Optimal Test Sets for Context-Free Languages

Mikaël Mayer  
EPFL  
mikael.mayer@epfl.ch

Jad Hamza  
EPFL/INRIA  
jad.hamza@epfl.ch

Abstract
A test set for a formal language (set of strings) $L$ is a subset $T$ of $L$ such that for any two string homomorphisms $f$ and $g$ defined on $L$, if the restrictions of $f$ and $g$ on $T$ are identical functions, then $f$ and $g$ are identical on the entire $L$. Previously, it was shown that there are context-free grammars for which smallest test sets are cubic in the size of the grammar, which gives a lower bound on tests set size. Existing upper bounds were higher degree polynomials; we here give the first algorithm to compute test sets of cubic size for all context-free grammars, settling the gap between the upper and lower bound.

Keywords test sets, context-free languages, context-free grammars

1 Introduction
It is known that given a context-free language $L$ (given by a context-free grammar $G$ of size $n$), one can construct a test set $T$ for $L$ whose size is $O(n^5)$ $[1][2][3].$

Moreover, it was shown $[1][2][3]$ that $O(n^3)$ is a lower bound, in the sense that there exists an infinite family of context-free grammars $G_1, G_2, \ldots$, such that the size of $G_n$ is $O(n)$ and the number of words contained in $G_n$ is $O(n^3)$ but $G_n$ does not contain a test set $T$ as a strict subset. The only test set for $G_n$ is $G_n$.

Our contribution is to prove that the $O(n^3)$ bound is in fact tight. More specifically, we give an algorithm that given a context-free grammar $G$ of size $n$, produces a test set $T$ whose size is $O(n^3)$. We thus greatly improve the original $O(n^5)$ upper bound $[1][2][3].$

2 Notations and Definitions
2.1 Grammars
A context-free grammar $G$ is a tuple $(N, \Sigma, R, S)$ where:
- $N$ is a set of non-terminals,
- $\Sigma$ is a set of terminals,
- $R \subseteq N \times (N \cup \Sigma)^*$ is a set of production rules,
- $S \in N$ is the starting non-terminal symbol.

A production $(A, \text{rhs}) \in R$ is denoted $A \rightarrow \text{rhs}$. The size of $G$, denoted $|G|$, is the sum of sizes of each production in $R$: $\sum_{A \rightarrow \text{rhs} \in R} (|\text{rhs}| + 1)$.

By an abuse of notation, we denote by $G$ the set of words produced by $G$.

A grammar is linear if for every for every production $A \rightarrow \text{rhs} \in R$, the $\text{rhs}$ string contains at most one occurrence from $N$.

2.2 Morphisms and Test Sets
Given a (partial) function from $f : A \rightarrow B$, and a set $C$, $f|_C$ denotes the (partial) function $g : A \cap C \rightarrow B$ such that $g(a) = f(a)$ for all $a \in A \cap C$.

A morphism $f : \Sigma^* \rightarrow \Gamma^*$ is a function such that $f(\epsilon) = \epsilon$ and for every $u, v \in \Sigma^*$, $f(u \cdot v) = f(u) \cdot f(v)$, where the symbol `$\cdot$' denotes the concatenation of words.

A subset $T \subseteq L$ of a language $L$ is a test set if for any two morphisms $f, g : \Sigma^* \rightarrow \Gamma^*$, $f|_T = g|_T$ implies $f|_L = g|_L$.

3 Test Sets for Context-Free Languages
3.1 Plandowski’s Test Set
The following lemma was originally used $[1][2]$ to show that, for any linear context-free grammar, there exists a test set containing at most $O(|R|^6)$ elements. We show in Section $3.2$ how this lemma can be used to show a $2|R|^3$ bound.

Let $\Sigma_4 = \{a_i, \overline{a}_i, b_i, \overline{b}_i | i \in \{1, 2, 3, 4\}\}$ be an alphabet. We define:

$$L_4 = \{x_1 x_3 x_2 x_1 : x_1 x_2 x_3 x_4 \}$$

$$\forall i \in \{1, 2, 3, 4\}, (x_i, \overline{x}_i) = (a_i, \overline{a}_i) \lor (x_i, \overline{x}_i) = (b_i, \overline{b}_i)$$

and $T_3 = L_4 \setminus \{b_4 b_3 b_2 b_1 \overline{b}_1 \overline{b}_2 \overline{b}_3 \overline{b}_4\}$.

The sets $L_4, T_3 \subseteq \Sigma_4$ have 16 and 15 elements respectively.

Lemma 1 $[1][2]$. $T_3$ is a test set for $L_4$.

3.2 Linear Context-Free Grammars
We now prove that for any context-free grammar $G$, there exists a test set whose size is $2|R|^3$. Like the original proof of $[1][2]$ that gave a $O(|R|^6)$ upper bound, our proof relies on Lemma 1. However, our proof uses a different construction to obtain the new, tight, bound.

Theorem 1. Let $G = (N, \Sigma, R, S)$ be a linear context-free grammar. There exists a test set $T \subseteq G$ for $G$ containing at most $2|R|^3$ elements.

Proof. Before building the test set, we introduce some notation.

Graph of $G$. Define the labeled graph $\text{graph}(G) = (V, E)$ where $V = N \cup \{\bot\}$, $E \subseteq V \times V$ such that:
- for non-terminals $A, B \in N$ and a rule $r \in R$, let $(A, r, B) \in E$ iff $r$ is of the form $A \rightarrow u B v$ where $u, v \in \Sigma^*$ (i.e., $B$ is the only non-terminal occurring in $\text{rhs}$).
- for a non-terminal $A \in N$ and $r \in R$, $(A, r, \bot) \in E$ if and only if $r = A \rightarrow \text{rhs}$ for some $\text{rhs} \in \Sigma^*$.

A path of $\text{graph}(G)$ is a (possibly cyclic) sequence of edges of $E$, of the form: $(A_1, r_1, A_2) \cdot (A_2, r_2, A_3) \cdots (A_n, r_n, A_{n+1})$. A path is accepting if $A_1 = S$ and $A_{n+1} = \bot$.
Figure 1: The four optimal subpaths $Q_1$, $Q_2$, $Q_3$, and $Q_4$ define 15 alternative paths from $S$ to $\perp$ which are all strictly smaller (with respect to order $<$) than $P_{e_1}P_{e_2}P_{e_3}P_{e_4}W_5$.

**Link between $\text{graph}(G)$ and $G$.** Given a rule $A \rightarrow uBv \in R$, where $A, B \in N$ and $u, v \in \Sigma^*$, we denote $\pi(r) = u$ and $\overline{\pi}(r) = v$. For a rule of the form $A \rightarrow u$ where $u \in \Sigma^*$ we denote $\pi(r) = u$ and $\overline{\pi}(r) = \epsilon$. For a path $P = (A_1, r_1, A_2) \cdot (A_2, r_2, A_3) \cdot \cdots (A_n, r_n, A_{n+1})$ we define $\pi(P) = \pi(r_1) \cdots \pi(r_n)$, and $\overline{\pi}(P) = \overline{\pi}(r_n) \cdots \overline{\pi}(r_1)$.

Each accepting path $P$ in $\text{graph}(G)$ corresponds to a word $\pi(P)$, and $\overline{\pi}(P)$ in $G$, and conversely, for any word $w \in G$, there exists an accepting path (not necessarily unique) in $\text{graph}(G)$ corresponding to $w$.

**Total order on paths.** We fix an arbitrary total order $< \in R$, and extend it to sequence of edges in $R^\ast$ as follows. Given paths $P_1, P_2 \in R^\ast$, we have $P_1 < P_2$ iff

- $|P_1| < |P_2|$ (length of $P_1$ is smaller than length of $P_2$), or
- $|P_1| = |P_2|$ and $P_1$ is smaller lexicographically than $P_2$.

A path $P$ is called **optimal** if it is the minimal path from the first vertex of $P$ to the last vertex of $P$.

**Test set for $G$.** Let $\Phi_b(G)$ be the set of words of $G$ corresponding to accepting paths of the form $P_{e_1}P_{e_2} \cdots P_{e_n}P_{e_{n+1}}$ where $P_i \in R^\ast$, $e_i \in E$, and for $i \in \{1, \ldots, n+1\}$, $P_i$ is optimal, and for $i \in \{1, \ldots, n\}$, $P_i$ is not optimal. By construction, a path in $\Phi_b(G)$ is uniquely determined (when it exists) by the choice of edges $e_1, \ldots, e_n$, as optimal paths between two vertices are unique. Therefore, $\Phi_b(G)$ contains at most $\sum_{k=0}^n |R|^k \leq 2|R|^n$ words.

We now show that $\Phi_b(G)$ is a test set for $G$ (which gives us the desired bound of the theorem: $2|R|^n$). Assume there exist two morphisms $f, g : \Sigma^* \rightarrow \Gamma^*$ such that $f|_{\Phi_b(G)} = g|_{\Phi_b(G)}$ and there exists $w \in G$ such that $f(w) \neq g(w)$.

By assumption, $w$ does not belong to $\Phi_b(G)$, and must correspond to a path $P = P_{e_1}P_{e_2} \cdots P_{e_n}P_{e_{n+1}}$ for $n \geq 4$, such that for $i \in \{1, \ldots, n+1\}$, $P_i$ is optimal, and $P_{e_i}$ is not optimal. We pick $w$ having the property $f(w) \neq g(w)$ such that the path $P$ is the smallest possible (according to the order $<$ defined above).

The path $P$ can be written $P_{e_1}P_{e_2}P_{e_3}P_{e_4}W_5$, where for $i \in \{1, 2, 3, 4\}$, $P_i$ is optimal, and $P_{e_i}$ is not optimal ($W_5$ is not necessarily optimal). For $i \in \{1, 2, 3\}$, we define $Q_i$ to be the optimal path from the source of $P_{e_i}$ to its target; hence, $Q_i < P_{e_i}$. Moreover, $Q_4$ is defined to be the optimal path from the source of $P_{e_4}W_5$ to its target, with $Q_4 < P_{e_4}W_5$. Effectively, as shown in Figure 1 this defines 15 paths that can be derived from $P$ by replacing subpaths by their corresponding optimal path ($Q_1$, $Q_2$, $Q_3$, $Q_4$).

Let $P'$ be one of those 15 paths (where at least one subpath has been replaced by its optimal counterpart $Q_1$, $Q_2$, $Q_3$, or $Q_4$), and let $w' \in G$ be the word corresponding to $P'$. By construction of $P'$, and by definition of the order $<$, we have $P' < P$. Since we have chosen $P$ to be the optimal path such that $f$ and $g$ are not equal on the corresponding word, we deduce that $f(w') = g(w')$.

To conclude, we show that we obtain a contradiction, thanks to Lemma 1. For this, we construct two morphisms $f', g' : \Sigma_4 \rightarrow \Gamma$ as follows ($i$ ranges over $\{1, 2, 3, 4\}$ and $j$ over $\{1, 2, 3\}$):

- $f'(a_i) = f(\pi(Q_i))$,
- $f'(b_i) = f(\pi(P_{e_i}))$,
- $f'(b_j) = f(\pi(P_{e_j}))$,
- $f'(b_4) = f(\pi(P_{e_4}W_5))$,
- $f'(b_4) = f(\overline{\pi}(P_{e_4}W_5))$.

The morphism $g'$ is defined similarly, using $g$ instead of $f$. We can then verify that $f'$ and $g'$ coincide on $T_4$, but are not equal on the word $b_4b_3b_2b_1b_2b_3b_4 \in L_4$, thus contradicting Lemma 1. $\square$

### 3.3 Context-Free Grammars

To obtain a test set for a context-free grammar $G$ which is not necessarily linear, $\Pi$ constructs from $G$ a linear context-free grammar $\text{Lin}(G)$ which produces a subset of $G$ which is a test set for $G$.

Formally, $\text{Lin}(G)$ is derived from $G$ as follows:

- For every productive non-terminal symbol $A$ in $G$, we choose a word $w_A$ that is produced by $A$.
- Every rule $r : A \rightarrow x_0 \cdots x_n$ in $G$, where for every $i$, $x_i \in \Sigma$ and $A_i$ in $N$ is productive, is replaced by $n$ different rules, each one obtained from $r$ by replacing all $A_i$ with $w_{A_i}$ except one.

Note that the definition of $\text{Lin}(G)$ is not unique, and depends on the choice of the words $w_A$. The following result holds for any choice of the words $w_A$.

**Lemma 2 (1.2).** $\text{Lin}(G)$ is a test set for $G$.

Using Theorem 1 we improve the $O(|G|^6)$ bound of Lemma 2 for the test set of $G$ to $2|G|^5$.

**Theorem 2.** Let $G = (N, \Sigma, R, S)$ be a context-free grammar. There exists a test set $T \subseteq G$ for $G$ containing at most $2|G|^3$ elements.

**Proof.** Follows from Theorem 1 and from the fact that $\text{Lin}(G)$ has at most $|G| = \sum_{A \rightarrow \text{rhs} \in R} |\text{rhs}| + 1$ rules, when constructing $\text{Lin}(G)$, each rule $A \rightarrow \text{rhs}$ of $G$ is duplicated at most $|\text{rhs}|$ times.) $\square$

### 3.4 Construction of $\Phi_3(G)$

To construct $\Phi_3(G)$ for a linear context-free grammar $G = (N, \Sigma, R, S)$, we precompute in time $O(|N|^2|R|$), for each pair of vertices $(A, B)$, the optimal path from $A$ to $B$ in $\text{graph}(G)$. Then for each possible choice of at most 3 edges $e_1 \leftarrow A_1 \rightarrow \cdots \rightarrow A_{n+1}$, with $0 \leq n \leq 3$, we construct the path $P = P_{e_1} \cdots P_{e_{n+1}}$ where each $P_i$ is the optimal path from $A_{i-1}$ to $B_i$ (if it exists) with $A_0 = S$ and $B_{n+1} = \perp$ by construction. We then add the word corresponding to $P$ to our result.

To conclude, since the length of each optimal path is bounded by $|N|$, we can construct $\Phi_3(G)$ in time $O(|N| \cdot |R|^3)$.

### 4 Acknowledgements

Thanks to Viktor Kuncak, Mukund Raghothaman, and Ravichandran Madhavan for the helpful talks.

**References**

[1] W. Plandowski. Testing equivalence of morphisms on context-free languages. In *European Symposium on Algorithms*, pages 460–470. Springer, 1994.

[2] W. Plandowski. The complexity of the morphism equivalence problem for context-free languages. 1995.

[3] W. Plandowski. Test sets for large families of languages. In *International Conference on Developments in Language Theory*, pages 75–94. Springer, 2003.