Separating Many Words
by Counting Occurrences of Factors

Aleksi Saarela

Department of Mathematics and Statistics
University of Turku, Finland

DLT 2019, Warsaw
Outline

1. Motivations
2. Separating sets of factors
3. Infinite words
4. Regular languages
5. Conclusion
Motivation 1: Separating words problem

Notation: \( \Sigma \) is the alphabet, \( |w|_x \) is the number of occurrences of \( x \) in \( w \).

**Question**

If \( \text{sep}(u, v) \) is the size of the smallest DFA that accepts one of the words \( u, v \in \Sigma^* \) and rejects the other, then what is

\[
\max\{\text{sep}(u, v) \mid u, v \in \Sigma^{\leq n}\}?
\]

- **Lower bound**: \( \Omega(\log n) \) (Goralčík, Koubek 1986).
- **Upper bound**: \( O(n^{2/5}(\log n)^{3/5}) \) (Robson 1989).
- **Example of other results**: If \( |u|_x \neq |v|_x \) for some factor \( x \), then \( \text{sep}(u, v) = O(|x| \log n) \) (Demaine, Eisenstat, Shallit, Wilson 2011).
Counting factors

How well can words be separated if we forget about automata and only consider the simple idea of counting occurrences of factors?

- For all $u, v \in \Sigma^n$, $u \neq v$, there exists $x \in \Sigma^*$ such that $|u|_x \neq |v|_x$ and $|x| \leq \lfloor n/2 \rfloor + 1$ (Manuch 2000).
- If we want to separate more than two words (possibly infinitely many) at once, and we can do this by counting the numbers of occurrences of more than one factor.

Question

Given a language $L$, does there exist a finite language $X$ such that for all distinct words $u, v \in L$, there exists $x \in X$ such that $|u|_x \neq |v|_x$?
Motivation 2: Old guessing game

From a given set of options, Alice secretly picks one. Bob is allowed to ask any yes-no questions, and he is trying to figure out what Alice picked.

- Famous versions: “Twenty Questions”, “Guess Who”.
- Theoretically, the required number of questions is logarithmic with respect to the number of options.
- Many more complicated variations exist.
Guessing a word

What if Alice picks a word $w$ from a given language and Bob can ask for the number $|w|_x$ for different factors $x$?

Example

If Alice has chosen $w \in \{ac, ad, be, bf\}$, then Bob can ask for the numbers $|w|_a, |w|_c, |w|_e$, and this will always reveal $w$.

(Two questions are enough if Bob can choose the second question after hearing the answer to the first one.)

Question

Given a language from which Alice has secretly picked one word $w$, can Bob find a finite language $X$ such that the answers to the questions “What is $|w|_x$?” for all $x \in X$ are guaranteed to reveal the correct word $w$?
Motivation 3: \( k \)-abelian complexity

For a positive integer \( k \), words \( u \) and \( v \) are \( k \)-abelian equivalent, denoted \( u \equiv_k v \), if \( |u|_x = |v|_x \) for all factors \( x \) of length at most \( k \).

**Example**

\( aabab \equiv_2 abaab \).

\( aba \not\equiv_2 bab \), even though \( |aba|_x = |bab|_x \) for all \( x \in \{a, b\}^2 \).

Let \( w \in \Sigma^\omega \). The *factor complexity* of \( w \) is the function

\[
P_w : \mathbb{Z}_{\geq 0} \to \mathbb{Z}_{\geq 0}, \quad P_w(n) = |\text{Fact}_n(w)|,
\]

and the *\( k \)-abelian complexity* of \( w \) is the function

\[
P_w^k : \mathbb{Z}_{\geq 0} \to \mathbb{Z}_{\geq 0}, \quad P_w^k(n) = |\text{Fact}_n(w)/\equiv_k|,
\]

where \( \text{Fact}_n(w) \) is the set of factors of \( w \) of length \( n \).
Motivations

$k$-abelian complexity

There are many results about the $k$-abelian complexities of some specific words, about the possible growth rates of $\mathcal{P}^k_w$ or of $\mathcal{P}^{k+1}_w/\mathcal{P}^k_w$ etc.

Question

Given an infinite word $w$, does there exist a number $k$ such that $\mathcal{P}^k_w = \mathcal{P}_w$?

The *growth function* of $L \subseteq \Sigma^*$ is the function

$$\mathcal{P}_L : \mathbb{Z}_{\geq 0} \rightarrow \mathbb{Z}_{\geq 0}, \quad \mathcal{P}_L(n) = |L \cap \Sigma^n|.$$ 

The *$k$-abelian growth function* of $L \subseteq \Sigma^*$ is the function

$$\mathcal{P}^k_L : \mathbb{Z}_{\geq 0} \rightarrow \mathbb{Z}_{\geq 0}, \quad \mathcal{P}^k_L(n) = |(L \cap \Sigma^n)/\equiv_k|.$$ 

Question

Given a language $L$, does there exist a number $k$ such that $\mathcal{P}^k_L = \mathcal{P}_L$?
Outline

1 Motivations
2 Separating sets of factors
3 Infinite words
4 Regular languages
5 Conclusion
SSFs

A language $X$ is a \textit{separating set of factors} (SSF) of a language $L$ if for all distinct words $u, v \in L$, there exists $x \in X$ such that $|u|_x \neq |v|_x$.

\textbf{Example}

The language $a^*$ has two inclusion-minimal SSFs: $\{\varepsilon\}$ and $\{a\}$. The language $\{aa, ab, ba, bb\}$ has eight inclusion-minimal SSFs:

$$\{a, ab\}, \{a, ba\}, \{b, ab\}, \{b, ba\},$$
$$\{aa, ab, ba\}, \{aa, ab, bb\}, \{aa, ba, bb\}, \{ab, ba, bb\}.$$ 

We study properties of SSFs and answer the following question for sets of factors of infinite words and for regular languages.

\textbf{Question}

Which languages have a finite SSF?
SSFs and $k$-abelian equivalence

Lemma

Let $L \subseteq \Sigma^*$.

1. $\Sigma \leq^k$ is an SSF of $L$ $\iff$ $\mathcal{P}_L^k = \mathcal{P}_L$.

2. $L$ has a finite SSF $\iff$ $\exists k : \mathcal{P}_L^k = \mathcal{P}_L$.

The condition $\mathcal{P}_L^k = \mathcal{P}_L$ is equivalent to the words in $L$ being pairwise $k$-abelian nonequivalent.
Example

In a list of \( \sim 140000 \) English words, there are no 4-abelian equivalent words. The only pairs of 3-abelian equivalent words are \textit{reregister}, \textit{registerer} and \textit{reregisters}, \textit{registerers}, and the other pairs of 2-abelian equivalent words are

\begin{align*}
\text{indenter, intender} & \quad \text{indenters, intenders} \\
\text{pathophysiologic, physiopathologic} & \quad \text{pathophysiologial, physiopathological} \\
\text{pathophysiology, physiopathology} & \quad \text{pathophysiologies, physiopathologies} \\
\text{tamara, tarama} & \quad \text{tamaras, taramas} \\
\text{tantarara, tarantara} & \quad \text{tantararas, tarantaras} \\
\text{tantaras, tarantas} & 
\end{align*}

It follows that \( \Sigma^{\leq 2} \cup \left\{ rere, hop, ind, tan, tar \right\} \) is an SSF of the language (\( \Sigma \) contains the 26 letters from \( a \) to \( z \) and many other symbols).
SSFs and rational operations

**Lemma**

Let $K$ and $L$ be languages.

1. $L$ has a finite SSF and $|K| < \infty \implies L \cup K$ has a finite SSF.
2. $L$ does not have a finite SSF $\implies L \cup K$ does not have a finite SSF.
3. $L$ has a finite SSF and $|K| = 1 \implies KL$ and $LK$ have finite SSFs.
4. $L$ does not have a finite SSF and $K \neq \emptyset$ $\implies KL$ and $LK$ do not have finite SSFs.
5. $L^*$ has a finite SSF $\iff L \subseteq w^*$ for some word $w$.

**Example**

Let $L = \{ a^k ba^{k-1} \mid k \in \mathbb{Z}_+ \}$. Then both $L$ and $Laa$ have the finite SSF $\{ \varepsilon \}$. On the other hand, $L\{ \varepsilon, aa \} = L \cup Laa$ does not have a finite SSF.
Outline

1. Motivations
2. Separating sets of factors
3. Infinite words
4. Regular languages
5. Conclusion
Result

Theorem

Let \( w \in \Sigma^\omega \). There exists \( k \) such that \( \mathcal{P}_w^k = \mathcal{P}_w \) iff \( w \) is ultimately periodic.

Proof (sketch).

If \( w \) is ultimately periodic, we can write \( w = uv^\omega \) and let \( k = |uv| + 1 \). It can be proved quite easily that \( \mathcal{P}_w^k = \mathcal{P}_w \).

If \( w \) is aperiodic and \( k \geq 2 \) is arbitrary, there exists words \( x \in \Sigma^{k-1} \) and \( y \in \Sigma^* \) such that \( xyz \) occurs infinitely often as a factor of \( w \). Then we can write \( w = z_0xyxz_1xyxz_2xyx \cdots \) for some words \( z_0, z_1, z_2, \ldots \). By aperiodicity, \( xy \) and \( xz_i \) have a different primitive root for some \( i \geq 1 \). Then \( xyxz_i x \neq xz_i xyx \), but \( xyxz_i x \equiv_k xz_i xyx \), so \( \mathcal{P}_w^k \neq \mathcal{P}_w \).

Corollary

The set of factors of \( w \in \Sigma^\omega \) has a finite SSF iff \( w \) is ultimately periodic.
Outline

1. Motivations
2. Separating sets of factors
3. Infinite words
4. Regular languages
5. Conclusion
Lemma

If a language $L$ has a subset of the form $xw^*yw^*z$ for some words $w, x, y, z$ such that $wy \neq yw$, then $L$ does not have a finite SSF.

Proof.

For all $k$, the words $xw^kyw^{k-1}z$ and $xw^{k-1}yw^kz$ are distinct but $k$-abelian equivalent.
Bounded languages

A language \( L \subseteq \Sigma^* \) is \textit{bounded} if it is a subset of a language of the form

\[
\nu_1^* \cdots \nu_n^*.
\]

A regular language is bounded iff it is a finite union of languages of the form

\[
u_0 \nu_1^* u_1 \cdots \nu_n^* u_n
\]

(Ginsburg, Spanier 1966).

\textbf{Lemma}

\textit{Every unbounded regular language has a subset of the form} \( xw^*yw^*z \) \textit{for some words} \( w, x, y, z \) \textit{such that} \( wy \neq yw \), \textit{and therefore does not have a finite SSF.}
Result

Theorem

A regular language $L$ has a finite SSF iff $L$ does not have a subset of the form $xw^*yw^*z$ for any words $w, x, y, z$ such that $wy \neq yw$.

Example

The language $K = a^*(abab)^*ba(ba)^*$ does not have a finite SSF:

$$K \supset (abab)^*ba(ba)^* = (abab)^*b(ab)^*a \supset (abab)^*b(abab)^*a.$$ 

The language

$$L = a^*(abab)^*aba(ba)^* = a^*(abab)^*(ab)^*aba = a^*(ab)^*aba.$$ 

has a finite SSF: It can be proved that if $L$ has a subset $xw^*yw^*z$ with $w \neq \varepsilon$, then the primitive root of $w$ is $a$ or $ab$ or $ba$, and $wy = yw$. 
Proof (idea).

It is sufficient to consider the case where $L$ is infinite, bounded, and does not have a subset of the specified form. We can write

$$L = \bigcup_{i=1}^{s} u_{i0} \prod_{j=1}^{r_i} v_{ij}^* u_{ij}$$

$$n = 2 \cdot \max \left\{ \left| u_{i0} \prod_{j=1}^{r_i} v_{ij} u_{ij} \right| \bigg| i \in \{1, \ldots, s\} \right\},$$

$$k = \max \left\{ \left| u_{i0} \prod_{j=1}^{r_i} v_{ij}^{n+2} u_{ij} \right| \bigg| i \in \{1, \ldots, s\} \right\}.$$ 

It can be proved that if some words in $L$ are $k$-abelian equivalent, then they are equal. The proof is quite long and technical.
Outline

1. Motivations
2. Separating sets of factors
3. Infinite words
4. Regular languages
5. Conclusion
Conclusion

- We have considered the question of whether a given language has a finite SSF. We have answered this question for sets of factors of infinite words and for regular languages. This question could be studied for other families of languages.

- Given a language with a finite SSF, what is the minimal size of an SSF of this language? For example, this could be considered for $\Sigma^n$.

- Given a language with no finite SSF, how “small” can the growth function of an SSF of this language be? For example, this could be considered for $\Sigma^*$. 
Conclusion

- We have considered the question of whether a given language has a finite SSF. We have answered this question for sets of factors of infinite words and for regular languages. This question could be studied for other families of languages.

- Given a language with a finite SSF, what is the minimal size of an SSF of this language? For example, this could be considered for $\Sigma^n$.

- Given a language with no finite SSF, how “small” can the growth function of an SSF of this language be? For example, this could be considered for $\Sigma^*$. 

Thank You!