Reduction Ratio of the IS-Algorithm: Worst and Random Cases

Vincent Jugé ✉
LIGM, CNRS, Univ Gustave Eiffel, F77454 Marne-la-Vallée, France

Abstract

We study the IS-algorithm, a well-known linear-time algorithm for computing the suffix array of a word. This algorithm relies on transforming the input word w into another word, called the *reduced* word of w, that will be at least twice shorter; then, the algorithm recursively computes the suffix array of the reduced word. In this article, we study the *reduction ratio* of the IS-algorithm, i.e., the ratio between the lengths of the input word and the word obtained after reducing k times the input word. We investigate both worst cases, in which we find precise results, and random cases, where we prove some strong convergence phenomena. Finally, we prove that, if the input word is a randomly chosen word of length n, we should not expect much more than $\log(\log(n))$ recursive function calls.

2012 ACM Subject Classification Theory of computation → Pattern matching

Keywords and phrases Word combinatorics, Suffix array, IS algorithm

Digital Object Identifier 10.4230/LIPIcs.CPM.2022.8

Related Version *Full Version*: https://arxiv.org/abs/2204.04422

1 Introduction

The suffix array of a word is the permutation of its suffixes that orders them for the lexicographic order. Suffix arrays were introduced in 1990 by Manber and Meyers [9] as a space-efficient alternative to suffix trees. Like suffix trees, they have been used since then in many applications [1, 3, 10]: data compression, pattern matching, plagiarism detection, ...

Suffix arrays were first constructed *via* the construction of suffix trees. Then, various algorithms were proposed to construct suffix arrays directly [4, 5, 6, 7]. A more comprehensive list of approaches towards constructing suffix trees can be found in [14]. In 2010, a new algorithm, called the *IS-algorithm*, was proposed for constructing suffix arrays [12]. This algorithm, which is extremely efficient in practice, is recursive: except if the letters of its input word w are pairwise distinct, in which case the suffix array of w is easy to compute directly, the algorithm transforms w into a shorter word w' and deduces the suffix array of w from the suffix array of w'.

Thus, the question of knowing the *reduction ratio* $|w'|/|w|$ between the lengths of the words w' and w, as well as the number of recursive calls, is critical to evaluating the efficiency of the algorithm. More generally, denoting by $\mathsf{is}^k(w)$ the word obtained after k recursive calls (with $\mathsf{is}^0(w) = w$), we wish to evaluate the ratio $|\mathsf{is}^k(w)|/|w|$ for all k, as well as computing the number of recursive calls that the algorithm will make, i.e., the maximal value of k.

In this article, we focus on these two questions in two different contexts. In Section 3, we consider worst cases, and prove that there exist arbitrarily long words w such that $|\mathsf{is}^k(w)| \approx 2^{-k}|w|$ for all $k \leqslant \log_2(|w|) - 3$, thereby extending results from [2].

Then, in Section 4, we refine the work of [11] and consider words whose letters are generated by a Markov chain of order 1. In this context, and under mild conditions about the Markov chain, we prove, for each integer $k \geqslant 0$, that the ratio $|\mathsf{is}^k(w)|/|w|$ almost surely tends to a given constant γ_k when $|w| \to +\infty$. Finally, in Section 5, we study the constant γ_1 (and,

in some cases, γ_2) when the letters of w are identically and independently generated and, in Section 6, we propose upper bounds on the number of recursive steps on the IS-algorithm when the letters of w are given by a finite Markov chain.

2 Preliminaries

2.1 Definitions and notations

Let $\mathcal{A}$ be a non-empty alphabet, endowed with a linear order $\leqslant$. For every integer $n \geqslant 0$, we denote by $\mathcal{A}^n$ the set of words of length n over $\mathcal{A}$, i.e., the set of sequences of n letters in $\mathcal{A}$. We also denote by $\mathcal{A}^*$ the set of all finite words over $\mathcal{A}$, i.e., the union $\bigcup_{n\geqslant 0} \mathcal{A}^n$, and by ε the empty word.

Let w be a finite word over $\mathcal{A}$. We denote by $|w|$ the length of w, and by $w_0, w_1, \ldots, w_{|w|-1}$ the letters of w. We may abusively denote by w_{-k} the letter $w_{|w|-k}$, i.e., the k^{th} rightmost letter of w. For all integers i and j such that $0 \leqslant i \leqslant j \leqslant |w| - 1$, we also denote by $w_{i\cdots j}$ the word $w_i w_{i+1} \cdots w_j$. Every such word is called a *factor* of w. If $j = |w| - 1$, this word is a suffix of w, and we also denote it by $w_{i\cdots}$. Finally, given two words u and v, we denote by $u \cdot v$ their concatenation, i.e., the word $u_0 u_1 \cdots u_{|u|-1} v_0 v_1 \cdots v_{|v|-1}$.

The *suffix array* [9] of a word $w \in \mathcal{A}^*$ is the unique permutation σ of $\{0, 1, \ldots, |w| - 1\}$ such that $w_{\sigma(0)\cdots} <_{\text{lex}} w_{\sigma(1)\cdots} <_{\text{lex}} \ldots <_{\text{lex}} w_{\sigma(|w|-1)\cdots}$, where $<_{\text{lex}}$ denotes the lexicographic ordering. The IS-algorithm [12] aims at computing the suffix array of its input word w in time linear in $|w|$, when the alphabet $\mathcal{A}$ is either a given finite set or a subset of $\{0, 1, \ldots, |w| - 1\}$.

2.2 Unimodal factors and one-step reduction

Let w be a finite word over $\mathcal{A}$, and let \$ be a fictitious letter, called the *sentinel*, that is defined to be smaller than all letters in $\mathcal{A}$. Below, we simply denote by $\mathcal{A}_\$$ the set $\mathcal{A} \cup \{\$\}$.

An integer $i \leqslant |w| - 1$ is said to be *w-non-decreasing* if there exists an integer j such that $i + 1 \leqslant j \leqslant |w| - 1$ and $w_i = w_{i+1} = \ldots = w_{j-1} < w_j$. If, in addition, $i \geqslant 1$ and $w_{i-1} > w_i$, we say that i is *w-locally minimal*.

Then, let $i_0 < i_1 < \ldots < i_{k-1}$ be the w-locally minimal integers (with $k \geqslant 0$). We also set $i_k = |w|$, and we abusively set $w_{|w|} = \$$. This amounts to replacing w by the word $w \cdot \$$, whose suffix array is the same as the one of w, except that we appended the letter \$ to every suffix and that \$ is now the least non-empty suffix of $w \cdot \$$.

We define the *unimodal factors* of w, also called *LMS factors* [11, 12], as the k words $w_{i_0 \cdots i_1}, w_{i_1 \cdots i_2}, \ldots, w_{i_{k-1} \cdots i_k}$, which belong to $\mathcal{A}^+ \cdot (\varepsilon + \$)$. We call these factors *unimodal* because each sequence $w_{i_\ell}, w_{i_\ell+1}, \ldots, w_{i_{\ell+1}}$ consists of a non-decreasing prefix followed by a non-increasing suffix, and we denote by $\mathsf{eis}(w)$ – for *expanded IS-reduction* of w – the word over the infinite alphabet $\mathcal{A}^+ \cdot (\varepsilon + \$)$ whose letters are the unimodal factors of w.

For instance, if w is the word `COMBINATORIAL` over the latin alphabet $\mathcal{A}$, its unimodal factors are `BINA`, `ATO`, `ORIA` and `AL$`, and thus $\mathsf{eis}(w)$ is the four-letter word `BINA·ATO·ORIA·AL$` over the alphabet $\mathcal{A}^+ \cdot (\varepsilon + \$)$.

In subsequent sections, we may extend to infinite words w (to which we append the letter \$ if w is left-infinite, but not if w is right-infinite) the notions of w-locally minimal integer, of unimodal factor, and of expanded IS-reduction.

The IS-algorithm roughly works as follows:

1. compute w-locally minimal integers and the associated unimodal factors, which form the letters of $\mathsf{eis}(w)$;
2. sort these factors;

3. if w has ℓ distinct unimodal factors, identify each factor with an integer $i \in \{0, 1, \dots, \ell-1\}$: factors f and f' such that $f <_{\text{lex}} f'$ are identified with integers i and i' such that $i < i'$;
4. identify the word $\mathsf{eis}(w)$ with a word $\mathsf{is}(w)$ over the alphabet $\{0, 1, \dots, \ell - 1\}$;
5. compute the suffix array of $\mathsf{is}(w)$, either directly (if the letters of $\mathsf{is}(w)$ are pairwise distinct) or recursively (if at least two letters of $\mathsf{is}(w)$ coincide with each other);
6. based on that array, sort all suffixes of w.

As mentioned by its authors [12], steps 1, 3 and 4 of the algorithm can clearly be performed in time $\mathcal{O}(|w|)$. If $\mathcal{A}$ is a given finite set, or a subset of $\{0, 1, \dots, |w| - 1\}$, bucket sorts allow sorting in linear time unimodal words whose rightmost letters are already sorted, thereby performing steps 2 and 6 in time $\mathcal{O}(|w|)$. Finally, no two consecutive integers $i \leqslant |w| - 1$ are w-locally minimal, and therefore $|\mathsf{is}(w)| \leqslant |w|/2$, thereby proving that the IS-algorithm works in time $\mathcal{O}(|w|)$.

Thus, a natural question would be that of evaluating the constant hidden in this $\mathcal{O}(|w|)$ running time. To that end, we could focus closely on how each of the steps 1 to 4 and 6 is performed. However, several variants might be considered for performing each of these steps. Consequently, we focus on the step 5 and study the behaviour of the ratio $|\mathsf{is}(w)|/|w|$ or, more generally, $|\mathsf{is}^k(w)|/|w|$.

2.3 Markov chains and ergodicity

In Sections 4 to 6, we consider *random* words, whose letters result from a probabilistic process, and are random variables that form a (homogeneous) *Markov chain*. Below, we focus exclusively on such Markov chains, and thus abandon the epithet "homogeneous".

Let $\mathcal{S}$ be a countable set, let $\mu : \mathcal{S} \mapsto \mathbb{R}$ be a probability distribution, and let $M : \mathcal{S} \times \mathcal{S} \mapsto \mathbb{R}$ be a function such that $\sum_{t \in \mathcal{S}} M(s, t) = 1$ for all $s \in \mathcal{S}$. A homogeneous Markov chain with *set of states* $\mathcal{S}$, *initial distribution* μ and *transition matrix* M is a sequence of random variables $(X_n)_{n \geqslant 0}$ with values in $\mathcal{S}$ such that $\mathbb{P}(X_0 = x) = \mu(x)$ for all $x \in \mathcal{S}$ and such that, for every integer $n \geqslant 1$ and every tuple $(x_0, x_1, \dots, x_n) \in \mathcal{S}^{n+1}$, we have

$$\mathbb{P}(X_n = x_n \mid X_0 = x_0, X_1 = x_1, \dots, X_{n-1} = x_{n-1}) = M(x_{n-1}, x_n)$$

whenever $\mathbb{P}(X_0 = x_0, X_1 = x_1, \dots, X_{n-1} = x_{n-1}) > 0$. Below, we identify the Markov chain with the pair (M, μ), or with the transition matrix M in contexts where the initial distribution is irrelevant and might need to be changed. We also abusively say that $(X_n)_{n \geqslant 0}$ is a *trajectory* of the Markov chain (M, μ) or, alternatively, is *generated* by (M, μ).

The *underlying graph* of (M, μ) is the weighted graph $G = (V, E, \pi)$ with vertex set $V = \mathcal{S}$, edge set $E = \{(s, t) \in \mathcal{S} \times \mathcal{S} \colon M(s, t) > 0\}$, and whose weight function $\pi : E \mapsto \mathbb{R}$ is defined by $\pi(s, t) = M(s, t)$. We say that (M, μ) is *irreducible* if G is strongly connected, and *aperiodic* when the lengths of its cycles have no common divisor $d \geqslant 2$.

These notions are connected to the *ergodicity* of a Markov chain, which can be defined as follows. Given a probability distribution ν on $\mathcal{S}$, we denote by $M\nu$ the probability distribution defined by $(M\nu)(x) = \sum_{y \in \mathcal{S}} M(y, x)\nu(y)$. Then, the L^1 distance between two distributions ν and θ is defined as the real number $\|\nu - \theta\|_1 = \sum_{x \in \mathcal{S}} |\nu(x) - \theta(x)|$. The Markov chain M is said to be *ergodic* if there exists a positive probability distribution ν on $\mathcal{S}$ (i.e., a probability distribution such that $\nu(x) > 0$ for all $x \in \mathcal{S}$) such that $\lim_{k \to +\infty} \|\nu - M^k\theta\|_1 = 0$ for all probability distributions θ on $\mathcal{S}$.

Such a distribution ν must be the unique *stationary distribution* of the Markov chain M, i.e., the unique probability distribution such that $\nu = M\nu$. Conversely, when M is irreducible and has a stationary distribution that is positive on $\mathcal{S}$, we say that M is *irreducible and positive recurrent*. This latter assumption relieves us from the need of aperiodicity, and yet retains some desirable properties of ergodic Markov chains.

A typical example of an ergodic Markov chain arises if $\mathbb{P}(X_n = t \,|X_{n-1} = s) = \nu(t)$ for all s and t in $\mathcal{S}$, i.e., if $M\theta = \nu$ for all probability distributions θ on $\mathcal{S}$. In that case, the random variables $(X_n)_{n\geqslant 0}$ are said to be *independent and identically distributed*.

We refer the reader to [8, 13] for a comprehensive review about Markov chains and their properties, from which we present three crucial results below.

▶ **Proposition 1** (Corollary 1.18 and Theorem 21.14 of [8]). *Every ergodic Markov chain is irreducible and aperiodic. Conversely, every irreducible and aperiodic Markov chain is ergodic, provided that its state set is finite or that it has a positive stationary distribution.*

We are particularly interested in Theorem 4.16 of [8], on which we will base Section 4. However, we will not necessarily handle ergodic Markov chains, and therefore we shall relax the notion of ergodicity to a less stringent, *ad hoc* notion that we call *almost surely eventually positive recurrent and irreducible* (or EPRI) Markov chains.

A Markov chain M with underlying graph $G = (\mathcal{S}, E, \pi)$, is said to be EPRI if there exists a set $\mathcal{X} \subseteq \mathcal{S}$ of states, called the *terminal component* of M, such that (i) $\mathcal{X}$ is a strongly connected component of G; (ii) M has stationary distribution ν, i.e., a probability distribution ν such that $M\nu = \nu$, that is positive on $\mathcal{X}$ and zero on $\mathcal{S} \setminus \mathcal{X}$; and (iii) for every initial distribution μ, the sequence generated by (M, μ) almost surely contains a vertex $x \in \mathcal{X}$.

Note that, since ν is positive on $\mathcal{X}$ and zero elsewhere, no edge of G can leave $\mathcal{X}$, i.e., the set E contains no edge (x, y) such that $x \in \mathcal{X}$ and $y \notin \mathcal{X}$.

In this notion, we completely abandon any requirement to be acyclic, which prevents the L^1 convergence that characterises ergodicity. However, when focusing on *average, long-term* behaviours of a Markov chain, such as the frequency of occurrence of a given vertex of sequence of consecutive vertices, whether the Markov chain is cyclic or acyclic is irrelevant. Thus, we may just focus on irreducible, positive recurrent Markov chains. Moreover, in EPRI Markov chains, the path followed before entering the terminal component quickly vanishes. Consequently, the following result, which is usually stated for irreducible, positive recurrent Markov chains only, can be generalised to all EPRI Markov chains whose state space is either finite or countably infinite.

▶ **Theorem 2** (Theorem 4.16 of [8], Theorem 2.1.1 of [13]). *Let $(M, \mu) = (X_n)_{n\geqslant 0}$ be an EPRI Markov chain with set of states $\mathcal{S}$ and stationary distribution ν. Let ℓ be a positive integer, let $f : \mathcal{S}^\ell \mapsto \mathbb{R}$ be a bounded function, and let*

$$\mathbb{E}_\nu[f] = \sum_{x_1,x_2,\ldots,x_\ell \in \mathcal{S}} \nu(x_1)M(x_1, x_2)M(x_2, x_3)\cdots M(x_{\ell-1}, x_\ell)f(x_1, x_2, \ldots, x_\ell).$$

We have

$$\mathbb{P}\left[\frac{1}{n}\sum_{k=0}^{n-1} f(X_k, X_{k+1}, \ldots, X_{k+\ell-1}) \xrightarrow{n\to+\infty} \mathbb{E}_\nu[f]\right] = 1.$$

Proof. It is well-known [13] that Theorem 2 holds when M is irreducible and positive recurrent, i.e., when its state space $\mathcal{S}$ coincides with its terminal component $\mathcal{X}$.

In the general case, trajectories of the Markov chain almost surely meet $\mathcal{X}$ after a finite number of steps, say p, that depends of the trajectory. Once it meets $\mathcal{X}$, the trajectory starts behaving like an irreducible, positive recurrent Markov chain with state space $\mathcal{X}$. Thus,

$$\frac{1}{n-p}\sum_{k=p}^{n-1} f(X_k, X_{k+1}, \ldots, X_{k+\ell-1})$$

converges almost surely (as $n \to +\infty$) to $\mathbb{E}_\nu[f]$. Theorem 2 follows. ◀

Finally, a crucial well-known property of irreducible, positive recurrent Markov chains whose initial distribution coincides with their stationary distribution is that they can be *reversed*.

▶ **Theorem 3** (Proposition 1.22 of [8]). *Let $(X_n)_{n\geqslant 0}$ be an irreducible, positive recurrent Markov chain with set of states $\mathcal{S}$, transition matrix M, and whose initial distribution coincides with the stationary distribution ν of M. For all integers $\ell \geqslant 0$, the sequence $(X_{\ell-n})_{0\leqslant n\leqslant \ell}$ contains the first $\ell+1$ elements of an irreducible, positive recurrent Markov chain, called the* reverse *Markov chain of (M,ν), with initial distribution ν and whose transition matrix $\hat{M}$ is defined by*

$$\hat{M}(x,y) = \frac{\nu(y)}{\nu(x)} M(y,x).$$

More generally, if M is EPRI, and provided that its initial distribution is ν, it already starts inside of its terminal component $\mathcal{X}$, which it cannot leave. Thus, up to deleting those states of M that do not belong to $\mathcal{X}$, the Markov chain M becomes irreducible and positive recurrent, and Theorem 3 applies, with the following caveat: the state space of its reverse Markov chain is restricted to $\mathcal{X}$, and needs not be extended to states outside of $\mathcal{X}$.

3 Deterministic worst case

By construction, no two consecutive integers $i \leqslant |w|-1$ are w-locally minimal, and all w-locally minimal integers belong to the set $\{1,2,\dots,|w|-2\}$. Hence, at most $(|w|-1)/2$ integers are w-locally minimal. This means that $|\mathsf{is}(w)|+1 \leqslant (|w|+1)/2$ and, more generally, that $|\mathsf{is}^k(w)|+1 \leqslant 2^{-k}(|w|+1)$ for every integer $k\geqslant 0$ and every word $w\in\mathcal{A}^*$ such that $\mathsf{is}^k(w)$ exists. A genuine question is then: can we do better? The answer, which was known to be negative [2] when we allow alphabets $\mathcal{A}$ with size $\log_2(|w|)$, remains negative for every fixed size $|\mathcal{A}|\geqslant 2$.

▶ **Theorem 4.** *Let $\mathcal{A}$ be an alphabet of cardinality at least 4. For every integer $n\geqslant 3$, there exists a word $w\in\mathcal{A}^{2^n-1}$ on which the IS-algorithm performs $n-2$ recursive calls, and $|\mathsf{is}^k(w)|+1 = 2^{-k}(|w|+1)$ for all $k\in\{0,1,\dots,n-2\}$.*

Proof. Without loss of generality, we assume that $\mathcal{A}=\{0,1,2,4\}$. Let also $\mathcal{B}=\{0,1,2,3,4\}$. Then, let $\varphi\colon \mathcal{B}^*\mapsto\mathcal{B}^*$ and $\psi\colon\mathcal{B}^*\mapsto\mathcal{A}^*$ be morphisms of monoids, uniquely defined by their values on $\mathcal{B}$: $\varphi(0)=02$, $\varphi(1)=04$, $\varphi(2)=12$, $\varphi(3)=13$ and $\varphi(4)=14$; $\psi(a)=a$ for all $a\in\mathcal{A}$, and $\psi(3)=4$. We prove below that the word $\psi(\varphi^n(3)_{1\cdots})$ satisfies the requirements of Theorem 4.

We say that a word $w=w_0w_1\cdots w_k\in\mathcal{B}^*$ is *balanced* if (1) its length $|w|=k+1$ is even, (2) its rightmost letter $w_k=3$, (3) its suffix $w_{1\cdots}$ contains each of the letters $0,1,2,3,4$, and (4) for all $i\leqslant k-1$, we have $w_i\in\{0,1\}$ if i is even and $w_i\in\{2,4\}$ if i is odd. The eight-letter word $\varphi^3(3)=02140413$ is balanced, and φ maps each balanced word to a balanced word.

Provided that w is balanced, the $\varphi(w)_{1\cdots}$-minimal integers are $1,3,5,\dots,2k-1$, and the associated unimodal factors are $\varphi(w_1)\cdot\varphi(w_2)_0, \varphi(w_2)\cdot\varphi(w_3)_0,\dots,\varphi(w_{k-1})\cdot\varphi(w_k)_0,\varphi(w_k)\cdot\$$. Since $\varphi(0)_1=\varphi(1)_1=0$ and $\varphi(2)_1=\varphi(3)_1=\varphi(4)_1=1$, this means that the unimodal factors of $\varphi(w)$ are $\theta(w_1),\theta(w_2),\dots,\theta(w_k)$, where we set $\theta(0)=021$, $\theta(1)=041$, $\theta(2)=120$, $\theta(3)=13\$$ and $\theta(4)=140$. The function θ is increasing, and thus, $\mathsf{is}(\varphi(w)_{1\cdots})=w_{1\cdots}$.

Moreover, if w is balanced, and since the rightmost letter of $\varphi(w)$ is its only occurrence of the letter 3, the words $\varphi(w)_{1\cdots}$ and $\psi(\varphi(w)_{1\cdots})$ have the same unimodal factors, except that their last factors are $13\$$ and $14\$$, respectively. Hence, $\mathsf{is}(\psi(\varphi(w)_{1\cdots}))=\mathsf{is}(\varphi(w)_{1\cdots})=w_{1\cdots}$. Thus, the map is successively sends $\psi(\varphi^n(3)_{1\cdots})$ to $\varphi^{n-1}(3)_{1\cdots},\varphi^{n-2}(3)_{1\cdots},\dots,\varphi^3(3)_{1\cdots}$, and observing that $\mathsf{is}(\varphi^3(3)_{1\cdots})=201$ completes the proof. ◀

Although the conclusions of Theorem 4 are not valid for alphabets of cardinality 2 or 3, it is still possible to find variants of this worst case. In these variants, the first step of the IS-algorithm is more efficient, with respective reduction ratios of 3 and 5/2, but every word considered after that first step belongs to an alphabet of cardinality 4, which explains why the reduction ratios we compute have similar orders of magnitude.

▶ **Corollary 5.** *Let $\mathcal{A}$ be an alphabet of cardinality 2. For every integer $n \geqslant 3$, there exists a word $w \in \mathcal{A}^{3\times 2^n-2}$ on which the IS-algorithm performs $n-1$ recursive calls, and $|\mathsf{is}^k(w)| + 1 = 2^{1-k}(|w|+2)/3$ for all $k \in \{1, 2, \ldots, n-1\}$.*

Proof. Let us assume that $\mathcal{A} = \{0, 1\}$, and let $\mathcal{B}$ and φ be the alphabet and the morphism defined in the proof of Theorem 4. An immediate induction on ℓ shows that, for all $\ell \geqslant 3$, the word $\varphi^\ell(3)$ starts with the letter 0, ends with the letter 3, and contains $2^{\ell-2}$ letters 0, $2^{\ell-2}$ letters 1, $2^{\ell-2}-1$ letters 2, one letter 3 (the rightmost one) and $2^{\ell-2}$ letters 4.

Then, we consider a new morphism $\psi_2 \colon \mathcal{B}^* \mapsto \mathcal{A}^*$, such that $\psi_2(0) = 0001$, $\psi_2(1) = 001$, $\psi_2(2) = 01$, and $\psi_2(3) = \psi_2(4) = 011$. Like in the proof of Theorem 4, we prove that $\mathsf{is}(1 \cdot \psi_2(w_{1\cdots})) = \mathsf{is}(\varphi(w)_{1\cdots}) = w_{1\cdots}$ when w is balanced, and having counted occurrences of each letter in $\varphi^n(3)$ allows us to conclude that the word $1 \cdot \psi_2(\varphi^n(3)_{1\cdots})$ satisfies the requirements of Corollary 5. ◀

▶ **Corollary 6.** *Let $\mathcal{A}$ be an alphabet of cardinality 3. For every integer $n \geqslant 3$, there exists a word $w \in \mathcal{A}^{5\times 2^n-3}$ on which the IS-algorithm performs n recursive calls, and $|\mathsf{is}^k(w)| + 1 = 2^{2-k}(|w|+3)/5$ for all $k \in \{1, 2, \ldots, n\}$.*

Proof. The proof is the same as that of Corollary 5, except that we have now $\mathcal{A} = \{0, 1, 2\}$ and that, instead of the morphism ψ_2, we use a new morphism $\psi_3 \colon \mathcal{B}^* \mapsto \mathcal{A}^*$, such that $\psi_3(0) = 001$, $\psi_3(1) = 01$, $\psi_3(2) = 012$, $\psi_3(3) = \psi_3(4) = 02$. Indeed, we also have $\mathsf{is}(1 \cdot \psi_3(w_{1\cdots})) = \mathsf{is}(\varphi(w)_{1\cdots}) = w_{1\cdots}$ when w is balanced, from which we conclude that the word $1 \cdot \psi_3(\varphi^n(3)_{1\cdots})$ satisfies the requirements of Corollary 6. ◀

4 Words generated by an ergodic Markov chain

Let $\mathcal{A}$ be a finite or countably infinite set. Below, we study the typical behaviour of the IS-algorithm on a word $w \in \mathcal{A}^n$ whose letters are the first n elements of an EPRI Markov chain (M, μ) with set of states $\mathcal{A}$. We prove below the following result, which is the main (and technically most demanding) result presented in this paper.

▶ **Theorem 7.** *Provided that w is generated by an EPRI Markov chain, and for all integers $k \geqslant 0$, there exist a constant γ_k and a sequence $(\varepsilon_n)_{n\geqslant 0}$ that tends to 0 such that*

$$\mathbb{P}\left[\left|\frac{|\mathsf{is}^k(w)|}{|w|} - \gamma_k\right| \geqslant \varepsilon_{|w|}\right] \leqslant \varepsilon_{|w|}.$$

A particular case of interest arises when w is a word over a finite alphabet generated by an ergodic Markov chain. However, even in that restricted case, studying the words $\mathsf{is}^k(w)$ for $k \geqslant 1$ will require us to consider words over infinite alphabets, which might be generated by Markov chains no longer ergodic, but only EPRI. That is why, facing the need to treat such a generalised setting, we chose to include it from the start in our study.

In addition, all finite-state Markov chains can be decomposed as a "sum" of EPRI Markov chains. Indeed, if the underlying graph of such a Markov chain (M, μ) has k terminal strongly connected components, the Markov chain will almost surely reach one of these

components. Thus, in order to study the Markov chain (M, μ), we may consider, one by one, its k terminal components; for each such component K, compute the probability that (M, μ) eventually reaches K; finally, simulate the behaviour of (M, μ) by first selecting at random which terminal component K it will reach, and then assuming that (M, μ) must reach that component, thereby transforming (M, μ) into an EPRI Markov chain. This allows us to obtain the following variant of Theorem 7.

▶ **Theorem 8.** *Let w be a word whose letters are generated by a finite-state Markov chain. There exist a constant κ and a probability law X over the set $\{1, 2, \dots, \kappa\}$ with the following property: For all integers $k \geqslant 0$, there exist constants $\gamma_{1,k}, \gamma_{2,k}, \dots, \gamma_{\kappa,k}$ and a sequence $(\varepsilon_n)_{n\geqslant 0}$ that tends to 0 such that, for all $i \leqslant \kappa$,*

$$\left| \mathbb{P}\left[\left| \frac{|\mathsf{is}^k(w)|}{|w|} - \gamma_{i,k} \right| \leqslant \varepsilon_{|w|} \right] - \mathbb{P}[X = i] \right| \leqslant \varepsilon_{|w|}.$$

4.1 Generating letters from right to left

In [11], the letters of w are generated from right to left, i.e., the letter w_{n-k} is the k^{th} element of the Markov chain. Here, we mainly focus on this case too. Generating the letters of w from right to left makes things easier because, although being w-non-decreasing is *not* a local property, it enjoys the following local, recursive characterization: an integer i is w-non-decreasing if and only if $i \leqslant |w| - 2$ and either (a) $w_i < w_{i+1}$, or (b) $w_i = w_{i+1}$ and $i + 1$ is w-non-decreasing.

Below, we wish to study the sequence $w, \mathsf{is}(w), \mathsf{is}^2(w), \dots$ and in particular the lengths of these words. In fact, it will be easier to study the sequence $w, \mathsf{eis}(w), \mathsf{eis}^2(w), \dots$ These two sequences differ from each other because they do not use the same alphabets. Yet, for all $k \geqslant 0$, the words $\mathsf{is}^k(w)$ and $\mathsf{eis}^k(w)$ are "isomorphic" to each other: they have the same length, and there exists an increasing mapping φ from the letters of $\mathsf{eis}^k(w)$ to those of $\mathsf{is}^k(w)$, such that $\varphi(\mathsf{eis}^k(w)_i) = \mathsf{is}^k(w)_i$ for all $i < |\mathsf{eis}^k(w)|$.

Following [11, 12], we transform the Markov chain (M, μ) into another Markov chain $(\overline{M}, \overline{\mu})$ that starts with the letter \$ and, in addition to telling which letter we produce, also tells whether the corresponding index is w-non-decreasing: instead of producing letters $a \in \mathcal{A}_\$$, this new Markov chain shall produce pairs $(a, \uparrow)$ or $(a, \downarrow)$, depending on whether the current position is w-non-decreasing or not: we produce a pair $(a, \uparrow)$ if the former case, and $(a, \downarrow)$ in the latter case. Formally, the Markov chain $(\overline{M}, \overline{\mu})$ is defined as follows. Its states form the set $\overline{\mathcal{S}} = \mathcal{A}_\$ \times \{\uparrow, \downarrow\}$. Its initial distribution is defined by $\overline{\mu}(\$, \uparrow) = 1$, and $\overline{\mu}(s) = 0$ whenever $s \neq (\$, \uparrow)$. Its transition matrix is then defined by

$$\begin{cases} \overline{M}\big((\$, \updownarrow), (y, \downarrow)\big) = \mu(y) & \text{if } y \in \mathcal{A}; \\ \overline{M}\big((x, \updownarrow), (y, \downarrow)\big) = M(x, y) & \text{if } (x, y) \in \mathcal{A}^2 \text{ and } x < y; \\ \overline{M}\big((x, \updownarrow), (y, \uparrow)\big) = M(x, y) & \text{if } (x, y) \in \mathcal{A}^2 \text{ and } x > y; \\ \overline{M}\big((x, \updownarrow), (y, \updownarrow)\big) = M(x, y) & \text{if } (x, y) \in \mathcal{A}^2,\ x = y \text{ and } \updownarrow = \updownarrow; \\ \overline{M}\big((x, \updownarrow), (y, \updownarrow)\big) = 0 & \text{otherwise.} \end{cases}$$

▶ **Proposition 9.** *Let (M, μ) be an EPRI Markov chain whose terminal component has size at least two. The Markov chain $(\overline{M}, \overline{\mu})$ defined above is EPRI.*

Proof. Let $G = (\mathcal{A}, E, \pi)$ be the underlying graph of the Markov chain (M, μ), let $\mathcal{X}$ be its terminal component, and let ν be its stationary distribution. In addition, for all $x \in \mathcal{A}$, let $x^\uparrow = \{y \in \mathcal{X} \colon x < y \text{ and } (y, x) \in E\}$ and $x^\downarrow = \{y \in \mathcal{X} \colon x > y \text{ and } (y, x) \in E\}$.

Since $M(x,x) < 1$ for all $x \in \mathcal{A}$, the distribution $\overline{\nu}$ on $\overline{\mathcal{S}}$ defined by $\overline{\nu}(\$, \updownarrow) = 0$ and by

$$\overline{\nu}(x, \updownarrow) = \frac{1}{1 - M(x,x)} \sum_{y \in x^{\updownarrow}} M(y,x)\nu(y)$$

for all $(x, \updownarrow) \in \mathcal{A} \times \{\uparrow, \downarrow\}$ is a probability distribution, because

$$\overline{\nu}(x,\uparrow) + \overline{\nu}(x,\downarrow) = \frac{1}{1 - M(x,x)} \sum_{y\,:\,x \neq y} M(y,x)\nu(y) = \frac{M\nu(x) - M(x,x)\nu(x)}{1 - M(x,x)} = \nu(x) \quad (1)$$

for all $x \in \mathcal{A}$. We further deduce from (1) that

$$\begin{aligned}\overline{M}\overline{\nu}(x, \updownarrow) - M(x,x)\overline{\nu}(x, \updownarrow) = \sum_{y \in x^{\updownarrow}} M(y,x)\big(\overline{\nu}(y,\uparrow) + \overline{\nu}(y,\downarrow)\big) &= \sum_{y \in x^{\updownarrow}} M(y,x)\nu(y) \\ &= (1 - M(x,x))\overline{\nu}(x, \updownarrow),\end{aligned}$$

i.e., that $\overline{M}\overline{\nu}(x, \updownarrow) = \overline{\nu}(x, \updownarrow)$, for all $(x, \updownarrow) \in \mathcal{A} \times \{\uparrow, \downarrow\}$. This means that $\overline{\nu}$ is a stationary distribution of $(\overline{M}, \overline{\mu})$.

This probability distribution is positive on the set

$$\overline{\mathcal{X}} \stackrel{\text{def}}{=\!=} \{(x,\uparrow) \colon x \in \mathcal{X}, x^{\uparrow} \neq \emptyset\} \cup \{(x,\downarrow) \colon x \in \mathcal{X}, x^{\downarrow} \neq \emptyset\}$$

and is zero outside of $\overline{\mathcal{X}}$. Since $\overline{\nu}$ is non-zero, it follows that $\overline{\mathcal{X}}$ is non-empty.

Then, let $\overline{G}$ be the underlying graph of $(\overline{M}, \overline{\mu})$. We shall prove that $\overline{\mathcal{X}}$ satisfies the requirements (i) and (iii) of EPRI Markov chains. Hence, consider some state $(x,\uparrow)$ in $\overline{\mathcal{X}}$, and let y be a state in $x^{\uparrow}$. For every state $(z, \updownarrow)$ in $\overline{\mathcal{X}}$, the graph G contains a finite path from z to x whose second-to-last vertex is y, and thus $\overline{G}$ contains a finite path from $(z, \updownarrow)$ to $(x,\uparrow)$. Similarly, every state $(x,\downarrow)$ in $\overline{\mathcal{X}}$ is accessible from every state $(z, \updownarrow)$ in $\overline{\mathcal{X}}$, and thus $\overline{\mathcal{X}}$ satisfies the requirement (i).

Finally, consider some trajectory $(\overline{X}_n)_{n \geqslant 0}$ of $(\overline{M}, \overline{\mu})$. Deleting its first vertex and removing the second component of each vertex transforms $(\overline{X}_n)_{n \geqslant 0}$ into a trajectory $(X_n)_{n \geqslant 1}$ of the Markov chain M, which almost surely contains a vertex $x \in \mathcal{X}$ and then almost surely meets a vertex distinct from x; let y be the first such vertex. The trajectory $(\overline{X}_n)_{n \geqslant 0}$ contains the vertex $(y,\uparrow)$ if $y < x$, or $(y,\downarrow)$ if $y > x$, and in both cases that vertex belongs to $\overline{\mathcal{X}}$. This shows that $\overline{\mathcal{X}}$ satisfies the requirement (iii). ◀

Using Theorem 2 for the function $f : \overline{\mathcal{S}} \times \overline{\mathcal{S}} \mapsto \mathbb{R}$ defined by

$$\begin{cases} f\big((x,\uparrow),(y,\downarrow)\big) = 1 & \text{for all } x, y \in \mathcal{A}; \\ f(u,v) = 0 & \text{in all other cases} \end{cases}$$

already allows us to prove a special case of Theorem 7 for $k = 1$, which was already proven in [11] in the case $\mathcal{A}$ is finite and (M, μ) is ergodic.

However, if the terminal component of M contains only one state z, the Markov chain $(\overline{M}, \overline{\mu})$ is no longer EPRI, since its graph contains two self-loops around $(z,\uparrow)$ and $(z,\downarrow)$, each one with weight 1. We overcome this difficulty by merging the two states $(z,\uparrow)$ and $(z,\downarrow)$ into one single state z, thereby recovering an EPRI Markov chain, and we modify the function f, redefining it by

$$\begin{cases} f\big((x,\uparrow),(y,\downarrow)\big) = 1 & \text{for all } x, y \in \mathcal{A} \setminus \{z\}; \\ f\big((x,\uparrow),z\big) = 1 & \text{for all } x \in z^{\downarrow}; \\ f(u,v) = 0 & \text{in all other cases.} \end{cases}$$

Tackling this special case allows us to derive the following result, whose validity does not depend on the size of the terminal component of M.

▶ **Corollary 10.** *If the letters of w are generated from right to left by an EPRI Markov chain, there exists a constant γ_1 such that $\mathbb{P}[|\mathsf{eis}(w)|/|w| \to \gamma_1] = 1$.*

Moreover, since $|\mathsf{is}^{k+1}(w)| \leqslant |\mathsf{is}^k(w)|$ for all words w and all integers $k \geqslant 0$, we already know that Theorem 7 holds, with $\gamma_k = 0$, when the terminal component of M has size one. Henceforth, we assume that this terminal component has size at least two.

Under this assumption, let us show that the letters of the word $\mathsf{eis}(w)$ are also generated by a Markov chain. In order to do so, we introduce the function $M^+ \colon \mathcal{A} \to \mathbb{R}$ defined by

$$M^+(x) = \sum_{y\colon x<y} M(x,y)$$

for every letter $x \in \mathcal{A}$, and the function $m \colon \mathcal{A}^+ \cdot (\varepsilon + \$) \to \mathbb{R}$ defined by

$$m(w_0 w_1 \cdots w_k) = M(w_1, w_0) M(w_2, w_1) \cdots M(w_k, w_{k-1})$$

and $m(w \cdot \$) = m(w)\mu(w_{-1})$ for every word $w = w_0 w_1 \cdots w_k$ in $\mathcal{A}^+$. We also define the set

$$\mathcal{U}^\wedge \stackrel{\text{def}}{=\!=} \{w_0 w_1 \cdots w_\ell \in \mathcal{A}^+ \cdot (\varepsilon + \$) \colon M^+(w_0) > 0 \text{ and } \exists k \leqslant \ell, w_0 \leqslant \ldots \leqslant w_{k-1} < w_k \geqslant \ldots \geqslant w_{\ell-1} > w_\ell\}.$$

▶ **Lemma 11.** *The letters of the word $\mathsf{eis}(w)$ are generated from right to left by the Markov chain $(\mathring{M}, \mathring{\mu})$ with set of states $\mathcal{U}^\wedge$, whose initial distribution is defined by*

$$\mathring{\mu}(w) = M^+(w_0) m(w) \mathbf{1}_{w_{-1}=\$}$$

for every word $w \in \mathcal{U}^\wedge$, and whose transition matrix is defined by

$$\mathring{M}(w, w') = \frac{M^+(w'_0)}{M^+(w_0)} \mathbf{1}_{w_0 = w'_{-1}} m(w').$$

Proof. Let $u^{(1)}, u^{(2)}, \ldots, u^{(k)}$ be unimodal words such that $u^{(i)}_{-1} = u^{(i+1)}_0$ for all $i \leqslant k-1$. These are the k rightmost letters of the word $\mathsf{eis}(w)$ if and only if there exists a letter $x \in \mathcal{A}$ such that $x > u^{(1)}_0$ and $w \cdot \$$ ends with the suffix $x \cdot u^{(1)} \cdot u^{(2)}_{1\ldots} \cdot u^{(3)}_{1\ldots} \cdots u^{(k)}_{1\ldots}$, which happens with probability

$$\mathbf{P}_x \stackrel{\text{def}}{=\!=} M\big(u^{(1)}_0, x\big) m\big(u^{(1)}\big) m\big(u^{(2)}\big) \cdots m\big(u^{(k-1)}\big) m\big(u^{(k)}\big) \mathbf{1}_{u^{(k)}_{-1}=\$}.$$

Summing these probabilities $\mathbf{P}_x$ for all $x > u^1_0$, we observe that $u^{(1)}, u^{(2)}, \ldots, u^{(k)}$ are the rightmost letters of $\mathsf{eis}(w)$ with probability

$$\begin{aligned}\mathbf{P} &= M^+\big(u^{(1)}_0\big) m\big(u^{(1)}\big) m\big(u^{(2)}\big) \cdots m\big(u^{(k-1)}\big) m\big(u^{(k)}\big) \mathbf{1}_{u^{(k)}_{-1}=\$} \\ &= \mathring{M}\big(u^{(2)}, u^{(1)}\big) \mathring{M}\big(u^{(3)}, u^{(2)}\big) \cdots \mathring{M}\big(u^{(k)}, u^{(k-1)}\big) \mathring{\mu}\big(u^{(k)}\big).\end{aligned}$$

Finally, Corollary 10 proves that, if w is a left-infinite word whose letters are generated by (M, μ) from right to left, the word $\mathsf{eis}(w)$ is almost surely infinite. It follows that $\mathring{\mu}$ is indeed a probability distribution and that M is indeed a transition matrix, i.e., that

$$\sum_{w' \in \mathcal{U}^\wedge} \mathring{\mu}(w') = 1 \text{ and } \sum_{w' \in \mathcal{U}^\wedge} \mathring{M}(w, w') = 1$$

for all words $w \in \mathcal{U}^\wedge$. ◀

Our next move consists in proving that the Markov chain $(\mathring{M}, \mathring{\mu})$ is EPRI, by exhibiting its stationary distribution. To that end, we first require the following result, which roughly states that "almost surely, every letter of a left-infinite word w generated by (M, μ) belongs to a unimodal factor of w", and whose formal proof can be found in Appendix A.1.

▶ **Lemma 12.** *For all letters $x \in \mathcal{A}$ such that $M^+(x) \neq 0$, we have*

$$\overline{\nu}(x, \uparrow) = \sum_{w \in \mathcal{U}^\wedge \colon x = w_0} m(w) \overline{\nu}(w_{-1}, \uparrow).$$

With this result in hand, we can now prove Proposition 13, following the same lines of the proofs used for Proposition 9.

▶ **Proposition 13.** *Let (M, μ) be an EPRI Markov chain whose terminal component has size at least two. The Markov chain $(\mathring{M}, \mathring{\mu})$ is EPRI.*

Proof. First, let γ_1 be the constant of Corollary 10. Theorem 2 proves that

$$\gamma_1 = \sum_{(x,\uparrow) \in \overline{\mathcal{X}}} \left(\sum_{y \in \mathcal{X} \colon x < y} M(x, y) \overline{\nu}(x, \uparrow) \right) = \sum_{(x,\uparrow) \in \overline{\mathcal{X}}} M^+(x) \overline{\nu}(x, \uparrow).$$

Then, consider the distribution $\mathring{\nu}$ defined by

$$\mathring{\nu}(w) = \frac{1}{\gamma_1} M^+(w_0) m(w) \overline{\nu}(w_{-1}, \uparrow)$$

Lemma 12 proves that

$$\sum_{w \in \mathcal{U}^\wedge} \mathring{\nu}(w) = \frac{1}{\gamma_1} \sum_{x \in \mathcal{A}} M^+(x) \sum_{w \in \mathcal{U}^\wedge \colon x = w_0} m(w) \overline{\nu}(w_{-1}, \uparrow) = \frac{1}{\gamma_1} \sum_{x \in \mathcal{A}} \overline{\nu}(x, \uparrow) M^+(x) = 1,$$

i.e., that $\mathring{\nu}$ is a probability distribution.

Moreover, for every word $w \in \mathcal{U}^\wedge$, Lemma 12 also proves that

$$\begin{aligned} \mathring{M}\mathring{\nu}(w) &= \frac{1}{\gamma_1} \sum_{w' \in \mathcal{U}^\wedge} \mathbf{1}_{w_{-1} = w'_0} M^+(w_0) m(w) m(w') \overline{\nu}(w'_{-1}, \uparrow) \\ &= \frac{1}{\gamma_1} M^+(w_0) m(w) \overline{\nu}(w_{-1}, \uparrow) = \mathring{\nu}(w). \end{aligned}$$

This means that $\mathring{\nu}$ is a stationary probability distribution of $(\mathring{M}, \mathring{\mu})$.

This probability distribution is positive on the set $\mathring{\mathcal{X}} \stackrel{\text{def}}{=\!=} \mathcal{U}^\wedge \cap \mathcal{X}^*$ and is zero outside of that set. Since $\mathring{\nu}$ is a probability distribution, it follows that $\mathring{\mathcal{X}} \neq \emptyset$.

Then, let G and $\mathring{G}$ be the respective underlying graphs of (M, μ) and $(\mathring{M}, \mathring{\mu})$. We shall prove that $\mathring{\mathcal{X}}$ satisfies the requirements (i) and (iii) of EPRI Markov chains.

Hence, consider two words w and w' in $\mathring{\mathcal{X}}$, and let us choose letters $x, y, z, t \in \mathcal{X}$ such that $x \in (w'_{-1})^\uparrow$, $w'_0 \in y^\downarrow$, $z \in w_{-1}^\uparrow$ and $w_0 \in t^\downarrow$. The graph G contains a finite path that starts with the letter x, then the letters of w' (listed from right to left) and then the letter y, and finishes with the letter z, the letters of w (listed from right to left), and then the letter t. Writing these letters from right to left, we obtain a word u whose leftmost unimodal factor is w and whose second rightmost unimodal factor is w'. This proves that $\mathring{G}$ contains a path from w' to w, i.e., that $\mathring{\mathcal{X}}$ satisfies the requirement (i).

Finally, consider some trajectory $(\mathring{X}_n)_{n \geqslant 0}$ of the Markov chain $(\mathring{M}, \mathring{\mu})$. Up to removing the first letter of every word (i.e., vertex) $w \in \mathcal{U}^\wedge$ encountered on this trajectory, reversing these shortened words, and then concatenating the resulting words, we obtain a trajectory

$(X_n)_{n\geqslant 0}$ of (M,μ). That trajectory almost surely contains a vertex $x\in\mathcal{X}$, and will then keep visiting vertices in $\mathcal{X}$. Thus, our initial trajectory almost surely contains a word $\mathring{X}_n$ that is a word with a letter $x\in\mathcal{X}$, and all states $\mathring{X}_m$ such that $m\geqslant n+1$ will then belong to the set $\mathcal{U}^\wedge\cap\mathcal{X}^*=\mathring{X}$, thereby showing that $\mathring{\mathcal{X}}$ satisfies the requirement (iii). ◀

▶ **Proposition 14.** *The conclusion of Theorem 7 holds, provided that the letters of w are generated by an EPRI Markov chain from right to left.*

Proof. Let ℓ be the smallest integer, if any, such that the letters of the word $\mathsf{eis}^\ell(w)$ are not generated, from right to left, by an EPRI Markov chain whose terminal component has size at least two.

If $\ell\geqslant k$, or if ℓ does not exist, applying Corollary 10 to the words $w,\mathsf{eis}(w),\dots,\mathsf{eis}^{k-1}(w)$ proves that, for all $i\leqslant k-1$, there exists a positive constant θ_i such that

$$\mathbb{P}[|\mathsf{eis}^{i+1}(w)|/|\mathsf{eis}^i(w)|\to\theta_i]=1$$

when $|\mathsf{eis}^i(w)|\to+\infty$. In that case, the constant $\gamma_k=\theta_0\theta_1\cdots\theta_{k-1}$ satisfies the requirements of Theorem 7.

However, if $\ell\leqslant k-1$, then $\mathsf{eis}^\ell(w)$ is generated by an EPRI Markov chain whose terminal component has size one, i.e., consists in an absorbing state. In that case, Corollary 10 proves that $|\mathsf{eis}^{\ell+1}(w)|/|\mathsf{eis}^\ell(w)|\to 0$ almost surely, and thus the constant $\gamma_k=0$ satisfies the requirements of Theorem 7. ◀

4.2 Generating letters from left to right

We focus now on the case where the letters of w are generated from left to right, i.e., the letter w_k is the $(k+1)^{\text{th}}$ element of a Markov chain $(\boldsymbol{M},\boldsymbol{\mu})$ – we use a bold-face version of those notations used in Section 4.1.

The two following phenomena make generating the letters of w from left to right harder. First, whether an integer k is w-non-decreasing depends on the letters w_ℓ for $\ell\geqslant k$, and not on the letters w_ℓ for $\ell\leqslant k$. Second, we defined w as the prefix of length n of a right-infinite word $\overline{w}$. However, whether a given integer $k\leqslant n-1$ is w-non-decreasing may depend on n since, for instance, $n-1$ is *never* w-non-decreasing. We overcome this second issue by generalising the notion of non-decreasing integer and of expanded IS-reduction to infinite words, which allows us to use the following result.

▶ **Lemma 15.** *Let $\overline{w}$ be a right-infinite word, let $n\geqslant 4$ be an integer, and let w be a word such that $n-4\leqslant|w|\leqslant n+6$ and $w_{0\cdots n-5}=\overline{w}_{0\cdots n-5}$. Finally, let λ be the number of $\overline{w}$-locally minimal integers that are smaller than n. We have $\lambda-4\leqslant|\mathsf{eis}(w)|\leqslant\lambda+6$, and $\mathsf{eis}(w)_{0\cdots\lambda-5}=\mathsf{eis}(\overline{w})_{0\cdots\lambda-5}$ if $\lambda\geqslant 4$.*

Proof. Let $i_0<i_1<\ldots<i_{\lambda-1}$ the $\overline{w}$-locally minimal integers smaller than n. By construction, we know that $i_j+2\leqslant i_{j+1}$ for all $j\leqslant\lambda-2$. This means that $i_{\lambda-3}\leqslant n-5$, and therefore an integer $j<i_{\lambda-3}$ is $\overline{w}$-locally minimal if and only if it is also w-locally minimal. Thus, the $\lambda-4$ first unimodal factors of both w and $\overline{w}$ are the words $\overline{w}_{i_j\dots i_{j+1}}$, where $0\leqslant j\leqslant\lambda-5$. This already proves that $|\mathsf{eis}(w)|\geqslant\lambda-4$ and that $\mathsf{eis}(w)_{0\cdots\lambda-5}=\mathsf{eis}(\overline{w})_{0\cdots\lambda-5}$.

Finally, if an integer $j\leqslant n-5$ is locally w-minimal but not locally $\overline{w}$-minimal, we know that $\overline{w}_{j-1}=w_{j-1}>w_j=\overline{w}_j$, and therefore j is w-non-decreasing but not $\overline{w}$-non-decreasing. This means that $w_{j-1}>w_j=w_{j+1}=\ldots=w_{n-5}$, and therefore there may be at most one such integer j. Furthermore, since no two consecutive integers may be w-minimal, the

interval $\{n-4, n-3, \ldots, n+5\}$ contains at most five w-locally minimal integers. Hence, there exist at most six w-locally minimal integers that do not belong to the set $\{i_0, i_1, \ldots, i_{\lambda-1}\}$. This means that $|\mathsf{eis}(w)| \leqslant \lambda + 6$. ◀

Lemma 15 allows us to approximate $\mathsf{eis}(w)$ with a prefix of length λ of the word $\mathsf{eis}(\overline{w})$, and proves that this approximation is of excellent quality. Indeed, if we set $\lambda_0 = n$, and inductively define λ_{i+1} as the number of $\mathsf{eis}^i(\overline{w})$-minimal integers smaller than λ_i, Lemma 15 ensures that $\lambda_i - 4 \leqslant |\mathsf{eis}^i(w)| \leqslant \lambda_i + 6$. Thus, evaluating $|\mathsf{eis}^i(w)|$ amounts to evaluating λ_i: this is the task on which we focus below, which allows us to identify w with an right-infinite word, thereby saving us from many technicalities.

The first hurdle we mentioned, which requires being able to "guess" whether a given integer will be w-non-increasing, is easy to overcome by proceeding as follows. When generating a new letter a, the corresponding position in the word has a given probability of being w-non-decreasing, which depends only on a. Thus, we can "guess" whether this position should be w-non-decreasing with the correct probability, and then stick to our guess. Hence, once again, we transform our Markov chain $(\boldsymbol{M}, \boldsymbol{\mu})$ into another Markov chain $(\overline{\boldsymbol{M}}, \overline{\boldsymbol{\mu}})$ that will generate pairs of the form $(w_i, \updownarrow_i)$, where w_i is the $(i+1)^{\text{th}}$ letter of our word w, whereas $\updownarrow_i = \uparrow$ if i is w-non-decreasing, and $\updownarrow_i = \downarrow$ otherwise. Note that, unlike its variant $(\overline{M}, \overline{\mu})$, this Markov chain never generates pairs of the form $(\$, \updownarrow)$, which means that its state space is simply a subset of $\mathcal{A} \times \{\uparrow, \downarrow\}$.

Using this technique allows us to follow the same lines of proof as in Section 4.1. Therefore, we will just mention some milestone constructions and results towards proving Theorem 7, and omit their proofs, which can be found in Appendix A.2.

Assume here that the terminal component of the EPRI Markov chain $(\boldsymbol{M}, \boldsymbol{\mu})$ has size at least two. Before defining the new Markov chain $(\overline{\boldsymbol{M}}, \overline{\boldsymbol{\mu}})$, we first define functions $\boldsymbol{M}^{\uparrow}$ and $\boldsymbol{M}^{\downarrow}$ by

$$\boldsymbol{M}^{\uparrow}(x) = \frac{1}{1 - \boldsymbol{M}(x,x)} \sum_{y\colon x<y} \boldsymbol{M}(x,y) \text{ and } \boldsymbol{M}^{\downarrow}(x) = \frac{1}{1 - \boldsymbol{M}(x,x)} \sum_{y\colon x>y} \boldsymbol{M}(x,y)$$

for all $x \in \mathcal{A}$. Then, the Markov chain $(\overline{\boldsymbol{M}}, \overline{\boldsymbol{\mu}})$ uses the set of states

$$\overline{\boldsymbol{\mathcal{S}}} \stackrel{\text{def}}{=\!=} \{(x, \updownarrow) \in \mathcal{A} \times \{\uparrow, \downarrow\} \colon \boldsymbol{M}^{\updownarrow}(x) \neq 0\},$$

the initial distribution defined by $\overline{\boldsymbol{\mu}}(x, \updownarrow) = \boldsymbol{\mu}(x)\boldsymbol{M}^{\updownarrow}(x)$ for all $(x, \updownarrow) \in \overline{\boldsymbol{\mathcal{S}}}$, and the transition matrix defined by

$$\begin{cases} \overline{\boldsymbol{M}}\big((x,\uparrow),(y,\Updownarrow)\big) = \frac{\boldsymbol{M}^{\Updownarrow}(y)}{\boldsymbol{M}^{\uparrow}(x)} \boldsymbol{M}(x,y) & \text{if } x < y; \\ \overline{\boldsymbol{M}}\big((x,\downarrow),(y,\Updownarrow)\big) = \frac{\boldsymbol{M}^{\Updownarrow}(y)}{\boldsymbol{M}^{\downarrow}(x)} \boldsymbol{M}(x,y) & \text{if } x > y; \\ \overline{\boldsymbol{M}}\big((x,\updownarrow),(y,\Updownarrow)\big) = \boldsymbol{M}(x,x) & \text{if } x = y \text{ and } \updownarrow = \Updownarrow; \\ \overline{\boldsymbol{M}}\big((x,\updownarrow),(y,\Updownarrow)\big) = 0 & \text{otherwise.} \end{cases}$$

As expected, when projecting every pair generated by $(\overline{\boldsymbol{M}}, \overline{\boldsymbol{\mu}})$ onto its first coordinate, we recover a realisation of the Markov chain $(\boldsymbol{M}, \boldsymbol{\mu})$. Furthermore, since the word w is now assumed to be infinite, the k^{th} pair generated by $(\overline{\boldsymbol{M}}, \overline{\boldsymbol{\mu}})$ is of the form $(a, \uparrow)$ if $k-1$ is w-non-decreasing, or $(a, \downarrow)$ otherwise, except if the Markov chain keeps looping around a state $(a, \uparrow)$, which happens with probability 0 since the terminal component has size at least two. In addition, this new Markov chain is, unsurprisingly, EPRI.

If the terminal component of our Markov chain contains only one state, say z, we need to adapt our construction. For all $x \in \mathcal{A} \setminus \{z\}$, we have $\boldsymbol{M}(x,x) < 1$, and thus the above construction is well-defined on such states. Then, we just merge the two states $(z,\uparrow)$ and $(z,\downarrow)$ into a single *sink* state, say $(z,\downarrow)$, and we set $\overline{\boldsymbol{M}}\big((z,\downarrow),(z,\downarrow)\big) = 1$.

Fortunately, the following result does not depend on the size of the terminal component of the Markov chain.

▶ **Proposition 9b.** *Let $(\boldsymbol{M},\boldsymbol{\mu})$ be an EPRI Markov chain. The Markov chain $(\overline{\boldsymbol{M}},\overline{\boldsymbol{\mu}})$ defined above is EPRI.*

Hence, let us consider the function $g\colon \overline{\boldsymbol{\mathcal{S}}} \times \overline{\boldsymbol{\mathcal{S}}} \mapsto \mathbb{R}$ defined by

$$\begin{cases} g\big((x,\downarrow),(y,\uparrow)\big) = 1 & \text{for all } x,y \in \mathcal{A}; \\ g(u,v) = 0 & \text{in all other cases.} \end{cases}$$

Given a realisation $(w_0,\updownarrow_0),(w_1,\updownarrow_1),\dots$ of the Markov chain $(\overline{\boldsymbol{M}},\overline{\boldsymbol{\mu}})$, and denoting by $w = w_0w_1\dots$ the word obtained by projecting these pairs onto their first coordinate, an integer $i \geqslant 1$ is w-locally minimal if and only if $\updownarrow_{i-1} = \downarrow$ and $\updownarrow_i = \uparrow$, i.e., if $g\big((w_{i-1},\updownarrow_{i-1}),(w_i,\updownarrow_i)\big) = 1$. Thus, using Theorem 2 for the function g and Lemma 15 allows us to prove a special case of Theorem 7 for $k = 1$, which consists in the following variant of Corollary 10.

▶ **Corollary 10b.** *If the letters of w are generated from left to right by an EPRI Markov chain, there exists a constant γ_1 such that $\mathbb{P}[\lambda_1/\lambda_0 \to \gamma_1] = 1$ when $\lambda_0 \to +\infty$.*

We focus below on the case where the Markov chain has a terminal component of size at least two. In that case, we show that the letters of $\mathsf{eis}(w)$ are also generated from left to right by an EPRI Markov chain. Mimicking Section 4.1, we introduce the function $\boldsymbol{m}$ defined by

$$\boldsymbol{m}(w_0w_1\cdots w_k) = \boldsymbol{M}(w_0,w_1)\boldsymbol{M}(w_1,w_2)\cdots\boldsymbol{M}(w_{k-1},w_k)$$

for every word $w_0w_1\cdots w_k$ in $\mathcal{A}^*$. We also define the sets

$$\begin{aligned} \boldsymbol{\mathcal{U}}^\wedge &\stackrel{\text{def}}{=} \{w_0w_1\cdots w_\ell \in \mathcal{A}^* \colon \boldsymbol{M}^\uparrow(w_\ell) > 0 \text{ and} \\ &\qquad\qquad \exists k \leqslant \ell-1, w_0 \leqslant \dots \leqslant w_{k-1} < w_k \geqslant w_{k+1} \geqslant \dots \geqslant w_{\ell-1} > w_\ell\} \\ \boldsymbol{\mathcal{V}}^\wedge &\stackrel{\text{def}}{=} \{w_0w_1\cdots w_\ell \in \mathcal{A}^* \colon \exists k \leqslant \ell-1, w_0 \leqslant \dots \leqslant w_{k-1} \leqslant w_k \geqslant w_{k+1} \geqslant \dots \geqslant w_{\ell-1} > w_\ell\}. \end{aligned}$$

▶ **Lemma 11b.** *The letters of the word $\mathsf{eis}(w)$ are generated from left to right by the Markov chain $(\mathring{\boldsymbol{M}},\mathring{\boldsymbol{\mu}})$ with set of states $\boldsymbol{\mathcal{U}}^\wedge$, whose initial distribution is defined by*

$$\mathring{\boldsymbol{\mu}}(w) = \sum_{w' \in \boldsymbol{\mathcal{V}}^\wedge} \mathbf{1}_{w'_{-1}=w_0}\boldsymbol{\mu}(w'_0)\boldsymbol{m}(w')\boldsymbol{m}(w)\boldsymbol{M}^\uparrow(w_{-1}),$$

and whose transition matrix is defined by

$$\mathring{\boldsymbol{M}}(w,w') = \frac{\boldsymbol{M}^\uparrow(w'_{-1})}{\boldsymbol{M}^\uparrow(w_{-1})}\boldsymbol{m}(w')\mathbf{1}_{w_{-1}=w'_0}.$$

▶ **Proposition 13b.** *Let $(\boldsymbol{M},\boldsymbol{\mu})$ be an EPRI Markov chain whose terminal component has size at least two. The Markov chain $(\mathring{\boldsymbol{M}},\mathring{\boldsymbol{\mu}})$ is EPRI.*

The above properties allow us to prove the following result.

▶ **Proposition 14b.** *The conclusion of Theorem 7 holds, provided that the letters of w are generated by an EPRI Markov chain from left to right.*

Proof. Let $\overline{w}$ be the right-infinite word whose letters are generated, from left to right, by our Markov chain. Then, let ℓ be the smallest integer, if any, such that the letters of the word $\mathsf{eis}^\ell(\overline{w})$ are *not* generated, from left to right, by an EPRI Markov chain whose terminal component has size at least two.

If $\ell \geqslant k$, or if ℓ does not exist, applying Corollary 10b to the words $\overline{w}, \mathsf{eis}(\overline{w}), \ldots, \mathsf{eis}^{k-1}(\overline{w})$ proves that, for all $i \leqslant k-1$, there exists a positive constant θ_i such that $\mathbb{P}[\lambda_{i+1}/\lambda_i \to \theta_i] = 1$ when $\lambda_i \to +\infty$. In that case, the constant $\gamma_k = \theta_0\theta_1\cdots\theta_{k-1}$ satisfies the requirements of Theorem 7.

However, if $\ell \leqslant k-1$, then $\mathsf{eis}^\ell(\overline{w})$ is generated by an EPRI Markov chain whose terminal component has size one. In that case, $\lambda_{\ell+1}/\lambda_\ell \to 0$ when $\lambda_\ell \to +\infty$, and therefore the constant $\gamma_k = 0$ satisfies the requirements of Theorem 7. ◀

5 Words with independent and identically distributed letters

Theorem 7 roughly states that, if the letters of a word w are generated (either from left to right or from right to left) by an EPRI Markov chain (M, μ), and provided that $|w|$ is large enough, the ratio $|\mathsf{is}^k(w)|/|w|$ should be approximately equal to a given constant γ_k depending only on k and on the Markov chain.

If we are out of luck, the Markov chain (M, μ) might generate one unique infinite word of the form $w \cdot w \cdot w \cdots$, where w is one of the worst-case words provided in Theorem 4. Consequently, and given an integer $k \geqslant 0$, it is possible to choose the Markov chain (M, μ) in order to have the equality $\gamma_k = 2^{-k}$. This is indeed a worst case, given that $\gamma_{\ell+1} \leqslant \gamma_\ell/2$ for every Markov chain and every integer $\ell \geqslant 0$.

A specific context that will shield us from such bad cases, while being natural, is that of words whose letters $w_0, w_1, \ldots, w_{n-1}$ are independent and identically distributed random variables with values in the alphabet $\mathcal{A}$. Let X be their common probability law. We first recall a result that concerns cases where $\mathcal{A}$ is finite and X is the uniform law over $\mathcal{A}$.

▶ **Proposition 16** (Lemma 3 of [11]). *Let w be a word over a finite alphabet $\mathcal{A}$, whose letters are sampled independently and uniformly over $\mathcal{A}$, i.e., $\mathbb{P}[w_i = a] = 1/|\mathcal{A}|$ for all integers $i \leqslant |w| - 1$ and all letters $a \in \mathcal{A}$. The constant γ_1 of Theorem 7 satisfies the equality*

$$\gamma_1 = \frac{1}{3} - \frac{1}{6|\mathcal{A}|}.$$

This shows that, in the most simple cases, the constant γ is bounded from above by $1/3$, although γ can be arbitrarily close to $1/3$ when the cardinality of $\mathcal{A}$ increases. We prove below that this upper bound is *universal*.

▶ **Proposition 17.** *Let $n \geqslant 1$ be an integer, and let $\mathcal{A}$ be a finite or countably infinite alphabet. Let X be a probability law on $\mathcal{A}$, let*

$$\Omega \stackrel{def}{=\!=} \{t \in [0,1] \colon \exists a \in \mathcal{A} \text{ such that } \mathbb{P}[X < a] < t < \mathbb{P}[X \leqslant a]\}$$

be a subset of $[0,1]$ of Lebesgue measure 1, and let $f : \Omega \mapsto \mathcal{A}$ be the function such that $f(t)$ is the letter $a \in \mathcal{A}$ for which $\mathbb{P}[X < a] < t < \mathbb{P}[X \leqslant a]$. We extend f to a partial function $[0,1]^n \mapsto \mathcal{A}^n$ by setting $f(u_0u_1\cdots u_{n-1}) = f(u_0)f(u_1)\cdots f(u_{n-1})$ if each letter u_i belongs to Ω, and not defining f over $[0,1]^n \setminus \Omega^n$.

For every word $u \in \Omega^n$, we have $|\mathsf{is}(u)| \geqslant |\mathsf{is}(f(u))|$. Furthermore, if the letters $u_0, u_1, \ldots, u_{n-1}$ are independent and distributed according to the uniform law $\mathbb{U}$ over $[0,1]$, they almost surely belong to Ω, and then the letters $f(u_0), f(u_1), \ldots, f(u_{n-1})$ are also independent and distributed according to the law X.

Proof. First, Ω is a disjoint union of countably many intervals whose lengths $\mathbb{P}[X = a]$ sum up to 1, and thus it has Lebesgue measure 1. The last sentence of Proposition 17 is then immediate. Hence, we focus on proving that $|\mathsf{is}(u)| \geqslant |\mathsf{is}(f(u))|$ when $u \in \Omega^n$.

Given a word w, we say that a sequence of integers $a_1 < b_1 \leqslant a_2 < b_2 \leqslant \ldots \leqslant a_{2k} < b_{2k}$ is *w-alternating of size k* if $b_{2k} < |w|$, $w_{a_i} > w_{b_i}$ for all *odd* indices i, and $w_{a_i} < w_{b_i}$ for all *even* indices i. One checks easily that $|\mathsf{is}(w)|$ is the largest size of a w-alternating sequence. Since every $f(u)$-alternating sequence is also u-alternating, Proposition 17 follows. ◀

Unfortunately, in general, the letters of the word $\mathsf{is}(u)$ are not independent, and both inequalities $|\mathsf{is}^2(u)| < |\mathsf{is}^2(f(u))|$ and $|\mathsf{is}^2(u)| > |\mathsf{is}^2(f(u))|$ may hold, which prevents us from designing simple bijection-flavoured variants of Proposition 17 for investigating the length of $\mathsf{is}^k(f(u))$. Yet, Proposition 17 still leads to the following result.

▶ **Theorem 18.** *For every alphabet $\mathcal{A}$ and every probability law X on $\mathcal{A}$, we have $\gamma_1 \leqslant 1/3$.*

Proof. Let u and w be n-letter words whose letters are independent random variables following the laws $\mathbb{U}$ and X, as described in the statement of Proposition 17. Each integer $i \in \{1, 2, \ldots, n-2\}$ is u-minimal if and only if $u_i = \min\{u_{i-1}, u_i, u_{i+1}\}$, which happens with probability $1/3$, while 0 and $n-1$ cannot be u-minimal. It follows that

$$\mathbb{E}[|\mathsf{is}(w)|] \leqslant \mathbb{E}[|\mathsf{is}(u)|] = (n-2)/3 \leqslant n/3$$

and, thanks to Theorem 7, that $\gamma_1 \leqslant 1/3$. ◀

In view of Proposition 16 and Theorem 18, proving that $\gamma_1 \leqslant 1/3 - 1/(6|\mathcal{A}|)$ even if X is *not* uniform might be tempting. Unfortunately, the inequality is invalid when $|\mathcal{A}| = 3$ and $(p_1, p_2, p_3) = (3/8, 1/4, 3/8)$, because in that case $\gamma_1 = 9/32 > 5/18 = 1/3 - 1/(6|\mathcal{A}|)$.

However, the case $|\mathcal{A}| = 2$ is still promising. Indeed, in that case, $\gamma_1 = p_1(1-p_1) \leqslant 1/4$, and the letters of the word $\mathsf{eis}(w)$ are independent and identically distributed, since the only constraints they are subject to is that they should begin with the letter 0 and end with the suffix 10. Thus, we can still use Theorem 18 to evaluate the ratio $|\mathsf{is}^2(w)|/|\mathsf{is}(w)|$, thereby deriving the following result, which suggests excellent performances of the IS-algorithm.

▶ **Proposition 19.** *If $|\mathcal{A}| = 2$, we have $\gamma_1 \leqslant 1/4$ and $\gamma_2 \leqslant 1/12$.*

6 Bounding the number of function calls

In this last section, we provide a short argument for proving that, if $\mathcal{A}$ is finite and if the letters of the word w are generated, either from left to right or from right to left, by a (non necessarily EPRI) Markov chain (M, μ), we should expect $\mathcal{O}(\log(\log(|w|)))$ recursive function calls. This is the object of the following result, whose formal proof can be found in Appendix A.3.

▶ **Theorem 20.** *Let $w \in \mathcal{A}^n$ be a word whose letters are generated by a Markov chain (M, μ). For all integers $\ell \geqslant 0$, and provided that n is large enough, the IS-algorithm has a probability $\mathbf{P} \leqslant n^{-2^\ell}$ of performing more than $2\log_2(\log_2(n)) + \ell$ recursive function calls.*

Proof idea. The probability that two independent trajectories of M (whose initial distributions may differ) coincide with each other on their k first steps decreases exponentially fast with k, unless they get trapped into a cycle from which they cannot escape. However,

every letter of the word $\mathsf{eis}^\ell(w)$ represents at least 2^ℓ letters from w. Thus, if two such letters coincide, the word w must contain two identical subwords of length 2^ℓ, an event whose probability decreases severely once 2^ℓ exceeds $\log(|w|)$.

It remains to treat the case where w gets trapped into a cycle from which it cannot escape. Again, the probability that it would take more than k steps to reach that cycle decreases exponentially fast with k, and, when $\ell \geqslant \log_2(k)$, these n steps (i.e., letters) will all be subsumed in the same letter of the word $\mathsf{eis}^\ell(w)$. However, all the other letters of $\mathsf{eis}^\ell(w)$ will coincide with each other, and thus $\mathsf{eis}^{\ell+1}(w)$ will contain at most one letter, thereby preventing subsequent recursive calls to the IS-algorithm. ◀

This result illustrates the fact that detecting as soon as possible special cases in which suffix arrays are easy to compute (here, observing that the letters of w are pairwise distinct) can result in dramatically decreasing the size of the recursive call stack. However, the notion of being a *large enough* integer n heavily depends on the Markov chain (M, μ), as illustrated by the worst cases studied in Section 3, which can be arbitrarily well approximated by Markov chains.

References

1 Mohamed Ibrahim Abouelhoda, Stefan Kurtz, and Enno Ohlebusch. Replacing suffix trees with enhanced suffix arrays. *Journal of discrete algorithms*, 2(1):53–86, 2004.

2 Timo Bingmann, Johannes Fischer, and Vitaly Osipov. Inducing suffix and LCP arrays in external memory. *Journal of Experimental Algorithmics (JEA)*, 21:1–27, 2016.

3 Maxime Crochemore, Lucian Ilie, and William F Smyth. A simple algorithm for computing the Lempel Ziv factorization. In *Data Compression Conference (DCC 2008)*, pages 482–488. IEEE, 2008.

4 Juha Kärkkäinen, Dominik Kempa, Simon J Puglisi, and Bella Zhukova. Engineering external memory induced suffix sorting. In *2017 Proceedings of the Ninteenth Workshop on Algorithm Engineering and Experiments (ALENEX)*, pages 98–108. SIAM, 2017.

5 Juha Kärkkäinen and Peter Sanders. Simple linear work suffix array construction. In *International Colloquium on Automata, Languages, and Programming (ICALP)*, pages 943–955. Springer, 2003.

6 Dong Kyue Kim, Jeong Seop Sim, Heejin Park, and Kunsoo Park. Linear-time construction of suffix arrays. In *Annual Symposium on Combinatorial Pattern Matching (CPM)*, pages 186–199. Springer, 2003.

7 Pang Ko and Srinivas Aluru. Space efficient linear time construction of suffix arrays. *Journal of Discrete Algorithms*, 3(2-4):143–156, 2005.

8 David Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Society, 2017.

9 Udi Manber and Gene Myers. Suffix arrays: a new method for on-line string searches. *SIAM Journal on Computing*, 22(5):935–948, 1993.

10 Maxim Mozgovoy, Kimmo Fredriksson, Daniel White, Mike Joy, and Erkki Sutinen. Fast plagiarism detection system. In *International Symposium on String Processing and Information Retrieval (SPIRE)*, pages 267–270. Springer, 2005.

11 Cyril Nicaud. A probabilistic analysis of the reduction ratio in the suffix-array IS-algorithm. In *Annual Symposium on Combinatorial Pattern Matching (CPM)*, pages 374–384. Springer, 2015.

12 Ge Nong, Sen Zhang, and Wai Hong Chan. Two efficient algorithms for linear time suffix array construction. *IEEE Transactions on Computers*, 60(10):1471–1484, 2010.

13 Ursula Porod. Dynamics of Markov chains for undergraduates, 2021. URL: `https://www.math.northwestern.edu/documents/book-markov-chains.pdf`.

14 Simon J Puglisi, William F Smyth, and Andrew H Turpin. A taxonomy of suffix array construction algorithms. *ACM Computing Surveys (CSUR)*, 39(2):4–es, 2007.

A Appendix

A.1 Proving Lemma 12

We focus here on formally proving Lemma 12, whose intuitive meaning was already given in Section 4.1. To that end, we first introduce new variants of the set $\mathcal{U}^{\wedge}$. These are the sets

$$\mathcal{U}^{\backslash} \stackrel{\text{def}}{=\!=} \{w_0 w_1 \cdots w_\ell \in \mathcal{A}^* \cdot (\varepsilon + \$) \colon w_0 \geqslant \ldots \geqslant w_{\ell-1} > w_\ell\}$$
$$\mathcal{U}' \stackrel{\text{def}}{=\!=} \{w_0 w_1 \cdots w_\ell \in \mathcal{A}^* \colon w_0 \leqslant \ldots \leqslant w_{\ell-1} < w_\ell\}$$

of non-increasing (respectively, non-decreasing) words in $\mathcal{A}^* \cdot (\varepsilon + \$)$ whose last two letters differ from each other. We can now prove the following auxiliary result, from which we will then deduce Lemma 12.

▶ **Lemma 11-1.** *For all letters $x \in \mathcal{A}$, we have*

$$\overline{\nu}(x,\downarrow) = \sum_{w \in \mathcal{U}^{\backslash} \colon x = w_0} m(w)\overline{\nu}(w_{-1},\uparrow) \text{ and } \overline{\nu}(x,\uparrow) = \sum_{w \in \mathcal{U}' \colon x = w_0} m(w)\overline{\nu}(w_{-1},\downarrow).$$

Proof. Up to reversing the order $\leqslant$ on $\mathcal{A}_\$$, both equalities are equivalent to each other. Hence, we focus on proving the left one. Let x be some element of $\mathcal{X}$, let $\hat{M}$ the *reverse* transition matrix of $\overline{M}$, such as described in Theorem 3, and let $(Y_n)_{n \geqslant 0}$ be the Markov chain with first element $Y_0 = x$ and with transition matrix $\hat{M}$. Then, let $\mathbf{T}$ be the stopping time defined as the smallest integer $n \geqslant 1$ such that Y_n belongs to the set $\{(y,\uparrow) \colon y \in \mathcal{X}\}$. Since $\hat{M}$ is EPRI, the stopping time $\mathbf{T}$ is almost surely finite.

For each word $w \in \mathcal{U}^{\backslash}$ such that $x = w_0$ and $w_{-1}^{\uparrow} \neq \emptyset$, i.e., $\overline{\nu}(w_{-1},\uparrow) \neq 0$, the Markov chain $(Y_n)_{n \geqslant 0}$ has a probability

$$\mathbf{P}_w \stackrel{\text{def}}{=\!=} \hat{M}\big((w_0,\downarrow),(w_1,\downarrow)\big)\hat{M}\big((w_1,\downarrow),(w_2,\downarrow)\big) \cdots \hat{M}\big((w_{-2},\downarrow),(w_{-1},\uparrow)\big)$$

of starting with the letters $(w_0,\downarrow),(w_1,\downarrow),\ldots,(w_{-2},\downarrow),(w_{-1},\uparrow)$, in which case $\mathbf{T} = |w| - 1$. Using Theorem 3 and the construction of $\overline{M}$, we have

$$\mathbf{P}_w = \frac{\overline{\nu}(w_{-1},\uparrow)}{\overline{\nu}(x,\downarrow)} M(w_1,w_0)M(w_2,w_1)\cdots M(w_{-1},w_{-2}) = \frac{m(w)\overline{\nu}(w_{-1},\uparrow)}{\overline{\nu}(x,\downarrow)}.$$

Conversely, whenever $\mathbf{T} < +\infty$, the Markov chain $(Y_n)_{n \geqslant 0}$ starts with such a sequence of letters. Consequently, the probabilities $\mathbf{P}_w$ sum up to 1, which completes the proof. ◀

▶ **Lemma 12.** *For all letters $x \in \mathcal{A}$ such that $M^+(x) \neq 0$, we have*

$$\overline{\nu}(x,\uparrow) = \sum_{w \in \mathcal{U}^{\wedge} \colon x = w_0} m(w)\overline{\nu}(w_{-1},\uparrow).$$

Proof. Let us associate every pair $(u,v) \in \mathcal{U}' \times \mathcal{U}^{\backslash}$ such that $u_{-1} = v_0$ with the word $w \stackrel{\text{def}}{=\!=} u \cdot v_{1\ldots} \in \mathcal{U}^{\wedge}$. Lemma 11-1 then proves that

$$\overline{\nu}(x,\uparrow) = \sum_{u \in \mathcal{U}' \colon x = u_0} \left(\sum_{v \in \mathcal{U}^{\backslash} \colon u_{-1} = v_0} m(u)m(v)\overline{\nu}(v_{-1},\uparrow) \right) = \sum_{w \in \mathcal{U}^{\wedge} \colon x = w_0} m(w)\overline{\nu}(w_{-1},\uparrow). \quad ◀$$

A.2 Proving Proposition 14b

We focus here on formally proving Proposition 14b, by providing complete proofs of the results mentioned in Section 4.2. This proofs had first been omitted because of their similarity to those of Section 4.1. Consequently, we list below results that were mentioned explicitly in Section 4.2 (sometimes adapting their wording) or were left implicit in Section 4.2 but whose variants had appeared in Section 4.1.

▶ **Proposition 9b.** *Let $(\boldsymbol{M}, \boldsymbol{\mu})$ be an EPRI Markov chain whose terminal component has size at least two. The Markov chain $(\overline{\boldsymbol{M}}, \overline{\boldsymbol{\mu}})$ defined in Section 4.2 is EPRI.*

Proof. Let $\boldsymbol{G} = (\mathcal{A}, \boldsymbol{E}, \boldsymbol{\pi})$ be the underlying graph of the Markov chain $(\boldsymbol{M}, \boldsymbol{\mu})$, let $\boldsymbol{\mathcal{X}}$ be its terminal component, and let $\boldsymbol{\nu}$ be its stationary distribution. In addition, for all $x \in \mathcal{A}$, let $x^{\uparrow} = \{y \in \mathcal{X} \colon x < y \text{ and } (x, y) \in E\}$ and $x^{\downarrow} = \{y \in \mathcal{X} \colon x > y \text{ and } (x, y) \in E\}$.

The distribution $\overline{\boldsymbol{\nu}}$ on $\overline{\boldsymbol{\mathcal{S}}}$ defined by $\overline{\boldsymbol{\nu}}(x, \updownarrow) = \boldsymbol{\nu}(x)\boldsymbol{M}^{\updownarrow}(x)$ is a probability distribution, because

$$\overline{\boldsymbol{\nu}}(x, \uparrow) + \overline{\boldsymbol{\nu}}(x, \downarrow) = \frac{1}{1 - \boldsymbol{M}(x, x)} \sum_{y \colon x \neq y} \boldsymbol{M}(x, y)\boldsymbol{\nu}(x) = \boldsymbol{\nu}(x) \tag{2}$$

for all $x \in \mathcal{A}$. We also deduce from (2) that

$$\begin{aligned}
\overline{\boldsymbol{M}}\overline{\boldsymbol{\nu}}(x, \updownarrow) - \boldsymbol{M}(x, x)\overline{\boldsymbol{\nu}}(x, \updownarrow) &= \textstyle\sum_{y \colon x < y} \frac{\boldsymbol{M}^{\updownarrow}(x)}{\boldsymbol{M}^{\downarrow}(y)} \boldsymbol{M}(y, x)\overline{\boldsymbol{\nu}}(y, \downarrow) + \sum_{y \colon x > y} \frac{\boldsymbol{M}^{\updownarrow}(x)}{\boldsymbol{M}^{\uparrow}(y)} \boldsymbol{M}(y, x)\overline{\boldsymbol{\nu}}(y, \uparrow) \\
&= \boldsymbol{M}^{\updownarrow}(x) \sum_{y \colon x \neq y} \boldsymbol{M}(y, x)\boldsymbol{\nu}(y) \\
&= \boldsymbol{M}^{\updownarrow}(x) \left(\boldsymbol{M}\boldsymbol{\nu}(x) - \boldsymbol{M}(x, x)\boldsymbol{\nu}(x)\right) = (1 - \boldsymbol{M}(x, x))\overline{\boldsymbol{\nu}}(x, \updownarrow),
\end{aligned}$$

i.e., that $\overline{\boldsymbol{M}}\overline{\boldsymbol{\nu}}(x, \updownarrow) = \overline{\boldsymbol{\nu}}(x, \updownarrow)$, for all $(x, \updownarrow) \in \mathcal{A} \times \{\uparrow, \downarrow\}$. This means that $\overline{\boldsymbol{\nu}}$ is a stationary distribution of $(\overline{\boldsymbol{M}}, \overline{\boldsymbol{\mu}})$.

This probability distribution is positive on the set $\overline{\boldsymbol{\mathcal{X}}} \stackrel{\text{def}}{=\!=} \{(x, \updownarrow) \in \overline{\boldsymbol{\mathcal{S}}} \colon x \in \boldsymbol{\mathcal{X}}\}$, and zero outside of $\overline{\boldsymbol{\mathcal{X}}}$. Since $\overline{\boldsymbol{\nu}}$ is non-zero, it follows that $\overline{\boldsymbol{\mathcal{X}}}$ is non-empty.

Now, let $\overline{\boldsymbol{G}}$ be the underlying graph of $(\overline{\boldsymbol{M}}, \overline{\boldsymbol{\mu}})$. We shall prove that $\overline{\boldsymbol{\mathcal{X}}}$ satisfies the requirements (i) and (iii) of EPRI Markov chains.

Consider two states $(x, \updownarrow)$ in $\overline{\boldsymbol{\mathcal{X}}}$ and $(z, \updownarrow)$ in $\overline{\boldsymbol{\mathcal{S}}}$. Let y and t be letters in $x^{\updownarrow}$ and $z^{\updownarrow}$, respectively. The graph $\boldsymbol{G}$ contains a finite path from z to y whose second vertex is t and whose second last vertex is x. Therefore, $\overline{\boldsymbol{G}}$ contains a finite path from $(z, \updownarrow)$ to $(x, \updownarrow)$, which shows that $\overline{\boldsymbol{\mathcal{X}}}$ satisfies the requirement (i).

Finally, consider a trajectory $(\boldsymbol{Y}_n)_{n \geqslant 0}$ of $\overline{\boldsymbol{M}}$. Its projection onto the first component is a trajectory in $\overline{\boldsymbol{G}}$, and almost surely contains a vertex $x \in \mathcal{X}$, followed by another vertex y. Thus, $(\boldsymbol{Y}_n)_{n \geqslant 0}$ contains the vertex $(x, \uparrow)$ if $x < y$, or $(x, \downarrow)$ if $x > y$, and in both cases that vertex belongs to $\overline{\boldsymbol{\mathcal{X}}}$. This shows that $\overline{\boldsymbol{\mathcal{X}}}$ satisfies the requirement (iii). ◀

▶ **Lemma 11b.** *The letters of the word $\mathsf{eis}(w)$ are generated from left to right by the Markov chain $(\mathring{\boldsymbol{M}}, \mathring{\boldsymbol{\mu}})$ with set of states $\boldsymbol{\mathcal{U}}^{\wedge}$, whose initial distribution is defined by*

$$\mathring{\boldsymbol{\mu}}(w) = \sum_{w' \in \boldsymbol{\mathcal{V}}^{\wedge}} \mathbf{1}_{w'_{-1} = w_0} \boldsymbol{\mu}(w'_0)\boldsymbol{m}(w')\boldsymbol{m}(w)\boldsymbol{M}^{\uparrow}(w_{-1}),$$

and whose transition matrix is defined by

$$\mathring{\boldsymbol{M}}(w, w') = \frac{\boldsymbol{M}^{\uparrow}(w'_{-1})}{\boldsymbol{M}^{\uparrow}(w_{-1})} \boldsymbol{m}(w') \mathbf{1}_{w_{-1} = w'_0}.$$

Proof. Let $u^{(1)}, u^{(2)}, \dots, u^{(k)}$ be unimodal words such that $u^{(i)}_{-1} = u^{(i+1)}_0$ for all $i \leqslant k-1$. These are the k leftmost letters of the word $\mathsf{eis}(w)$ if and only if there exists a word $v \in \boldsymbol{\mathcal{V}}^\wedge$, two letters $x, y \in \mathcal{A}$ and an integer $\ell \geqslant 0$ such that $v_{-1} = u^{(1)}_0$, $u^{(k)}_{-1} = x < y$, and w begins with the prefix $v \cdot u^{(1)}_{1\dots} \cdot u^{(2)}_{1\dots} \cdots u^{(k)}_{1\dots} \cdot x^\ell \cdot y$. This happens with probability

$$\mathbf{P}_{v,x^{\ell-1}\cdot y} \stackrel{\text{def}}{=\!=} \boldsymbol{\mu}(v_0)\boldsymbol{m}(v)\boldsymbol{m}(u^{(1)})\boldsymbol{m}(u^{(2)})\cdots\boldsymbol{m}(u^{(k)})\boldsymbol{M}(x,x)^\ell\boldsymbol{M}(x,y).$$

Summing these probabilities for all v, y and ℓ, we observe that $u^{(1)}, u^{(2)}, \dots, u^{(k)}$ are the left letters of $\mathsf{eis}(w)$ with probability

$$\begin{aligned}\mathbf{P} &= \sum_{v\in\boldsymbol{\mathcal{V}}^\wedge} \mathbf{1}_{v_{-1}=w_0}\boldsymbol{\mu}(v_0)\boldsymbol{m}(v)\boldsymbol{m}\big(u^{(1)}\big)\boldsymbol{m}\big(u^{(2)}\big)\cdots\boldsymbol{m}\big(u^{(k)}\big)\boldsymbol{M}^\uparrow\big(u^{(k)}_{-1}\big)\\ &= \mathring{\boldsymbol{\mu}}\big(u^{(1)}\big)\mathring{\boldsymbol{M}}\big(u^{(1)},u^{(2)}\big)\mathring{\boldsymbol{M}}\big(u^{(2)},u^{(3)}\big)\cdots\mathring{\boldsymbol{M}}\big(u^{(k-1)},u^{(k)}\big).\end{aligned}$$

Finally, Corollary 10b proves that, if w is a right-infinite word whose letters are generated by $(\boldsymbol{M}, \boldsymbol{\mu})$ from left to right, the word $\mathsf{eis}(w)$ is almost surely infinite. It follows that $\mathring{\boldsymbol{\mu}}$ is indeed a probability distribution that $\mathring{\boldsymbol{M}}$ is indeed a transition matrix, i.e., that

$$\sum_{w'\in\boldsymbol{\mathcal{U}}^\wedge}\mathring{\boldsymbol{\mu}}(w') = 1 \text{ and } \sum_{w'\in\boldsymbol{\mathcal{U}}^\wedge}\mathring{\boldsymbol{M}}(w,w') = 1$$

for all words $w \in \boldsymbol{\mathcal{U}}^\wedge$. ◀

Then, we adapt Lemma 11-1, which requires introducing variants of the sets $\mathcal{U}^\wedge$ and $\mathcal{U}'$ of Section 4.1. These variants are the sets

$$\begin{aligned}\boldsymbol{\mathcal{U}}^\wedge &\stackrel{\text{def}}{=\!=} \{w_0w_1\cdots w_\ell \in \mathcal{A}^* \colon w_0 < w_1 \geqslant w_2 \geqslant \dots \geqslant w_{\ell-1}\}\\ \boldsymbol{\mathcal{U}}' &\stackrel{\text{def}}{=\!=} \{w_0w_1\cdots w_\ell \in \mathcal{A}^* \colon w_0 > w_1 \leqslant w_2 \leqslant \dots \leqslant w_{\ell-1}\}.\end{aligned}$$

▶ **Lemma 11-1b.** *For all letters $x \in \mathcal{A}$, we have*

$$\boldsymbol{\nu}(x) = \sum_{w\in\boldsymbol{\mathcal{U}}^\wedge\colon\, x=w_{-1}}\boldsymbol{\nu}(w_0)\boldsymbol{m}(w) \text{ and } \boldsymbol{\nu}(x) = \sum_{w\in\boldsymbol{\mathcal{U}}'\colon\, x=w_{-1}}\boldsymbol{\nu}(w_0)\boldsymbol{m}(w).$$

Proof. Up to reversing the order $\leqslant$ on $\mathcal{A}$, both equalities are equivalent to each other. Hence, we focus on proving the left one. Let x be some element of $\boldsymbol{\mathcal{X}}$, let $\hat{\boldsymbol{M}}$ be the *reverse* transition matrix of $\boldsymbol{M}$, such as described in Theorem 3, and let $(\boldsymbol{Y}_n)_{n\geqslant 0}$ be the Markov chain with first element $\boldsymbol{Y}_0 = x$ and with transition matrix $\hat{\boldsymbol{M}}$. Finally, let $\mathbf{T}$ be the stopping time defined as the smallest integer $n \geqslant 1$ such that $\boldsymbol{Y}_n < \boldsymbol{Y}_{n-1}$. Since $\hat{\boldsymbol{M}}$ is EPRI, $\mathbf{T}$ is almost surely finite.

For each word $w \in \boldsymbol{\mathcal{U}}^\wedge$ such that $x = w_{-1}$, the Markov chain $(\boldsymbol{Y}_n)_{n\geqslant 0}$ has a probability

$$\mathbf{P}_w \stackrel{\text{def}}{=\!=} \hat{\boldsymbol{M}}(w_{-1},w_{-2})\cdots\hat{\boldsymbol{M}}(w_2,w_1)\hat{\boldsymbol{M}}(w_1,w_0)$$

of starting with the letters $w_{-1}, \dots, w_2, w_1, w_0$, in which case $\mathbf{T} = |w| - 1$. Theorem 3 thus proves that

$$\mathbf{P}_w = \frac{\boldsymbol{\nu}(w_0)}{\boldsymbol{\nu}(w_{-1})}\boldsymbol{M}(w_0,w_1)\boldsymbol{M}(w_1,w_2)\cdots\boldsymbol{M}(w_{-2},w_{-1}) = \frac{\boldsymbol{m}(w)\boldsymbol{\nu}(w_0)}{\boldsymbol{\nu}(x)}.$$

Conversely, whenever $\mathbf{T} < 0$, the Markov chain $(\boldsymbol{Y}_n)_{n\geqslant 0}$ starts with such a sequence of letters. Consequently, the probabilities $\mathbf{P}_w$ sum up to 1, which completes the proof. ◀

Let us now introduce the function $\boldsymbol{\nu}^+ \colon \mathcal{A} \to \mathbb{R}$ defined by

$$\boldsymbol{\nu}^+(x) = \sum_{y \colon x<y} \boldsymbol{\nu}(y)\boldsymbol{M}(y,x)$$

for every letter $x \in \mathcal{A}$.

▶ **Lemma 12b.** *For all letters $x \in \mathcal{A}$, we have*

$$\boldsymbol{\nu}^+(x) = \sum_{w \in \boldsymbol{\mathcal{U}}^\wedge \colon x = w_{-1}} \boldsymbol{\nu}^+(w_0)\boldsymbol{m}(w).$$

Proof. We associate every pair $(u,v) \in \mathcal{U}' \times \mathcal{U}^\setminus$ such that $u_{-1} = v_0$ and $v_{-1} > x$ with the pair $(y,w) \stackrel{\text{def}}{=\!=} (u_0, u_{1\dots} \cdot v_{1\dots} \cdot x) \in \mathcal{A} \times \mathcal{U}^\wedge$, which is such that $y > w_0$. This association is bijective, and thus Lemma 11-1b proves that

$$\begin{aligned}\boldsymbol{\nu}^+(x) &= \sum_{y \colon x<y} \boldsymbol{\nu}(y)\boldsymbol{M}(y,x) = \sum_{y \colon x<y} \left(\sum_{v \in \boldsymbol{\mathcal{U}}^\setminus \colon v_{-1}=y} \left(\sum_{u \in \boldsymbol{\mathcal{U}}' \colon u_{-1}=v_0} \boldsymbol{\nu}(u_0)\boldsymbol{m}(u)\boldsymbol{m}(v) \right) \right) \\ &= \sum_{w \in \boldsymbol{\mathcal{U}}^\wedge \colon x = w_{-1}} \boldsymbol{\nu}^+(w_0)\boldsymbol{m}(w). \end{aligned}$$ ◀

▶ **Proposition 13b.** *Let $(\boldsymbol{M}, \boldsymbol{\mu})$ be an EPRI Markov chain whose terminal component has size at least two. The Markov chain $(\mathring{\boldsymbol{M}}, \mathring{\boldsymbol{\mu}})$ is EPRI.*

Proof. First, let γ_1 be the constant of Corollary 10b. Theorem 2 proves that

$$\gamma_1 = \sum_{x \in \mathcal{A}} \boldsymbol{\nu}^+(x)\boldsymbol{M}^\uparrow(x).$$

Then, consider the distribution $\mathring{\boldsymbol{\nu}}$ defined by

$$\mathring{\boldsymbol{\nu}}(w) = \frac{1}{\gamma_1}\boldsymbol{\nu}^+(w_0)\boldsymbol{m}(w)\boldsymbol{M}^\uparrow(w_{-1}).$$

Lemma 12b proves that

$$\sum_{w \in \boldsymbol{\mathcal{U}}^\wedge} \mathring{\boldsymbol{\nu}}(w) = \sum_{y \in \mathcal{A}} \left(\sum_{w \in \boldsymbol{\mathcal{U}}^\wedge \colon y = w_{-1}} \mathring{\boldsymbol{\nu}}(w) \right) = \frac{1}{\gamma_1} \sum_{y \in \mathcal{A}} \boldsymbol{\nu}^+(y)\boldsymbol{M}^\uparrow(y) = 1,$$

i.e., that $\mathring{\boldsymbol{\nu}}$ is a probability distribution.

Moreover, for every word $w \in \boldsymbol{\mathcal{U}}^\wedge$, Lemma 12b proves that

$$\mathring{\boldsymbol{M}}\mathring{\boldsymbol{\nu}}(w) = \frac{1}{\gamma_1} \sum_{w' \in \mathcal{U}^\wedge \colon w'_{-1} = w_0} \boldsymbol{\nu}^+(w'_0)\boldsymbol{m}(w' \cdot w)\boldsymbol{M}^\uparrow(w_{-1}) = \frac{1}{\gamma_1}\boldsymbol{\nu}^+(w_0)\boldsymbol{m}(w)\boldsymbol{M}^\uparrow(w_{-1}) = \mathring{\boldsymbol{\nu}}(w).$$

This means that $\mathring{\boldsymbol{\nu}}$ is a stationary probability distribution of $(\mathring{\boldsymbol{M}}, \mathring{\boldsymbol{\mu}})$.

This probability distribution is positive on the set

$$\mathring{\boldsymbol{\mathcal{X}}} \stackrel{\text{def}}{=\!=} \{w \in \boldsymbol{\mathcal{U}}^\wedge \cap \boldsymbol{\mathcal{X}}^* \colon \exists x \in \boldsymbol{\mathcal{X}}, x > w_0 \text{ and } \boldsymbol{m}(x \cdot w) \neq 0\}$$

and zero outside of that set. Since $\mathring{\boldsymbol{\nu}}$ is a probability distribution, it follows that $\mathring{\boldsymbol{\mathcal{X}}} \neq \emptyset$.

Then, let $\boldsymbol{G}$ and $\mathring{\boldsymbol{G}}$ be the respective underlying graphs of $(\boldsymbol{M}, \boldsymbol{\mu})$ and $(\mathring{\boldsymbol{M}}, \mathring{\boldsymbol{\mu}})$. We shall prove that $\mathring{\boldsymbol{\mathcal{X}}}$ satisfies the requirements (i) and (iii) of EPRI Markov chains.

Hence, consider two words w and w' in $\mathring{\boldsymbol{\mathcal{X}}}$, and let us choose letters $x, y, z, t \in \mathcal{X}$ such that $x \in (w'_{-1})^{\uparrow}$, $w'_0 \in y^{\downarrow}$, $z \in w^{\uparrow}_{-1}$ and $w_0 \in t^{\downarrow}$. The graph $\boldsymbol{G}$ contains a finite path that starts with the letter t, then the letters of w (listed from left to right) and then the letter z, and finishes with the letter y, the letters of w' (listed from left to right), and then the letter x. This path forms a word u whose leftmost unimodal factor is w and whose second rightmost unimodal factor is w'. This proves that $\mathring{\boldsymbol{G}}$ contains a path from w to w', i.e., that $\mathring{\boldsymbol{\mathcal{X}}}$ satisfies the requirement (i).

Finally, consider some trajectory $(\mathring{\boldsymbol{Y}}_n)_{n\geqslant 0}$ of the Markov chain $(\mathring{\boldsymbol{M}}, \mathring{\boldsymbol{\mu}})$. Up to removing the first letter of every word (i.e., vertex) $w \in \boldsymbol{\mathcal{U}}^{\wedge}$ encountered on this trajectory, and then concatenating the resulting words, we obtain a trajectory $(\boldsymbol{Y}_n)_{n\geqslant 0}$ of $\boldsymbol{M}$ (for an initial distribution that may differ from $\boldsymbol{\mu}$). That trajectory almost surely contains a vertex $x \in \boldsymbol{\mathcal{X}}$, and will then keep visiting vertices in $\boldsymbol{\mathcal{X}}$. Thus, our initial trajectory almost surely contains a word $\mathring{\boldsymbol{Y}}_n$ that is a word with a letter $x \in \boldsymbol{\mathcal{X}}$, and all states $\mathring{\boldsymbol{Y}}_m$ such that $m \geqslant n+1$ will then belong to the set $\boldsymbol{\mathcal{U}}^{\wedge} \cap \boldsymbol{\mathcal{X}}^* = \mathring{\boldsymbol{\mathcal{X}}}$, thereby showing that $\mathring{\boldsymbol{\mathcal{X}}}$ satisfies the requirement (iii). ◀

A.3 Proving Theorem 20

▶ **Theorem 20.** *Let $w \in \mathcal{A}^n$ be a word whose letters are generated by a Markov chain (M, μ). For all integers $\ell \geqslant 0$, and provided that n is large enough, the IS-algorithm has a probability $\mathbf{P} \leqslant n^{-2^\ell}$ of performing more than $2\log_2(\log_2(n)) + \ell$ recursive function calls.*

Proof. Given a finite word v with v-locally minimal integers $i_0 < i_1 < \ldots < i_{k-1}$, we abusively set $i_{k+1} = |v|$ and $v_{|v|} = \$$, so that $\mathsf{eis}(v)_\ell = v_{i_\ell \cdots i_{\ell+1}}$ for all $\ell \leqslant k-1$. Then, let the *source* of a word $v' = \mathsf{eis}(v)_{a\cdots b}$ be the word $v_{i_a \cdots i_{b+1}-1}$, which we also denote by $\mathsf{src}(v')$, and which is a factor of $v_{1\cdots}$. If two factors of $\mathsf{eis}(v)$ coincide with each other, so do their sources, and if they do not overlap with each other, neither do their sources. Moreover, the word $\mathsf{src}(v')$ is at least twice longer than v'.

More generally, the ℓ^{th} *source* of a factor v' of $\mathsf{eis}^\ell(v)$, which we denote by $\mathsf{src}^\ell(v')$, is just v' itself if $\ell = 0$, or the $(\ell-1)^{\text{th}}$ source of $\mathsf{src}(v')$ if $\ell \geqslant 1$. Thus, if two letters of $\mathsf{eis}^\ell(v)$ coincide with each other, so do their ℓ^{th} sources, which are non-overlapping factors of $v_{2^\ell-1\cdots}$ of length at least 2^ℓ. Moreover, since the last letter of $\mathsf{eis}^\ell(v)$ is the only one that ends with the character \$, it cannot coincide with any other letter of $\mathsf{eis}^\ell(v)$. Therefore, the ℓ^{th} sources of our two equal letters are in fact factors of the word $v_{2^\ell-1\cdots|v|-2^\ell}$.

In addition, we say that the word v is *k-periodic except at borders of length b* if $v_j = v_{j+k}$ whenever $b \leqslant j < j+k \leqslant |v| - b$. If the factor $v_{b\cdots|v|-b}$ has exactly one letter, none of the integers $b+1, \ldots, |v|-b$ is locally v-minimal, and thus $|\mathsf{eis}(v)| \leqslant b$, thereby proving that the word $\mathsf{eis}^\ell(v)$ cannot exist whenever $\ell \geqslant \log_2(b)+1$. This case occurs in particular when $k = 1$.

Similarly, if $|v| \leqslant 2b+3k$, the word $\mathsf{eis}^\ell(v)$ cannot exist whenever $\ell \geqslant \log_2(\max\{b,k\})+3$.

If, on the contrary, the factor $v_{b\cdots|v|-b}$ has at least two letters and is of length at least $3k$, there exists a factor $\mathbf{f}$ of $\mathsf{eis}(v)$ whose source is a word of the form $v_{j\ldots j+k-1}$ for some j such that $b \leqslant j < j+k \leqslant |v|-b$. Let us then write v as a concatenation of the form $u \cdot \mathsf{src}(\mathbf{f})^t \cdot u'$ where u and u' have length at most $b+k$, and t is a positive integer. We can also write $\mathsf{eis}(v)$ as a word of the form $\mathbf{a} \cdot \mathbf{f}^t \cdot \mathbf{a}'$ such that $\mathsf{src}(\mathbf{a})$ is a suffix of u and $\mathsf{src}(\mathbf{b}) = u'$. By construction, we have

$$|\mathbf{a}| \leqslant |u|/2 \leqslant (b+k)/2, \quad |\mathbf{f}| \leqslant |\mathsf{src}(\mathbf{f})|/2 = k/2 \quad \text{and} \quad |\mathbf{a}'| \leqslant |u'|/2 \leqslant (b+k)/2,$$

which means that $\mathsf{eis}(v)$ is k'-periodic except at borders of length b' for some integers $k' \leqslant k/2$ and $b' \leqslant (b+k)/2 \leqslant \max\{b,k\}$. Thus, an immediate induction on k proves that the word $\mathsf{eis}^\ell(v)$ cannot exist whenever $\ell \geqslant \log_2(\max\{b,k\}) + \log_2(k) + 3$.

Now, let $G = (\mathcal{S}, E)$ be the underlying graph of the Markov chain (M, μ), and let $s = |\mathcal{S}|$ be the number of states of the Markov chain. Let $\mathcal{X}$ (respectively, $\mathcal{Y}$) be the set of states $x \in E$ that belong to a cyclic (respectively, non-cyclic) terminal connected component of G. Finally, let ε be the smallest non-zero edge weight in G, i.e., $\varepsilon = \min\{M(x,y) \colon M(x,y) > 0\}$, and let $\eta = -\log_2(1-\varepsilon^s)/s > 0$.

From each state $x \in E$, there is a path starting at x and ending in $\mathcal{X} \cup \mathcal{Y}$. Furthermore, the shortest such path is of length at most s. It follows, for all $k \geqslant 0$, that

$$\mathbb{P}[X_{k+s} \in \mathcal{X} \cup \mathcal{Y} \mid X_k = x] \geqslant \varepsilon^s$$

and, more generally, that

$$\mathbb{P}[X_m \notin \mathcal{X} \cup \mathcal{Y}] \leqslant (1-\varepsilon^s)^{m/s-1} = 2^{-(m-s)\eta}$$

for all $m \geqslant 0$.

Similarly, assume that $\mathcal{Y} \neq \emptyset$. Consider some state $x \in \mathcal{Y}$, and let $y \in \mathcal{Y}$ be a state accessible from x and with at least two outgoing edges (y,z) and (y,z'). Then, let p be a path from x to y. The shortest such path has length at most $s-1$. Therefore, provided that $X_k = x$ for some integer $k \geqslant 0$, the trajectory $(X_i)_{i \geqslant k}$ has a probability at least ε^s of starting with the path p and then going to z, and a probability at least ε^s of starting with the path p and then going to z'. In particular, for each finite sequence $\mathbf{q}$ consisting of $s+1$ states in $\mathcal{Y}$, we have

$$\mathbb{P}[(X_i)_{k \leqslant i \leqslant k+s} = \mathbf{q} \mid X_k] \leqslant 1 - \varepsilon^s$$

and, more generally, if $\mathbf{q}$ is a sequence consisting of $m+1$ states in $\mathcal{Y}$, we have

$$\mathbb{P}[(X_i)_{k \leqslant i \leqslant k+m} = \mathbf{q} \mid X_k] \leqslant (1-\varepsilon^s)^{m/s-1} = 2^{-(m-s)\eta}.$$

Finally, assume that w is a word of length $n \geqslant 2^{16s^2(s+1)+64s^2/\eta}$, and set $u = \log_2(n)/(4s)$, $t = 2\lfloor \log_2(u) \rfloor + \ell$ and $m = 2^t - 1$. Since $m \geqslant 2^{\ell-2}u^2 - 1$ and $2^\ell u \geqslant 1$, we have

$$2^{-(m-s)\eta} \leqslant 2^{-(2^{\ell-2}u^2-s-1)\eta} \leqslant 2^{-(2^\ell us(s+1)+2^{\ell+2}us/\eta-(s+1))\eta} \leqslant 2^{-2^{\ell+2}us} = n^{-2^{\ell+2}}.$$

In conclusion, let us consider several (non mutually exclusive) events:

- the event $\mathcal{E}_1$, which occurs if $X_m \notin \mathcal{X} \cup \mathcal{Y}$;
- the event $\mathcal{E}_2$, which occurs if $X_m \in \mathcal{X}$;
- for all integers u and v such that $m \leqslant u$, $u + m < v$ and $v + m < n - m$, the event $\mathcal{F}_{u,v}$, which occurs if $X_m \in \mathcal{Y}$ and $X_{u+i} = X_{v+i}$ whenever $0 \leqslant i \leqslant m$.

If $\mathcal{E}_2$ happens, the word w is k-periodic except at borders of length m, where $k \leqslant s$ is the length of the cycle of G to which X_m belongs. Thus, in that case, the IS-algorithm cannot make more than

$$\begin{aligned}\log_2(\max\{s,m\}) + \log_2(s) + 2 &= \log_2(m) + \log_2(4s) \\ &\leqslant 2\log_2(u) + \log_2(4s) + \ell \leqslant 2\log_2(\log_2(n)) + \ell\end{aligned}$$

recursive function calls.

Then, if the IS-algorithm makes more than $2\log_2(\log_2(n)) + \ell \geqslant t$ recursive function calls, two letters of the word $\mathsf{eis}^t(w)$ must coincide with each other. This means that two non-overlapping length-m factors of the word $w_{m \cdots |w|-m-1}$ must coincide with each other, and therefore that either $X_m \notin \mathcal{Y}$ or that one of the events $\mathcal{F}_{u,v}$ must have occurred. If $X_m \notin \mathcal{Y}$, and since $\mathcal{E}_2$ may not have occurred, this means that $\mathcal{E}_1$ occurred.

Moreover, the events $\mathcal{E}_1$ and $\mathcal{F}_{u,v}$ are rare: our above study proves that $\mathbb{P}[\mathcal{E}_1] \leqslant n^{-2^{\ell+2}}$; then, for all u and v, the sequence $(X_i)_{u \leqslant i \leqslant u+m}$ being fixed, the event $\mathcal{F}_{u,v}$ also occurs with probability $\mathbb{P}_{u,v} \leqslant n^{-2^{\ell+2}}$.

In conclusion, the IS-algorithm makes more than $2\log_2(n) + \ell$ recursive function calls with a probability $\mathbf{P} \leqslant \mathbb{P}[\mathcal{E}_1] + \sum_{u,v} \mathbb{P}[\mathcal{F}_{u,v}] \leqslant n^2 \times n^{-2^{\ell+2}} \leqslant n^{-2^\ell}$. ◀