=0}^d \alpha_i x^i \) for \( \alpha_i \in \mathbb{Z} \). If \( s \) is a multiple of \( qk^m \), 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-\( k \) stay different. Since \( f(x + s) = \sum_{i=0}^d \alpha_i (x + s)^i = \alpha_d s^d + O(s^{d-1}(dN)^d) \), 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 \( |a_i - b_i| \leq 2^D \), but this is impossible for non-linear polynomials. 

We end with the omitted proof.

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 \( i \)-th \( d \)-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 \( (2t + 3) \)\( d \)-parent (excluding the node, but including its \( (2t + 3) \)\( d \)-parent), or if it does not have a \( (2t + 3) \)\( d \)-parent. Because of the binary structure of the \( d \)-branch nodes, the root and each \( e \)-branch node can cause at most \( 2^{2t+3} \) bad \( d \)-branch nodes, so there is a \( d \)-branch node that is not bad. Consider the path from the \( (2t + 3) \)\( d \)-parent to a non-bad \( d \)-branch node. Note that the nodes on this path might have an excluded descendant, but since there is no \( e \)-branching node along the path, 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.

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 expression \( \alpha A \beta \) is derived from \( A \).
Apply the corresponding rules from the derivation tree to $\alpha$ and $\beta$ to obtain the expression $vAx$ 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.

\[ \square \]

Remark

This note started as a CSTheory.SE answer\[2\]

References

[1] Yehoshua Bar-Hillel, Micha A. Perles, and Eli Shamir, On formal properties of simple phrase-structure grammars. Zeitschrift für Phonetik, Sprachwissenschaft, und Kommunikationsforschung. 14 (2): 143172, 1961.

[2] Pál Dömösi, Manfred Kudlek, Strong iteration lemmata for regular, linear, context-free, and linear indexed languages, in: Fundamentals of Computation Theory (FCT) 1999, LNCS 1684.

[3] William Ogden, Rockford J. Ross, and Karl Winklmann, An “Interchange Lemma” for Context-Free Languages. SIAM J. Comput. 14 (2): 410–415, 1982.

[4] Jeffrey Shallit, A Second Course in Formal Languages and Automata Theory (1 ed.), 2008. Cambridge University Press, New York, NY, USA.

\[2\]See my two answers for \url{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.