CONTEXT–FREE LANGUAGES OF SUB–EXPONENTIAL GROWTH

MARTIN R. BRIDSON AND ROBERT H. GILMAN

Abstract. There do not exist context–free languages of intermediate growth.

The function γ whose value at each non–negative integer n is the number of words on length n in a fixed formal language L is called the growth function of L. Flajolet [3] asked if there are context–free languages of intermediate growth; that is, such that γ is not bounded above by a polynomial, but $\limsup \gamma(n)/r^n = 0$ for all $r > 1$. The answer to this question is a corollary to the following theorem.

Theorem 1. *If L is a context–free language with growth function γ, then either there is a number $r > 1$ and integer n_0 such that $\gamma(n) \geq r^n$ for all $n \geq n_0$, or else L is a bounded language.*

A bounded language is one which is a subset of $w_1^* \cdots w_n^*$ for some words $\{w_1, \ldots, w_n\}$. Since it is clear that the growth of a bounded language is bounded above by a polynomial, we have the desired corollary.

Corollary 2. *There do not exist context–free languages of intermediate growth.*

We note that by a recent result of Grigorchuk and Machì there are indexed languages of intermediate growth [4].

Corollary 2 was obtained independently by Roberto Incitti [6]. Theorem 1 occurs in our previous work [1] as a remark that the proof given there of the weaker result [1, Proposition 1.3] suffices for Theorem 1. In this note we give a quicker proof of Theorem 1 based on work of Ginsburg and Spanier [5], who also obtain a corresponding decidability result.

Theorem 3 ([5, Theorem 5.2]). *It is decidable whether or not the language L generated by a given context–free grammar is bounded; and if L is bounded, one can effectively find words $\{w_1, \ldots, w_n\}$ such that $L \subset w_1^* \cdots w_n^*$.*

1991 *Mathematics Subject Classification.* 68Q45.

Key words and phrases. context-free languages, growth functions.

Bridson's research is supported by an EPSRC Advanced Fellowship.

Now we prove Theorem 1. Suppose the language L over the finite alphabet Σ is generated by a context–free grammar G with start symbol S. Without loss of generality we may assume that each non–terminal A of G participates in a derivation of some word in L. Write $A \overset{*}{\rightarrow} \alpha$ to indicate that A derives the sentential form α. Following [5] we define for each non–terminal A

$$\begin{aligned} Y_A &= \{u \mid A \overset{*}{\rightarrow} uAv \text{ for some } v \in \Sigma^*\} \\ Z_A &= \{v \mid A \overset{*}{\rightarrow} uAv \text{ for some } w \in \Sigma^*\}. \end{aligned}$$

Theorem 4 ([5, Theorem 5.1])**.** *A necessary and sufficient condition that the non–empty language L generated by a context–free grammar G be bounded is that Y_A and Z_A both be commutative for every non–terminal A of G.*

To complete the proof of Theorem 1 it suffices to show that if some Y_A or Z_A is not commutative, then L_A, the language of all words derivable from A, has growth function bounded below by an exponential. Indeed since A occurs in the derivation of at least one word in L, there is a derivation $S \overset{*}{\rightarrow} uAv$ for some words $u, v \in \Sigma^*$, and it follows easily that the growth function of L is bounded below by an exponential once the growth function for L_A is.

Suppose a particular Y_A is not commutative (the argument is similar for Z_A) and pick two derivations

$$A \overset{*}{\rightarrow} u_1Av_1 \quad A \overset{*}{\rightarrow} u_2Av_2 \quad u_1u_2 \neq u_2u_1.$$

Choose an integer m with $m \geq |u_i|$, $m \geq |v_i|$ and $m \geq |w|$ for some $w \in L_A$, and then choose two more integers integers d, e such that u_1^d and u_2^e have the same length. Each derivation may be used to expand the non–terminal A occuring in u_1Av_1 and in u_2Av_2. By iterating these expansions, one sees that for every word $W = W(x_1, x_2)$ in the free monoid $\{x_1, x_2\}^*$, L_A contains $W(u_1^d, u_2^e)w\bar{W}(v_1^d, v_2^e)$, where $\bar{W}$ is W written backwards.

If two distinct words W and W' of the same length k yield the same element of L_A, then one of $W(u_1^d, u_2^e)$, $W'(u_1^d, u_2^e)$ must be a prefix of the other. Because u_1^d and u_2^e have the same length, $W(u_1^d, u_2^e) = W'(u_1^d, u_2^e)$ in this case. As W and W' are distinct, we must have $u_1^d = u_2^e$. According to [2, Corollary 4.1] or [5, Lemma 5.1], this implies that u_1 and u_2 commute. Since u_1 and u_2 do not commute, we conclude that the 2^k words W of length k yield 2^k distinct words $W(u_1^d, u_2^e)w\bar{W}(v_1^d, v_2^e)$ of length at most $(2k+1)mf$ in L_A. Here f is the maximum of $\{d, e\}$.

Hence the growth function γ_A of L_A, satisfies $\gamma_A((2k+1)mf) \geq 2^k$ if $k \geq 1$.

Suppose $n \geq 6mf$. Dividing n by mf we obtain $n = (2k+1)mf + r$ for some $k \geq 1$ and r with $0 \leq r < 2mf$. Thus $\gamma_A(n) \geq 2^k$ and $k \geq (n-3mf)/(2mf) \geq n/(4mf)$. Hence $n \geq 6mf$ implies $\gamma_A(n) \geq r^n$ for $r = 2^{-4mf}$.

References

[1] M. Bridson and R. Gilman, Formal language theory and the geometry of 3-manifolds, Commentarii Math. Helv., **71**, 1996, 525–555.

[2] C. Choffrut and J. Karhumäki, Combinatorics of words, in *Handbook of Formal Languages*, vol. 1, G. Rozenberg and A. Salomaa eds., Springer Verlag, 1997, 329–438

[3] P. Flajolet, Analytic models and ambiguity of context–free languages, Theoret. Comput. Sci. 49 (1987), 283–309.

[4] R. I. Grigorchuk and A. Machì, An example of an indexed language of intermediate growth. Theoret. Comput. Sci. 215 (1999), 325–327.

[5] S. Ginsburg and E. Spanier, Bounded ALGOL–like languages, Trans. Amer. Math. Soc. 113 (1964), 333-368.

[6] R. Incitti, The growth of context–free languages, to appear.

Martin R. Bridson, Mathematical Institute, 24–29 St. Giles, Oxford OX1 3LB, UK

E-mail address: `bridson@maths.ox.ac.uk`

Robert H. Gilman, Department of Mathematical Sciences, Stevens Institute of Technology, Hoboken, NJ 07030, USA

E-mail address: `rgilman@stevens-tech.edu`