CFGs for Restricted Histogram Languages

Yuval Filmus

December 2010

1 Introduction

The following question was asked on `math.stackexchange.com`. Given a finite alphabet Σ, consider the language of all words containing an even number of each symbol $\sigma \in \Sigma$. The language is clearly regular, but a DFA for it requires $2^{|\Sigma|}$ states, by Nerode's theorem. Does the language have a context-free grammar (CFG) of size polynomial in $|\Sigma|$? In this note, we answer this question in the negative.

2 Problem statement

We will consider the following generalization of the problem alluded to in the introduction.

Definition 2.1. Let $\Lambda \subset \mathbb{N}$ be an arbitrary subset, and let Σ be a finite alphabet. The language L_Λ consists of all words in which the number of occurrences of each symbol $\sigma \in \Sigma$ belongs to Λ.

In other words, if we make a histogram for an arbitrary word $w \in \Sigma^*$, then $w \in L_\Lambda$ iff the histogram is supported by Λ. We may call these languages *restricted histogram languages*.

Note that in general, the language L_Λ need not be context-free, or even computable (there are uncountably many choices for Λ). However, if Λ is finite then the language is *regular*, with a minimal DFA having $|\Lambda|^{|\Sigma|}$ states. We get similar results if Λ is cyclic or eventually cyclic.

For some Λ we have very simple DFAs.

Definition 2.2. A subset $\Lambda \subset \mathbb{N}$ is *trivial* if it is one of

$$\emptyset, \{0\}, \mathbb{N} \setminus \{0\}, \mathbb{N}.$$

The languages $L_\emptyset, L_\mathbb{N}$ are accepted by DFAs with a single state, whereas $L_{\{0\}}, L_{\mathbb{N}\setminus\{0\}}$ require two states. All of these have linear size CFGs. Our goal is to provide an exponential lower bound for the size of CFGs of L_Λ for all non-trivial Λ.

3 Chomsky normal form

It will be convenient to work with a grammar in Chomsky normal form (CNF).

Definition 3.1. A grammar is in Chomsky normal form if all its productions are either of the form $A \to BC$ or of the form $A \to a$, where A, B, C are non-terminals and a is a terminal. In addition, we also allow the production $S \to \epsilon$, where S is the starting symbol.

Every CFG can be put into Chomsky normal form (CNF) with at most a quadratic blowup.

Lemma 3.2. *If G is a CFG then there is an equivalent CNF G' with $|G'| = O(|G|^2)$.*

Proof. Check any standard text. □

A derivation of a word using a CNF grammar can be viewed as a binary tree where each node either has two non-leaf children or one leaf child. This implies the following useful lemma.

Lemma 3.3. *Let L be a context-free language with CNF grammar G. For each word w and each positive $\ell \leq |w|$ there is a subword x of w generated by a non-terminal of G of size $\ell \leq |x| < 2\ell$.*

Proof. Consider the derivation tree of w. For a node v, let $w(v)$ be the subword generated by v. We find the required subword using an iterative process. The starting point v_0 is the root. We stop the process at v_t if $|w(v_t)| < 2\ell$. Otherwise, we choose v_{t+1} as the child of v_t generating the bigger subword.

The process must eventually stop. If it stops at v_0 then w is the required subword. If it stops at v_{t+1} then $|w(v_{t+1})| \geq |w(v_t)|/2 \geq \ell$. □

4 Main theorem

Theorem 4.1. *Let $\Lambda \subset \mathbb{N}$ be non-trivial. There is a constant $c > 1$ such that any CFG grammar for L_Λ on alphabet Σ is of size $\Omega(c^{|\Sigma|})$.*

Proof. Put $n = |\Sigma|$. We given an exponential lower bound for the number of non-terminals in a CNF grammar for L_Λ; the result follows from Lemma 3.2.

Since Λ is non-trivial, there is an $\ell > 0$ such that $\ell \in \Lambda$ and either $\ell - 1 \notin \Lambda$ or $\ell + 1 \notin \Lambda$. For $\pi \in S(\Sigma)$, the set of all $n!$ permutation of Σ, define $w_\pi = \pi^\ell \in L_\Lambda$. For each π we use Lemma 3.3 to find a subword x_π of length $n/3 \leq |x_\pi| < 2n/3$ generated by some non-terminal s_π. Note that since $|x_\pi| \leq n$, x_π has no repeated symbols.

If $s_\alpha = s_\beta$ then we can replace x_α with x_β in w_α, and x_β with x_α in w_β. If $\ell - 1 \notin \Lambda$ then $w_\alpha(x_\alpha = x_\beta) \in L$ implies that as sets $x_\alpha \subset x_\beta$. Using $w_\beta(x_\beta = x_\alpha) \in L$ we conclude that as sets $x_\alpha = x_\beta$. If $\ell + 1 \notin \Lambda$ then the inclusions are reversed. Thus x_α, x_β are permutations of each other. Given x_α, for how many permutations is it true that x_β is a permutation of x_α? We have

$$|x_\alpha|(n - |x_\alpha|) < (n/3)!(2n/3)!$$

choices for x_β and for the rest of β; however, this defines β only up to cyclic rotation, so that $x_\alpha = x_\beta$ for at most $n(n/3)!(2n/3)!$ permutations β. Since there are $n!$ permutations, the grammar must contain at least these many symbols:

$$\frac{n!}{n(n/3)!(2n/3)!} = \Omega\left(\frac{c'^n}{n^{3/2}}\right), \quad c' = \frac{3}{2^{2/3}}.$$

The approximation can be obtained using Stirling's formula. Applying Lemma 3.2, we get a lower bound of

$$\Omega\left(\frac{c^n}{n^{3/4}}\right), \quad c = \frac{3^{1/2}}{2^{1/3}}.$$

Note that $3^3 > 2^2$ and so $c > 1$. □