Streaming Property Testing of Visibly Pushdown Languages*

Nathanaël François[†1], **Frédéric Magniez**[‡2], **Michel de Rougemont**[§3], **and Olivier Serre**[¶2]

1 **Fakultät für Informatik, TU Dortmund, Germany**
2 **CNRS, IRIF, Univ Paris Diderot, Sorbonne Paris-Cité, France**
3 **University of Paris II and IRIF, CNRS, France**

Abstract

In the context of formal language recognition, we demonstrate the superiority of streaming property testers against streaming algorithms and property testers, when they are not combined. Initiated by Feigenbaum *et al.*, a streaming property tester is a streaming algorithm recognizing a language under the property testing approximation: it must distinguish inputs of the language from those that are ε-far from it, while using the smallest possible memory (rather than limiting its number of input queries). Our main result is a streaming ε-property tester for visibly pushdown languages (VPL) with memory space $\mathrm{poly}((\log n)/\varepsilon)$.

Our construction is done in three streps. First, we simulate a visibly pushdown automaton in one pass using a stack of small height but whose items can be of linear size. In a second step, those items are replaced by small sketches. Those sketches rely on a notion of suffix-sampling we introduce. This sampling is the key idea for taking benefit of both streaming algorithms and property testers in the third step. Indeed, the last step relies on a (non-streaming) property tester for weighted regular languages based on a previous tester by Alon *et al*. This tester can directly be used for streaming testing special cases of instances of VPL that are already hard for both streaming algorithms and property testers. We then use it to decide the correctness of completed items, given their sketches, before removing them from the stack.

Keywords and phrases Streaming Algorithm, Property Testing, Visibly Pushdown Languages

1 Introduction

We focus on streams representing data with both a linear ordering and a hierarchically nested matching of items. Data with such dual linear-hierarchical structure arise in various context, *e.g.* in semi-structured data management when handling HTML/XML documents or in program analysis when considering executions of recursive programs. Regular languages, as recognised by finite state automata, revealed a natural and successful tool to express properties of streams but lack the ability to handle the hierarchical structure. Context-free languages easily capture the latter but turn out to be too expressive hence, quickly lead to intractable complexity. In contrast, visibly pushdown languages (VPL) [6] while encompassing regular languages, enjoy most of its good properties and permit to handle data with both a linear and a hierarchical structure. In the context of semi-structured documents, they are closely related with regular languages of unranked trees as captured by hedge automata: indeed, a well-known result [3] states that, when the tree is given by its depth-first traversal, such automata correspond to visibly pushdown automata (VPA) (see *e.g.* [19] for an overview on automata and logic for unranked trees). In databases, this word encoding of XML document is known as SAX

* Partially supported by the French ANR projects ANR-12-BS02-005 (RDAM) and ANR-14-CE25-0017 (AGREG)
† nathanael.francois@tu-dortmund.de
‡ frederic.magniez@cnrs.fr
§ mdr@liafa.univ-paris-diderot.fr
¶ Olivier.Serre@cnrs.fr

Conference title on which this volume is based on.
Editors: Billy Editor and Bill Editors; pp. 1–25

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

representation: the document is a linear sequence of text characters, along with a hierarchically nested matching of open-tags with closing tags. Numerous popular subclasses of XML documents (*e.g.* those satisfying a given DTD specifications) are subclasses of VPL. In program analysis, VPA permit to capture natural properties of execution traces of recursive finite-state programs. For such programs, desirable specifications are expressed on the call-stack (*e.g.* "a module A should be invoked only if the module B belongs to the call-stack"): such properties can be expressed in the temporal logic of calls and returns (CaRet) [5, 4] that itself is captured by VPA. Hence, the analysis of execution traces boils down to check membership in a VPL. Therefore, the study of VPL is central to understand how massive semi-structured data (*e.g.* large semi-structured documents or execution traces) can be analyzed by sublinear algorithms, such as streaming algorithms and property testers.

Historically, VPL got several names such as input-driven languages or, more recently, languages of nested words. Intuitively, a VPA is a pushdown automaton whose actions on stack (push, pop or nothing) are solely decided by the currently read symbol. As a consequence, symbols can be partitioned into three groups: push, pop and neutral symbols. The complexity of VPL recognition has been addressed in various computational models. The first results go back to the design of logarithmic space algorithms [11] as well as NC^1-circuits [13]. Later on, other models motivated by the context of massive data were considered, such as streaming algorithms and property testers (described below).

Streaming algorithms (see *e.g.* [23]) have only a sequential access to their input, on which they can perform a single pass, or sometimes a small number of additional passes. The size of their internal (random access) memory is the crucial complexity parameter, which should be sublinear in the input size, and even polylogarithmic if possible. The area of streaming algorithms has experienced tremendous growth in many applications since the late 1990s. The analysis of Internet traffic [2], in which traffic logs are queried, was one of their first applications. Nowadays, they have found applications with big data, notably to test graphs properties, and more recently in language recognition on very large inputs. The streaming complexity of language recognition has been firstly considered for languages that arise in the context of memory checking [8, 12], of databases [29, 28], and later on for formal languages [21, 7]. However, even for simple VPL, any randomized streaming algorithm with p passes requires memory $\Omega(n/p)$, where n is the input size [18].

As opposed to streaming algorithms, (standard) property testers [9, 10, 16] have random access to their input but in the query model. They must query each piece of the input they need to access. They should sample only a sublinear fraction of their input, and ideally make a constant number of queries. In order to make the task of verification possible, decision problems need to be approximated as follows. Given a distance on words, an ε-tester for a language L distinguishes with high probability the words in L from those ε-far from L, using as few queries as possible. Property testing of regular languages was first considered for the Hamming distance [1]. When the distance allows sufficient modifications of the input, such as moves of arbitrarily large factors, it has been shown that any context-free language becomes testable with a constant number of queries [20, 15]. However, for more realistic distances, property testers for simple languages require a large number of queries, especially if they have one-sided error only. For example the complexity of an ε-tester for well-parenthesized expressions with two types of parentheses is between $\Omega(n^{1/11})$ and $\mathrm{O}(n^{2/3})$ [26], and it becomes linear, even for one type of parentheses, if we require one-sided error [1]. The difficulty of testing regular tree languages was also addressed when the tester can directly query the tree structure [24, 25].

Faced by the intrinsic hardness of VPL in both streaming and property testing, we study the complexity of *streaming property testers* of formal languages, a model of algorithms combining both approaches. Such testers were historically introduced for testing specific problems (groupedness) [14] relevant for network data. They were later studied in the context of testing the insert/extract-sequence of a priority-queue structure [12]. We extend these studies to classes of problems. A streaming property tester is a streaming algorithm recognizing a language under the property testing approximation:

it must distinguish inputs of the language from those that are ε-far from it, while using the smallest possible memory (rather than limiting its number of input queries). Such an algorithm can simulate any standard non-adaptive property tester. Moreover, we will see that, using its full scan of the input, it can construct better sketches than in the query model.

In this paper, we consider a natural notion of distance for VPL, the *balanced-edit distance*, which refines the edit distance on *balanced words* (where for each push symbol there is a matching pop symbol at the same height of the stack, and conversely). It can be interpreted as the edit distance on trees when trees are encoded as balanced words. Neutral symbols can be deleted/inserted, but any push symbol can only be deleted/inserted together with its matching pop symbol. Since our distance is larger than the standard edit distance, our testers are also valid for the edit distance.

In Section 3, we first design an exact algorithm that maintains a small stack but whose items can be of linear size as opposed to the standard simulation of a pushdown automaton which usually has a stack of possible linear size but with constant size items. In our algorithm, stack items are prefixes of some peaks (which we call unfinished peaks), where a *peak* is a balanced factor whose push symbols appear all before the first pop symbol. Our algorithm compresses an unfinished peak $u = u_+v_-$ when it is followed by a long enough sequence. More precisely, the compression applies to the peak v_+v_- obtained by disregarding part of the prefix of push sequence u_+. Those peaks are then inductively replaced, and therefore compressed, by the state-transition relation they define on the given automaton. The relation is then considered as a single symbol whose weight is the size of the peak it represents. In addition, to maintain a stack of logarithmic depth, one of the crucial properties of our algorithm (**Proposition 5**) is rewriting the input word as a peak formed by potentially a linear number of intermediate peaks, but with only a logarithmic number of nested peaks.

In Section 4, for the case of a single peak, we show how to sketch the current unfinished peak of our algorithm. The simplicity of those instances will let us highlight our first idea. Moreover, they are already expressive enough in order to demonstrate the superiority of streaming testers against streaming algorithms and property testers, when they are not combined. We first reduce the problem of streaming testing such instances to the problem of testing regular languages in the standard model of property testing (**Theorem 15**). Since our reduction induces weights on the letters of the new input word, we need a tester for weighted regular languages (**Theorem 26**). Such a property tester has previously been devised in [25] extending constructions for unweighted regular languages [1, 24]. However, we consider a slightly simpler construction that could be of independent interest. As a consequence we get a streaming property tester with polylogarithmic memory for recognizing peak instances of any given VPL (**Theorem 16**), a task already hard for streaming algorithms and property testers (**Fact 7**).

In Section 5, we construct our main tester for a VPL L given by some VPA. For this we introduce a more involved notion of sketches made of a polylogarithmic number of samples. They are based on a new notion of suffix sampling (**Definition 17**). This sampling consists in a decomposition of the string into an increasing sequence of suffixes, whose weights increase geometrically. Such a decomposition can be computed online on a data stream, and one can maintain samples in each suffix of the decomposition using a standard reservoir sampling. This suffix decomposition will allow us to simulate an appropriate sampling on the peaks we compress, even if we do not yet know where they start. Our sampling can be used to perform an approximate computation of the compressed relation by our new property tester of weighted regular languages which we also used for single peaks. We first establish a result of stability which basically states that we can assume that our algorithm knows in advance where the peak it will compress starts (**Lemma 22**). Then we prove the robustness of our algorithm: words that are ε-far from L are rejected with high probability (**Lemma 24**). As a consequence, we get a one-pass streaming ε-tester for L with one-sided error η and memory space $\mathrm{O}(m^5 2^{3m^2} (\log n)^6 (\log 1/\eta)/\varepsilon^4)$, where m is the number of states of a VPA

Algorithm 1 Reservoir Sampling

```
Input: Data stream u, Integer t > 1 standing for the number of samples
Data structure:
  σ ← 0 // Current weight of the processed stream
  S ← empty multiset // Multiset of sampled letters
Code:
a ← Next(u), σ ← |a|
S ← t copies of a
While u not finished
   a ← Next(u), σ ← σ + |a|
   For each b ∈ S
      Replace b by a with probability |a|/σ
Output S
```

recognizing L (**Theorem 20**).

2 Definitions and Preliminaries

Let $\mathbb{N}^*$ be the set of positive integers, and for any $n \in \mathbb{N}^*$, let $[n] = \{1, 2, \ldots, n\}$. A t-subset of a set S is any subset of S of size t. For a finite alphabet Σ we denote the set of finite words over Σ by Σ^*. We denote by $u \cdot v$ (or simply uv) the word obtained by concatenating u and v. For a word $u = u(1)u(2)\cdots u(n)$, we call n the *length* of u, and $u(i)$ the ith letter in u. A *factor* of u is a word $u[i,j] = u(i)u(i+1)\cdots u(j)$ with $1 \leq i \leq j \leq n$. When we mention letters and factors of u we implicitly also mention their positions in u. We say that v is a *sub-factor* of v', denoted $v \leq v'$, if $v = u[i,j]$ and $v' = u[i',j']$ with $[i,j] \subseteq [i',j']$. Similarly we say that $v = v'$ if $[i,j] = [i',j']$. If $i \leq i' \leq j \leq j'$ we say that the *overlap* of v and v' is $u[i',j]$. If v is a sub-factor of v' then the overlap of v and v' is v. Given two multisets of factors S and S', we say that $S \leq S'$ if for each factor $v \in S$ there is a corresponding factor $v' \in S'$ such that $v \leq v'$.

2.1 Weighted Words and Sampling

A *weight function* on a word u with n letters is a function $\lambda : [n] \to \mathbb{N}^*$ on the letters of u, whose value $\lambda(i)$ is called the *weight of* $u(i)$. A *weighted word* over Σ is a pair (u, λ) where $u \in \Sigma^*$ and λ is a weight function on u. We define $|u(i)| = \lambda(i)$ and $|u[i,j]| = \lambda(i) + \lambda(i+1) + \ldots + \lambda(j)$. The length of (u, λ) is the length of u. For simplicity, we will denote by u the weighted word (u, λ). Weighted letters will be used to substitute factors of same weights.

Our algorithms will be based on sampling of small factors according to their weights. We introduce a very specific notion adapted to our setting. For a weighted word u, we denote by *k-factor sampling on* u the sampling over factors $u[i, i+l]$ with probability $|u(i)|/|u|$, where $l \geq 0$ is the smallest integer such that $|u[i, i+l]| \geq k$ if it exists, otherwise l is such that $i + l$ is the last letter of u. More generally, we call *k-factor* such a factor. For the special case of $k = 1$, we call this sampling a *letter sampling on* u. In fact the general case $k > 1$ simply reduces to $k = 1$. Indeed, simply observe that k-factor sampling can be obtained from letter sampling by sampling on the first letters of the factors and online completing any sampled letter to produce its associated k-factor. Therefore, from now on, we only focus how to perform letter samplings, that we implicitly extend to samplings on k-factors when required. In particular, without further constraints, a letter sampling can be implemented using a standard reservoir sampling (see Algorithm 1).

Even if our algorithm will require several samples from a k-factor sampling, we will often only

be able to simulate this sampling by sampling either larger factors, more factors, or both. Let $\mathcal{W}_1$ be a sampler producing a random multiset S_1 of factors of some given weighted word u. Then $\mathcal{W}_2$ *over-samples* $\mathcal{W}_1$ if it produces a random multiset S_2 of factors of u such that for each factor v of u, we have $\Pr(\exists v' \in S_2$ such that v is a factor of $v') \geq \Pr(\exists v' \in S_1$ such that v is a factor of $v')$.

2.2 Finite State Automata and Visibly Pushdown Automata

A *finite state automaton* is a tuple of the form $\mathcal{A} = (Q, \Sigma, Q_{in}, Q_f, \Delta)$ where Q is a finite set of control states, Σ is a finite input alphabet, $Q_{in} \subseteq Q$ is a subset of initial states, $Q_f \subseteq Q$ is a subset of final states and $\Delta \subseteq Q \times \Sigma \times Q$ is a transition relation. We write $p \xrightarrow{u} q$, to mean that there is a sequence of transitions in $\mathcal{A}$ from p to q while processing u, and we call (p, q) a *u-transitions*. A word u is accepted if $q_{in} \xrightarrow{u} q_f$ for some $q_{in} \in Q_{in}$ and $q_f \in Q_f$. The language $L(\mathcal{A})$ of $\mathcal{A}$ is the set of words accepted by $\mathcal{A}$, and we refer to such a language as a *regular language*. For $\Sigma' \subseteq \Sigma$, the *Σ'-diameter* (or simply *diameter* when $\Sigma' = \Sigma$) of $\mathcal{A}$ is the maximum over all possible pairs $(p, q) \in Q^2$ of $\min\{|u| : p \xrightarrow{u} q$ and $u \in \Sigma'^*\}$, whenever this minimum is not over an empty set. We say that $\mathcal{A}$ is *Σ'-closed*, when $p \xrightarrow{u} q$ for some $u \in \Sigma^*$ if and only if $p \xrightarrow{u'} q$ for some $u' \in \Sigma'^*$.

A *pushdown alphabet* is a triple $\langle \Sigma_+, \Sigma_-, \Sigma_= \rangle$ that comprises three disjoint finite alphabets: Σ_+ is a finite set of *push symbols*, Σ_- is a finite set of *pop symbols*, and $\Sigma_=$ is a finite set of *neutral symbols*. For any such triple, let $\Sigma = \Sigma_+ \cup \Sigma_- \cup \Sigma_=$. Intuitively, a *visibly pushdown automaton* [27] over $\langle \Sigma_+, \Sigma_-, \Sigma_= \rangle$ is a pushdown automaton restricted so that it pushes onto the stack only on reading a push, it pops the stack only on reading a pop, and it does not modify the stack on reading a neutral symbol. Up to coding, this notion is similar to the one of input driven pushdown automata [22] and of nested word automata [6].

▶ **Definition 1.** A *visibly pushdown automaton* (VPA) over $\langle \Sigma_+, \Sigma_-, \Sigma_= \rangle$ is a tuple $\mathcal{A} = (Q, \Sigma, \Gamma, Q_{in}, Q_f, \Delta)$ where Q is a finite set of states, $Q_{in} \subseteq Q$ is a set of initial states, $Q_f \subseteq Q$ is a set of final states, Γ is a finite stack alphabet, and $\Delta \subseteq (Q \times \Sigma_+ \times Q \times \Gamma) \cup (Q \times \Sigma_- \times \Gamma \times Q) \cup (Q \times \Sigma_= \times Q)$ is the transition relation.

To represent stacks we use a special bottom-of-stack symbol $\perp$ that is not in Γ. A *configuration* of a VPA $\mathcal{A}$ is a pair (σ, q), where $q \in Q$ and $\sigma \in \perp \cdot \Gamma^*$. For $a \in \Sigma$, there is an *a-transition* from a configuration (σ, q) to (σ', q'), denoted $(\sigma, q) \xrightarrow{a} (\sigma', q')$, in the following cases:

- If a is a push symbol, then $\sigma' = \sigma\gamma$ for some $(q, a, q', \gamma) \in \Delta$, and we write $q \xrightarrow{a} (q', \mathsf{push}(\gamma))$.
- If a is a pop symbol, then $\sigma = \sigma'\gamma$ for some $(q, a, \gamma, q') \in \Delta$, and we write $(q, \mathsf{pop}(\gamma)) \xrightarrow{a} q'$.
- If a is a neutral symbol, then $\sigma = \sigma'$ and $(q, a, q') \in \Delta$, and we write $q \xrightarrow{a} q'$.

For a finite word $u = a_1 \cdots a_n \in \Sigma^*$, if $(\sigma_{i-1}, q_{i-1}) \xrightarrow{a_i} (\sigma_i, q_i)$ for every $1 \leq i \leq n$, we also write $(\sigma_0, q_0) \xrightarrow{u} (\sigma_n, q_n)$. The word u is *accepted* by a VPA if there is $(p, q) \in Q_{in} \times Q_f$ such that $(\perp, p) \xrightarrow{u} (\perp, q)$. The language $L(\mathcal{A})$ of $\mathcal{A}$ is the set of words accepted by $\mathcal{A}$, and we refer to such a language as a *visibly pushdown language* (VPL).

At each step, the height of the stack is pre-determined by the prefix of u read so far. The *height* $\mathrm{height}(u)$ of $u \in \Sigma^*$ is the difference between the number of its push symbols and of its pop symbols. A word u is *balanced* if $\mathrm{height}(u) = 0$ and $\mathrm{height}(u[1, i]) \geq 0$ for all i. We also say that a push symbol $u(i)$ *matches* a pop symbol $u(j)$ if $\mathrm{height}(u[i, j]) = 0$ and $\mathrm{height}(u[i, k]) > 0$ for all $i < k < j$. By extension, the height of $u(i)$ is $\mathrm{height}(u[1, i-1])$ when $u(i)$ is a push symbol, and $\mathrm{height}(u[1, i])$ otherwise.

For all balanced words u, the property $(\sigma, p) \xrightarrow{u} (\sigma, q)$ does not depend on σ, therefore we simply write $p \xrightarrow{u} q$, and say that (p, q) is a *u-transition*. We also define similarly to finite automata the *Σ'-diameter* of $\mathcal{A}$ (or simply diameter) and the notion $\mathcal{A}$ being *Σ'-closed* on balanced words only.

Our model is inherently restricted to input words having no prefix of negative stack height, and we defined acceptance with an empty stack. This implies that only balanced words can be accepted. From now on, we assume that the input is balanced as verifying this in a streaming context is easy.

2.3 Streaming Property Testers

Assume we have, for any $\varepsilon > 0$, a criterion to declare that an input u is ε-far from a language L. An ε-tester for L accepts all inputs in L with probability 1 and rejects with high probability all inputs ε-far from L. Two-sided error testers have also been studied but in this paper we stay with the notion of one-sided testers, that we adapt in the context of streaming algorithm as in [14].

▶ **Definition 2.** Let $\varepsilon > 0$ and let L be a language. A *streaming ε-tester* for L with one-sided error η and memory $s(n)$ is a randomized algorithm A such that, for any input u of length n given as a data stream:

- If $u \in L$, then A accepts with probability 1;
- If u is ε-far from L, then A rejects with probability at least $1 - \eta$;
- A processes u within a single sequential pass while maintaining a memory space of $O(s(n))$ bits.

Even if we only focus on the space complexity of streaming testers, all our streaming testers have polylogarithmic (in n/ε) time per processing letter.

For a distance d between words, we say that a word u is ε-far from a language L if $d(u, v) > \varepsilon|u|$ for every $v \in L$, *i.e.* the ε-neighbourhood of u does not intersect L. Hence, any distance on words leads to a notion of streaming property tester. Remark that any ε-tester for some distance d_1 turns out to be also a $(c\varepsilon)$-tester for any other distance d_2 such that $d_2 \leq cd_1$, where $c > 0$ is some constant.

2.4 Balanced/Standard Edit Distance

The usual distance between words in property testing is the Hamming distance. In this work, we consider an easier distance to manipulate in property testing but still relevant for most applications, which is the edit distance, that we adapt to weighted words.

Given a word u, we define two possible *edit operations*: the *deletion* of a letter in position i with corresponding cost $|u(i)|$, and its converse operation, the *insertion* where we also select a weight for the new $u(i)$. Note that, for simplicity, we drop the usual substitution operation, leading to a possible multiplicative factor of 2 in the resulting distance. This is not an issue when designing streaming property testers as observed above. The *(standard) edit distance* $\mathsf{dist}(u, v)$ between two weighted words u and v is defined as the minimum total cost of a sequence of edit operations changing u to v. All letters that have not been inserted nor deleted must keep the same weight. For a restricted set of letters Σ', we define $\mathsf{dist}_{\Sigma'}(u, v)$ when insertions are restricted to letters in Σ'.

We will also consider a restricted version of this distance for balanced words, motivated by our study of VPL. Similarly, *balanced-edit operations* can be deletions or insertions of letters, but each deletion of a push symbol (resp. pop symbol) requires the deletion of the matching pop symbol (resp. push symbol). Similarly for insertions: if a push (resp. pop) symbol is inserted, then a matching pop (resp. push) symbol must also be inserted simultaneously. The cost of these operations is the weight of the affected letters, as with the edit operations. We define the *balanced-edit distance* $\mathsf{bdist}(u, v)$ between two balanced words as the total cost of a sequence of balanced-edit operations changing u to v. Similarly to $\mathsf{dist}_{\Sigma'}(u, v)$ we define $\mathsf{bdist}_{\Sigma'}(u, v)$. We omit Σ' when $\Sigma' = \Sigma$.

When dealing with a visibly pushdown language, we will always use the balanced-edit distance, whereas we will use the standard-edit distance for regular languages. Note that since balanced-edit distance is larger than the standard edit distance, our testers will also be valid for that distance.

3 Exact Algorithm

Fix a VPA $\mathcal{A}$ recognizing some VPL L on $\Sigma = \Sigma_+ \cup \Sigma_- \cup \Sigma_=$. In this section, we design an exact streaming algorithm that decides whether an input belongs to L. Algorithm 2 maintains a stack of small height but whose items can be of linear size. In Section 5, we replace stack items by appropriated small sketches

3.1 Notations and Algorithm Description

Call a *peak* a sequence of push symbols followed by an equal number of pop symbols, with possibly intermediate neutral symbols, *i.e.* an element of the language $\Lambda = \bigcup_{j\geq 0}((\Sigma_=)^* \cdot \Sigma_+)^j \cdot (\Sigma_=)^* \cdot (\Sigma_- \cdot (\Sigma_=)^*)^j$. One can compress any peak $v \in \Lambda$ by the set $R_v = \{(p,q) : p \xrightarrow{v} q\}$ of the v-transitions, and consider R_v as a new neutral symbol with weight $|v|$. In fact, for the purpose of the analysis of our algorithm, we augment neutral symbols by many more relations for which $\mathcal{A}$ remains Σ-closed. Indeed, we allow any relation R of any weight such that, when $(p,q) \in R$, there is a $v \in \Lambda$ such that $p \xrightarrow{v} q$, but that v could be different for every $(p,q) \in R$. For the rest of the paper, they will be the only symbols with weight potentially larger than 1.

▶ **Definition 3.** Let Σ_Q be $\Sigma_=$ augmented by all letters 'R' encoding a relation $R \subseteq Q \times Q$ such that for every $(p,q) \in R$ there is a balanced word $u \in \Sigma^*$ with $p \xrightarrow{u} q$. In addition we allow any weight $|R| \geq 1$ for those letters. Let Λ_Q be Λ where $\Sigma_=$ is replaced by Σ_Q.

We then write $p \xrightarrow{R} q$ whenever $(p,q) \in R$, and extend $\mathcal{A}$ and L accordingly. Of course, our notion of distance will be solely based on the initial alphabet Σ. If $R_1, R_2 \subseteq Q \times Q$ are two relations on Q we define their composition $R_1 \circ R_2$ to be $\{(x,z) \mid \exists y \text{ s.t. } (x,y) \in R_1 \text{ and } (y,z) \in R_2\}$.

A general balanced input instance u will consist of many nested peaks. However, we will recursively replace each factor $v \in \Lambda_Q$ by R_v with weight $|v|$.

Denote by $\mathrm{Prefix}(\Lambda_Q)$ the language of prefixes of words in Λ_Q. While processing the prefix $u[1,i]$ of the data stream u, Algorithm 2 maintains a suffix $u_0 \in \mathrm{Prefix}(\Lambda_Q)$ of $u[1,i]$, that is an unfinished peak, with some simplifications of factors v in Λ_Q by their corresponding relation R_v. Therefore u_0 consists of a sequence of push symbols and neutral symbols possibly followed by a sequence of pop symbols and neutral symbols. The algorithm also maintains a subset $R_{\texttt{temp}} \subseteq Q \times Q$ that is the set of transitions for the maximal prefix of $u[1,i]$ in Λ_Q. When the stream is over, the set $R_{\texttt{temp}}$ is used to decide whether $u \in L$ or not.

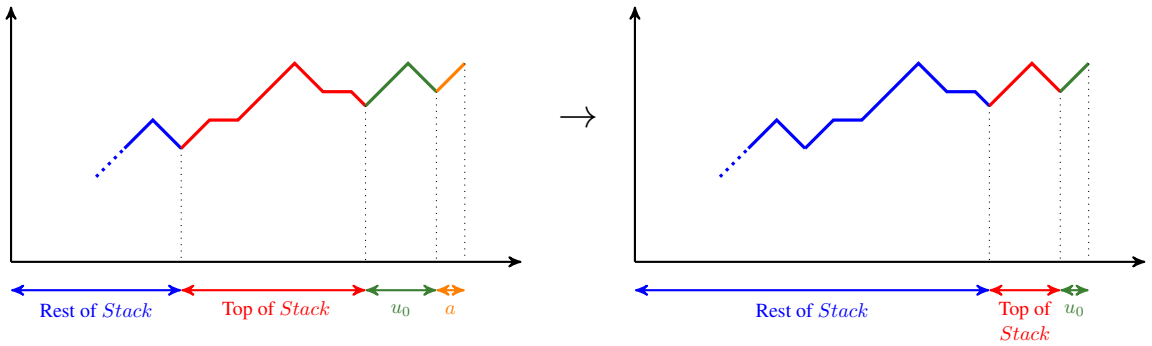
When a push symbol a comes after a pop sequence, $u_0 \cdot a$ is no longer in $\mathrm{Prefix}(\Lambda_Q)$ hence, Algorithm 2 puts u_0 on the stack of unfinished peaks (see lines 10 to 11 and Figure 1a) and u_0 is reset to a. In other situations, it adds a to u_0. In case u_0 becomes a word in Λ_Q (see lines 13 to 17 and Figure 1b), Algorithm 2 computes the set of u_0-transitions $R_{u_0} \in \Sigma_Q$, and adds R_{u_0} to the previous unfinished peak that is retrieved on top of the stack and becomes the current unfinished peak; in the special case where the stack is empty it simply updates $R_{\texttt{temp}}$ by taking its composition with R_{u_0}.

3.2 Algorithm Analysis

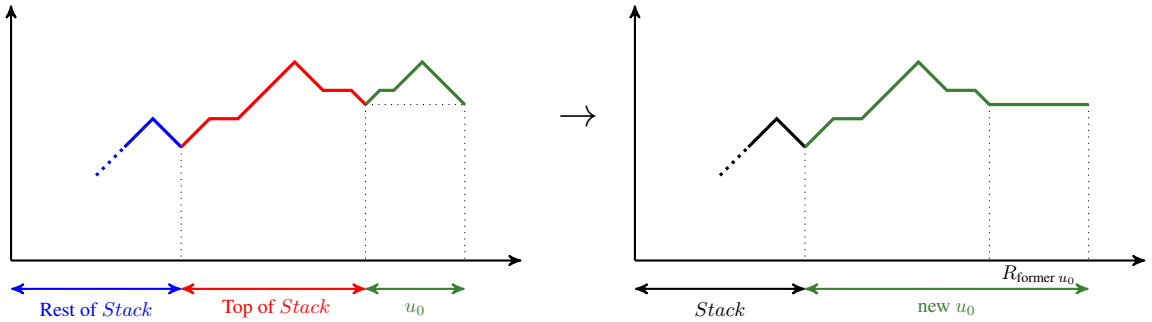
We now introduce the quantity $\mathrm{Depth}(v)$ for each factor v constructed in Algorithm 2. It quantifies the number of processed nested peaks in v as follows:

▶ **Definition 4.** For each factor constructed in Algorithm 2, Depth is defined dynamically by $\mathrm{Depth}(a) = 0$ when $a \in \Sigma$, $\mathrm{Depth}(v) = \max_i \mathrm{Depth}(v(i))$ and $\mathrm{Depth}(R_v) = \mathrm{Depth}(v) + 1$.

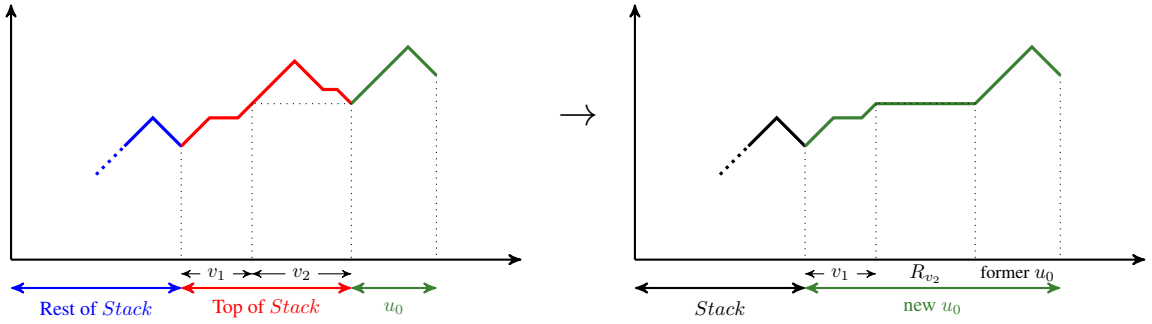
In order to bound the size of the stack, Algorithm 2 considers the maximal balanced suffix v_2 of the topmost element $v_1 \cdot v_2$ of the stack and, whenever $|u_0| \geq |v_2|/2$, it computes the relation



(a) Illustration of lines 10 to 11 from Algorithm 2



(b) Illustration of lines 13 to 17 from Algorithm 2



(c) Illustration of lines 18 to 20 from Algorithm 2

Figure 1 Illustration of Algorithm 2.

Algorithm 2 Exact Tester for a VPL

```
1  Input: Balanced data stream u
2  Data structure:
3  Stack ← empty stack // Stack of items v with v ∈ Prefix(Λ_Q)
4  u_0 ← ∅ // u_0 ∈ Prefix(Λ_Q) is a suffix of the processed part u[1,i] of u
5        // with possibly some factors v ∈ Λ_Q replaced by R_v
6  R_temp ← {(p,p)}_{p∈Q} // Set of transitions for the max. prefix of u[1,i] in Λ_Q
7  Code:
8  While u not finished
9    a ← Next(u) //Read and process a new symbol a
10   If a ∈ Σ_+ and u_0 has a letter in Σ_- // u_0 · a ∉ Prefix(Λ_Q)
11     Push u_0 on Stack, u_0 ← a
12   Else u_0 ← u_0 · a
13   If u_0 is balanced // u_0 ∈ Λ_Q: compression
14     Compute R_{u_0} the set of u_0-transitions
15     If Stack = ∅, then R_temp ← R_temp ∘ R_{u_0}, u_0 ← ∅
16        // where ∘ denotes the composition of relations
17     Else Pop v from Stack, u_0 ← v · R_{u_0}
18   Let (v_1 · v_2) ← top(Stack) s.t. v_2 is maximal and balanced // v_2 ∈ Λ_Q
19   If |u_0| ≥ |v_2|/2  // u_0 is big enough and v_2 can be replaced by R_{v_2}
20     Compute R_{v_2} the v_2-transitions, Pop v from Stack, u_0 ← (v_1 · R_{v_2}) · u_0
21 If (Q_in × Q_f) ∩ R_temp ≠ ∅, Accept; Else Reject // R_temp = R_u
```

R_{v_2} and continues with a bigger current peak starting with v_1 (see lines 18 to 20 and Figure 1c). A consequence of this compression is that the elements in the stack have geometrically decreasing weight and therefore the height of the stack used by Algorithm 2 is logarithmic in the length of the input stream. This can be proved by a direct inspection of Algorithm 2.

▶ **Proposition 5.** *Algorithm 2 accepts exactly when $u \in L$, while maintaining a stack of at most $\log |u|$ items.*

We state that Algorithm 2, when processing an input u of length n, considers at most $\mathrm{O}(\log n)$ nested peaks, that is $\mathrm{Depth}(v) = \mathrm{O}(\log n)$ for all factors constructed in Algorithm 2.

▶ **Lemma 6.** *Let v be the factor used to compute R_v at line either 14 or 20 of Algorithm 2. Then $|v(i)| \leq 2|v|/3$, for all i. Moreover, for any factor w constructed by Algorithm 2 it holds that $\mathrm{Depth}(w) = \mathrm{O}(\log |w|)$.*

Proof. One only has to consider letters in Σ_Q. Hence, let R_w belongs to v for some w: either w was simplified into R_w at line 14 or at line 20 of Algorithm 2.

Let us first assume that it was done at line 20. Therefore, there is some $v' \in \mathrm{Prefix}(\Lambda_Q)$ to the right of w with total weight greater than $|w|/2 = |R_w|/2$. This factor v' is entirely contained within v: indeed, when R_w is computed v includes v'. Therefore $|R_w| \leq 2|v|/3$.

If R_w comes from line 14, then $w = u_0$ and this u_0 is balanced and compressed. We claim that at the previous round the test in line 19 failed, that is $|u_0| - 1 \leq |v_2|/2$ where v_2 is the maximal balanced suffix of $\mathrm{top}(Stack)$. Indeed, when performing the sequence of actions following a positive test in line 19, the number of unmatched push symbols in the new u_0 is augmented at least by 1 from the previous u_0: hence, it cannot be equal to 1 as the elements in the stack have unmatched push symbols and therefore in the next round u_0 cannot be balanced. Therefore one has $|u_0| - 1 \leq |v_2|/2$. Now when $R_w = R_{u_0}$ is created, it is contains in a factor that also contains v_2 and at least one unmatched push symbols before v_2. Hence, $|R_w| \leq 2|v|/3$.

Finally, the fact that for any factor w constructed by Algorithm 2, $\mathrm{Depth}(w) = \mathrm{O}(\log |w|)$ derives from the fact that if $\mathrm{Depth}(w) = k$, then $|w| \geq (3/2)^k$. This can in turn be shown by induction on the depth. Obviously any factor will have weight at least 1. Let us assume all factors of depth k have weight at least $(3/2)^k$, and let $w(i)$ be a letter such that $\mathrm{Depth}(w(i)) = k+1$. By definition, $w(i) = R_v$ for some factor v with $\mathrm{Depth}(v) = k$. This means v contains at least one letter $v(j)$ of depth k. By our induction hypothesis, $|v(j)| \geq (3/2)^k$, and therefore $|w(i)| = |v| \geq (3/2)|v(j)| \geq (3/2)^{k+1}$. ◀

4 The Special Case Of Peaks

We now consider restricted instances consisting of a *single peak*. For these instances, Algorithm 2 never uses its stack but u_0 can be of linear size. We show how to replace u_0 by a small random sketch in order to get a streaming property tester using polylogarithmic memory. In Section 5, this notion of sketch will be later extended to obtain our final streaming property tester for general instances.

4.1 Hard Peak Instances

Peaks are already hard for both streaming algorithms and property testers. Indeed, consider the language Disj $\subseteq \Lambda$ over alphabet $\Sigma = \{0, 1, \overline{0}, \overline{1}, a\}$ and defined as the union of all languages $a^* \cdot x(1) \cdot a^* \cdot \ldots \cdot x(j) \cdot a^* \cdot \overline{y(j)} \cdot a^* \cdot \ldots \cdot \overline{y(1)} \cdot a^*$, where $j \geq 1$, $x, y \in \{0,1\}^j$, and $x(i)y(i) \neq 1$ for all i.

Then Disj can be recognized by a VPA with 3 states, $\Sigma_+ = \{0,1\}$, $\Sigma_- = \{\overline{0}, \overline{1}\}$ and $\Sigma_= = \{a\}$. However, the following fact states its hardness for both models. The hardness for non-approximation streaming algorithms comes for a standard reduction to Set-Disjointness. The hardness for property testing algorithms is a corollary of a similar result due to [26] for parenthesis languages with two types of parentheses.

▶ **Fact 7.** *Any randomized p-pass streaming algorithm for* Disj *requires memory space $\Omega(n/p)$, where n is the input length. Moreover, any (non-streaming) (2^{-6})-tester for* Disj *requires to query $\Omega(n^{1/11}/\log n)$ letters of the input word.*

Proof. We start with the hardness for exact streaming algorithms using a reduction to Set-Disjointness. The Set-Disjointness problem is defined as follows. Two players have respectively x and y from $\{0,1\}^n$ and they decide whether there is some $i \in [n]$ such that $x(i) = y(i) = 1$. The randomized communication complexity of this problem is well known: $\Omega(n)$ bits need to be exchanged between the two players. In addition, the problem is highly connected to our language Disj, but in the communication setting, just like if one player has got the push sequence, and the second one the pop sequence, without any neutral symbols a. Indeed, given a p-pass streaming algorithm with memory $s(n)$ for Disj, the two players can solve Set-Disjointness using $\mathrm{O}(p \times s(n))$ bits of communications. First they simulate a stream, whose first part correspond to the push sequence generated from x, and the second part to the pop sequence generated from y. Then, they simply simulate the streaming algorithm on each part of the stream they control, and send the current memory state when the algorithm changes from one part of the stream to the other one. Thus we get that $s(n) = \Omega(n/p)$.

We now prove hardness of testing Disj in the query model, and for that we use a result from [26]. Let PAR be the language of those well-parenthesized words on the alphabet $\{(, [,],), a\}$, where a is a neutral symbol, that additionally belong to Λ. It is known from [26] (Theorem 2) that any (2^{-6})-tester for PAR in the query model for Hamming distance requires $\Omega(n^{1/11}/\log n)$ queries.

We claim that PAR can be reduced to Disj: for that it suffices to replace (by 01,) by $\overline{01}$, [by 10,] by $\overline{10}$, and a by aa. For instance the word $(a(aa[a])a]a \notin$ PAR becomes

$01aa01aaaa10aa\overline{1001}aa\overline{10}aa$. This word is indeed not in Disj as $x(2) = y(2) = 1$. The previous reduction is a valid reduction in the sense that instances in (resp. not in) PAR are mapped to instances in (resp. not in) Disj, and that any query in PAR can be simulated by two queries in the Disj. Indeed, to simulate a query to position $2i-1$ (resp. $2i$) in DISJ one simply queries position i in PAR; this is for that reason that we mapped a to aa.

Hence, it means that the lower bound from from [26] also applies to Disj for Hamming distance. Now, to conclude that it also applies to *balanced edit distance*, it suffices to remark the following two things: (1) any tester for the balanced edit distance is also a tester for the edit distance; (2) the results of [26] remain valid for edit distance. The latter comes from the fact that the Hamming distance and the edit distance of any word u to PAR are identical. Indeed, one can first remark that there is no need to insert a to bring a word to PAR. Then, if a sequence of (parenthesis) insertions brings some word u inside PAR, then the same sequence where any insertion of a parenthesis is replaced by the deletion of the matching parenthesis also brings u to PAR with the same cost. Hence, one can safely assume that only deletions are performed in the edit sequence. Now, noting that any deletion can similarly be replaced by a substitution of the character being deleted with a we obtain from any optimal sequence for the edit distance a sequence (of same cost) that only uses substitutions thus we get the announced property. ◀

Surprisingly, for every $\varepsilon > 0$, we will show that languages of the form $L \cap \Lambda$, where L is a VPL, become easy to ε-test by streaming algorithms. This is mainly because, given their full access to the input, streaming algorithms can perform an input sampling which makes the property testing task easy, using only a single pass and few memory.

4.2 Slicing Automaton

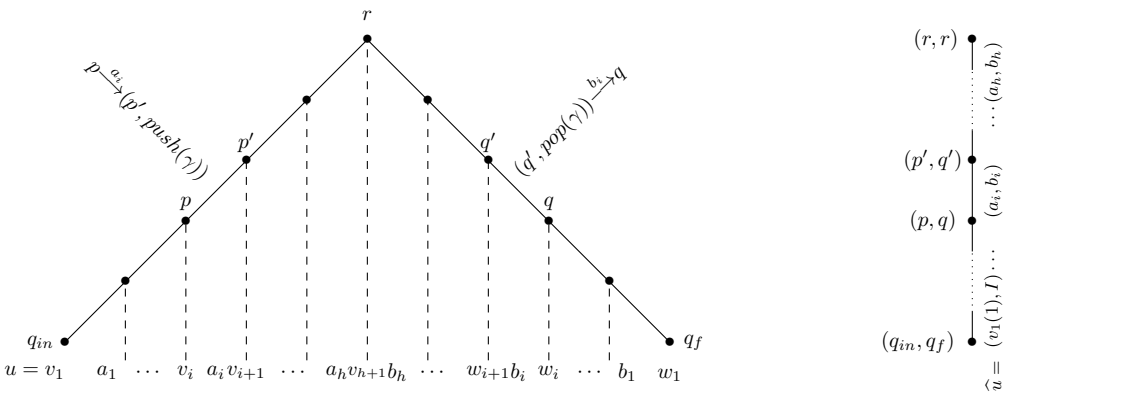
Observe that Algorithm 2 will never use the stack in the case of a single peak. After Algorithm 2 has processed the i-th letter of the data stream, u_0 contains $u[1,i]$ where the eventual initial sequence of neutral symbols has been removed. We will show how to compute R_{u_0} at line 14 using a standard finite state automaton without any stack.

Indeed, for every VPL L, one can construct a regular language $\widehat{L}$ such that testing whether $u \in L \cap \Lambda$ is equivalent to test whether some other word $\widehat{u}$ belongs to $\widehat{L}$. For this, let I be a special symbol not in $\Sigma_=$ encoding the relation set $\{(p,p) : p \in Q\}$. For a word $v \in \Sigma_=^l$, write $[v, \mathrm{I}]$ for the word $(v(1), \mathrm{I}) \cdot (v(2), \mathrm{I}) \cdots (v(l), \mathrm{I})$, and similarly $[\mathrm{I}, v]$. Consider a weighted word of the form $u = \left(\prod_{i=1}^{j} v_i \cdot a_i\right) \cdot v_{j+1} \cdot \left(\prod_{i=j}^{1} b_i \cdot w_i\right)$, where $a_i \in \Sigma_+$, $b_i \in \Sigma_-$, and $v_i, w_i \in \Sigma_=^*$. Then the *slicing* of u (see Figure 2) is the word $\widehat{u}$ over the alphabet $\widehat{\Sigma} = (\Sigma_+ \times \Sigma_-) \cup (\Sigma_= \times \{\mathrm{I}\}) \cup (\{\mathrm{I}\} \times \Sigma_=)$ defined by $\widehat{u} = \left(\prod_{i=1}^{j} [v_i, \mathrm{I}] \cdot [\mathrm{I}, w_i] \cdot (a_i, b_i)\right) \cdot [v_{j+1}, \mathrm{I}]$.

▶ **Definition 8.** Let $\mathcal{A} = (Q, \Sigma, \Gamma, Q_{in}, Q_f, \Delta)$ be a VPA. The *slicing* of $\mathcal{A}$ is the finite automaton $\widehat{\mathcal{A}} = (\widehat{Q}, \widehat{\Sigma}, \widehat{Q_{in}}, \widehat{Q_f}, \widehat{\Delta})$ where $\widehat{Q} = Q \times Q$, $\widehat{Q_{in}} = Q_{in} \times Q_f$, $\widehat{Q_f} = \{(p,p) : p \in Q\}$, and the transitions $\widehat{\Delta}$ are:

1. $(p,q) \xrightarrow{(a,b)} (p', q')$ when $p \xrightarrow{a} (p', \mathsf{push}(\gamma))$ and $(q', \mathsf{pop}(\gamma)) \xrightarrow{b} q$ are both transitions of Δ.
2. $(p,q) \xrightarrow{(c,\mathrm{I})} (p', q)$, resp. $(p,q) \xrightarrow{(\mathrm{I},c)} (p, q')$, when $p \xrightarrow{c} p'$, resp. $q \xrightarrow{c} q'$, is a transition of Δ.

This construction will be later used in Section 5 for weighted languages. In that case, we define the weight of a letter in $\widehat{u}$ by $|(a,b)| = |a| + |b|$, with the convention that $|I| = 0$. Moreover, we write $\widehat{\Sigma_Q}$ for the alphabet obtained similarly to $\widehat{\Sigma}$ using Σ_Q instead of $\Sigma_=$. Note that the slicing automaton $\widehat{\mathcal{A}}$ defined on $\widehat{\Sigma_Q}$ is $\widehat{\Sigma}$-closed and has $\widehat{\Sigma}$-diameter at most $2m^2$ where $m = |Q|$. Indeed, the slicing automaton has m^2 states and every letter in $\widehat{\Sigma}$ has weight at most 2, hence the shortest path from two states (when exists) has weight at most $2m^2$. In particular, it directly implies the following.



Run in the VPA $\mathcal{A}$ on u Run in the slicing automaton $\widehat{\mathcal{A}}$ on $\widehat{u}$

Figure 2 Slicing of a word $u \in \Lambda$ and evolution of the stack height for u.

▶ **Proposition 9.** *Let $v \in \Lambda$ be s.t. $(p,q) \xrightarrow{\widehat{v}} (p',q')$. There is $w \in \Lambda$ s.t. $|w| \leq 2m^2$ and $(p,q) \xrightarrow{\widehat{w}} (p',q')$.*

▶ **Lemma 10.** *If $\mathcal{A}$ is a* VPA *accepting L, then $\widehat{\mathcal{A}}$ is an automaton accepting $\widehat{L} = \{\widehat{u} : u \in L \cap \Lambda\}$.*

Proof. Because transitions on push symbols do not depend on the top of the stack, transitions in $\widehat{\Delta}$ correspond to slices that are valid for Δ (see Figure 2). Finally, $\widehat{Q_{in}}$ ensures that a run for L must start in Q_{in} and end in Q_f, and $\widehat{Q_f}$ that a state at the top of the peak is consistent from both sides. ◀

4.3 Random Sketches

We are now ready to build a tester for $L \cap \Lambda$. To test a word u we use a property tester for the regular language $\widehat{L}$. Regular languages are known to be ε-testable for the Hamming distance with $O((\log 1/\varepsilon)/\varepsilon)$ non-adaptive queries on the input word [1], that is queries that can all be made simultaneously. Those queries define a small random sketch of u that can be sent to the tester for approximating R_u. Since the Hamming distance is larger than the edit distance, those testers are also valid for the latter distance. Observe also that, for $v_1, v_2 \in \Lambda_Q$, we have $\mathsf{bdist}(v_1, v_2) \leq 2\mathsf{dist}(\widehat{v_1}, \widehat{v_2})$. The only remaining difficulty is to provide to the tester an appropriate sampling on $\widehat{u}$ while processing u.

We will proceed similarly for the general case in Section 5, but then we will have to consider weighted words. Therefore we show how to sketch u in that general case already. Indeed, the tester of [1] was simplified for the edit distance in [24], and later on adapted for weighted words in [25]. We consider here an alternative approach that we believe simpler, but slightly less efficient than the tester of [25]. In particular, we introduce in Appendix A a new criterion, κ-saturation, that permits to significantly simplify the correctness proof of the tester compared to the one in [1] and in [25].

Our tester for weighted regular languages is based on k-factor sampling on $\widehat{u}$ that we will simulate by an over-sampling built from a letter sampling on u, that is according to the weights of the letters of u only. This new sampling can be easily performed given a stream of u using a standard reservoir sampling.

Let $u \in \Lambda$ and let $u[i, i+k]$ be a factor that contains at least one push symbol. Call i_1 (resp. i_2) the smallest (resp. largest) integer such that $i_1 \geq i$ (resp. $i_2 \leq i+k$) and $u(i_1)$ (resp. $u(i_2)$) is a push symbol. Then the *matching pop sequence* of $u[i, i+k]$ is defined as $u[j_1, j_2]$ where $u(j_1)$ (resp. $u(j_2)$) is the matching pop symbol of $u(i_1)$ (resp. $u(i_2)$).

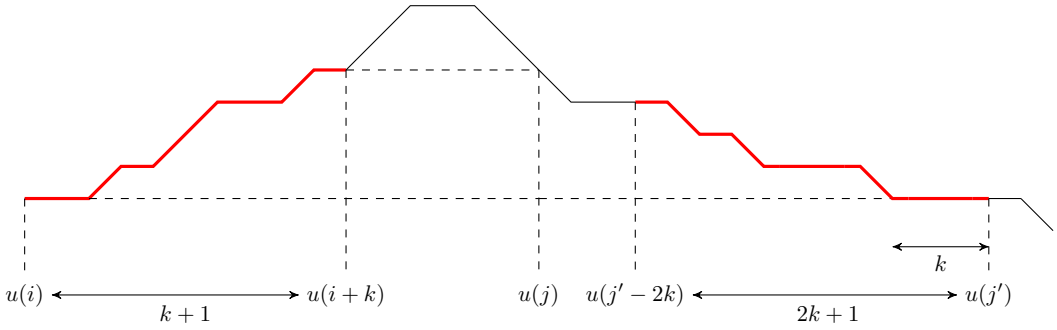


Figure 3 The sampling $\mathcal{W}_k(u)$ from Definition 11: sample is in red.

▶ **Definition 11.** For a weighted word $u \in \Lambda_Q$, denote by $\mathcal{W}_k(u)$ the sampling over subwords of u constructed as follows (see Figure 3):

(1) Sample a factor $u[i, i+k]$ of u with probability $|u(i)|/|u|$.

(2) If $u(i)$ is before the first pop symbol of u, let $u[j, j']$ be the matching pop sequence of $u[i, i+k]$, extended by the first k neutral symbols after the last pop symbol, if any. Add $u[\max(j, j'-2k), j']$ to the sample (hence, some matching pops of $u[i, i+k]$ may not belong to $u[\max(j, j'-2k), j']$).

▶ **Fact 12.** *There is a randomized streaming algorithm with memory* $\mathrm{O}(k + \log n)$ *which, given* k *and* u *as input, samples* $\mathcal{W}_k(u)$.

Proof. (1) can easily be obtained using reservoir sampling. If the sampling enters the pop sequence as the current candidate is part of the push sequence, then (2) can be done for that candidate, and forgotten if the sampling eventually picks another one. That eventual candidate will not be part of the push sequence, so we are done. ◀

▶ **Lemma 13.** *Let* u *be a weighted word, and let* k *be such that* $4k \leq |u|$. *Then* $4k$ *independent copies of* $\mathcal{W}_k(u)$ *over-sample the* k*-factor sampling on* $\widehat{u}$.

Proof. Denote by $\widehat{\mathcal{W}}$ the k-factor sampling on $\widehat{u}$, and by $\mathcal{W}$ some $4k$ independent copies of $\mathcal{W}_k(u)$. For any k-factor v of $\widehat{u}$, we will show that the probability that $\widehat{v}$ is sampled by $\widehat{\mathcal{W}}$ is at most the probability that $\widehat{v}$ is a factor of an element sampled by $\mathcal{W}$. For that, we distinguish the following three cases:

- $\widehat{v}$ contains only letters in $\{I\} \times \Sigma_Q$. Then the probability that $\widehat{v}$ is sampled by $\widehat{\mathcal{W}}$ is equal to the probability that it is sampled by $\mathcal{W}_k(u)$ in step (1).
- $\widehat{v}$ starts by a letter (a, b) in $\Sigma_+ \times \Sigma_-$ or by a letter in $\Sigma_Q \times \{I\}$. Then the probability that the $u(i)$ selected by $\mathcal{W}_k(u)$ is a is at least half of the probability that $\mathcal{W}_k(u)$ samples $\widehat{v}$, as a (push,pop) pair in $\widehat{u}$ has weight 2 while a push has weight 1 in u. Because $\widehat{v}$ is a k-factor, it is contained in $(u[i, i+k], u[j'-2k, j'])$. Hence, the probability that $\widehat{v}$ is sampled by $\widehat{\mathcal{W}}$ is at most the probability that $\widehat{v}$ is a factor of an element sampled by $\mathcal{W}_k(u)$ in step (2).
- $\widehat{v}$ starts by a letter in $\{I\} \times \Sigma_Q$ but also contains letters outside of this set. Since $|\widehat{u}| \geq |u|/2$, we get

$$\Pr(\mathcal{W}_k(u) \text{ samples } \widehat{v}) \geq 1/|u| \quad \text{and} \quad \Pr(\widehat{\mathcal{W}} \text{ samples } \widehat{v}) \leq k/|\widehat{u}| \leq 2k/|u|.$$

Thus the probability that one of the $4k$ samples of $\mathcal{W}$ has the factor $\widehat{v}$ is at least $1 - (1 - 1/|u|)^{4k}$. As $1 - (1 - 1/|u|)^{4k} \geq 1 - \frac{1}{1+4k/|u|} = \frac{4k}{|u|+4k} \geq 2k/|u|$ when $|u| \geq 4k$, we conclude again that the probability that $\widehat{v}$ is sampled by $\widehat{\mathcal{W}}$ is at most the probability that $\widehat{v}$ is a factor of an element sampled by $\mathcal{W}_k(u)$ in step (2).

◀

We can now give an analogue of the property tester for weighted regular languages in $L \cap \Lambda_Q$. For that, we use the following notion of approximation.

▶ **Definition 14.** Let $R \subseteq Q^2$. Then R (ε, Σ)-*approximates* a balanced word $u \in (\Sigma_+ \cup \Sigma_- \cup \Sigma_Q)^*$ on $\mathcal{A}$, if for all $p, q \in Q$:
(1) If $p \xrightarrow{u} q$, then $(p, q) \in R$;
(2) If $(p, q) \in R$, there is a word v such that $\mathsf{dist}_\Sigma(u, v) \leq \varepsilon|u|$ and $p \xrightarrow{v} q$.

Our tester is going to be robust enough in order to consider samples that do not exactly match the peaks we want to compress.

▶ **Theorem 15.** *Let $\mathcal{A}$ be a* VPA *with $m \geq 2$ states and Σ-diameter $d \geq 2$. Let $\varepsilon > 0$, $\eta > 0$, $t = 2\lceil 4dm^3(\log 1/\eta)/\varepsilon \rceil$, $k = \lceil 4dm/\varepsilon \rceil$ and $T = 4kt$. There is an algorithm that, given T random subwords $z_1, \ldots, z_T$ of some weighted word $v \in \Lambda_Q$, such that each z_i comes from an independent sampling $\mathcal{W}_k(v)$, outputs a set $R \subseteq Q \times Q$ that (ε, Σ)-approximates v on $\mathcal{A}$ with bounded error η. Let v' be obtained from v by at most $\varepsilon|v|$ balanced deletions. Then, the conclusion is still true if the algorithm is given an independent $\mathcal{W}_k(v')$ for each z_i instead, except that R now provides a $(3\varepsilon, \Sigma)$-approximation. Last, each sampling can be replaced by an over-sampling.*

Proof. The argument uses as a subroutine the algorithm of Theorem 26 for $\widehat{\mathcal{A}}$, where $\mathcal{A}$ has been extended to Σ_Q. Recall that $\mathcal{A}$ is Σ-closed and its Σ-diameter is also the $\widehat{\Sigma}$-diameter of $\widehat{\mathcal{A}}$. Also observe that $\mathsf{bdist}_\Sigma(u, v) \leq 2\mathsf{dist}_{\widehat{\Sigma}}(\widehat{u}, \widehat{v})$.

By Lemma 13, the T independent samplings $\mathcal{W}_k(v)$ provide us the sampling we need for Theorem 26.

For the case where we do not have an exact k-factor sampling on v however, we need to compensate for the prefix of v of size $\varepsilon|v|$ that may not be included in the sampling. This introduces potentially an additional error of weight $2\varepsilon|v|$ on the approximation R. ◀

As a consequence we get our first streaming tester for $L \cap \Lambda$.

▶ **Theorem 16.** *Let $\mathcal{A}$ be a* VPA *for L with $m \geq 2$ states, and let $\varepsilon, \eta > 0$. Then there is a streaming ε-tester for $L \cap \Lambda$ with one-sided error η and memory space $\mathrm{O}((m^8 \log(1/\eta)/\varepsilon^2)(m^3/\varepsilon + \log n))$, where n is the input length.*

Proof. We use Algorithm 2 where we replace the current factor u_0 by $T = 4kt$ independent samplings $\mathcal{W}_k(u_0)$. We know that such samplings can be computed using memory space $\mathrm{O}(k + \log n)$ by Fact 12. By Proposition 9, the slicing automaton has $\widehat{\Sigma}$-diameter d at most $2m^2$. Therefore, from Theorem 15, taking $t = 4\lceil 4dm^3(\log 1/\eta)/\varepsilon \rceil$ and $k = \lceil 4dm/\varepsilon \rceil$ leads to the desired conclusion. ◀

5 Algorithm With Sketching

5.1 Sketching Using Suffix Samplings

We now describe the sketches used by our main algorithm. They are based on the generalization of the random sketches described in Section 4.3. Moreover, they rely on a notion of suffix sampling, that ensures a good letter sampling on each suffix of a data stream. Recall that the letter sampling on a weighted word u samples a random letter $u(i)$ (with its position) with probability $|u(i)|/|u|$.

Recall, as explained in the preliminaries, that we can easily derive from a letter sampling a sampling on k-factors: this will permit us to use (α, t)-suffix sampling to sample k-factors.

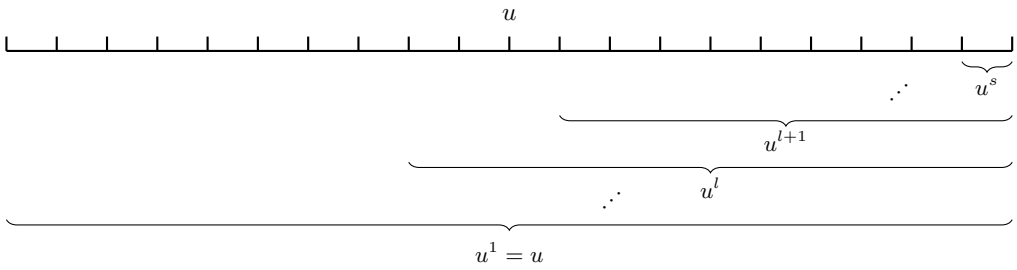


Figure 4 An α suffix decomposition of u of size s. For every l, either $|u^l| \leq \alpha|u^{l+1}|$, or $u^l = a \cdot u^{l+1}$ where a is a letter.

▶ **Definition 17.** Let u be a weighted word and let $\alpha > 1$. An *α-suffix decomposition of u of size s* (see Figure 4) is a sequence of suffixes $(u^l)_{1 \leq l \leq s}$ of u such that: $u^1 = u$, u^s is the last letter of u, and for all l, u^{l+1} is a strict suffix of u^l and if $|u^l| > \alpha|u^{l+1}|$ then $u^l = a \cdot u^{l+1}$ where a is a single letter. An *(α, t)-suffix sampling on u of size s* is an α-suffix decomposition of u of size s with t letter samplings on each suffix of the decomposition.

We observe that (α, t)-suffix samplings can be either concatenated or compressed as stated below.

▶ **Proposition 18.** *Given an (α, t)-suffix sampling D_u on u of size s_u and another one D_v on v of size s_v, there is an algorithm **Concatenate**(D_u, D_v) computing an (α, t)-suffix sampling on the concatenated word $u \cdot v$ of size at most $s_u + s_v$ in time $\mathrm{O}(s_u)$.*
*Moreover, given an (α, t)-suffix sampling D_u on u of size s_u, there is an algorithm **Simplify**(D_u) computing an (α, t)-suffix sampling on u of size at most $2\lceil \log |u| / \log \alpha \rceil$ in time $\mathrm{O}(s_u)$.*

Proof. We sketch those procedures. They are fully described in Algorithm 3. The correctness of those procedures is immediate. For **Concatenate**, it suffices to do the following. For each suffix u^l of D_u: (1) replace u^l by $u^l \cdot v$; and (2) replace the i-th sampling of u^l by the i-th sampling of v with probability $|v|/(|u| + |v|)$, for $i = 1, \ldots, t$.

For **Simplify**, do the following. For each suffix u^l of D_u, from $l = s_u$ (the smallest one) to $l = 1$ (the largest one): (1) replace all suffixes $u^{l-1}, u^{l-2}, \ldots, u^m$ by the largest suffix u^m such that $|u^m| \leq \alpha|u^l|$; and (2) suppress all samples from deleted suffixes. ◀

Using this proposition, one can easily design a streaming algorithm constructing online a suffix decomposition of polylogarithmic size. Starting with an empty suffix-sampling S, simply concatenate S with the next processed letter a of the stream, and then simplify it. We formalize this, together with functions **Concatenate** and **Simplify**, in Algorithm 3.

▶ **Lemma 19.** *Given a weighted word u as a data stream and a parameter $\alpha > 1$, **Online-Suffix-Sampling** in Algorithm 3 constructs an α-suffix sampling on u of size at most $1 + 2\lceil \log |u| / \log \alpha \rceil$.*

One can then slightly modify Algorithm 3 so that within each suffix of the decomposition it simulates t letter samplings in order to construct an (α, t)-suffix sampling.

5.2 Final Algorithm

Our final algorithm is a modification of Algorithm 2: in particular it approximates relations R_v (in the spirit of Definition 14) by elements in Σ_Q, instead of exactly computing them. Let us stress that even

Algorithm 3 α-Suffix Sampling

Data structure:
  // D, D_u, D_v, D_{temp} stacks of items (σ, b), one for each suffix
  // of the decomposition where σ encodes the weight and b the t samples
Code:
Concatenate(D_u, D_v)
  $D \leftarrow D_u$
  $(c_1, \ldots, c_t) \leftarrow$ all t samples on v (the largest suffix in D_v)
  For each $(\sigma, b) \in D$ where $b = (b_1, \ldots, b_t)$
    Replace each b_i by c_i with probability $|v|/(|v| + \sigma)$
    Replace (σ, b) by $(\sigma + |v|, b)$
  Append D_v to the top of D
  Return D
Simplify(D_u)
  $D \leftarrow D_u$
  For each $(\sigma, b) \in D$ from top to bottom
    $D_{\text{temp}} \leftarrow$ elements $(\tau, c) \in D$ below (σ, b) with $\tau \leq \alpha\sigma$
    Replace D_{temp} in D by the bottom most element of D_{temp}
  Return D
Online-Suffix-Sampling
  $D \leftarrow \emptyset$
  While u not finished
    $a \leftarrow \text{Next}(u)$
    Concatenate(D, a) where a encodes the suffix sampling $(|a|, (a, \ldots, a))$
    Simplify(D)
  Return D

if some R_v is approximated by an R that does not correspond to any R_u, one has $R \in \Sigma_Q$, which means that for any $(p, q) \in R$, there is a balanced word $u \in \Sigma^*$ depending on (p, q) with $p \xrightarrow{u} q$.

To mimic Algorithm 2 we need to encode (compactly) each unfinished peak v of the stack and u_0: for that we use the data structure described in Data Structure 4. Our final algorithm, Algorithm 5, is simply Algorithm 2 with this new data structure and corresponding adapted operations, where $\varepsilon' = \varepsilon/(6 \log n)$, $T = 4608 m^4 2^{2m^2} (\log^2 n)(\log 1/\eta)/\varepsilon^2$ and $k = 24 m 2^{m^2} (\log n)/\varepsilon$.

The methods are described in Algorithm 5, where we implicitly assume that each letter processed by the algorithm comes with its respective height and (exact or approximate) weight. They use functions **Concatenate** and **Simplify** described in Proposition 18 (and in details in Algorithm 3), while adapting them.

In the next section, we show that the samplings S_{v^l} are close enough to an $(1+\varepsilon')$-suffix sampling

Data Structure 4 Sketch for an unfinished peak

Parameters: real $\varepsilon' > 0$, integers $T \geq 1$ and $k \geq 1$.
Data structure for a weighted word $v \in \text{Prefix}(\Lambda_Q)$
  Weights of v and of its first letter $v(1)$
  Height of $v(1)$
  Boolean indicating whether v contains a pop symbol
  $(1+\varepsilon')$-suffix decomposition $v^1, \ldots, v^s$ of v encoded for every $l = 1, \ldots, s$ by
    Estimates $|v^l|_{\text{low}}$ and $|v^l|_{\text{high}}$ of $|v^l|$
    T independent samplings S_{v^l} on k-factors of v^l // see details below
      with corresponding weights and heights

Algorithm 5 Adaptation of Algorithm 2 using sketches

Run Algorithm 2 using Data structure 4 with the following adaptations:
Adaption of functions from Proposition 18
- **Concatenate**(D_u, D_v) with an exact estimate of $|v|$ is modified s.t. the replacement probability is now $|v|/(|u|_{\texttt{high}} + |v|)$ and $|u^l \cdot v|_z \leftarrow |u^l|_z + |v|$, for $z = \texttt{low}, \texttt{high}$
- **Simplify**(D_u) with $\alpha = 1 + \varepsilon'$ has now relaxed condition $|u^m|_{\texttt{high}} \leq (1+\varepsilon')|u^l|_{\texttt{low}}$.
- **Online-Suffix-Sampling** is unchanged except for doing k-factor sampling.

Adaption of operations on factors used in Algorithm 2
- **Compute relation:** R_v
  - Run the algorithm of Theorem 15 using samples in D_v
- **Decomposition:** $v_1 \cdot v_2 \leftarrow v$
  - Find largest suffix v^i in D_v s.t. $v^i \in \mathrm{Prefix}(\Lambda_Q)$ *// i.e. s.t. v^i is in v_2*
  - $D_{v|v_1} \leftarrow$ suffixes $(v^l)_{l<i}$ with their samples
  - $D_{v_2} \leftarrow$ suffix v^i with its samples and weight estimates *// to compute R_{v_2}*
    - $(|v^i|_{\texttt{high}}, |v^i|_{\texttt{low}})$ when v^{i-1} and v^i differ by a single letter (then $v^i = v_2$)
    - $(|v^{i-1}|_{\texttt{high}}, |v^i|_{\texttt{low}})$ otherwise
- **Test:** $|u_0| \geq |v_2|/2$ using $|v_2|_{\texttt{low}}$ instead of $|v_2|$
- **Concatenation:** $u_0 \leftarrow (v_1 \cdot R_{v_2}) \cdot u_0$
  - $D_{v'} \leftarrow (D_{v|v_1}, R_{v_2})$ replacing each samples of $D_{v|v_1}$ in v_2 by R_{v_2}
  - *// The height of a sample determines whether it is in v_2*
  - $D_{u_0} \leftarrow$ **Simplify**(**Concatenate**$(D_{v'}, D_{u_0})$)

on v^l. This lets us build an over-sampling of an $(1+\varepsilon')$-suffix sampling. We also show that it only requires a polylogarithmic number of samples. Then, we explain how to recursively apply the tester from Theorem 15 (with ε') in order to obtain the compressions at line 14 and 20 while keeping a cumulative error below ε. We now state our main result whose proof relies on Lemmas 22 and 24.

▶ **Theorem 20.** *Let $\mathcal{A}$ be a* VPA *for L with $m \geq 2$ states, and let $\varepsilon, \eta > 0$. Then there is an ε-streaming algorithm for L with one-sided error η and memory space $\mathrm{O}(m^5 2^{3m^2} (\log^6 n)(\log 1/\eta)/\varepsilon^4)$, where n is the input length.*

Proof. We use Algorithm 5, which uses the tester from Theorem 15 for the compressions at lines 14 and 20 of Algorithm 2. We know from Lemma 24 that it is enough to choose $\varepsilon' = \varepsilon/(6 \log n)$, $\eta' = \eta/n$, and Fact 21 gives us $d = 2^{m^2}$. Therefore we need to sample k-factors from a $(1+\varepsilon', t)$-suffix sampling, where Theorem 15 gives us that $t = 2304 m^4 2^{2m^2} (\log^2 n)(\log 1/\eta)/\varepsilon^2$ and $k = 24 m 2^{m^2} (\log n)/\varepsilon$. Lemma 22 tells us that using $2t$ samples from our algorithm is enough.

Because of the sampling variant we use, the size of each suffix decomposition is at most $144(\log n)^2/\varepsilon + \mathrm{O}(\log n)$ by Lemma 22. The samples in each element of the decomposition use memory space k, and there are $2t$ of them. Furthermore, each element of the stack has its own sketch, and the stack is of height at most $\log n$. Multiplying all those together gives us the upper bound on the memory space used by Algorithm 5. ◀

5.3 Final Analysis

As Algorithm 5 may fail at various steps, the relations it considers may not correspond to any word. However, each relation R that it produces is still in Σ_Q. Furthermore, the slicing automaton $\widehat{\mathcal{A}}$ over $\widehat{\Sigma_Q}$ is $\widehat{\Sigma}$-closed. Fact 21 below bounds the $\widehat{\Sigma}$-diameter of $\widehat{\mathcal{A}}$ (which is equal to the Σ-diameter of $\mathcal{A}$) by 2^{m^2}. For simpler languages, as those coming from a DTD, this bound can be lowered to m.

▶ **Fact 21.** *Let $\mathcal{A}$ be a* VPA *with m states. Then the Σ-diameter of $\mathcal{A}$ is at most 2^{m^2}.*

Proof. A similar statement is well known for any context-free grammar given in Chomsky normal form. Let N be the number of non-terminal symbols used in the grammar. If the grammar produces one balanced word from some non-terminal symbol, then it can also produce one whose length is at most 2^N from the same non-terminal symbol. This is proved using a pumping argument on the derivation tree. We refer the reader to the textbook [17].

Now, in the setting of visibly pushdown languages one needs to transform $\mathcal{A}$ into a context-free grammar in Chomsky normal form. For that, consider first an intermediate grammar whose non-terminal symbols are all the X_{pq} where p and q are states from $\mathcal{A}$: such a non-terminal symbol will produce exactly those words u such that $p \xrightarrow{u} q$, hence our initial symbol will be those of the form $X_{q_0 q_f}$ where q_0 is an initial state and q_f is a final state. The rewriting rules are the following ones:

- $X_{pp} \to \varepsilon$
- $X_{pq} \to X_{pr} X_{rq}$ for any state r
- $X_{pq} \to a X_{p'q'} b$ whenever one has in the automaton $p \xrightarrow{a} (p', \mathsf{push}(\gamma))$ and $(q', \mathsf{pop}(\gamma)) \xrightarrow{a} q$ for some push symbol a, pop symbol b and stack letter γ.
- $X_{pq} \to a X_{p'q}$ whenever one has in the automaton $p \xrightarrow{a} p'$ for some neutral symbol a.
- $X_{pq} \to X_{pq'} a$ whenever one has in the automaton $q' \xrightarrow{a} q$ for some neutral symbol a.

Obviously, this grammar generates language $L(\mathcal{A})$.

As we are here interested only in the length of the balanced words produced by the grammar, we can replace any terminal symbol by a dummy symbol $\sharp$. Now, once this is done we can put the grammar into Chomsky normal form by using an extra non-terminal symbol (call it $X_\sharp$ as it is used to produce the $\sharp$ terminal). As we have $m^2 + 1$ non-terminal in the resulting grammar we are almost done. To get to the tight bound announced in the statement, one simply removes the extra non-terminal symbol $X_\sharp$ and reasons on the length of the derivation directly. ◀

We first show that the decomposition, weights and sampling we maintain are close enough to an $(1 + \varepsilon')$-suffix sampling with the correct weights. Recall that $\varepsilon' = \varepsilon/(6 \log n)$.

▶ **Lemma 22** (Stability lemma). *Let v be an unfinished peak with $\mathcal{W}_1, \mathcal{W}_2$ two of the T samplings maintained by Algorithm 5. Then $(\mathcal{W}_1, \mathcal{W}_2)$ over-samples an $(1 + \varepsilon')$-suffix sampling on v, and the decomposition has size at most $144(\log |v|)(\log n)/\varepsilon + \mathrm{O}(\log n)$.*

Before proving the stability lemma, we first prove that Algorithm 5 maintains a strucutre that is not too far from $(1 + \varepsilon')$-suffix sampling.

▶ **Proposition 23.** *Let v be an unfinished peak, and let $v^1, \ldots, v^s$ be the suffix decomposition maintained by the algorithm. The following is true:*
(1) $v^1, \ldots, v^s$ is a valid $(1 + \varepsilon')$-suffix decomposition of v.
(2) For each letter a of every v^l, and for every sample s, $\Pr[S_{v^l} = a] \geq |a|/|v^l|_{\mathtt{high}}$.
(3) Each v^l satisfies $|v^l|_{\mathtt{high}} - |v^l|_{\mathtt{low}} \leq 2\varepsilon' |v^l|_{\mathtt{low}}/3$.

Proof. Property (1) is guaranteed by the (modified) **Simplify** function used in Algorithm 5, which preserves even more suffixes than the original algorithm.

Properties (2) and (3) are proven by induction on the last letter read by Algorithm 5. Both are true when no symbol has been read yet.

We start with property (2). Let us first consider the case where we concatenate after the last letter was read. Then for all v^l, the (modified) **Concatenate** function ensures S_{v^l} becomes a with probability $1/|v^l|_{\mathtt{high}}$. Otherwise, S_{v^l} remains unchanged and by induction $S_{v^l} = b$ with probability at least $(1 - 1/|v^l|_{\mathtt{high}})|b|/(|v^l|_{\mathtt{high}} - 1) = |b|/|v^l|_{\mathtt{high}}$, for each other letter b of v^l.

The other case is that some R_{v_2} is computed at line 20 of Algorithm 2. In this case, v is equal to some $(v_1 \cdot R_{v_2}) \cdot u_0$ concatenation. For each suffix $(v_1 \cdot v_2)^l$ in $D_{(v_1 \cdot v_2)}$ containing R_{v_2}, we proceed

in the same way with the **Concatenate** function, replacing any sample in v_2 with R_{v_2}. Now consider v_2^i the largest suffix of $D_{(v_1 \cdot v_2)}$ contained in v_2, and $v^l = R_{v_2} \cdot u_0$. We use the fact that **Concatenate** looks at $|v^l|_{\texttt{high}} \geq |u_0| + |R_{v_2}|$ for replacing samples. This means that we choose R_{v_2} as a sample for v^l with probability $(|v^l|_{\texttt{high}} - |u_0|)/|v^l|_{\texttt{high}} \geq |R_{v_2}|/|v^l|_{\texttt{high}}$, and therefore the property is verified.

We now prove property (3). If v^l has just been created, it contains only one letter of weight 1, and obviously $|v^l|_{\texttt{low}} = |v^l|_{\texttt{high}} = |v^l|$. In addition, unless some R_{v_2} has been computed at line 20 of Algorithm 2 when the last letter was read, then $|v^l|$ is only augmented by some exactly known $|a|$ or $|u_0|$ compared to the previous step. Therefore the difference $|v^l|_{\texttt{high}} - |v^l|_{\texttt{low}}$ does not change, and by induction it remains smaller than $2\varepsilon'|v^l|_{\texttt{low}}/3$ which can only increase. Now consider R_{v_2} computed at line 20 and $v^l = R_{v_2} \cdot u_0$. We again consider v_2^i for the largest suffix in the decomposition of $v_1 \cdot v_2$ that is contained within v_2, as used in Algorithm 5, and v_2^{i-1} is the suffix immediately preceding v_2^i in that decomposition.

If $|v_2^{i-1}|_{\texttt{high}} > (1+\varepsilon')|v_2^i|_{\texttt{low}}$, then from the **Simplify** function, the difference between those two suffixes cannot be more than one letter, and then $v_2^i = v_2$. Therefore, we have $|R_{v_2} \cdot u_0|_{\texttt{high}} = |v_2|_{\texttt{high}} + |u_0|$ and $|R_{v_2} \cdot u_0|_{\texttt{low}} = |v_2|_{\texttt{low}} + |u_0|$. We conclude by induction on $|v_2|$.

We end with the case $|v_2^{i-1}|_{\texttt{high}} \leq (1+\varepsilon')|v_2^i|_{\texttt{low}}$. By definition, $|R_{v_2} \cdot u_0|_{\texttt{high}} = |v_2^{i-1}|_{\texttt{high}} + |u_0|$ and $|R_{v_2} \cdot u_0|_{\texttt{low}} = |v_2^i|_{\texttt{low}} + |u_0|$. Therefore the difference $|v^l|_{\texttt{high}} - |v^l|_{\texttt{low}}$ is at most $\varepsilon'|v_2^i|_{\texttt{low}}$. Since the test at line 19 of Algorithm 2 (modified by ALgorithm 5) was satisfied, we know that $|v_2^i|_{\texttt{low}} \leq 2|u_0|$, and finally $\varepsilon'|v_2^i|_{\texttt{low}} \leq 2\varepsilon'(|v_2^i|_{\texttt{low}} + |u_0|)/3 \leq 2\varepsilon'|v^l|_{\texttt{low}}/3$, which concludes the proof. ◀

We can now prove the stability lemma.

Proof of Lemma 22. The first property is a direct consequence of property (1) and (2) in Proposition 23, as in the proof of Lemma 13.

The second is a consequence of the (modified) **Simplify** used in Algorithm 5: $D_{\texttt{temp}}$ is defined as the set of suffixes below with $m < l$ such that $|v^m|_{\texttt{high}} \leq (1+\varepsilon')|v^l|_{\texttt{low}}$. Because **Simplify** deletes all but one elements from $D_{\texttt{temp}}$, it follows that $|v^{l-2}|_{\texttt{high}} > (1+\varepsilon')|v^l|_{\texttt{low}}$. Now, from property (3) of Proposition 23 we have that $|v^l|_{\texttt{low}} \geq |v^l|_{\texttt{high}} - 2\varepsilon'|v^l|_{\texttt{low}}/3 \geq (1 - 2\varepsilon'/3)|v^l|_{\texttt{high}}$. Therefore we have that $|v^{l-2}|_{\texttt{high}} > (1+\varepsilon')(1-2\varepsilon'/3)|v^l|_{\texttt{high}}$

By successive applications, we obtain $|v^{l-6}|_{\texttt{high}} > (1+\varepsilon')^3(1-2\varepsilon'/3)^3|v^l|_{\texttt{high}}$. Now, as $|v^l|_{\texttt{high}} > |v^l|$ and $|v^l| \geq |v^l|_{\texttt{low}} \geq (1-2\varepsilon'/3)|v^l|_{\texttt{high}}$ we have: $|v^{l-6}|/(1-2\varepsilon'/3) > (1+\varepsilon')^3(1-2\varepsilon'/3)^3|v^l|$. Equivalently, $|v^{l-6}| > (1+\varepsilon')^3(1-2\varepsilon'/3)^4|v^l|$.

Thus, the size of the suffix decomposition is at most $6\log_{(1+\varepsilon')^3(1-2\varepsilon'/3)^4}|v| \leq 6\log|v|/\log(1+\varepsilon'/3+\mathrm{O}(\varepsilon'^2)) \leq 144(\log|v|)(\log n)/\varepsilon + \mathrm{O}(\log(n))$. ◀

Using the tester from Theorem 15 for computing each R, we can then prove the robustness lemma.

▶ **Lemma 24** (Robustness lemma). *Let $\mathcal{A}$ be a* VPA *recognizing L and let $u \in \Sigma^n$. Let $R_{\texttt{final}}$ be the final value of $R_{\texttt{temp}}$ in Algorithm 5 If $u \in L$, then $R_{\texttt{final}} \in L$; and if $R_{\texttt{final}} \in L$, then $\mathsf{bdist}_\Sigma(u, L) \leq \varepsilon n$ with probability at least $1-\eta$.*

Proof. One way is easy. A direct inspection reveals that each substitution of a factor w by a relation R enlarges the set of possible w-transitions.Therefore $R_{\texttt{final}} \in L$ when $u \in L$.

For the other way, consider some word u such that $R_{\texttt{final}} \in L$. Since the tester of Theorem 15 has bounded error $\eta' = \eta/n$ and was called at most than n times, none of the calls fails with probability at least $1-\eta$. From now on we assume that we are in this situation.

Let $h = \mathrm{Depth}(R_{\texttt{final}})$. We will inductively construct sequences $u_0 = u, \ldots, u_h = R_{\texttt{final}}$ and $v_h = R_{\texttt{final}}, \ldots, v_0$ such that for every $0 \leq l \leq h$, $u_l, v_l \in (\Sigma_+ \cup \Sigma_- \cup \Sigma_Q)^*$, $\mathsf{bdist}_\Sigma(u_l, v_l) \leq 3(h-l)\varepsilon'|u_l|$ and $v_l \in L$. Furthermore, each word u_l will be the word u with some substitutions of

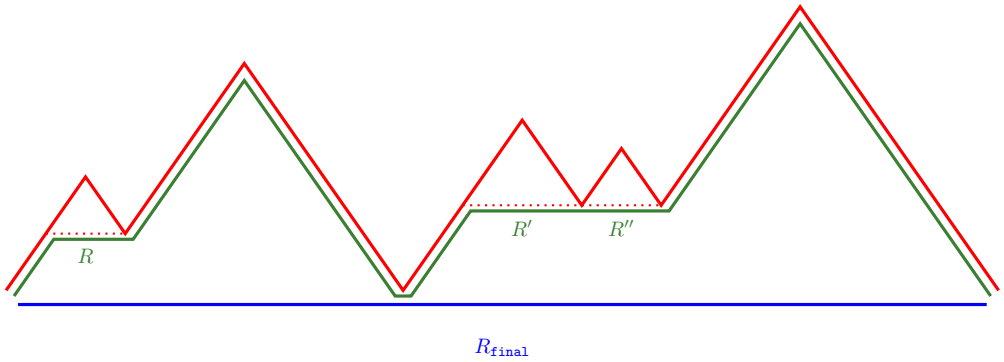


Figure 5 Constructing the words u_0, u_1 and u_2 as in Lemma 24 where $\mathrm{Depth}(R_{\mathtt{final}}) = 2$

factors by relations R computed by the tester. Therefore, $\mathrm{Depth}(u_l)$ is well defined and will satisfy $\mathrm{Depth}(u_l) = l$. This will conclude the proof using that $\mathrm{Depth}(R_{\mathtt{final}}) \leq \log_{3/2} n$ from Lemma 6. This will give us $\mathsf{bdist}_\Sigma(u, v_0) \leq 6\varepsilon' n \log n \leq \varepsilon n$.

We first define the sequence $(u_l)_l$ (see Figure 5 for an illustration). Starting from $u_0 = u$, let u_{l+1} be the word u_l where some factors in Λ_Q have been replaced by a $(3\varepsilon', \Sigma)$-approximation in Σ_Q. These correspond to all the approximations eventually performed by the algorithm that did not involve a symbol already in Σ_Q. Observe that after this collapse, the symbol is still a $(3\varepsilon', \Sigma)$-approximation. In particular, $u_h = R_{\mathtt{final}}$, $u_l \in (\Sigma_+ \cup \Sigma_- \cup \Sigma_Q)^*$ and $\mathrm{Depth}(u_l) = l$ by construction.

We now define the sequence $(v_l)_l$ such that $v_l \in L$. Each letter of v_l will be annotated by an accepting run of states for $\mathcal{A}$. Set $v_h = R_{\mathtt{final}}$ with an accepting run from p_{in} to q_f for some $(p_{in}, q_f) \in R_{\mathtt{final}} \cap (Q_{in} \times Q_f)$. Consider now some level $l < h$. Then v_l is simply v_{l+1} where some letters $R \in \Sigma_Q$ in common with u_{l+1} are replaced by some factors in $w \in (\Lambda_Q)^*$ as explained in the next paragraph. Those letters are the ones that are present in u_l but not u_{l+1}, and are still present in v_{l+1} (i.e. they have not been further approximated down the chain from u_{l+1} to u_h, or deleted by edit operations moving up from v_h to v_{l+1}).

Let $w \in (\Lambda_Q)^*$ be one of those factors and $R \in \Sigma_Q$ its respective $(3\varepsilon', \Sigma)$-approximation. By hypothesis R is still in v_{l+1} and corresponds to a transition (p, q) of the accepting run of v_{l+1}. We replace R by a factor w' such that $p \xrightarrow{w'} q$ and $\mathsf{bdist}_\Sigma(w, w') \leq 3\varepsilon'|w|$, and annotate w' accordingly. By construction, the resulting word v_l satisfies $v_l \in L$ and $\mathsf{bdist}_\Sigma(u_l, v_l) \leq 3(h - l)\varepsilon'|u_l|$. ◀

References

1 N. Alon, M. Krivelevich, I. Newman, and M. Szegedy. Regular languages are testable with a constant number of queries. *SIAM Journal on Computing*, 30(6), 2000.

2 N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences*, 58(1):137–147, 1999.

3 R. Alur. Marrying words and trees. In *Proc. of 26th ACM Symposium on Principles of Database Systems*, pages 233–242, 2007.

4 R. Alur, M. Arenas, P. Barceló, K. Etessami, N. Immerman, and L. Libkin. First-order and temporal logics for nested words. In *Proc. of 22nd IEEE Symposium on Logic in Computer Science*, pages 151–160, 2007.

5 R. Alur, K. Etessami, and P. Madhusudan. A temporal logic of nested calls and returns. In *Proc. of 10th International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, pages 467–481, 2004.

6 R. Alur and P. Madhusudan. Adding nesting structure to words. *Journal of the ACM*, 56(3), 2009.

7 A. Babu, N. Limaye, and G. Varma. Streaming algorithms for some problems in log-space. In *Proc. of 7th Conference on Theory and Applications of Models of Computation*, pages 94–104, 2010.

8 M. Blum, W. Evans, P. Gemmell, S. Kannan, and M. Naor. Checking the correctness of memories. *Algorithmica*, pages 90–99, 1995.

9 M. Blum and S. Kannan. Designing programs that check their work. *Journal of the ACM*, 42(1):269–291, 1995.

10 M. Blum, M. Luby, and R. Rubinfeld. Self-testing/correcting with applications to numerical problems. *Journal of Computer and System Sciences*, 47(3):549–595, 1993.

11 B. von Braunmühl and R. Verbeek. Input-driven languages are recognized in log n space. In *Proc. of 4th Conference on Fundamentals of Computation Theory*, volume 158, pages 40–51, 1983.

12 M. Chu, S. Kannan, and A. McGregor. Checking and spot-checking the correctness of priority queues. In *Proc. of 34th International Colloquium on Automata, Languages and Programming*, pages 728–739, 2007.

13 P. Dymond. Input-driven languages are in $\log n$ depth. *Information Processing Letters*, 26(5):247–250, 1988.

14 J. Feigenbaum, S. Kannan, M. Strauss, and M. Viswanathan. Testing and spot-checking of data streams. *Algorithmica*, 34(1):67–80, 2002.

15 E. Fischer, F. Magniez, and M. de Rougemont. Approximate satisfiability and equivalence. *SIAM Journal on Computing*, 39(6):2251–2281, 2010.

16 O. Goldreich, S. Goldwasser, and D. Ron. Property testing and its connection to learning and approximation. In *Proc. of 37th IEEE Symposium on Foundations of Computer Science*, pages 339–348, 1996.

17 J. Hopcroft, R. Motwani, and J. Ullman. *Introduction to Automata Theory, Languages, and Computation (3rd Edition)*. Addison-Wesley, 2006.

18 C. Konrad and F. Magniez. Validating XML documents in the streaming model with external memory. *ACM Transactions on Database Systems*, 38(4):27, 2013. Special issue of ICDT'12.

19 L. Libkin. Logics for unranked trees: An overview. *Logical Methods in Computer Science*, 2(3), 2006.

20 F. Magniez and M. de Rougemont. Property testing of regular tree languages. *Algorithmica*, 49(2):127–146, 2007.

21 F. Magniez, C. Mathieu, and A. Nayak. Recognizing well-parenthesized expressions in the streaming model. *SIAM Journal on Computing*, 43(6):1880–1905, 2014.

22 K. Mehlorn. Pebbling mountain ranges and its application to dcfl-recognition. In *Proc. of 7th International Colloquium on Automata, Languages, and Programming*, pages 422–435, 1980.

23 S. Muthukrishnan. Data streams: Algorithms and applications. *Foundations and Trends in Theoretical Computer Science*, 1(2):117–236, 2005.

24 A. Ndione, A. Lemay, and J. Niehren. Approximate membership for regular languages modulo the edit distance. *Theoretical Computer Science*, 487:37–49, 2013.

25 A. Ndione, A. Lemay, and J. Niehren. Sublinear DTD validity. In *Proc. of 19th International Conference on Language and Automata Theory and Applications*, pages 739–751, 2015.

26 M. Parnas, D. Ron, and R. Rubinfeld. Testing membership in parenthesis languages. *Random Structures & Algorithms*, 22(1):98–138, 2003.

27 A. Rajeev and P. Madhusudan. Visibly pushdown languages. In *Proc. of 36th ACM Symposium on Theory of Computing*, pages 202–211, 2004.

28 L. Segoufin and C. Sirangelo. Constant-memory validation of streaming XML documents against DTDs. In *Proc. of 11th International Conference on Database Theory*, pages 299–313, 2007.

29 L. Segoufin and V. Vianu. Validating streaming XML documents. In *Proc. of 11th ACM Symposium on Principles of Database Systems,*, pages 53–64, 2002.

A A Tester for Weighted Regular Languages

We design a non-adaptive property tester for weighted regular languages that serves as a basic routine of our main algorithm. Property testing of regular languages was first considered in [1] for the Hamming distance and we adapt this tester to weighted words for the simple case of edit distance. Such a property tester has been already constructed first for edit distance in [24], and later on for weighted words in [25], with an approach based on [1].

In this work, we take an alternative approach that we believe simpler, but slightly less efficient than the tester of [25]. We consider the graph of components of the automaton and focus on paths in this graph; we however introduce a new criterion, κ-saturation (for some parameter $0 < \kappa \leq 1$), that permits to significantly simplify the correctness proof of the tester compared to the one in [1] and in [25]. In particular Lemma 29 permits to design a non-adaptive tester for L and also to approximate the action of u on $\mathcal{A}$ as follows.

▶ **Definition 25.** Let $\Sigma' \subseteq \Sigma$ and $R \subseteq Q \times Q$. Then R *(ε, Σ')-approximates a word u on $\mathcal{A}$ (or simply ε-approximates when $\Sigma' = \Sigma$)*, if for all $p, q \in Q$: (1) $(p, q) \in R$ when $p \xrightarrow{u} q$; (2) u is (ε, Σ')-close to some word v satisfying $p \xrightarrow{v} q$ when $(p, q) \in R$.

Our main contribution is the following one.

▶ **Theorem 26.** *Let $\mathcal{A}$ be an automaton with $m \geq 2$ states and diameter $d \geq 2$. Let $\varepsilon > 0$, $\eta > 0$, $t \geq 2\lceil 2dm^3(\log 1/\eta)/\varepsilon \rceil$ and $k \geq \lceil 2dm/\varepsilon \rceil$. There is an algorithm that, given t random factors of $v_1, \ldots, v_t$ of some weighted word u, such that each v_i comes from an independent k-factor sampling on u, outputs a set $R \subseteq Q \times Q$ that ε-approximates u on $\mathcal{A}$ with one-sided error η.*
This is still true with any combination of the following generalization:
- *The algorithm is given an over-sampling of each of factors v_i instead.*
- *When $\mathcal{A}$ is Σ'-closed, and d is the Σ'-diameter of $\mathcal{A}$, then R also (ε, Σ')-approximates u on $\mathcal{A}$.*

The rest of this section is devoted to the proof of Theorem 26 and therefore we fix a regular language L recognized by some finite state automaton $\mathcal{A}$ on Σ with a set of states Q of size $m \geq 2$, and a diameter $d \geq 2$. Define the directed graph $G_{\mathcal{A}}$ on vertex set Q whose edges are pairs (p, q) when $p \xrightarrow{a} q$ for some $a \in \Sigma$.

A *component* C of $G_{\mathcal{A}}$ is a maximal subset (w.r.t. inclusion) of vertices of $G_{\mathcal{A}}$ such that for every p_1, p_2 in C one has a path in $G_{\mathcal{A}}$ from p_1 to p_2. The *graph of components* $\mathcal{G}_{\mathcal{A}}$ of $G_{\mathcal{A}}$ describes the transition relation of $\mathcal{A}$ on components of $G_{\mathcal{A}}$: its vertices are the components and there is a directed edge (C_1, C_2) if there is an edge of $G_{\mathcal{A}}$ from a vertex in C_1 toward a vertex in C_2.

▶ **Definition 27.** Let C be a component of $G_{\mathcal{A}}$, let $\Pi = (C_1, \ldots, C_l)$ be a path in $\mathcal{G}_{\mathcal{A}}$.
- A word u is *C-compatible* if there are states $p, q \in C$ such that $p \xrightarrow{u} q$.
- A word u is *Π-compatible* if u can be partitioned into $u = v_1 a_1 v_2 \ldots a_{l-1} v_l$ such that $p_i \xrightarrow{v_i} q_i$ and $q_i \xrightarrow{a_i} p_{i+1}$, where v_i is a factor, a_i a letter, and $p_i, q_i \in C_i$.
- A sequence of factors $(v_1, \ldots, v_t)$ of a word u is *Π-compatible* if they are factors of another Π-*compatible* word with the same relative order and same overlap.

Note that the above properties are easy to check. Indeed, C-compatibility is a reachability property while the two others easily follow from C-compatibility checking.

We now give a criterion that characterizes those words u that are ε-far to every Π-compatible word. Note that it will not be used in the tester that we design in Theorem 26 for weighted regular languages, but only in Lemma 29 which is the key tool to prove its correctness.

For a component C and a C-incompatible word v, let $v_1 \cdot a$ be the shortest C-incompatible prefix of v. We define and denote the *C-cut* of v as $v = v_1 \cdot a \cdot v_2$. When v_1 is not the empty word, we say that v_1 is a *C-factor* and a is a *C-separator for* v_1, otherwise we say that a is a *strong C-separator*.

Fix a path $\Pi = (C_1, \dots, C_l)$ in $\mathcal{G}_\mathcal{A}$, a parameter $0 < \kappa \leq 1$, and consider a weighted word u. We define a natural partition of u according to Π, that we call the Π-*partition* of u. For this, start with the first component $C = C_1$, and consider the C_1-cut $u_1 \cdot a \cdot u_2$ of u. Next, we inductively continue this process with either the suffix $a \cdot u_2$ if a is a C_1-separator, or the suffix u_2 if a is a strong C_1-separator. Based on some criterion defined below we will move from the current component C_i to a next component C_j of Π, where most often $j = i + 1$, until the full word u is processed. If we reach $j = l + 1$, we say that u *κ-saturates* Π and the process stops. We now explain how we move on in Π. We stay within C_i as long as both the number of C_i-factors and the total weight of strong C_i-separators are at most $\kappa|u|$ each. Then, we continue the decomposition with some fresh counting and using a new component C_j selected as follows. One sets $j = i + 1$ except when the transition is the consequence of a strong C_i-separator a of weight greater than $\kappa|u|$, that we call a *heavy strong separator*. In that case only, one lets $j \geq i + 1$, if exists, to be the minimal integer such that $q \xrightarrow{a} q'$ with $q \in C_{j-1} \cup C_j$ and $q' \in C_j$, and $j = l + 1$ otherwise.

▶ **Proposition 28.** *Let $0 < \kappa \leq \varepsilon/(2dl)$. If u is ε-far to every Π-compatible word, then u κ-saturates Π.*

Proof. The proof is by contraposition. For this we assume that u does not κ-saturate Π and we correct u to a Π-compatible word as follows.

First, we delete each strong separator of weight less that $\kappa|u|$. Their total weight is at most $2l\kappa|u|$. Because u does not saturate, each strong separator of weight larger than $\kappa|u|$ fits in the Π-partition, and does not need to be deleted.

We now have a sequence of consecutive C_i-factors and of heavy strong C_i-separators, for some $1 \leq i \leq l$, in an order compatible with Π. However, the word is not yet compatible with Π since each factor may end with a state different than the first state of the next factor. However, for each such pair there is a path connecting them. We can therefore bridge all factors by inserting a factor of weight at most d, the diameter of $\mathcal{A}$.

The resulting word is then Π-compatible by construction, and the total cost of the edit operations is at most $(2l + dl)\kappa|u| \leq \varepsilon|u|$, since $d \geq 2$. ◀

For a weighted word u, we remind that the k-factor sampling on u is defined in Section 2.1. The following lemma is the key lemma for the tester for weighted regular languages.

▶ **Lemma 29.** *Let u be a weighted word, let $\Pi = C_1 \dots C_l$ be a path in $\mathcal{G}_\mathcal{A}$. Let $0 < \kappa \leq \varepsilon/(2dl)$ and let $\mathcal{W}$ denote the $\lceil 2/\kappa \rceil$-factor sampling on u. Then for every $0 < \eta < 1$ and $t \geq 2l(\log 1/\eta)/\kappa$, the probability $P(u, \Pi) = \Pr_{(v_1,\dots,v_t)\sim\mathcal{W}^{\otimes t}}[(v_1, \dots, v_t)$ is Π-compatible$]$ satisfies $P(u, \Pi) = 1$ when u is Π-compatible, and $P(u, \Pi) \leq \eta$ when u is ε-far for from being Π-compatible.*

Proof. The first part of the theorem is immediate. For the second part, assume that u is ε-far from any Π-compatible word. For simplicity we assume that $2/\kappa$ and $\kappa|u|/2$ are integers. We first partition u according to Π and κ. Then, Proposition 28 tells us that u κ-saturates Π. For each C_i, we have three possible cases.

1. There are $\kappa|u|$ disjoint C_i-factors in u. Since they have total weight at most $|u|$, there are at least $\kappa|u|/2$ of them whose weight is at most $2/\kappa$ each. Since each letter has weight at least 1, the total weight of the first letters of each of those factors is at least $\kappa|u|/2$. Therefore one of them together with its C_i-separator is a sub-factor of some sampled factor v_j with probability at least $1 - (1 - \kappa/2)^t$.
2. The total weight of strong C_i-separators of u is at least $\kappa|u|$. Therefore one of them is the first letter of some sampled factor v_j with probability at least $1 - (1 - \kappa)^t$.

3. There is not any C_i-factor and any C_i-separator of u, because of a strong $C_{i'}$-separator of weight greater than $\kappa|u|$, for some $i' < i$. This separator is the first letter of some sampled factor v_j with probability at least $1 - (1-\kappa)^t$.

By union bound, the probability that one of the above mentioned samples fails to occurs is at most $l(1-\kappa)^t \leq \eta$. We assume now that they all occur, and we show that they form a Π-incompatible sequence. For each i, let w_i be the above described sub-factors of those samples. Each w_i appears in u after w_{i-1} or, in the case of a strong separator of heavy weight, $w_i = w_{i-1}$. Moreover each factor w_i which is distinct from w_{i-1} forces next factors to start from some component $C_{i'}$ with $i' > i$. As a result $(w_1, \ldots, w_l)$ is not Π-compatible, and as a consequence $(v_1, \ldots, v_t)$ neither, so the result. ◀

We can now conclude with the proof of Theorem 26.

Proof of Theorem 26. The algorithm is very simple:

1. Set $R = \emptyset$
2. For all states $p, q \in Q$
   a. Check if factors $v_1, \ldots, v_t$ could come from a word v such that $p \xrightarrow{v} q$
      // Step (a) is done using the graph $\mathcal{G}_\mathcal{A}$ of connected components of $\mathcal{A}$
   b. If yes, then add (p, q) to R
3. Return R

It is clear that this R contains every (p, q) such that $p \xrightarrow{u} q$. Now for the converse, we will show that, with bounded error η, the output set R only contains pairs (p, q) such that there exists a path $\Pi = C_1, \ldots, C_l$ on $\mathcal{G}_\mathcal{A}$ such that $p \in C_1$, $q \in C_l$, and u is Π-compatible. In that case, there is an ε-close word v satisfying $p \xrightarrow{v} q$.

Indeed, using $l \leq m$ and Lemma 29 with t, $\kappa = \varepsilon/(2dm)$ and $\eta' = \eta/2^m$, the samples satisfy $P(u, \Pi) \leq \eta/2^m$, when u is not Π-compatible. Therefore, we can conclude using a union bound argument on all possible paths on $\mathcal{G}_\mathcal{A}$, which have cardinality at most 2^m, that, with probability at least $1 - \eta$, there is no Π such that the samples are Π-compatible but u is not Π-compatible.

The structure of the tester is such that it has only more chances to reject a word that is not Π-compatible given an over-sampling as input instead. Words u such that $p \xrightarrow{u} q$ will always be accepted no matter the amount and length of samples. Therefore the theorem still holds with an over sampling.

Last, $\mathcal{A}$ being Σ'-closed ensures that the notions of compatibility and saturation remain unchanged. Using the Σ'-diameter in Lemma 29 (and therefore in Proposition 28) let us use bridges in Σ'^* instead of Σ^* with weight at most d. ◀