Factorization Forests of Finite Height

Jérémie Chalopin

September 16, 2002

Training Period done with Hing LEUNG

Department of Computer Science,
New Mexico State University
Las Cruces, NM 8803, U.S.A.

Introduction

The factorization forests have been introduced by Imre Simon in [Si90] to describe factorizations of words over a given alphabet. In the same paper, Simon has prooved that every morphism from a free semigroup to a finite one S admits a Ramseyan factorization forest of height at most $9|S|$. Later, Simon has given in [Si92] a simpler proof of the same problem but he obtained an exponential bound instead of a linear one.

In our paper, the main result is the theorem 5 in which we found the same result as Simon with a bound of $7|S|$. To proove this result, we present an algorithm to build factorization forests which is close from Simon's one, but our proof is done in a direct way.

We first give a presentation of the problem and a few basic results on finite semigroups that will be useful later. Then we show the result in three steps: we first work on two particular cases and then we show the result for the general case. In the last part, we want to show that the algorithm cannot be improved significantly. We describe different kinds of examples for which we will need a factorization tree of linear height.

The main result can be used to proove Brown's lemma on locally finite semigroups in a constructive way. Simon has also used this result to proove the limitedness problem on distance automata in [Si94] and to find a double-exponential bound to this problem.

Contents

1 Presentation of the problem

Given a set X, we will note X^+ or $\mathcal{F}(X)$ the free semigroup generated by X. A factorization forest over an alphabet A can be defined by a function d from A^+ into $\mathcal{F}(A^+)$ such that for every $x \in A^+$, $d(x) = (x_1, x_2, \dots x_p)$ implies that $x = x_1 x_2 \dots x_p$: $d(x)$ is a factorization of x.

We call this a factorization forest, because for each word $x \in A^+$, we can associate a rooted tree $T(x)$ such that the nodes are labelled by words in A^+. If $|d(x)| = 1$, $T(x)$ consists just of a root labelled x; if $d(x) = (x_1, x_2, ... x_p)$ with $p \geq 2$, then the root of $T(x)$ is labelled by x and has p sons who are the $T(x_i)$ for $1 \leq i \leq p$.

We define the degree of x to be the number of sons of the root of $T(x)$: if $|d(x)| = 1$, the degree of x is 0 and it's $|d(x)|$ otherwise. We will call the height of x the height of the tree $T(x)$: if $|d(x)| = 0$, $h(x) = 0$ and if $d(x) = (x_1, x_2, \dots x_p)$, $h(x) = 1 + max\{h(x_i); 1 \leq i \leq p\})$. The height of a factorization forest F is $h(F) = sup\{h(x); x \in A^+\}$.

Let f be a morphism from a free semigroup A^+ to a finite one S. A factorization forest F is ramseyan mod f if for every x of degree $p \geq 3, d(x) = (x_1, x_2, \dots, x_p)$ implies that there exists an idempotent e such that $e = f(x) = f(x_1) = f(x_2) = \cdots = f(x_p)$. We say that f admits a ramseyan factorization forest if it admits a factorization forest F over A, which is ramseyan mod f and such that the only words such that $d(x) = x$ are the elements of A.

2 A few things about semigroups

In this part, we will present a few things about the theory of semigroups that will be useful later. All the topics discussed here can be found in the chapter 2 (Green's relations) of the book of Gérard Lallement [La79].

Given S a finite semigroup, we define S^1 as follows: if S has an identity, then $S^1 = S$; otherwise we add a new element 1 to S to obtain S^1 such that this new element is an identity for S^1. We define four relations $\mathcal{R}$, $\mathcal{L}$, $\mathcal{D}$[1] and $\mathcal{H}$ on S as follows:

$$\begin{aligned} a\mathcal{R}b &\iff aS^1 = bS^1 \\ a\mathcal{L}b &\iff S^1a = S^1b \\ a\mathcal{D}b &\iff S^1aS^1 = S^1bS^1 \\ a\mathcal{H}b &\iff a\mathcal{R}b \wedge a\mathcal{L}b \end{aligned}$$

From the definition, it is clear that the former relations are equivalences; the class of an element a for the relation $\mathcal{R}$ (resp. $\mathcal{L}$, $\mathcal{D}$, $\mathcal{H}$) is denoted by R_a (resp. L_a, D_a, H_a). We can also note that given $a, b \in S$, we have $a\mathcal{R}b$ if and only if there exist $u, v \in S^1$ such that $au = b$ and $a = bv$; we have a similar result for the relations $\mathcal{L}$ and $\mathcal{D}$. We say that a $\mathcal{D}$-class D is regular if there is an idempotent belonging to D.

In a semigroup S, there is a natural partial ordering of the classes of the relation $\mathcal{D}$: given $a, b \in S$, $D_a \leq_{\mathcal{D}} D_b$ if and only if $S^1aS^1 \subseteq S^1bS^1$. We can consequently represent the set of $\mathcal{D}$-clases by a directed graph as shown in figure

[1] In the literature, the definition given here for $\mathcal{D}$ is called $\mathcal{J}$, but it is known that in a finite semigroup $\mathcal{J} = \mathcal{D}$

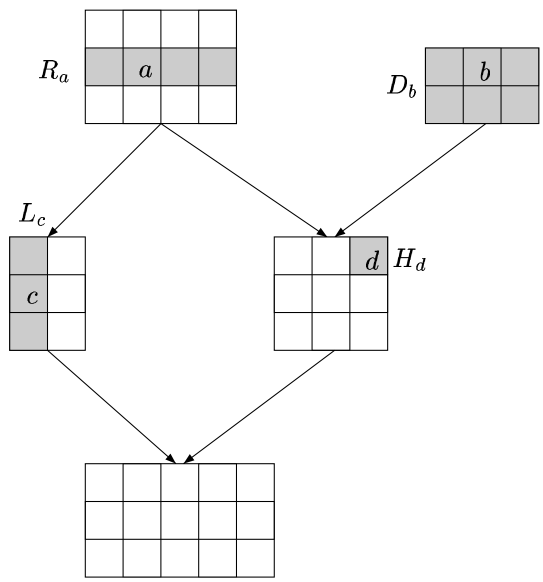

Figure 1: An example of decomposition diagram for a semigroup

1. Furthermore, it is obvious to see that two elements in a same $\mathcal{R}$-class (resp. $\mathcal{L}$-class) are in the same $\mathcal{D}$-class. We can also see that the equivalence classes for the relation $\mathcal{H}$ are included in classes for the $\mathcal{R}$ and $\mathcal{L}$ relations. Consequently, in each $\mathcal{D}$-class, we can split the elements between the differents $\mathcal{H}$-classes and we obtain an "egg-box" whose cells are the different $\mathcal{H}$-classes: the elements whose $\mathcal{H}$-classes are in the same row (resp. column) are in the same $\mathcal{R}$-class (resp. $\mathcal{L}$-class).

Lemma 1 *In a finite semigroup S, given $a, b \in S$, $a\mathcal{D}ab$ (resp. $b\mathcal{D}ab$) if and only if $a\mathcal{R}ab$ (resp. $b\mathcal{L}ab$).*

Proof:
It is obvious to see that $a\mathcal{D}ab$ if $a\mathcal{R}ab$. Suppose now that $a\mathcal{D}ab$: there exist $u, v \in S^1$ such that $a = uabv$. Consequently, for all $m \geq 1, a = u^m a(bv)^m$ and since S is finite, there exists $m_0 \in \mathbb{N}$ such that $(bv)^{m_0}$ is an idempotent.

$$
\begin{aligned}
a &= u^{m_0} a(bv)^{m_0} \\
&= u^{m_0} a(bv)^{2m_0} \\
&= u^{m_0} a(bv)^{m_0}(bv)^{m_0} \\
&= ab(vb)^{m_0-1}v
\end{aligned}
$$

And consequently $aS^1 \subseteq abS^1$ and since it is obvious to see that $abS^1 \subseteq aS^1$, we have prooved that $a\mathcal{R}ab$. $\square$

Lemma 2 *(Green's Lemma) Given a, b in a semigroup S such that $a\mathcal{R}b$, let $u, v \in S^1$ such that $au = b$ and $bv = a$ and ρ_u, ρ_v be the inner right translations* *defined by u and v. Then ρ_u is a bijection from L_a to L_b and ρ_v is the inverse bijection. These two bijections preserve the $\mathcal{H}$-classes: for all $x, y \in L_a$ (resp. L_b), $x\mathcal{H}y$ if and only if $xu\mathcal{H}yu$ (resp. $xv\mathcal{H}yv$).*

Proof:
Given $x \in L_a$, since $x\mathcal{L}a$, we have that $xu\mathcal{L}au = b$; consequently, ρ_u is a function from L_a to L_b. Furthermore, there exists $t \in S^1$ such that $x = ta$ and consequently, $\rho_v(\rho_u(x)) = xuv = tauv = tbv = ta = x$. A similar proof shows that ρ_v is a function from L_b to L_a and that $\rho_v\rho_u$ is the identity on L_b. Consequently, we can deduce that ρ_u and ρ_v are inverse bijections from L_a to L_b and from L_b to L_a respectively, establishing the first statement of the lemma.

For every $x \in L_a$, we have $x\mathcal{R}xu$, and consequently $x\mathcal{H}y$ implies $xu\mathcal{H}yu$; similarly $xu\mathcal{H}yu$ implies $x = xuv\mathcal{H}yuv = y$. With the same idea, we can proove the same results for two elements of L_b with v and therefore it establishes the second statement of the lemma. □

Remarks:

1. We have the same result between the $\mathcal{R}$-classes of two elements in the same $\mathcal{L}$-class with inner left translations λ_u and λ_v.

2. It is now easy to see that two $\mathcal{L}$-classes (resp. $\mathcal{R}$-classes and $\mathcal{H}$-classes) lying in the same $\mathcal{D}$-class have the same cardinality; therefore the cells in the egg-box have the same size.

Theorem 1 *(D. D. Miller and A. H. Clifford) For any two elements a, b in a semigroup S, $ab \in R_a \cap L_b$ if and only if $R_b \cap L_a$ contains an idempotent.*

Proof:
Suppose $ab \in R_a \cap L_b$, we have that $a\mathcal{R}ab$ and by lemma 2, ρ_b defines a bijection from L_a onto $L_{ab} = L_b$. There exists $c \in L_a$ such that $\rho_b(c) = cb = b$; moreover we know that $c\mathcal{R}cb = b$ and consequently $c \in R_b \cap L_a$ and there exists $u \in S^1$ such that $c = bu$. Since $c^2 = cbu = bu = c$, we have found an idempotent in $R_b \cap L_a$.

Conversely, if there exist $e \in R_b \cap L_a$ such that $e^2 = e$ then there exist $s, t \in S^1$ such that $b = et$ and $a = se$ and consequently $eb = eet = et = b$ and $ae = see = se = a$. Since $e\mathcal{R}b$ (resp. $e\mathcal{L}a$), we deduce $a = ae\mathcal{R}ab$ (resp. $b = eb\mathcal{L}ab$): we have shown that $ab \in R_a \cap L_b$. □

Remark: Given an $\mathcal{H}$-class H such that there exists an idempotent $e \in H$, for all $a, b \in H$, $ab \in H$. Moreover, given $a \in H$, there exist $s, t \in S^1$ such that $a = se = et$ and consequently, $ae = see = se = a$ and $ea = eet = et = a$. Therefore, we can easily see that there is at most one idempotent in each $\mathcal{H}$-class.

Theorem 2 *In a finite semigroup S, given D a $\mathcal{D}$-class of S, H a $\mathcal{H}$-class of D and $a, b \in H$, the following statements are equivalents.*

(i) $ab \in H$

(ii) *H is a group*

(iii) $ab \in D$

Proof:

- $(i) \Rightarrow (ii)$ Since $ab \in H$, from the theorem 1, we can find an idempotent $e \in H$ and we know that there is only one idempotent. We also know that for each $x \in H, xe = ex = x$ and moreover, we know that for all $c, d \in H$, $cd \in H$: H is a monoid.

  For each $x \in H$, $x^n \in H$ and since H is finite there exists m such that a^m is an idempotent. But there is only one idempotent in H and consequently $a^m = a^{m-1}a = aa^{m-1} = e$: each element in H has an inverse; H is a group.

- $(ii) \Rightarrow (iii)$ Since H is a group, $ab \in H$ and therefore $ab \in D$.

- $(iii) \Rightarrow (i)$ $ab \in D$ and S is finite; consequently, from lemma 1 $ab \in R_a \cap L_b = H$.

□

3 The group case

We consider in this section a morphism f from a free monoid A^+ to a finite group G. Let e be the identity of G; e is the only idempotent element of G, and consequently the only nodes with an outdegree greater than 2 will have a label x such that $f(x) = e$.

Theorem 3 *Every morphism $f : A^+ \to G$, where G is a finite group, admits a Ramseyan factorization forest of height at most* $3|G|$.

Proof:
Given a word x, let $Pref(x) = \{f(u); u \text{ is a proper prefix of } x\}$. For all $x, v \in A^+$ and $u, w \in A^*$ such that $x = uvw$, we have $f(u)Pref(v) \subseteq Pref(x)$. We will show by induction on $|Pref(x)|$ that we can find a tree for x of height at most $3|Pref(x)|$.
If $|Pref(x)| = 0$, we have $|x| = 1$, and we can put $d(x) = x$; we have a tree of height 0. Suppose now that $|Pref(x)| \geq 1$, and let $b \in Pref(x)$. We can write $x = a_1a_2 \dots a_p$ and let $1 \leq i_1 < i_2 < \dots < i_k \leq p$ be all the $1 \leq i \leq p$, such that $f(a_1 \dots a_i) = b$. Let $u = a_1 \dots a_{i_1}$, let $v = a_{i_k+1} \dots a_p$ and let $y_j = a_{i_j+1} \dots a_{i_{j+1}}$ for each $1 \leq j \leq k-1$; we will note $y = y_1 \dots y_{k-1}$. We know that $f(uy_1 \dots y_{i-1}) = f(uy_1 \dots y_i) = f(uy_1 \dots y_{i-1})f(y_i)$; and consequently for each $y_i, f(y_i) = e$; consequently we also know that $f(y) = e$. As described in the figure 2, we will put $d(x) = (uy, v)$, $d(uy) = (u, y)$ and we will put $d(y) = (y_1, \dots, y_{k-1})$. The only node that can have an outdegree greater than 2 is the node labelled by y but we know that $f(y) = f(y_1) = \dots = f(y_{k-1}) = e$: the construction is ramseyan.

We will now show that for all the leaves of the tree we have just built, the size of $Pref$ has decreased. We know that $Pref(u) \subseteq Pref(x)$ and that $b \in Pref(x) \setminus Pref(u)$ (i_1 is the smallest i such that $f(a_1 \dots a_i) = b$): consequently $|Pref(u)| < |Pref(x)|$. We also know that for each y_j, $f(uy_1 \dots y_{j-1})Pref(y_i) = bPref(y_j) \subseteq Pref(x)$, but we also know that $b \in Pref(x) \setminus bPref(y_j)$ (there

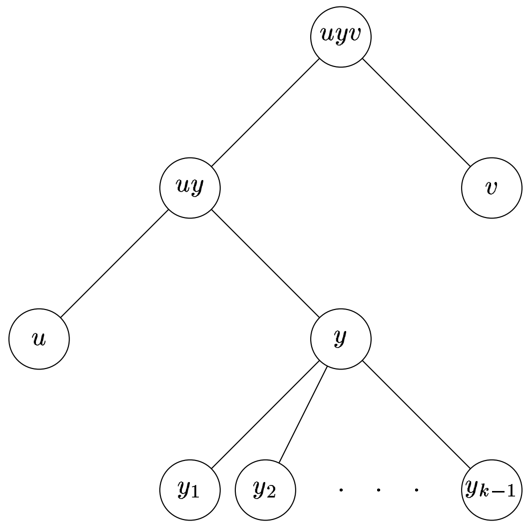

Figure 2: How to build a tree for the group case

isn't any i between i_j and i_{j+1} such that $f(a_1 \dots a_i) = b$): since we are in a group, we can conclude that $|Pref(y_j)| < |Pref(x)|$. With the same argument, we can see that $|Pref(v)| < |Pref(x)|$.
By the induction hypothesis, we know that we can find a tree for u, v and each y_i of height at most $3(|Pref(x)| - 1)$ and consequently we can build a tree of height lower than $3|Pref(x)|$.

Consequently, since the size of $Pref(x)$ is bounded by the size of G, we have found a way to build a factorization forest of height at most $3|G|$. □

4 The general case

In this section we will describe the steps of an algorithm to build a factorization forest in the general case. In a first time, we will suppose that through the morphism, for each letter a of the word x, $f(a)$ and $f(x)$ are in the same $\mathcal{D}$-class, and we will work on the $\mathcal{H}$-classes and use the precedent result that we have found for the group case. Then, for the general case, we will work on the $\mathcal{D}$-classes and use the result found for the particular case.

4.1 In a same $\mathcal{D}$-class

We will work on words that have the following property P: for each letter a of the word x, $f(a)$ and $f(x)$ are in the same D-class. If a word u is a factor of a word x that satisfies the property P, $f(u)$ is in the same $\mathcal{D}$-class than x and u satisfies the property P. This kind of word can only appear if the $\mathcal{D}$-class is

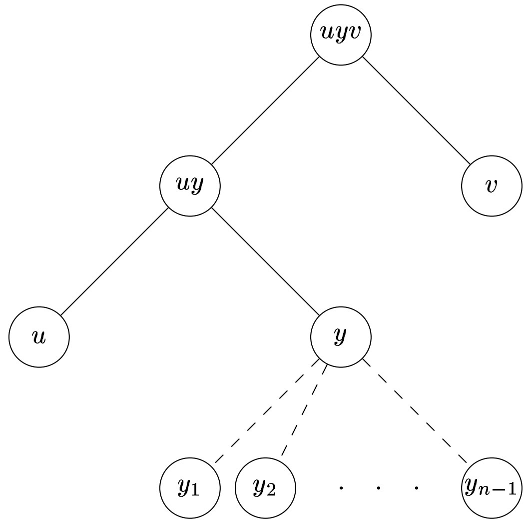

Figure 3: How to build a tree for a word which satisfies P

regular (given two elements a, b in a non regular $\mathcal{D}$-class, the product ab doesn't belong to the $\mathcal{D}$-class).

Theorem 4 *Given a word x with the property P, we can find a ramseyan factorization tree for x of height at most $5|D_{f(x)}|$.*

Proof:
For each word $x = a_1 a_2 \dots a_p$, we will note $Int_H(x) = \{(L_{f(a_i)}, R_{f(a_{i+1})}; 1 \leq i \leq p-1\}$. We can easily see that for each factor v of x, we have $Int_H(v) \subseteq Int_H(x)$.

Let q be the size of each $\mathcal{H}$-class (it is known that all the $\mathcal{H}$-class in a same $\mathcal{D}$-class have the same size). We will show by induction on $|Int_H(x)|$ that we can find for each word x a factorization tree of height at most $5qInt_H(x)$.

If $Int_H(x) = \emptyset$, we know that the length of x is one and therefore we can put $d(x) = x$ to get a tree of height 0.

Assume now that $Int_H(x) \geq 1$ and let (l, r) be an element of $Int_H(x)$. We can write $x = a_1 \dots a_p$ and let $1 \leq i_1 < i_2 < \dots < i_n \leq p-1$ be all the $1 \leq i \leq p-1$ such that $(L_{f(a_i)}, R_{f(a_{i+1})}) = (l, r)$. Let $u = a_1 \dots a_{i_1}$, $v = a_{i_n+1} \dots a_p$ and $y_j = a_{i_j+1} \dots a_{i_{j+1}}$ for each $1 \leq j \leq n-1$; we will note $y = y_1 \dots y_{n-1}$ (if $n = 1$, $y = \epsilon$). From lemma 1, we know that each y_j is in $R_{f(a_{i_j+1})} \cap L_{f(a_{i_{j+1}})} = r \cap l$ and y belongs also to this $\mathcal{H}$-class. We will build a tree as described in the figure 3: we put $d(x) = (uy, v), d(uy) = (u, y)$. Since y and all of the y_j are in the same $\mathcal{H}$-class, we know from theorem 2 that this $\mathcal{H}$-class is a group and therefore we can find a factorization tree of root y and leaves $y_1, \dots, y_{n-1}$ with a height at most $3|H_{f(y)}| = 3q$.

Since u, v, and all the y_i are factors of x, we already know that the sets $Int_H(u)$, $Int_H(v)$ and all the $Int_H(y_i)$ are included in $Int_H(x)$. Moreover, by the definition of u, v and the y_i, we know that (l,r) belongs to $Int_H(x)$ but not to $Int_H(u)$, $Int_H(v)$ or any of the $Int_H(y_i)$. We also know that u, v, and all the y_i satisfy the property P and consequently, by the induction hypothesis, we can find a factorization tree for u, v and each y_i of height at most $5q(Int_H(x)-1)$, and consequently, there exist a factorization tree for x of height at most $2+3q+5q(Int_H(x)-1) \leq 5q(Int_H(x))$.

Therefore, for each word x which satisfies the property P, we can find a factorization tree of height at most $5q|Int_H(x)|$. But $|Int_H(x)|$ is lower than the number of different $\mathcal{H}$-classes that we can find in the $\mathcal{D}$-class of $f(x)$ and q is the size of each one of these $\mathcal{H}$-classes: the height of the tree is lower than 5 times the size of the $\mathcal{D}$-class of $f(x)$. $\square$

4.2 For a general semigroup

Theorem 5 *Every morphism $f : A^+ \to S$, from a free semigroup to a finite one, admits a Ramseyan factorization forest of height at most $7|S|$.*

Proof:
Given a word x, we will work on the position of $D_{f(x)}$ in the partial ordering of the $\mathcal{D}$-classes. We will show by induction on this partial ordering that we can find a tree of height at most $7 \sum_{D \geq_{\mathcal{D}} D_{f(x)}} |D|$.

Suppose that $D_{f(x)}$ is one of the maximal element for the partial ordering $\leq_{\mathcal{D}}$. For each letter a of x, we know that $D_{f(x)} \leq_{\mathcal{D}} D_{f(a)}$, but since $D_{f(x)}$ is a maximal element, we have that $D_{f(x)} = D_{f(a)}$, and therefore, we can apply what we have proved before: there exist a factorization tree for x of height at most $5|D_{f(x)}|$, which is lower than $7 \sum_{D \geq_{\mathcal{D}} D_{f(x)}} |D|$.

For the general case, if $|x| = 1$, we put $d(x) = x$ and we have a tree of height 0 and we have nothing to prove.

Suppose now that $|x| \geq 1$, we will consider two cases: either x has a proper prefix u such that $D_{f(u)} = D_{f(x)}$, or it hasn't. If x hasn't any prefix of this kind, let $y \in A^+$ and $a \in A$ such that $x = ya$, $f(y)$ is in a different $\mathcal{D}$-class than $f(x)$ (y is a proper factor of x), and since y is a factor of x, $f(y)$ is in a higher $\mathcal{D}$-class than $f(x)$. Let $d(x) = (y, a)$ as shown in the figure 4 and by the induction hypothesis,

$$h(x) \leq \left(7 \sum_{D \geq_{\mathcal{D}} D_{f(y)}} |D|\right) + 1 \leq \left(7 \sum_{D >_{\mathcal{D}} D_{f(x)}} |D|\right) + 1 \leq 7 \sum_{D \geq_{\mathcal{D}} D_{f(x)}} |D|$$

If x has a proper prefix u such that $f(u)$ and $f(x)$ are in the same $\mathcal{D}$-class, let y_1 be the shortest prefix of x of this kind and let x_1 be the factor of x such that $y_1 x_1 = x$. If x_1 is in the same $\mathcal{D}$-class as x, we can find y_2 and x_2 with the same method (if x_1 hasn't any proper prefix of this kind, we put $y_2 = x_1$ and $x_2 = \epsilon$). By repeating this procedure, we can find a factorization of x of this kind: $x = y_1 y_2 \ldots y_n x_n$ such that each y_i is such that $f(y_i) \in D_{f(x)}$ and y_i hasn't any proper factor u such that $f(u) \in D_{f(y_i)}$ and such that $f(x_n)$ is in a higher $\mathcal{D}$-class than $f(x)$. Let $y = y_1 \ldots y_n$ and let factorize x as described

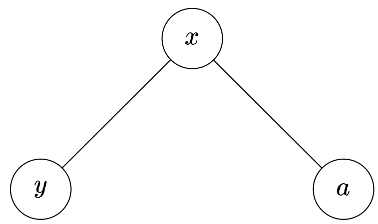

Figure 4: How to build a tree in the first case

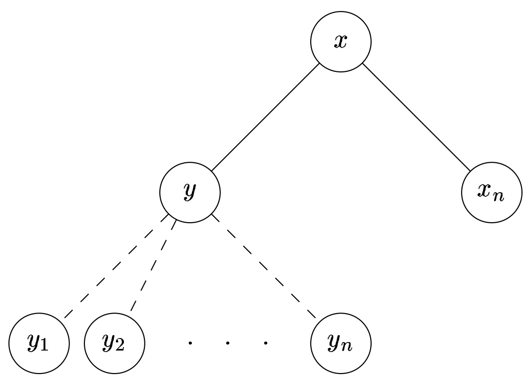

Figure 5: How to build a tree in the second case

in figure 5: we put $d(x) = (y, x_n)$ and since y and $y_1, \dots y_n$ are in the same $\mathcal{D}$-class, we know that we can find a factorization tree with a root labelled by y such that its leaves are labelled by $y_1, y_2, \dots, y_n$ of height at most $5|D_f(x)|$.

We know that we can find a factorization tree for each y_i (as what has been done in the first case) of height at most $\left(7 \sum_{D \geq_{\mathcal{D}} D_{f(y_i)}} |D|\right) + 1$ and that for x_n we can find a tree of height at most $\left(7 \sum_{D \geq_{\mathcal{D}} D_{f(x_n)}} |D|\right)$. Therefore, we can find a tree for x of height lower than $\left(7 \sum_{D >_{\mathcal{D}} D_{f(x)}} |D|\right) + 1 + 5|D_{f(x)}| + 1$, which is lower than $7 \sum_{D \geq_{\mathcal{D}} D_{f(x)}} |D|$.

Consequently, for each word x we can find a factorization tree of height at most $7 \sum_{D \geq_{\mathcal{D}} D_{f(x)}} |D|$, which is lower than $7|S|$. $\square$

5 A lower bound for a few examples

We have shown that for each semigroup S and each morphism f from a free monoid A^* to S, we can find a way to factorize each word over A so as to get a ramseyan factorization tree for this word of linear height. In this section, we will present three different examples showing that the result cannot be significantly improved: for each step of the algorithm, we will show that we can find a semigroup and a morphism such that there exist a word which will need a ramseyan factorization tree of linear height.

5.1 An example when the number of $\mathcal{D}$-class is the size of the semigroup

For each n, let S_n be the semigroup $\{\alpha_1, \ldots \alpha_n\}$ with the following associative operation: $\alpha_i \alpha_j = \alpha_j \alpha_i = \alpha_{max\{i,j\}}$. We can easily see that from each α_i, by multiplying on the both sides, we can reach $\{\alpha_j; j \geq i\}$: $D_{\alpha_i} = \{\alpha_i\}$ and if $i < j$, $D_i <_{\mathcal{D}} D_j$.

Let $A = \{a_1, \ldots a_n\}$ and let f be the morphism such that for each $1 \leq i \leq n$, $f(a_i) = \alpha_i$. For each word $x \in A^+$, we can easily see that $f(x) = \alpha_{i_m}$ where $i_m = max\{i; a_i$ is a letter of $x\}$. Let x_1 be the word $a_1 a_1$ and let $x_{i+1} = (x_i a_{i+1})^2$ for each $1 \leq i \leq n-1$: consequently, $x_i = ((a_1^2 a_2)^2 \ldots a_i)^2$.

We will need a lemma to prove that the minimal height of a factorization tree for x_n is at least n. For each word x, we will note $h_{min}(x)$ the minimal height of a factorization tree for x.

Lemma 3 *For each $k \in \mathbb{N}$, for all x in $\{a_1, \ldots, a_{k-1}\}^+$ and for all $u, v \in A^*$, we have the three following properties*

1. $h_{min}(x) \leq h_{min}(x a_k v)$
2. $h_{min}(x) \leq h_{min}(u a_k x)$
3. $h_{min}(x) \leq h_{min}(u a_k x a_k v)$

Proof:

1. We will show the first result by induction on $|x|$; if $|x| = 1$, then $h(x) = 0$ and there is nothing to prove. Suppose now that $|x| > 1$, we will do an induction on $|v|$. If $|v| = 0$ then there exists $p \in \mathbb{N}$ and $x_1, x_2, \ldots, x_p \in A^+$, such that $d(xa_k) = (x_1, \ldots x_p, a_k)$ or $d(xa_k) = (x_1, \ldots x_{p-1}, x_p a_k)$. In the first case, if $p \geq 2$, then $f(x_1) = f(a_k) = \alpha_k$, but since x_1 is a factor of x, we know that a_k is not a letter of x_1 and therefore $f(x_1) \neq \alpha_k$; consequently $d(xa_k) = (x, a_k)$ and $h_{min}(x) \leq h_{min}(xa_k)$. In the second case, if $p \geq 3$ we have exactly the same problem since $f(x_p a_k) = \alpha_k$ whereas $f(x_1) \neq \alpha_k$; consequently $d(xa_k) = (x_1, x_2 a_k)$ and by the

induction hypothesis on $|x|$, we know that $h_{min}(x_2) \leq h_{min}(x_2 a_k)$ (since $|x_2| < |x|$); consequently $h_{min}(x) \leq 1 + max\{h_{min}(x_1), h_{min}(x_2)\} \leq 1 + max\{h_{min}(x_1), h_{min}(x_2 a_k)\} = h_{min}(x a_k)$. Assume now that $|v| \geq 1$; if x is a factor of an element of $d(x a_k v)$, either x is an element of $d(x)$ and we can easily see that $h_{min}(x) \leq h_{min}(x a_k v)$, or the first element of $d(x a_k v)$ is of this kind $x a_k v'$, with $|v'| < |v|$ and we can apply the induction hypothesis on $|v|$ to show that $h_{min}(x) \leq h_{min}(x a_k v)$. If x is not a factor of any element of $d(x a_k v)$, for the same reasons as before, we can show that the only way to factorize $x a_k v$ is to find $x_1, x_2 \in \{a_1, \ldots, a_{k-1}\}^+$ such that $d(x a_k v) = (x_1, x_2 a_k v)$. And by the induction hypothesis on $|x|$, we know that $h_{min}(x_2) \leq h_{min}(x_2 a_k v)$ and therefore $h_{min}(x) \leq 1 + max\{h_{min}(x_1), h_{min}(x_2)\} \leq h_{min}(x a_k v)$. Consequently we have prove the first statement of the lemma.

2. The proof of this statement is exactly the same as before.

3. We will show the last statement by induction on $|u| + |v|$. If $|u| + |v| = 0$, we will wonder how can $a_k x a_k$ be factorized. If x is a factor of an element of $d(a_k x a_k)$, this element is either x, $a_k x$ or $x a_k$, and by what we have seen before, we know that $h(x) \leq h(a_k x a_k)$. If x isn't a factor of any element of $d(a_k x a_k)$, there exist $p \in \mathbb{N}$ and $x_1, \ldots, x_p$ such that $d(a_k x a_k) = (a_k, x_1, \ldots, x_p, a_k)$, $d(a_k x a_k) = (a_k x_1, \ldots, x_p, a_k)$, $d(a_k x a_k) = (a_k, x_1, \ldots, x_p a_k)$ or $d(a_k x a_k) = (a_k x_1, \ldots, x_p a_k)$. For the first three cases, if $p \geq 2$ (what is supposed since x is not supposed to be a factor of any of the element of $d(a_k x a_k)$), we have more than three components and therefore we must have $f(a_k) = \alpha_k = f(x_1)$ (resp. $f(a_k x_1) = \alpha_k = f(x_p)$ and $f(a_k x_p) = \alpha_k = f(x_1)$), but we know that for each $1 \leq j \leq p$, $f(x_j) \neq \alpha_k$ for the same reasons as before. In the fourth case, if $p \geq 3$, we have $f(a_k x_1) = \alpha_k = f(x_2)$, but we know that a_k is not a letter of x_2 and consequently $f(x_2) \neq \alpha_k$ and therefore $p = 2$ and $d(a_k x a_k) = (a_k x_1, x_2 a_k)$. By the others statements of the lemma, we know that $h_{min}(x_1) \leq h_{min}(a_k x_1)$ and $h_{min}(x_2) \leq h_{min}(x_2 a_k)$; consequently, $h_{min}(x) \leq 1 + max\{h_{min}(x_1), h_{min}(x_2)\} \leq 1 + max\{h_{min}(a_k x_1), h_{min}(x_2 a_k)\} = h_{min}(a_k x a_k)$. Suppose now that $|u| + |v| \geq 1$, if x is a factor of an element of $d(u a_k x a_k v)$, this element is of one of this kind: x and there is nothing to prove, $u' a_k x$ or $x a_k v'$ and we already know that the result is true, or $u' a_k x_a n v'$ with $|u'| + |v'| < |u| + |v|$ and we know the result by the induction hypothesis. If x is not a factor of an element of $d(x)$, we show for the same reason as before that the only way to factorize $u a_k x a_k v$ is to find $q, r \in \mathbb{N}$, $x_1, x_2 \in A^+$, $u_1, \ldots, u_q, v_1, \ldots, v_r \in A^*$ (we only authorize u_q and v_1 to be ϵ) such that $x = x_1, x_2$, $u = u_1 \ldots u_q$, $v = v_1 \ldots v_r$ and $d(u a_k x a_k v) = (u_1, \ldots, u_{q-1}, u_q a_k x_1, x_2 a_k v_1, v_2, \ldots, v_r)$. By the precedent results, we know that $h_{min}(x_1) \leq h_{min}(u_q a_k x_1)$ and $h_{min}(x_2) \leq h_{min}(x_2 a_k v_1)$; consequently $h_{min}(x) \leq 1 + max\{h_{min}(x_1), h_{min}(x_2)\} \leq h_{min}(u a_k x a_k v)$. Therefore we have proove the last statement of the lemma.

□

We want to show that for each k, $h(x_k) \geq k$, and we will do it by induction on k. For $k = 1$, $x_1 = a_1 a_1$ and it's obvious that we need a tree of height at least 1. Suppose now that $k = n > 1$, $x_k = x_{k-1}a_kx_{k-1}a_k$; if there exists $u_1, u_2, \ldots, u_p$ with $p \geq 3$ such that $d(x_k) = (u_1, \ldots, u_p)$, there exist $1 \leq i, j \leq p$ such that a_k is a letter of u_i but not of u_j. Consequently $f(u_i) = a_k \neq f(u_j)$ and therefore, this factorization is not ramseyan. Consequently, the only kind of factorization we can find is $d(x) = (u_1, u_2)$. If a_k is a letter of u_1, then u_1 is of this kind $x_{k-1}a_ku_1'$ and $h(u_1) \geq h(x_{k-1}) \geq k-1$: $h(x) \geq k$. If a_k is not a letter of u_1, u_2 is of this kind $u_2'a_kx_{k-1}a_k$ and $h(u_2) \geq h(x_{k-1}) \geq k-1$: $h(x) \geq k$. Therefore, for each k, $h(x_k) \geq k$ and consequently we can find a word in A^+ such that the size of the minimal ramseyan factorization tree for x is $|S|$.

5.2 A new kind of factorization forest

We will introduce a new kind of factorization forest, so as to show that the bound we have for each step of the algorithm cannot be really improved. We will now consider a fonction d' from A^+ to $\mathcal{F}(A^+)$ such that $d'(x) = (x_1, x_2, \ldots, x_k)$ implies that there exist an idempotent e such that for each $2 \leq i \leq k-1$, $f(x_i) = e$. As in the former definition, the only words x such that $d'(x) = x$ must be the elements of A. We can easily remark that if we can find an usual ramseyan factorization tree for a word x, this tree will also be a factorization tree for the new kind of factorization, and reversly if we have a factorization tree of height n for a word x, we can easily find an usual ramseyan factorization tree of height at most $3n$. We will note $h'_{min}(x)$ the minimal height of a new kind of factorization tree for x.

With this kind of factorization, we can proove the following lemma that will be useful later.

Lemma 4 *For all $x \in A^+$, for all $u, v \in A^*$, $h'_{min}(x) \leq h'_{min}(uxv)$.*

Proof:
We will do the proof by induction on $|x|$; if $|x| = 1$, $h'_{min}(x) = 0$ and there is nothing to proove. Suppose now that $|x| > 1$, we will do an induction on $|u| + |v|$; if $|u| + |v| = 0$, $uxv = x$ and there is nothing to prove. If $|u| + |v| \geq 1$, either x is a factor of an element of $d(x)$ which will be of this kind: $u'xv'$ with $|u'| + |v'| < |u| + |v|$; by induction, we know that $h'_{min}(x) \leq h'_{min}(u'xv')$ and consequently $h'_{min}(x) \leq h'_{min}(uxv)$. Otherwise, we can find $p \geq 2$, $q, r \geq 1$, $x_1, \ldots, x_p \in A^+$, $u_1, \ldots, u_q, v_1, \ldots, v_r \in A^*$ (we only authorize u_q and v_1 to be ϵ) such that $x = x_1 \ldots x_p$, $u = u_1 \ldots u_q$, $v = v_1 \ldots v_r$ and $d(uxv) = (u_1, \ldots, u_{q-1}, u_qx_1, x_2, \ldots, x_{p-1}, x_pv_1, v_2, \ldots v_r)$. We know that $(x_1, x_2, \ldots, x_p)$ can be a correct factorization for x: $f(x_2) = \cdots = f(x_{p-1})$ is an idempotent, since $(u_1, \ldots, u_{q-1}, u_qx_1, x_2, \ldots, x_{p-1}, x_pv_1, v_2, \ldots v_r)$ is a good factorization for uxv. And since $|x_1|$ and $|x_2|$ are lower than $|x|$, by induction hypothesis, we know that $h'_{min}(x_1) \leq h'_{min}(u_qx_1)$ and $h'_{min}(x_2) \leq h'_{min}(x_2v_1)$. Therefore $h'_{min}(x) \leq 1 + max_{1 \leq i \leq p}\{h'_{min}(x_i)\} = h'_{min}(uxv)$. □

5.3 The group case

Given a group G of size n, we want to show that there exist a word such that the height of all new factorization tree for this word will be at least n; and consequently a ramseyan factorization tree for this word will also be of height at least n.

Given a word x, we will note $x(m)$ the prefix of x of length m and $P(x) = \{f(x(m)); 1 \leq m \leq |x|\}$. Let $\alpha_1, \ldots, \alpha_n$ be the element of G such that $\alpha_1 = e$ is the identity of G and let $A = \{a_1, \ldots, a_n\}$. We define f to be the morphism such that for each $1 \leq i \leq n$, $f(a_i) = \alpha_i$.

Lemma 5 *Given two words x and y satisfying the following statement*

- $|y| = |x|$,
- *there exists a bijection g between $P(x)$ and $P(y)$,*
- *for all $1 \leq m \leq |x|$, $g(f(x(m))) = f(y(m))$,*

we have that $h'_{min}(x) = h'_{min}(y)$.

Proof:
We will show that $h'_{min}(x) \leq h'_{min}(y)$ by induction on x; if $|x| = 1$, $h'_{min}(x) = 0$ and there is nothing to proove. Suppose that $|x| > 1$, let $p \in N$ and $y_1, \ldots, y_p \in A^+$ be such that $d'(y) = (y_1, \ldots, y_p)$; let $x_1, \ldots, x_p \in A^+$ such that $x = x_1 \ldots x_p$ and for each $1 \leq i \leq p$, $|x_i| = |y_i|$ (it is possible thanks to the first hypothesis). For each $2 \leq i \leq p-1$, $f(y_i) = e$ and $f(y_1 \ldots y_i) = f(y_1)$; consequently, for each $2 \leq i \leq p-1$, $f(x_1 \ldots x_i) = g(f(y_1))$ and therefore $f(x_i) = e$: $(x_1, \ldots, x_p)$ is a good factorization of x. Moreover for each i, let g_i be the bijection such that for each z, $g_i(z) = f(y_1 \ldots y_{i-1})^{-1} g(x_1 \ldots x_{i-1} z)$. For all $1 \leq m \leq |x_i|$, $g(f(x_1 \ldots x_{i-1} x_i(m))) = f(y_1 \ldots y_{i-1} y_i(m))$ and so, $g_i(x_i(m)) = y_i(m)$. Since for each $1 \leq i \leq p$, $|x_i| = |y_i|$, g_i is a bijection (G is a group and g is a bijection) between $P(x_i)$ and $P(y_i)$ such that for all $1 \leq m \leq |x|$, $g_i(f(x_i(m))) = f(y_i(m))$, we can apply the induction hypothesis on each x_i ($|x_i| < |x|$). Therefore, $h'_{min}(x) \leq 1 + max_{1 \leq i \leq p}\{h'_{min}(x_i)\} \leq 1 + max_{1 \leq i \leq p}\{h'_{min}(y_i)\} = h'_{min}(y)$. □

For each $1 \leq k \leq n$, we will construct recursively a word x_k such that $P(x_k) = \{\alpha_1 \ldots \alpha_k\}$ and $h'_{min}(x_k) \geq k$.

If $k = 1$, let $x_1 = a_1 a_1$; we could easily see that $P(x_1) = \{\alpha_1\}$ and that the height for any new factorization tree will be at least 1.

Suppose now that $k \geq 1$, we will construct x_k from the structure of x_{k-1}. For each $1 \leq i \leq k$, let y_i be the word such that for each $1 \leq m \leq |x_{k-1}|$, if $f(x_{k-1}(m)) = \alpha_i$, $f(y_i(m)) = \alpha_k$ and $f(y_i(m)) = f(x_{k-1}(m))$ otherwise. Since $\alpha_k \notin P(x_{k-1})$, $y_k = x_{k-1}$ and consequently $P(y_k) = P(x_{k-1})$; moreover for each $1 \leq i \leq k-1$, $P(y_i) = P(x_{k-1}) \setminus \{\alpha_i\} \cup \{\alpha_k\}$. We can easily see that x_{k-1} satisfy the previous lemma with each y_i, and therefore for each $1 \leq i \leq k$, $h'_{min}(y_i) \geq k-1$.

We will construct $(b_i)_{1 \leq i \leq k-1} \in A^{k-1}$ recursively such that for each $1 \leq i \leq k-1$, $f(b_i) = f(y_1 b_1 \ldots y_i)^{-1}$; which is possible since for each $\alpha \in G$, there exists $a \in A$ such that $f(a) = \alpha$. We are now ready to create x_k: we put $x_k = y_1 b_1 y_2 b_2 \ldots y_{k-1} b_{k-1} y_k$.

$$P(x_k) = \bigcup_{1 \leq i \leq p} \{f(y_1 b_1 \ldots y_{i-1} b_{i-1}) P(y_i) \cup f(y_1 b_1 \ldots y_i b_i)\} = \{\alpha_1, \ldots, \alpha_k\}$$

Let $p \geq 2$ and $(z_1, \ldots, z_p) \in A^+$ such that $d(x) = (z_1, \ldots, z_p)$ and let $\beta = f(z_1) = \cdots = f(z_1 \ldots z_{q-1}) \in P(x_k)$. There exists $1 \leq i_0 \leq k$ such that $\beta \notin P(y_{i_0})$ and there exists $1 \leq j \leq p$ such that y_{i_0} is a factor of z_j: otherwise, there exist y'_{i_0}, y''_{i_0} such that $y_{i_0} = y'_{i_0} y''_{i_0}$ and there exist $1 \leq l < p$ such that $z_1 \ldots z_l = y_1 b_1 \ldots b_{i-1} y'_{i_0}$ and therefore $f(y'_{i_0}) = f(y_1 b_1 \ldots b_{i-1} y'_{i_0}) = f(z_1 \ldots z_l) = \beta$, which is false. Therefore, $h'_{min}(z_j) \geq h'_{min}(y_{i_0}) \geq k-1$ and therefore $h'_{min}(x_k) \geq k$.

We have consequently built a word x such that $P(x_k) = \{\alpha_1 \ldots \alpha_k\}$ and $h'_{min}(x_k) \geq k$.

Consequently, for each group G, we can find a word which will need a ramseyan factorization tree of height at least $|G|$.

5.4 The case of rectangular bands

Given two nonempty sets I and J, we can define an associative multiplication on the set $I \times J$ as follows:

$$\forall i, i' \in I, \forall j, j' \in J, (i,j)(i',j') = (i,j')$$

A semigroup S is called a rectangular band if there exist I and J such that S is isomorphic to $I \times J$ with the previous multiplication. We will note lp and rp the two functions such that $\forall (i,j) \in I \times J, lp(i,j) = i$ and $rp(i,j) = j$.

In a rectangular band, it is obvious to see that each element is an idempotent. Therefore, for each new factorization $(x_1, x_2, \ldots, x_n)$ of a word x, $x_2 = \cdots = x_{n-1}$ can take all the values possible in S and there exist $(i', j') \in I \times J$ such that for each $2 \leq l \leq n-2, rp(x_l) = i'$ and $lp(x_{l+1}) = j'$. We can also easily see that all the element are in the same $\mathcal{D}$-class and that each element is the only one belonging to its $\mathcal{H}$-class. Consequently, the number of $\mathcal{H}$-classes is the size of the rectangular band. Let $A = \{a_{ij}; i \in I, j \in J\}$ and f be the morphism such that for each $(i,j) \in I \times J$, $f(a_{ij}) = (i,j)$. Given a word $x = a_{i_1 j_1} \ldots a_{i_p j_p}$, we will use $Int(x) = \{(j_k, i_{k+1}); 1 \leq k \leq p-1\}$ and for each $1 \leq k \leq p-1$,we will define $betw_x(k) = (j_k, i_{k+1})$.

Given a rectangular band S defined by the sets I (of size n) and J (of size m), we want to show that there exist a word such that the height of all new factorization tree for this word will be at least $n \times m$; and consequently a ramseyan factorization tree for this word will also be of height at least $|S|$.

As for the group cases, we will construct recursively a word x_k such that $|Int(x_k)| = k$ and $h'_{min}(x_k) \geq k$. For $k = 1$, let $(i_0, j_0) \in S$ and we can put $x_1 = a_{i_0 j_0} a_{i_0 j_0}$ which will need a tree of height 1 and $|Int(x_1)| = |\{(j_0, i_0)\}| = 1$.

Suppose now that $k > 1$, we will construct x_k from the structure of x_{k-1}. Let $(j_1, i_1) \notin Int(x_{k-1})$; for each $(j,i) \in Int(x_{k-1})$, we can construct a word y_{ji} by replacing the letters of x_{k-1} such that for each $1 \leq p \leq |x|-1$, $betw_{y_{ji}}(p) = (j_1, i_1)$ if $betw_{x_{k-1}}(p) = (j,i)$ and $betw_{y_{ji}}(p) = betw_{x_{k-1}}(p)$ otherwise. Consequently, it is obvious to see that the minimal height of a new factorization tree for each y_{ij} is at least $k-1$ and that $\forall (i,j) \in Int(x_{k-1}), Int(y_{ji}) = \{(j_1, i_1)\} \cup Int(x_{k-1}) \setminus \{(j,i)\}$. Moreover, we can assume that $\forall (j,i) \in Int(x_{k-1}), f(y_{ji}) = (i_0, j_0)$; we will define y to be the concatenation of all the y_{ij} that have been defined previously and we will define $x_k = x_{k-1} x_{k-1} y x_{k-1}$. We can easily see that $f(x_k) = (i_0, j_0)$ and that $Int(x_k) = Int(x_{k-1}) \cup \{(j_1, i_1)\}$.

We will now proove that for each new factorization tree we can have for x_k, it will be a tree of height at least k. Let $d'(x_k) = (u_1, \ldots, u_r)$ be a new factorization for x_k. Either x_{k-1} is a factor of u_1 or u_r and in this case we know that u_1 or u_r will need a tree of height at least $k-1$ and therefore, we will have a tree of height at least k. In the other case, we know that $r \geq 3$; if $r = 3$, x_{k-1} will be a factor of u_2 and the height of a tree for u_2 will be at least $k - 1$ and consequently, the height of the tree for x_k will be at least k. If $r \geq 4$, there exist $i' \in I$ and $j' \in J$ such that for each $2 \leq l \leq r - 2$, we have a $rp(u_l) = j'$ and $rp(u_{l+1}) = i'$. If $(i', j') = (i_1, j_1)$, x_{k-1} will be a factor of u_2 and therefore the height of the tree for x_k will be at least k. In the other case, $(j', i') \in Int(x_{k-1})$ and there exist $2 \leq l \leq r - 1$ such that $y_{j'i'}$ is a factor of u_l; consequently the height of a tree for u_l is at least $k - 1$ and the height of the tree for x_k is at least k.

We have proove that for each $k \leq nm$ we can find a word x_k such that $h'_{min}(x_k) \geq k$. Consequently, for each rectangular band S, we can find a word which will need a ramseyan factorization tree of height at least $|S|$. .

5.5 About the algorithm

In the former examples, we have proove that we cannot improve significantly the algorithm described in the previous part. Indeed for each part of the algorithm, we have found an example which will need a ramseyan factorization tree of linear height. When the number of $\mathcal{D}$-classes is greater than one, we have found a family of semigroup such that the number of $\mathcal{D}$-classes is the size of the semigroup and for each semigroup of this family we can find a word which will need a ramseyan factorization tree of height greater or equal to the size of the semigroup. When the image of all the factors of a word are in the same $\mathcal{D}$-class than the word, we have found a kind of semigroup (the rectangular bands) in which we can always find a word which will need a ramseyan factorization tree of height greater or equal to the size of the semigroup. For the group case, we have prooved that for each group G, we can find a word which will need a ramseyan factorization tree of height at least $|G|$.

References

[Ha00] Kosaburo HASHIGUCHI. *New upper bounds to the limitdness of distance automata*, Theoretical Computer Science 233 (2000). 19-32

[La79] Gérard LALLEMENT. *Semigroups and Combinatorial Applications*, John Wiley and Sons, New York (1979)

[Le98] Hing LEUNG. *On finite automata with limited nondeterminism*, Acta Informatica 35 (1998). 595-624

[Le98'] Hing LEUNG. *The topological approach to the limitedness problem on distance automata*, Idempotency, Cambridge University Press (1998). 88-111

[LP02] Hing LEUNG and Viktor PODOLSKIY. *The limitedness problem on distance automata: Hashiguchi's method revisited*, submitted to Theorical Computer Science.

[Si78] Imre SIMON. *Limited subsets of a Free Monoid*, Proc. Nineteenth Annual IEEE Symposium on Foundations of Computer Science. 143-150

[Si90] Imre SIMON. *Factorization Forests of Finite Height*, Theoretical Computer Science 72 (1990). 65-94

[Si92] Imre SIMON. *A Short Proof of the Factorization Forest Theorem*, Tree Automata and Languages (1992). 433-438

[Si94] Imre SIMON. *On Semigroups of Matrices over the Tropical Semiring*, Informatique Théorique 28 (1994). 277-294