

# Combinatoire des mots

V. Berthé

berthe@irif.fr

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Un peu de vocabulaire</b>	<b>2</b>
<b>3</b>	<b>Autour du théorème de Fine et Wilf</b>	<b>3</b>
3.1	Énoncé et preuve . . . . .	3
3.2	Mot de Christoffel . . . . .	5
3.3	Mot central . . . . .	7
3.4	Mot de Christoffel dual . . . . .	10
<b>4</b>	<b>Systèmes dynamiques symboliques</b>	<b>10</b>
<b>5</b>	<b>Substitutions</b>	<b>13</b>
<b>6</b>	<b>Fonction de Complexité</b>	<b>15</b>
6.1	Définition . . . . .	15
6.2	Exemples . . . . .	17
<b>7</b>	<b>Suites sturmiennes</b>	<b>18</b>
7.1	Premières définitions . . . . .	18
7.2	Équilibre . . . . .	21
7.3	Facteurs spéciaux . . . . .	22
7.4	Graphe des mots . . . . .	23
7.5	Graphes des mots et fréquences . . . . .	25
7.6	Substitutions sturmiennes . . . . .	26

# 1 Introduction

La combinatoire des mots remonte historiquement au début du siècle avec les travaux d'A. Thue (1906). Pour plus de références, voir le cours [2]. Ses domaines d'application couvrent, par exemple, la bioinformatique ou le traitement du langage naturel, comme décrit dans l'ouvrage l'ouvrage [7]. La combinatoire des mots entretient des liens étroits avec la géométrie discrète, les mathématiques discrètes, l'algèbre, la théorie des nombres, la dynamique symbolique, les groupes libres, comme illustré dans [8]. Nous allons considérer comme objets dans ce cours les substitutions, les suites automatiques, les suites sturmiennes et leurs liens avec la géométrie discrète, le théorème de Fine et Wilf, la dynamique symbolique.

## 2 Un peu de vocabulaire

Soit  $\mathcal{A}$  un ensemble fini appelé *alphabet*. Une *mot* est une chaîne finie d'éléments de  $\mathcal{A}$ . Le mot vide est noté  $\varepsilon$ . L'ensemble des mots est noté  $\mathcal{A}^*$ ; on a  $\varepsilon \in \mathcal{A}^*$ .

**Définition 1** (Concaténation). La *concaténation* de deux mots  $V = v_1 \cdots v_r$  et  $W = w_1 \cdots w_s$  est le mot, noté  $VW$ , égal à  $v_1 \cdots v_r w_1 \cdots w_s$ .

Cette opération est associative et admet pour élément neutre le mot vide  $\varepsilon$ .

**Définition 2** (Monoïde). L'ensemble des mots finis  $\mathcal{A}^*$  muni de la concaténation est le *monoïde libre* engendré par  $\mathcal{A}$ .

**Définition 3** (Mots infinis). On considère des *mots infinis unilatéraux* (encore appelés *suites*)  $u = (u_n)_{n \in \mathbb{N}} \in \mathcal{A}^{\mathbb{N}}$ , et des *mots infinis bilatéraux* (encore appelés *suites biinfinies*)  $u = (u_n)_{n \in \mathbb{Z}} \in \mathcal{A}^{\mathbb{Z}} : \cdots u_{-l} \cdots u_{-1} \cdot u_0 u_1 \cdots u_k$ .

**Définition 4** (Facteur). Soit  $u$  un mot fini ou infini. Le mot  $V = v_1 \cdots v_r$  est un *facteur* de  $u$  s'il existe  $n$  tel que

$$u_n = v_1, \cdots, u_{n+r-1} = v_r.$$

On dit alors que  $V$  apparaît à l'indice  $n$ . Le mot  $u_n \cdots u_{n+r-1}$  est appelé *occurrence* de  $V$ .

Le mot  $V$  est un *préfixe* de  $u$  s'il existe un mot  $s$  tel que  $u = Vs$ .

Le mot  $V$  est un *suffixe* de  $u$  s'il existe un mot  $p$  tel que  $u = pV$ .

L'*image miroir* du mot  $v = v_1 \cdots v_r$  est le mot, noté  $\bar{V}$ , égal à  $v_r \cdots v_1$ .

Un *palindrome* est un mot égal à son image miroir.

**Notation 1.** On note  $|W|$  la longueur du mot  $W$  : si  $W = w_1 \cdots w_s$ , alors  $|W| = s$ . On note  $|W|_a$  le nombre d'occurrences de la lettre  $a$  dans  $W$ .

**Définition 5** (Langage). Soit  $u$  un mot fini ou infini. Le langage  $\mathcal{L}(u)$  de  $u$  est égal à l'ensemble des facteurs de  $u$ . On note  $\mathcal{L}_n(u)$  l'ensemble des facteurs de longueur  $n$  de  $u$ .

**Définition 6** (Système dynamique discret). Un *système dynamique discret* est défini par la donnée d'un ensemble  $X$  sur lequel agit une application  $T$ ; on le note  $(X, T)$ . L'*orbite*  $\mathcal{O}(x)$  du point  $x \in X$  est définie par  $\mathcal{O}(x) = \{T^n x, n \in \mathbb{N}\}$ .

En général, on suppose  $X$  compact et  $T$  continue surjective.

**Exemple 1.** • **Rotation** Soit  $\alpha \in \mathbb{R}$ . Soit  $X = \mathbb{R}/\mathbb{Z}$  et  $R_\alpha : \mathbb{R}/\mathbb{Z} \rightarrow \mathbb{R}/\mathbb{Z}$ ,  $x \mapsto x + \alpha$ .

• **Machine de Turing** On considère une machine de Turing sur un alphabet  $\Sigma$  d'ensemble d'états internes  $Q$ . L'état de la machine à l'instant  $t$  correspond à la donnée du contenu du ruban (qui appartient à  $\Sigma^{\mathbb{N}}$ ), de la position de la tête de lecture et de l'état interne de la machine (qui appartient à  $Q$ ), ce qui se code par un élément de l'espace  $X = \Sigma^{\omega} \times \mathbb{Z} \times Q$ . On considère alors une transformation de  $T$  de  $X$  dans  $X$ , qui décrit l'évolution de  $t$  à  $t + 1$  de la machine.

• **Triangle de Pascal modulo 2** On considère sur l'ensemble des suites  $\{0, 1\}^{\mathbb{Z}}$  l'application

$$T : (u_n)_{n \in \mathbb{Z}} \mapsto (v_n)_{n \in \mathbb{Z}}, \text{ avec } v_n = u_n + u_{n-1}, \forall n.$$

### 3 Autour du théorème de Fine et Wilf

Pour plus de détails sur les résultats de ce paragraphe, voir [3] et [5].

#### 3.1 Énoncé et preuve

**Définition 7.** Soit  $W = w_1 \cdots w_n \in \mathcal{A}^*$ . Un entier naturel  $k$  est une *période* de  $W$  si  $k \geq 1$  et  $w_i = w_{i+k}$  pour  $1 \leq i \leq n - k$ .

**Théorème 1** (Fine et Wilf). *Si un mot  $W$  a deux périodes  $p$  et  $q$ , et si  $|W| \geq p + q - \text{pgcd}(p, q)$ , alors  $\text{pgcd}(p, q)$  est aussi une période de  $W$ .*

*Preuve* Il suffit de démontrer le théorème pour  $d = \text{pgcd}(p, q) = 1$ . En effet, on prend une lettre sur  $d$ . On a alors  $d$  mots de longueur  $p/d + q/d - 1$  constants, d'où la  $d$ -périodicité.

Il suffit aussi de montrer le résultat pour une longueur égale à  $p + q - 1$ .

On suppose donc  $|W| = p + q - 1$ , avec  $\text{pgcd}(p, q) = 1$ .

On considère l'application  $\rho$  de  $\{0, 1, \dots, p + q - 1\}$  dans lui-même définie sur  $[0, q - 1]$  par  $\rho(x) = x + p$  et sur  $[q, p + q - 1]$ , par  $\rho(x) = x - q$ .

- L'application  $\rho$  est en fait une permutation (bijection de  $\{0, 1, \dots, p + q - 1\}$  dans lui-même). En effet, la surjectivité provient de  $\rho(\{0, 1, \dots, q - 1\}) = \{p, \dots, p + q - 1\}$  et  $\rho(\{q, \dots, p + q - 1\}) = \{0, \dots, p - 1\}$ . On déduit la bijectivité de l'égalité des cardinaux des ensembles.
- Montrons que l'orbite de 0, c'est-à-dire  $\{\rho^k(0), k \in \mathbb{N}\}$ , décrit les  $p + q$  points de  $\{0, 1, \dots, p + q - 1\}$ . Soit  $i \in \mathbb{N}^*$  tel que  $\rho^i(0) = 0$ . Alors il existe  $a, b \in \mathbb{N}$  avec  $i = a + b$  tel que  $ap - bq = 0$ . Comme  $p$  et  $q$  sont premiers entre eux, on en déduit que  $i \geq p + q$ . On en déduit que  $\rho^i(x) = x$  pour  $i \in \mathbb{N}^*$  implique  $i \geq p + q$ , et que

$$\{\rho^i(p), 0 \leq i \leq p + q - 2\} = \{1, \dots, p + q - 1\}.$$

On en déduit le théorème. ■

**Exercice 1.** Montrer que si deux mots  $U$  et  $V$  satisfont  $UV = VU$ , alors il existe un mot  $W$  et  $i, j \in \mathbb{N}$  tels que  $U = W^i$ ,  $V = W^j$ . Donner une preuve par récurrence sur  $\max(|U|, |V|)$ , et une preuve basée sur le théorème de Fine et Wilf.

**Exemple 2.** On considère  $p = 4$  et  $q = 7$ . On a  $p + q - \text{pgcd}(p, q) = 10$ . Un mot de longueur  $p + q - 1 = 10$  qui admet  $p$  et  $q$  comme période est constant. On le voit en considérant l'orbite de 4 sous l'action des opérations  $+4$  et  $-7$  dans  $\{1, \dots, 10\}$  : les 10 premiers termes de l'orbite décrivent l'ensemble  $\{1, \dots, 10\}$ .

Notons que le mot de longueur  $9 = p + q - 2$

001000100

a pour périodes 4 et 7.

### 3.2 Mot de Christoffel

Nous allons associer à l'application  $\rho$  définie dans la preuve précédente un mot fini en codant l'orbite de 0. Il s'agit d'une démarche classique en systèmes dynamiques discrets, à savoir, coder les orbites des points par rapport à une partition finie, ce qui donne des mots infinis. Nous y reviendrons lors de l'étude des mots sturmiens.

**Définition 8** (Mot de Christoffel). Soient  $p$  et  $q$  deux entiers naturels premiers entre eux et soit  $n = p + q$ .

On se donne l'alphabet ordonné  $\{x < y\}$ . Le mot de Christoffel  $W$  de pente  $\frac{p}{q}$  sur cet alphabet est défini par  $W = w_1 \cdots w_n$ , avec

$$w_i = \begin{cases} x & \text{si } (i-1)p \in \{0, 1, \dots, q-1\} \pmod{n} \\ y & \text{si } (i-1)p \in \{q, q+1, \dots, n-1\} \pmod{n} \end{cases}$$

pour  $i = 1, \dots, n$ .

**Notation 2.** La notation  $k \pmod{n}$  désigne le reste dans la division euclidienne de  $k$  par  $n$ .

**Remarque 1.** • On a

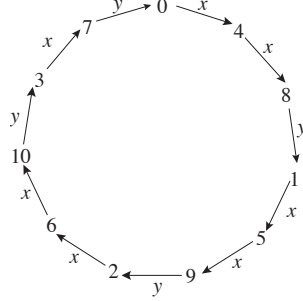
$$\begin{aligned} (i-1)p \in \{0, 1, \dots, q-1\} \pmod{n} & \text{ iff } ip \pmod{n} > (i-1)p \pmod{n} \\ & \text{ iff } ip \in \{p, 1, \dots, p+q-1\} \end{aligned}$$

- Soit  $W$  le mot de Christoffel sur l'alphabet  $\{x < y\}$  de pente  $\frac{p}{q}$ . On a  $|W|_x = q$  et  $|W|_y = p$ .
- Le terme pente vient de l'interprétation géométrique des mots de Christoffel que nous ferons plus tard, en lien avec l'étude des mots sturmiens.

On considère le graphe orienté suivant :

- Sommets :  $\{1, 2, \dots, p+q-1\}$ .
- On a une flèche de  $i \rightarrow j$  si  $i+p \equiv j \pmod{n}$  étiquetée  $x$ , si  $i < j$ , et  $y$ , si  $i > j$ .

Ce graphe est appelé *graphe de Cayley* du mot de Christoffel  $W$ .



**Figure 1**

Le mot de Christoffel  $w = xyxxyxyxyxy$   
de pente  $\frac{4}{7}$ , sur l'alphabet  $\{x < y\}$ ,  $p = 4$ ,  $q = 7$ ,  $n = 11$

**Définition 9** (Mot de Christoffel propre). Un mot de Christoffel est dit *propre* s'il n'est pas réduit à une lettre.

**Théorème 2.** Soient  $p$  et  $q$  deux entiers naturels premiers entre eux, et soit  $n = p + q$ . Soient  $p^*$  et  $q^*$  dans  $\{0, 1, \dots, p + q - 1\}$  définis par  $pp^* \equiv 1 \pmod{n}$  et  $qq^* \equiv 1 \pmod{n}$ .

Soit  $W = w_1 \cdots w_n$  le mot de Christoffel de pente  $\frac{p}{q}$  sur l'alphabet  $\{x, y\}$ . Soit  $U$  le mot défini par  $U = w_2 \cdots w_{n-1}$ . Le mot  $U$  a pour périodes  $p^*$  et  $q^*$ .

*Preuve* Notons que  $p^*$  et  $q^*$  existent, car  $p$  et  $q$  étant premiers entre eux, on a  $p$  premier avec  $n = p + q$ , et de même  $q$  premier avec  $n = p + q$ . De plus, l'égalité  $p + q = n$  implique  $(n - p^*)q \equiv (n - p^*)(n - p) \equiv 1 \pmod{n}$ , et donc que  $(n - p^*)$  est congru à l'inverse de  $q$  modulo  $n$ . Comme  $n - p^* \in \{0, 1, \dots, n - 1\}$ , on a  $q^* = n - p^*$ . On en déduit donc  $n = p^* + q^*$ .

Il suffit de montrer que  $U$  a pour période  $p^*$ . En effet, soit  $W'$  le mot de Christoffel de pente  $\frac{q}{p}$  sur l'alphabet  $\{x < y\}$ , et soit  $U'$  le mot obtenu en privant  $W'$  de ses première et dernière lettres. On obtient  $U'$  en échangeant  $x$  et  $y$  dans  $U$  et en prenant son image miroir. En effet, le graphe de Cayley de  $W'$  est obtenu en inversant l'orientation et en échangeant  $x$  and  $y$  dans le graphe de Cayley de  $W$ .

On a  $U = w_2 \cdots w_{n-1}$ . Il est suffisant de montrer que pour  $i, j \in \{2, \dots, n - 1\}$  et  $j = i + p^*$ , on a  $w_i = x \Leftrightarrow w_j = x$ . Notons que puisque  $pp^* \equiv 1 \pmod{n}$ , on a  $jp \equiv ip + 1 \pmod{n}$  et  $(j - 1)p \equiv (i - 1)p + 1 \pmod{n}$ .

- Supposons que  $x_i = x$ . On a alors  $ip \pmod{n} > (i - 1)p \pmod{n}$ . On ne peut avoir  $ip \pmod{n} = n - 1$ , sinon  $i = q^*$  (puisque  $pq^* \equiv -1 \pmod{n}$ ) et

on aurait alors  $j = i + p^* = p^* + q^* = n$ , ce qui est exclus. Par conséquent,  $(ip \bmod n) + 1 = (ip + 1) \bmod n$ . De même,  $(i - 1)p \bmod n \neq n - 1$ , sinon  $i - 1 = q^*$  impliquerait  $j = i + p^* = n + 1$ , ce qui est aussi exclus. On a donc encore  $((i - 1)p \bmod n) + 1 = ((i - 1)p + 1) \bmod n$ . Finalement,  $jp \bmod n = (ip \bmod n) + 1 > ((i - 1)p \bmod n) + 1 = (j - 1)p \bmod n$ , et donc  $x_j = x$ .

- Inversement, supposons que  $x_j = x$ . Alors,  $jp \bmod n > (j - 1)p \bmod n$ . Puisque  $j$  et  $j - 1$  ne sont pas égaux à  $n$ , on a  $jp \bmod n$  et  $(j - 1)p \bmod n$  non nuls. Par conséquent,  $(jp \bmod n) - 1 = (jp - 1) \bmod n$  et  $((j - 1)p \bmod n) - 1 = ((j - 1)p - 1) \bmod n$ . On en déduit que  $ip \bmod n = (jp \bmod n) - 1 > ((j - 1)p \bmod n) - 1 = (i - 1)p \bmod n$ , et donc  $x_i = x$ .

■

Notons que l'on a montré au cours de la preuve la propriété suivante : soit  $W$  le mot de Christoffel de pente  $\frac{p}{q}$  sur l'alphabet  $\{x < y\}$ ,  $W'$  le mot de Christoffel de pente  $\frac{q}{p}$  sur l'alphabet  $\{x < y\}$ . On obtient  $W'$  à partir de  $W$  en échangeant  $x$  et  $y$  dans  $W$  puis en prenant son image miroir.

**Théorème 3.** Soit  $W$  un mot de Christoffel propre sur l'alphabet  $\{x < y\}$ . On a alors  $W = xUy$ , où  $U$  est un palindrome.

*Preuve* On a  $w_1 = x$  car  $0 \in \{0, 1, \dots, q - 1\}$  et  $w_n = y$  car  $(n - 1)p \equiv -p \equiv q \bmod n$  et  $q \in \{q, \dots, n - 1\}$ .

Montrons que  $U$  est un palindrome sur le graphe de Cayley de  $W$ . On considère l'application définie sur le graphe qui envoie le sommet 0 sur lui-même, et  $k \mapsto n - k$  si  $k \in \{1, \dots, n - 1\}$ , et qui renverse l'orientation des flèches. Cette application laisse le graphe invariant sauf les étiquettes des deux flèches qui font intervenir le sommet 0. En effet, si  $i, j \neq 0$  et  $i \xrightarrow{x} j$ , alors  $i < j$ , et on a  $n - i \xleftarrow{x} n - j$ , et de même pour l'étiquette  $y$ . ■

### 3.3 Mot central

**Définition 10** (Mot central). Un mot est  $W$  est dit *mot central* si et seulement s'il admet deux périodes  $p$  et  $q$  avec  $\text{pgcd}(p, q) = 1$  et  $|W| = p + q - 2$ .

**Remarque 2.** Soit  $W$  mot de Christoffel de pente  $\frac{p}{q}$  sur l'alphabet  $\{x < y\}$ . Soit  $U$  tel que  $W = xUy$  (voir le théorème 2). Le mot  $U$  est un mot central de périodes  $p^*$  et  $q^*$ .

**Théorème 4.** *Un mot central est défini sur un alphabet à au plus deux lettres. Soient  $p, q$  deux entiers naturels premiers entre eux. On suppose  $p \geq 2$  et  $q \geq 2$ . Il existe exactement deux mots centraux ayant pour périodes  $p$  et  $q$ . Ces deux mots se correspondent à échange des lettres près.*

*Preuve*

On considère le graphe de Cayley du mot de Christoffel de pente  $\frac{p}{q}$  sur l'alphabet  $\{x < y\}$ . On supprime le sommet  $n-1 = p+q-1$ . On a deux classes de sommets dans le graphe en scindant l'orbite de 0 sous l'action de  $\rho$  en deux

$$0 \rightarrow k \rightarrow \dots \rightarrow p+q-1 \rightarrow \dots \rightarrow l \rightarrow 0,$$

à savoir les antécédents de 0 et les images de 0. Il suffit d'écrire un  $a$  pour toutes les lettres du premier groupe, et un  $b$  pour les autres : on obtient un mot de longueur  $p+q-2$  qui est à la fois  $p$  et  $q$ -périodique mais sans être constant.

Réciproquement, tout mot de longueur  $p+q-2$  de périodes  $p$  et  $q$  est obtenu de cette façon au choix des lettres près. ■

On en déduit que :

**Théorème 5.** *Un mot  $U \in \{x < y\}^*$  est central si et seulement si le mot  $xUy$  ou le mot  $x\bar{U}y$  est un mot de Christoffel propre sur l'alphabet  $\{x < y\}$ , où  $\bar{U}$  est le mot obtenu en échangeant les lettres  $x$  et  $y$ .*

**Exercice 2.** Montrer qu'un mot  $W \in \{0, 1\}^*$  est central si et seulement si

$$W \in 0^* \cup 1^* \cup (P \cap P10P)$$

où  $P$  est l'ensemble des mots palindromes sur  $\{0, 1\}$ .

On peut aussi en donner la preuve suivante par récurrence (voir [6] pour plus de détails). Soit  $W$  un mot central. Supposons que  $W$  contient au moins deux lettres. Soient  $p, q$ , avec  $2 \leq p < q$  deux périodes de  $W$  avec  $\text{pgcd}(p, q) = 1$ . On a  $|W| = p+q-2$ .

Comme  $W$  a pour période  $p$ , alors il existe  $X$  de longueur  $q-2$  tel que  $X$  est préfixe et suffixe de  $W$ . De même, il existe  $Y$  de longueur  $p-2$  tel que  $Y$  est préfixe et suffixe de  $W$ . Il existe donc deux mots  $U$  et  $V$  tels que  $|U| = |V| = 2$  et

$$W = YUX = XVY.$$

Montrons que l'on ne peut avoir  $p = q - 1$ . Sinon, alors il existe deux lettres  $a$  et  $b$  telles que  $X = Ya = bY$ . Ceci implique que  $a = b$  et que  $W$  est puissance d'une même lettre, ce que nous avons exclu.

Montrons par récurrence sur  $|W|$  que



- $X, Y, W$  sont des palindromes,
- $U$  contient deux lettres différentes, disons 0 et 1,
- $V$  contient éalement 0 et 1,
- $W \in \{0, 1\}^*$ .

Si  $p = 2$ , alors  $Y$  est le mot vide. On a donc  $UX = XV$  et  $q$  impair. Donc il existe deux lettres  $a \neq b$  telles que  $U = ab, V = ba, X = (ab)^n a, W = (ab)^{n+1} a$ . L'hypothèse de récurrence est satisfaite. Nous supposons donc  $p \leq q - 2$ . On a donc  $YU$  préfixe de  $X$ . Soit  $Z$  tel que  $YUZ = X$ . Alors

$$X = YUZ = ZVY.$$

On en déduit que  $X$  a deux périodes  $p = |YU|$  et  $q - p = |UZ|$ . Comme  $\gcd(p, q - p) = 1$ , et  $|X| = q - 2 = k + (q - p) - 2$ , on peut appliquer l'hypothèse de récurrence à  $X$  : on obtient alors  $X, Y, Z$  palindromes,  $U = ab$  avec  $a \neq b$ , et  $\tilde{U} = V$ . On en déduit que  $W$  est central et que  $W$  ne contient que les lettres  $a$  et  $b$ . ■

**Exemple 3.** On considère la suite de mots  $(F_n)$  définie sur  $\{0, 1\}^*$  par

$$F_0 = 0, F_1 = 01, F_{n+1} = F_n F_{n-1}, n \geq 1.$$

Pour tout  $n$ ,  $|F_n|$  et  $|F_{n+1}|$  sont premiers entre eux. Pour  $n \geq 2$ , soit  $G_n$  le préfixe de  $F_n$  de longueur  $|F_n| - 2$ .

$$G_5 = 01001010010 = F_3 F_3 0 = F_4 010.$$

On montre que pour tout  $n \geq 4$ ,

$$G_{n+1} = F_{n-1}^2 G_{n-2}.$$

(On montre par récurrence que  $F_n G_{n+1} = F_{n+1} G_n$ .)

On a pour  $n \geq 5$ ,  $F_{n+1}$  préfixe de  $F_n^2$  (car  $F_{n+1} = F_n F_{n-1}$  et  $F_{n-1}$  préfixe de  $F_n$ ), et  $G_{n+1}$  préfixe de  $F_{n-1}^3$  (car  $G_{n+1} = F_{n-1}^2 G_{n-2}$  et  $G_{n-2}$  préfixe de  $F_{n-1}$ ).

Le mot  $G_{n+1}$  est un préfixe commun à  $F_n^2$  et  $F_{n-1}^3$  de longueur  $|F_n| + |F_{n-1}| - 2$ . Ce mot a pour périodes  $|F_n|$  et  $|F_{n-1}|$ . Il n'est pas constant.

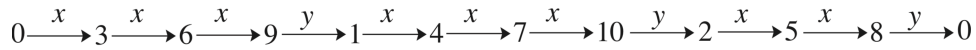
### 3.4 Mot de Christoffel dual

**Définition 11.** Soient  $p$  et  $q$  deux entiers naturels premiers entre eux, et soit  $n = p + q$ . Soient  $p^*$  et  $q^*$  dans  $\{0, 1, \dots, p + q - 1\}$  définis par  $pp^* \equiv 1 \pmod n$  et  $qq^* \equiv 1 \pmod n$ .

Le mot de Christoffel dual du mot de Christoffel sur l'alphabet  $\{x, y\}$  de pente  $p/q$  est défini comme le mot de Christoffel sur l'alphabet  $\{x, y\}$  de pente  $p^*/q^*$ .

**Remarque 3.** On a vu que  $n = p^* + q^*$  dans la preuve du Théorème 2.

**Exemple 4.** Prenons  $p = 4, q = 7$ ; on a  $p^* = 3, q^* = 8$ .



**Figure 2**

Le mot de Christoffel dual de pente  $\frac{3}{8}$  sur l'alphabet  $\{x < y\}$ ,  $p^* = 3, q^* = 8, n = 11$

D'après la preuve du théorème 4, le dual d'un mot de Christoffel se lit également sur le graphe de Cayley de ce mot de Christoffel. Par exemple, prenons le graphe de la Figure 2, et enlevons les sommets 0 et  $n - 1 = 10$ , les flèches et l'orientation. On obtient le graphe de la Figure 3.



**Figure 3**

Égalité des positions dans un mot de longueur 9 de périodes 3 et 8

Ce graphe exprime l'égalité des lettres selon leurs positions dans un mot de longueur 9 de périodes 3 et 8. Le mot central du mot de Christoffel dual de la Figure 2 est  $xyxxyxyxyx$  (les  $x$  sont en positions 3, 6, 9, 1, 4, 7 et les  $y$  en positions 2, 5, 8). On retrouve le mot de Christoffel  $xyxxyxyxyxy$  de la Figure 1.

## 4 Systèmes dynamiques symboliques

Soit  $\mathcal{A}$  un ensemble fini appelé *alphabet*. On peut faire agir sur l'ensemble des suites unidirectionnelles  $\mathcal{A}^{\mathbb{N}}$  ou bidirectionnelles  $\mathcal{A}^{\mathbb{Z}}$  l'application de *décalage*  $T$  qui à la suite  $(u_n)_n$  associe la suite  $(u_{n+1})_n$ . Travaillons par exemple avec  $\mathcal{A}^{\mathbb{N}}$ . Munissons  $\mathcal{A}^{\mathbb{N}}$  du produit des topologies discrètes. Cet ensemble est alors



l'est. Le mot  $u_3 \in \{a, b\}$  défini par

$$abaabbaaabbbaaaabbbb\dots$$

n'est pas récurrent.

**Exercice 3.** Montrer que le mot constant  $aaaaa \dots$  appartient à  $\overline{\mathcal{O}(u_3)}$ .

Construire à partir du mot  $u_3$  un mot récurrent.

**Exercice 4 (Mot de Chacon).** On considère le mot de Chacon défini comme le mot sur  $\{0, 1\}$  qui commence par la suite  $(B_n)_{n \in \mathbb{N}} \in \{0, 1\}^*$  de mots suivante :

$$B_0 = 0, \forall n \in \mathbb{N}, B_{n+1} = B_n B_n 1 B_n.$$

Montrer que le mot de Chacon est récurrent.

**Exemple 6 (Mot de Champernowne).** On considère le mot infini sur  $\{0, 1\}$  obtenu en concaténant les représentations binaires des entiers naturels.

$$1101100101110111\dots$$

**Définition 14 (Uniforme récurrence).** Soit  $u$  un mot infini. Le mot  $u$  est *uniformément récurrent* si chaque facteur apparaît infiniment souvent et les lacunes entre deux occurrences consécutives de chaque facteur sont bornées.

**Proposition 1.** Soit  $u$  un mot infini. Le mot  $u$  est uniformément récurrent si et seulement si pour tout  $n$ , il existe  $N$  tel que tout facteur de longueur  $N$  contient tous les facteurs de longueur  $n$  de la suite  $u$ .

**Définition 15 (Périodicité).** Un mot infini  $u$  est *périodique* s'il existe  $T > 0$  tel que  $\forall n \in \mathbb{N}, u_n = u_{n+T}$ . La période de  $u$  est définie comme  $\min\{T > 0, \forall n \in \mathbb{N}, u_n = u_{n+T}\}$ .

**Exercice 5.** 1. Montrer que

$$\overline{\mathcal{O}(u)} = \{x \in \mathcal{A}^{\mathbb{N}}, \mathcal{L}(x) \subset \mathcal{L}(u)\},$$

où  $\mathcal{L}(x)$  est l'ensemble des facteurs de la suite  $x$ .

2. Montrer que  $u$  est récurrente si et seulement s'il existe une suite strictement croissante  $(n_k)$  telle que

$$u = \lim_{k \rightarrow +\infty} T^{n_k} u.$$

En déduire que  $u$  est récurrente si et seulement si  $T$  est surjective sur  $\overline{\mathcal{O}(u)}$ .

## 5 Substitutions

**Définition 16** (Substitution). On appelle *substitution* ou *morphisme* un endomorphisme non effaçant pour la concaténation du monoïde libre  $\mathcal{A}^*$ , formé des mots finis sur  $\mathcal{A}$ .

**Exemple 7** (Substitution de Fibonacci). Un exemple classique est donné par la substitution définie sur  $\{a, b\}^*$  par  $\sigma(a) = ab$ ,  $\sigma(b) = a$ .

On peut étendre naturellement la définition d'une substitution à l'ensemble des suites  $\mathcal{A}^{\mathbb{N}}$ .

**Définition 17** (Point fixe). Soit  $\sigma$  une substitution. Un *point fixe* de  $\sigma$  est un mot infini  $u$  tel que  $\sigma(u) = u$ . Un *point périodique* de  $\sigma$  est un mot infini  $u$  tel qu'il existe  $k \in \mathbb{N}$  tel que  $\sigma^k(u) = u$ .

**Proposition 2.** Soit  $\sigma$  une substitution telle qu'il existe une lettre  $a$  telle que  $\sigma(a)$  commence par  $a$  et  $|\sigma(a)| \geq 2$ . Alors, la suite de mots  $a, \sigma(a), \sigma^2(a), \dots, \sigma^n(a), \dots$  converge vers un mot infini que l'on note  $\sigma^\infty(a)$ . Ce mot est un point fixe de  $\sigma$ .

*Preuve* Soit  $W$  tel que  $\sigma(a) = aW$ . On a  $\sigma^{j+1}(a) = aW\sigma(W)\sigma^2(W) \dots \sigma^j(W)$ , pour  $j \geq 0$ . Le mot  $\sigma^j(a)$  est donc un préfixe non trivial de  $\sigma^{j+1}(a)$ , ce qui donne la convergence. ■

**Exemple 8.** Le mot de Fibonacci est le point fixe commençant par  $a$  de la substitution de Fibonacci définie sur  $\{a, b\}^*$  par  $a \mapsto ab$ ,  $b \mapsto a$  :

*abaababaabaababaababaabaababaabaababaabaababaabaababaaba...*

Le mot de Thue-Morse sequence est le point fixe commençant par  $a$  de la substitution de Thue-Morse définie sur  $\{a, b\}^*$  par  $a \mapsto ab$  et  $b \mapsto ba$  :

*abbabaabbaababbabaababbaabbabaabbaababbaabbabaababbabaabbaabab...*

De nombreuses propriétés des suites substitutives sont décrites par leur *matrice d'incidence* :

**Définition 18** (Matrice d'incidence). On associe à la substitution  $\sigma$  définie sur l'alphabet  $\mathcal{A}$  la matrice  $M = [m_{i,j}]_{(i,j) \in \mathcal{A}^2}$  dont le terme général  $m_{i,j}$  compte le nombre d'occurrences de la lettre  $j$  dans  $\sigma(i)$ .

**Exemple 9.** La matrice d'incidence de la substitution de Fibonacci est  $\begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$ .

Celle de la substitution de Thue-Morse est  $\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ .

**Définition 19** (Matrice primitive). Une matrice carrée dont les entrées sont supérieures ou égales à 0 est dite *primitive* s'il existe une puissance de cette matrice dont toutes les entrées sont strictement positives.

**Définition 20** (Substitution primitive). Une substitution est dite *primitive* si la matrice associée l'est.

En d'autres termes, il existe un itéré de la substitution tel que l'image de toute lettre contient toutes les lettres de l'alphabet.

**Exemple 10.** La substitution de Fibonacci est primitive.

**Définition 21** (Système dynamique substitutif). Un *système dynamique substitutif* est un système dynamique engendré par un mot infini  $u$  point fixe d'une substitution  $\sigma$  primitive (c'est-à-dire tel que  $\sigma(u) = u$ ).

**Exercice 6.** Une matrice carrée  $M$  dont les entrées sont supérieures ou égales à 0 est dite *irréductible* si pour tout  $(i, j)$ , il existe une puissance  $k \in \mathbb{N}$  telle que le coefficient d'indice  $(i, j)$  de  $M^k$  soit strictement positif.

Donner un exemple de matrice irréductible et non primitive.

**Exercice 7.** Montrer que si  $\sigma$  est une substitution primitive alors il existe un mot infini  $u$  et un entier  $k$  tel que  $\sigma^k(u) = u$ .

**Théorème 6.** Soit  $\sigma$  une substitution primitive. Alors tout point fixe de  $\sigma$  est uniformément récurrent.

*Preuve* Soit  $u = \sigma(u)$  un point fixe de  $\sigma$ . Ce point fixe est obtenu comme  $\lim_{n \rightarrow \infty} \sigma^n(u_0)$ . Il existe  $k$  tel que toute lettre  $b$  de  $\mathcal{A}$  apparaisse dans  $\sigma(a)$ . Donc  $a$  apparaît à lacunes bornées dans  $u$ , de même que tout facteur de  $u$ . ■

**Exercice 8** (Triangle de Pascal). On considère la substitution bidimensionnelle

$$\sigma: 0 \mapsto \begin{matrix} 0 & 0 \\ 0 & 0 \end{matrix}, 1 \mapsto \begin{matrix} 1 & 1 \\ 1 & 0 \end{matrix}$$

Montrer que  $\sigma^\infty(1)$  est le triangle de Pascal réduit modulo 2.

## 6 Fonction de Complexité

### 6.1 Définition

La fonction de complexité est un outil très utile à l'étude des mots infinis et des systèmes dynamiques symboliques.

**Définition 22** (Fonction de complexité). Soit  $u = (u_n)_n$  un mot infini à valeurs dans l'alphabet fini  $\mathcal{A}$ . On appelle *fonction de complexité* de  $u$ , et l'on note  $p$ , la fonction (définie sur les entiers) qui compte le nombre de facteurs de  $u$  de longueur donnée :

$$p(n) = \text{Card}\{w; w \text{ est facteur de } u \text{ et } |w| = n\}.$$

Il est facile de voir que la fonction de complexité est croissante et que pour tout entier  $n$ , on a  $1 \leq p(n) \leq d^n$ , où  $d$  est le cardinal de l'alphabet.

Cette fonction peut être considérée comme une *mesure de la prédictibilité* du mot infini. La différence première de la fonction de complexité  $p(n+1) - p(n)$  compte le nombre d'extensions possibles dans la suite des facteurs de longueur  $n$ . Notons que la différence première de la fonction de complexité  $p(n+1) - p(n)$  est une version discrète de la dérivée : il s'agit d'un taux d'accroissement.

**Définition 23** (Prolongements). Soit  $u$  un mot infini. On appelle *prolongement à droite* (respectivement *prolongement à gauche*) d'un facteur  $W$  une lettre  $x$  telle que  $Wx$  (respectivement  $xW$ ) est un facteur de  $u$ .

Un prolongement est aussi appelé *extension*.

**Notation 3.** On note  $W^+$  le nombre de prolongements à droite du mot  $W$ .

**Théorème 7.** Soit  $u$  un mot infini. On a alors

$$p(n+1) = \sum_{W \in \mathcal{L}_n(u)} W^+,$$

et

$$p(n+1) - p(n) = \sum_{W \in \mathcal{L}_n(u)} (W^+ - 1).$$

Nous verrons au paragraphe 6.2 que la notion de facteur spécial est un outil efficace pour déterminer la différence première de la fonction de complexité. En effet, dans le cas d'une complexité basse, le nombre de facteurs spéciaux est en général relativement aisé à déterminer.

Soit  $u$  un mot infini à valeurs dans l'alphabet  $\mathcal{A}$ . Soit  $W^+$  le nombre d'extensions à droite.

**Définition 24** (Facteur spécial à gauche). Un facteur est dit *facteur spécial à droite* (respectivement *facteur spécial à gauche*) s'il admet plus d'une extension à droite (respectivement à gauche).

La fonction de complexité permet de caractériser les mots infinis périodiques. On a en effet le résultat classique suivant

**Théorème 8** (Coven-Hedlund). *Soit  $u$  un mot infini. Le mot  $u$  est ultimement périodique si et seulement si l'une des conditions suivantes équivalentes est vérifiée :*

1.  $\exists n, p(n) \leq n$
2.  $\exists C, \forall n p(n) \leq C$ .

*Preuve* Supposons qu'il existe  $n$  tel que  $p(n) \leq n$ . On suppose que le mot  $u$  n'est pas constant. On a alors  $p(1) \geq 2$ . Il existe donc  $k$  tel que  $p(k+1) = p(k)$ . Pour chaque mot  $W$  de longueur  $k$  qui apparaît dans  $u$ , il existe au moins un facteur de la forme  $Wa$ , où  $a \in \mathcal{A}$ . Comme  $p(k+1) = p(k)$ , il n'existe qu'un seul tel mot. Donc si  $u_i \cdots u_{i+k-1} = u_j \cdots u_{j+k-1} = W$ , alors  $u_{i+k} = u_{j+k}$ . Comme l'ensemble  $\mathcal{L}_k(u)$  des facteurs de  $u$  de longueur  $k$  est fini, il existe  $j > i$  tel que  $u_i \cdots u_{i+k-1} = u_j \cdots u_{j+k-1}$ , et donc  $u_{i+p} = u_{j+p}$ , pour tout  $p \in \mathbb{N}$ , et le mot est ultimement périodique. ■

La fonction de complexité permet d'exprimer simplement l'entropie topologique du système  $(\overline{\mathcal{O}(u)}, T)$ .

**Définition 25** (Entropie topologique). *L'entropie topologique  $h(u)$  d'un mot infini  $u$ , ou plus généralement du système dynamique symbolique associé  $(\overline{\mathcal{O}(u)}, T)$  est définie comme :*

$$h(u) = \lim_{n \rightarrow +\infty} \frac{\log_d(p(n))}{n},$$

où  $d$  est le cardinal de l'alphabet sur lequel le mot est défini.

Cette limite existe du fait de la sous-additivité de la fonction  $n \mapsto \log_d(p(n))$  :

$$\forall m, n, \log_d(p_u(n+m)) \leq \log_d(p_u(m)) + \log_d(p_u(n)).$$

Les mots infinis d'entropie topologique nulle sont dits *déterministes*. Les mots que nous étudierons dans ce cours sont déterministes. Nous considérons principalement deux familles de mots infinis : les mots engendrés par substitution et les mots sturmiens.

L'étude de la complexité conduit en particulier aux trois questions suivantes :



- Comment calculer la complexité d'un mot infini ?
- Quelles fonctions peuvent être réalisées comme des fonctions de complexité ?
- Peut-on caractériser des familles de mots infinis par leur complexité ? Peut-on déduire de la complexité une représentation géométrique de certains mots infinis ?

Nous allons voir au paragraphe 6.2 comment résoudre la première question en considérant les facteurs spéciaux d'un mot infini. En particulier, une question naturelle est de savoir si toutes les fonctions affines peuvent être réalisées (éventuellement ultimement) comme fonctions de complexité. La réponse est affirmative. Néanmoins, la seconde question est loin d'être résolue, surtout dans le cas de l'entropie positive. Bien que la fonction de complexité soit en général insuffisante pour décrire des mots infinis, nous allons voir que dans le cas des mots sturmiens (paragraphe 7), beaucoup d'informations peuvent se déduire de la connaissance de la fonction de complexité : les mots sturmiens sont les mots infinis de complexité  $n + 1$  indexés par  $\mathbb{N}$ .

## 6.2 Exemples

**Exercice 9.** Le mot infini de Fibonacci  $u$  est défini comme le point fixe commençant par  $a$  de la substitution suivante :  $\sigma(a) = ab$  and  $\sigma(b) = a$ .

1. Montrer que pour tout  $n$ , on a  $\sigma^{n+1}(a) = \sigma^n(a)\sigma^{n-1}(a)$ .
2. Montrer que pour tout  $n$ ,  $\sigma^n(a)$  apparaît à lacunes bornées. En déduire l'uniforme récurrence de  $u$ .
3. Montrer que le langage de  $u$  est stable par image miroir (utiliser l'exercice 3).
4. Montrer que tout facteur  $w$  du mot infini de Fibonacci peut être décomposé uniquement de la manière suivante :

$$w = r_1\sigma(v)r_2,$$

où  $v$  est un facteur du mot de Fibonacci,  $r_1 \in \{\varepsilon, b\}$ , et  $r_2 = a$ , si la dernière lettre de  $w$  est  $a$ , et  $r_2 = \varepsilon$ , sinon.

5. Montrer que si  $w$  un facteur spécial gauche non vide, alors il existe un unique facteur spécial gauche non vide  $v$  tel que  $w = \sigma(v)r_2$ , où  $r_2 = a$ , si la dernière lettre de  $W$  est  $a$ , et  $s = \varepsilon$ , sinon. En déduire la forme générale des facteurs spéciaux gauche.
6. Montrer que le mot infini de Fibonacci n'est pas ultimement périodique.
7. En déduire que la fonction de complexité du mot infini de Fibonacci satisfait  $p(n) = n + 1$  pour tout  $n$ .

**Exercice 10.** Soit  $u$  la suite de Thue-Morse définie comme le point fixe commençant par 0 de la substitution suivante :  $\sigma(0) = 01$  and  $\sigma(1) = 10$ .

1. Montrer que chaque facteur  $w$  peut être décomposé comme  $w = r_1\sigma(x)r_2$ , où  $x$  est un facteur et  $r_i \in \{\varepsilon, a, b\}$ . Si  $|w| \geq 5$ , alors cette décomposition est unique.
2. Montrer que  $p(2n) = p(n) + p(n + 1)$  et que  $p(2n + 1) = 2p(n + 1)$ , pour  $n \geq 1$ . En déduire une expression de la fonction de complexité.

**Exercice 11.** Soit  $u$  point fixe d'une substitution primitive. Montrer qu'il existe  $C$  tel que

$$\forall n \in \mathbb{N}, p(n) \leq Cn.$$

## 7 Suites sturmiennes

### 7.1 Premières définitions

On a vu que si la complexité d'une suite est telle qu'il existe un entier  $n$  pour lequel  $p(n) \leq n$ , alors cette suite  $u$  est périodique. Il apparaît alors naturel de s'intéresser aux suites de complexité  $n + 1$ , c'est-à-dire telles que  $p(n) = n + 1$ , pour tout  $n$ . De telles suites existent sur  $\mathbb{N}$ , par exemple, la suite de Fibonacci (voir l'exercice 9), et sur  $\mathbb{Z}$ ; la suite suivante a pour complexité  $n + 1$

...000010000...

**Définition 26.** Les suites de complexité  $n + 1$  indexées par  $\mathbb{N}$  sont appelées *suites sturmiennes*.

Les suites sturmiennes sont donc les suites de complexité minimale parmi les suites non ultimement périodiques. Cette définition implique que les suites sturmiennes sont définies sur un alphabet à deux lettres ( $p(1) = 2$ ).

**Théorème 9.** *Montrer que pour toute suite sturmiennne, chaque préfixe apparaît au moins deux fois dans la suite. Toute suite sturmiennne est récurrente.*

*Preuve* Raisonnons par l'absurde et supposons que  $u$  soit une suite sturmiennne non récurrente. Il existe un mot  $U$  qui apparaît un nombre fini de fois. Soit  $n$  sa longueur. Il existe  $N$  tel que le facteur  $U$  n'apparaisse plus après l'indice  $N$ . On considère la suite  $v = T^N(u)$ , où  $T$  désigne le décalage. La suite  $v$  a au plus  $n$  facteurs de longueur  $n$ . D'après le théorème 8, la suite  $v$  est ultimement périodique, et donc la suite  $u$  aussi, ce qui conduit à une contradiction. ■

**Proposition 3.** *Si  $u$  est une suite sturmiennne, alors un et un seul des mots 00 ou 11 apparaît dans  $u$ .*

*Preuve* On a  $p_u(2) = 3$ . Chaque mot apparaît infiniment souvent, en particulier les lettres 0 et 1, donc 01 et 10 aussi. ■

En conséquence, on distingue deux types de mots sturmiens.

**Définition 27.** Une suite sturmiennne est dite de *type* 0 (resp. 1) si la lettre 1 (resp. 0) est isolée, c'est-à-dire si 11 (resp. 00) n'apparaît pas.

L'exemple le plus classique de suite sturmiennne est la suite de Fibonacci, point fixe de la substitution  $\sigma$  définie par  $\sigma(a) = ab$  et  $\sigma(b) = a$  (voir exercice 9). Les suites sturmiennes sont donc définies de manière purement combinatoire, mais ce qui est remarquable, c'est qu'elles peuvent être également représentées de manière géométrique : les suites sturmiennes sont exactement les suites obtenues en codant l'orbite d'un point  $\rho$  du cercle unité sous la rotation d'angle irrationnel  $\alpha$ , par rapport à des intervalles complémentaires du cercle unité de longueurs  $\alpha$  et  $1 - \alpha$ .

*Dans tout ce qui suit  $R_\alpha$  désigne la rotation définie sur le tore  $\mathbb{T}_1 = \mathbb{R}/\mathbb{Z}$  de dimension 1, par  $R_\alpha x = x + \alpha$  modulo 1. S'il n'y a pas d'ambiguïté, nous omettrons de préciser que nous travaillons modulo 1.*

**Théorème 10** (Morse-Hedlund). *Soit  $u$  une suite sturmiennne à valeurs dans  $\{0, 1\}$ . Il existe alors  $\alpha$  irrationnel dans  $]0, 1[$  et  $\rho \in \mathbb{R}$  tels que l'on ait soit*

$$\forall n, (u_n = 0 \iff R_\alpha^n(\rho) = \rho + n\alpha \in [0, 1 - \alpha]),$$

soit

$$\forall n, (u_n = 0 \iff R_\alpha^n(\rho) = \rho + n\alpha \in ]0, 1 - \alpha]).$$

Pour une preuve de ce théorème, voir [6] ou [8].

Le lien avec la représentation par droites discrètes se déduit de la remarque suivante : on suppose que  $\forall n, (u_n = 0 \iff R_\alpha^n(\rho) = \rho + n\alpha \in [0, 1 - \alpha])$ ; On vérifie alors que  $u_n = 0$  si et seulement si  $\lfloor (n+1)\alpha + \rho \rfloor = \lfloor n\alpha + \rho \rfloor$ . De même,  $u_n = 1$  si et seulement si  $\lfloor (n+1)\alpha + \rho \rfloor = \lfloor n\alpha + \rho \rfloor + 1$ . De manière analogue, si l'on suppose que  $\forall n, (u_n = 0 \iff R_\alpha^n(\rho) = \rho + n\alpha \in ]0, 1 - \alpha])$ , on vérifie alors que  $u_n = 0$  si et seulement si  $\lceil (n+1)\alpha + \rho \rceil = \lceil n\alpha + \rho \rceil$ , et de même,  $u_n = 1$  si et seulement si  $\lceil (n+1)\alpha + \rho \rceil = \lceil n\alpha + \rho \rceil + 1$ . Une suite sturmiennne code donc la ligne brisée reliant les entiers les plus proches de la forme  $\lfloor n\alpha + \rho \rfloor$  (resp.  $\lceil n\alpha + \rho \rceil$ ), pour  $n \in \mathbb{N}$ , de la droite d'équation  $y = \alpha x + \rho$ , en codant les pas horizontaux par des 0 et les pas diagonaux par des 1.

On appelle *angle* d'une suite sturmiennne le réel  $\alpha$  qui lui est ainsi associé.

Un des intérêts de la représentation géométrique des suites sturmiennes indiquée dans le théorème 10, est qu'elle fournit une description simple, en termes d'intervalles du cercle unité, des facteurs de longueur donnée.

On rappelle les propriétés suivantes de la suite  $(n\alpha)_{n \in \mathbb{N}}$ .

**Théorème 11.** *Soit  $\alpha \notin \mathbb{Q}$  et soit  $\rho \in \mathbb{R}$ . La suite  $(n\alpha + \rho) \bmod 1$  est dense : pour tout intervalle  $I$  non vide de  $\mathbb{R}/\mathbb{Z}$ , il existe  $n \in \mathbb{N}$  tel que  $n\alpha + \rho \in I$ .*

*De plus la suite  $(n\alpha + \rho) \bmod 1$  est uniformément distribuée : pour tout intervalle  $I$  non vide de  $\mathbb{R}/\mathbb{Z}$ ,*

$$\lim_{N \rightarrow +\infty} \frac{1}{N} \text{Card}\{n \in \mathbb{N}, n\alpha + \rho \in I\} = |I|,$$

où  $|I|$  désigne la mesure de  $I$ .

**Lemme 1.** *Soit  $u$  une suite sturmiennne d'angle  $\alpha$ . Il existe une bijection entre les facteurs de longueur  $n$  de la suite  $u$  et les intervalles de la partition du cercle unité par les points  $-k\alpha, 0 \leq k \leq n$ .*

*Preuve* Supposons que  $u$  code l'orbite du point  $\rho$  par rapport à  $I_0 = [0, 1 - \alpha[$  et  $I_1 = [1 - \alpha, 1[$ . Un mot fini  $w_1 \cdots w_n$  défini sur l'alphabet  $\{0, 1\}$  est un facteur de la suite  $u$  si et seulement s'il existe un entier  $k$  tel que

$$\rho + k\alpha \in I(w_1, \dots, w_n) = \bigcap_{j=0}^{n-1} R^{-j}(I_{w_{j+1}}).$$

Comme  $\alpha$  est irrationnel, la suite  $(\rho + n\alpha)_{n \in \mathbb{N}}$  est dense, ce qui implique que  $w_1 w_2 \dots w_n$  est un facteur de  $u$  si et seulement si  $I(w_1, \dots, w_n) \neq \emptyset$ . ■

En particulier, l'ensemble des facteurs ne dépend pas du point initial  $\rho$ . On vérifie de plus que les ensembles  $I(w_1, \dots, w_n)$  sont connexes et bornés par les points  $-k\alpha \bmod 1$ , pour  $0 \leq k \leq n - 1$ . Il y a  $n + 1$  tels intervalles ( $\alpha$  est irrationnel) et alors  $n + 1$  facteurs de longueur  $n$  : la suite  $u$  est bien sturmiennne.

## 7.2 Équilibre

Notons que les suites sturmiennes ont de nombreuses autres caractérisations, tant géométriques que combinatoires. En particulier, les suites sturmiennes sont les suites équilibrées sur un alphabet à deux lettres qui sont non ultimement périodiques. Une suite *équilibrée* est telle que la différence entre le nombre d'occurrences d'une lettre dans deux de ses facteurs de même longueur, est bornée par 1 en valeur absolue (voir Définition 29).

**Définition 28** (Équilibre). Un mot infini  $u \in \{0, 1\}^{\mathbb{N}}$  est dit *équilibré* si pour tous facteurs  $U, V$  de même longueur de  $u$ , on a

$$||U|_1 - |V|_1| \leq 1.$$

Un langage  $L$  sur l'alphabet  $\{0, 1\}$  est dit *équilibré* si pour tous mots  $U, V$  de même longueur de  $L$ , on a

$$||U|_1 - |V|_1| \leq 1.$$

Notons qu'il est très facile de construire des suites sur  $\{0, 1\}$  2-équilibrées, c'est-à-dire telle que pour tous facteurs  $U, V$  de même longueur de  $u$ , on a

$$||U|_1 - |V|_1| \leq 2.$$

Il suffit de prendre l'image par la substitution

$$0 \mapsto 01, 1 \mapsto 10$$

de n'importe quelle suite sur l'alphabet  $\{0, 1\}$ .

### 7.3 Facteurs spéciaux

Soit  $u$  une suite à valeurs dans l'alphabet  $\mathcal{A}$ . On rappelle que  $W^+$  (resp.  $W^-$ ) désigne le nombre de prolongements à droite (resp. à gauche) de  $W$  dans  $u$ . Rappelons que

$$p(n+1) - p(n) = \sum_{W \in \mathcal{L}_n(u)} (W^+ - 1);$$

et qu'un facteur est dit *facteur spécial à droite* (respectivement *facteur spécial à gauche*) s'il admet plus d'une extension à droite (respectivement à gauche).

De la complexité d'une suite sturmienne on déduit que pour tout  $n$  il existe un unique facteur prolongeable à droite, noté  $D_n$  et un unique facteur prolongeable à gauche, noté  $G_n$ . En effet, on a

$$1 = p(n+1) - p(n) = 1 = \sum_{W \in \mathcal{L}_n(u)} (W^+ - 1) = \sum_{W \in \mathcal{L}_n(u)} (W^- - 1) = 1.$$

**Définition 29** (Équilibre). Un mot infini  $u \in \{0, 1\}^{\mathbb{N}}$  est dit *équilibré* si pour tous facteurs  $U, V$  de même longueur de  $u$ , on a

$$||U|_1 - |V|_1| \leq 1.$$

Un langage  $L$  sur l'alphabet  $\{0, 1\}$  est dit *équilibré* si pour tous mots  $U, V$  de même longueur de  $L$ , on a

$$||U|_1 - |V|_1| \leq 1.$$

**Définition 30.** Un langage est dit *factoriel* si pour tout mot de ce langage, tout facteur de ce mot appartient au langage.

**Théorème 12.** Soit  $L$  un langage factoriel et équilibré. Alors pour tout  $n$ , il existe au plus  $n + 1$  facteurs de longueur  $n$  dans  $L$ .

Voir par exemple [6] pour une preuve.

**Proposition 4.** Le langage d'une suite sturmienne est stable par image miroir. En particulier, on a pour tout  $n$ ,  $G_n = \tilde{D}_n$ .

*Preuve* Soit  $u$  une suite sturmienne. Soit  $\mathcal{L}(u)$  l'ensemble des facteurs de la suite  $u$  et  $\tilde{\mathcal{L}}(u)$  l'ensemble des images miroir des facteurs de la suite  $u$ . On a  $\mathcal{L}(\tilde{u}) \cup \tilde{\mathcal{L}}(u)$  est un langage factoriel équilibré. Donc pour tout  $n$ , il y a au plus  $n + 1$  facteurs de longueur  $n$  dans  $\mathcal{L}(\tilde{u}) \cup \tilde{\mathcal{L}}(u)$ . ■

**Proposition 5.** Soit  $u$  une suite sturmiennne d'angle  $\alpha$ . Pour tout  $n$ ,  $D_n$  est un suffixe de  $D_{n+1}$  et  $G_n$  est un préfixe de  $G_{n+1}$ . Pour tout  $n$ , on a  $G_n$  est égal au préfixe de longueur  $n$  de la suite sturmiennne, appelée suite caractéristique définie par

$$\forall n \in \mathbb{N}, u_n = 0 \text{ si et seulement si } n\alpha + \alpha \in [0, 1 - \alpha[.$$

**Exercice 12.** Montrer que deux suites sturmiennnes ayant le même facteur spécial droite de longueur  $n - 1$  ont les mêmes facteurs de longueur  $n$ .

*Preuve* On montre ce lemme par récurrence. On vérifie qu'il est vrai pour  $m = 2$ . Supposons que deux suites sturmiennnes ayant le même facteur expansif de longueur  $m - 1$  ont les mêmes facteurs de longueur  $m$ . Considérons alors deux suites sturmiennnes ayant le même facteur expansif  $D_m$  de longueur  $m$  et par conséquent le même facteur biprolongeable à gauche  $G_m$  (on a  $\widetilde{G}_m = D_m$ ). En particulier, par hypothèse de récurrence, ces deux suites ont mêmes facteurs de longueur  $m$ , car elles ont le même facteur expansif de longueur  $m - 1$ . Montrons que les facteurs de longueur  $m$  ont les mêmes extensions dans les deux suites.

Supposons que  $G_{m-1} \neq D_{m-1}$ . Le facteur  $D_m$  a pour extensions  $a$  et  $b$ , dans les deux suites. Les facteurs de longueur  $m$  différents de  $D_m$  ont une unique extension droite. Or le suffixe de longueur  $m - 1$  d'un facteur de longueur  $m$  différent de  $D_m$  est différent de  $D_{m-1}$ , car  $G_{m-1} \neq D_{m-1}$ ; on conclut alors en notant que ce suffixe a donc une unique extension droite, qui est la même dans les deux suites, par hypothèse de récurrence.

Supposons maintenant que  $G_{m-1} = D_{m-1}$ . On note  $D_m = xD_{m-1}$ . On a :  $G_m = G_{m-1}x$ . Notons de plus  $\bar{x} = a$ , si  $x = b$  et  $\bar{x} = b$ , si  $x = a$ . Le facteur  $xD_{m-1}$  a pour extensions droites  $a$  et  $b$ , dans les deux suites, par définition. De même, le facteur  $D_{m-1}x$  a pour extensions gauches  $a$  et  $b$ . Par conséquent, le facteur  $\bar{x}D_{m-1}$  a pour unique extension droite  $x$ , dans les deux suites. Le raisonnement est le même que précédemment pour les facteurs de longueur  $m$  restants. ■

## 7.4 Graphe des mots

Un outil très utile pour l'étude des suite sturmiennnes (et en particulier pour l'étude des fréquences des facteurs) est le *graphe des mots* de Rauzy. Soit  $u$  une suite définie sur l'alphabet fini  $\mathcal{A}$  (de cardinal  $d$ ). Le graphe des mots  $\Gamma_n$  des facteurs

de longueur  $n$  de la suite  $u$  est un graphe orienté qui est un sous-graphe du graphe des mots de de Bruijn<sup>1</sup>.

Le graphe  $\Gamma_n$  a pour sommets les facteurs de longueur  $n$  de la suite, avec une arête de  $U$  vers  $V$ , s'il existe un mot  $W$  de longueur  $n - 1$  tel que :

$$U = xW \text{ et } V = Wy, \text{ avec } x, y \in \mathcal{A},$$

et tel que  $xWy$  soit un facteur de la suite.

**Proposition 6.** *Soit  $u \in \mathcal{A}^{\mathbb{N}}$ . Les propriétés suivantes sont équivalentes :*

1. *Pour tout  $n$  le graphe des mots  $\Gamma_n$  d'ordre  $n$  associé à  $u$  est fortement connexe.*
2. *Tous les préfixes de  $u$  apparaissent au moins deux fois.*
3. *La suite  $u$  est récurrente.*

*Preuve* Montrons que 1) implique 2). Soit  $W$  préfixe de longueur  $n$  de la suite  $u$ . Comme  $\Gamma_n$  est fortement connexe, il existe une flèche qui arrive en  $W$ , donc  $W$  admet une autre occurrence que celle où  $W$  est préfixe.

On sait déjà que 2) implique 3).

Montrons que 3) implique 1). Soient  $V, W$  deux facteurs de longueur  $n$  de  $u$ . Il existe une occurrence de  $W$  qui apparaît après celle de  $V$ , donc il existe un chemin issu du sommet  $W$  qui arrive en  $V$ . De même, il existe une occurrence de  $V$  qui apparaît après celle de  $W$ , donc il existe un chemin issu du sommet  $V$  qui arrive en  $W$ . ■

Soit  $U$  un sommet de  $\Gamma_n$ . On note  $U^+$  le nombre d'arêtes de  $\Gamma_n$  d'origine  $U$  et  $U^-$  le nombre d'arêtes d'extrémité  $U$ .

On appelle *branche* un chemin de longueur maximale  $U_1 \rightarrow U_2 \cdots \rightarrow U_n$  tel que  $U_i^{-1} = 1$  pour  $i \geq 2$ , et  $U_i^+ = 1$  pour  $i < n$ .

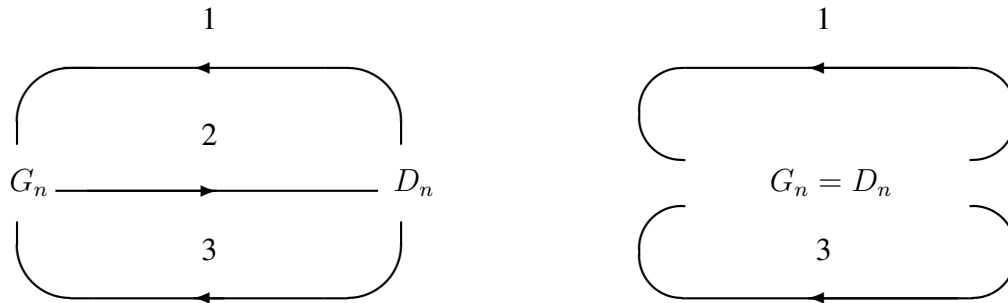
Supposons le graphe  $\Gamma_n$  fortement connexe et la suite non périodique, alors les extrémités d'une branche sont des facteurs spéciaux.

---

<sup>1</sup>Le graphe des mots de de Bruijn correspond au graphe des mots d'une suite de complexité maximale ( $\forall n, p(n) = d^n$ ) et a été introduit par de Bruijn dans le but de construire des suites finies circulaires de longueur  $d^n$  à valeurs dans  $\{0, 1, \dots, d - 1\}$ , telles que tout facteur de longueur  $n$  apparaît une fois et une seule : une telle suite correspond à un chemin Hamiltonien fermé dans le graphe de de Bruijn.



Une suite sturmienne présente deux types de graphes selon que  $G_n = D_n$  ou que  $G_n \neq D_n$ .



**Exercice 13.** Montrer qu'une suite sturmienne est uniformément récurrente.

## 7.5 Graphes des mots et fréquences

Cette forme simple du graphe des mots permet de déduire des informations sur les fréquences des facteurs des suites sturmiennes.

**Définition 31.** On appelle *fréquence* du facteur  $U$  dans le mot infini  $u$  la limite suivante si elle existe : de

$$\lim_{N \rightarrow +\infty} \frac{1}{N} \text{Card}\{k \in \mathbb{N}, U \text{ apparaît à l'indice } k \text{ dans } u\}.$$

D'après le théorème 11 et le lemme 1, on déduit l'existence des fréquences dans tout mot sturmien.

**Lemme 2.** *Supposons que  $U \rightarrow V$  et que  $U^+ = 1 = V^- = 1$ , alors les facteurs  $U$  et  $V$  ont même fréquence*

*Preuve* En effet, écrivons  $U = xW$  et  $V = Wy$ , où  $x$  et  $y$  sont des lettres. Comme  $U^+ = 1$ , le facteur  $U$  a pour unique extension droite  $y$ ; de même, le

facteur  $V$  a pour unique extension gauche  $x$ . Par conséquent, nous avons les égalités suivantes entre les fréquences :

$$f(U) = f(Uy) = f(xWy) = f(xV) = f(V).$$

■

On déduit alors du lemme précédent que les mots d'une même branche ont même fréquence. On associera donc à une branche la fréquence des mots de cette branche.

Revenons au cas sturmien.

Par branche (1) ou (3), représentées sur la figure ci-dessus, on entend tous les mots de ce chemin,  $D_n$  et  $G_n$  exclus. En revanche,  $D_n$  et  $G_n$  seront inclus dans la branche (2).

On déduit alors de ce lemme que tous les mots de la branche (1) (voir la figure ci-dessous), privé de  $D_n$  et  $G_n$  ont même fréquence, que, de même, tous les mots de la branche (3), privé de  $D_n$  et  $G_n$  ont même fréquence et enfin, que tous les mots de la branche (2),  $D_n$  et  $G_n$  inclus, ont même fréquence. On en déduit donc que les fréquences des facteurs de même longueur d'une suite sturmiennne prennent au plus 3 valeurs. De plus, si  $G_{n-1} = D_{n-1}$  alors l'une des deux branches (1) ou (3) est vide, c'est-à-dire que l'on a une arête de  $D_n$  vers  $G_n$ . On en déduit donc la proposition suivante.

**Proposition 7.** *Les fréquences des facteurs de même longueur d'une suite sturmiennne prennent au plus 3 valeurs. Si  $G_{n-1} = D_{n-1}$ , les fréquences des facteurs de longueur  $n$  prennent au plus 2 valeurs.*

On en déduit ainsi que du lemme 1 le théorème suivant connu sous le nom de théorème des trois longueurs :

**Théorème 13.** *Soit  $\alpha$  fixé. Les points  $k\alpha$  modulo 1, pour  $0 \leq k \leq N$ , divisent l'intervalle  $[0, 1]$  en intervalles dont les longueurs prennent trois valeurs au plus, l'une étant la somme des deux autres.*

## 7.6 Substitutions sturmiennes

**Définition 32** (substitution inversible). On peut étendre la définition d'une substitution définie sur un alphabet à  $d$  lettres au groupe libre  $F_d$  en posant

$$\sigma(s^{-1}) = (\sigma(s))^{-1}.$$

Une substitution est dite *inversible* s'il existe un morphisme  $\nu: F_d \rightarrow F_d$  tel que

$$\nu\sigma(a) = \sigma\nu(a),$$

pour toute lettre  $a \in F_d$ .

**Exemple 11.** La substitution de Fibonacci est inversible.

**Définition 33** (Automorphisme positif). Un automorphisme est dit positif si l'image de toute lettre à une puissance positive ne contient aucune puissance négative.

**Théorème 14** (Substitution sturmienne). *Soit  $\sigma$  une substitution sur un alphabet à deux lettres. Les propriétés suivantes sont équivalentes :*

- *il existe un mot sturmien  $u$  tel que  $\sigma(u)$  est sturmien*
- *l'image de tout mot sturmien par  $\sigma$  est sturmien.*

*On a alors que tout point fixe d'une substitution sturmienne a pour complexité  $n + 1$ .*

**Théorème 15** (Wen-Wen). *Les automorphismes positifs sur le groupe libre sur deux lettres  $F_2$  sont exactement les substitutions sturmiennes.*

On en déduit que si une substitution sur un alphabet à deux lettres est inversible, alors tout point fixe de cette substitution (ou de son carré) a pour complexité  $n + 1$ , pour tout  $n$ .

**Définition 34** (Substitutions conjuguées). Deux mots  $X, Y$  sont dits *conjugués* s'il existe des mots  $V, W$  tels que

$$X = VW, Y = WV.$$

Soient  $\sigma$  et  $\mu$  deux substitutions. La substitution  $\mu$  est un conjugué à droite de  $\sigma$  s'il existe un mot  $W$  tel que

$$\sigma(a)W = W\mu(a), \text{ pour toute lettre } a \in \mathcal{A}.$$

En particulier, les mots  $\sigma(a)$  et  $\mu(a)$  sont conjugués pour toute lettre  $a$ .

Si  $W$  est non vide dans

$$\sigma(a)W = W\mu(a), \text{ pour toute lettre } a \in \mathcal{A},$$

alors les mots  $\sigma(a)$  pour  $a \in \mathcal{A}$  commencent par la même lettre.

**Exemple 12.** Soit  $\sigma$  la substitution définie par

$$\sigma(0) = 01010, \sigma(1) = 01.$$

Les substitutions suivantes sont des conjuguées de  $\sigma$  :

$$\sigma_0(0) = 01010, \sigma_0(1) = 01,$$

$$\sigma_1(0) = 10100, \sigma_1(1) = 10,$$

$$\sigma_2(0) = 01001, \sigma_2(1) = 01,$$

$$\sigma_3(0) = 10010, \sigma_3(1) = 10,$$

$$\sigma_4(0) = 00101, \sigma_4(1) = 01,$$

$$\sigma_5(0) = 01010, \sigma_5(1) = 10.$$

Toutes ces substitutions sont sturmiennes.

**Théorème 16.** Soit  $\sigma$  une substitution sturmiennne sur un alphabet à deux lettres. Soit  $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$  sa matrice d'incidence. Il existe exactement  $a + b + c + d - 1$  substitutions sturmiennes de matrice d'incidence  $M$ . Ces substitutions sont toutes conjuguées.

## References

- [1] J.-P. Allouche, J. Shallit, *Automatic sequences: Theory, Applications, Generalizations*, Cambridge University Press, Cambridge (2003).
- [2] J. Berstel, J. Karhumäki, *Combinatorics on words— A tutorial*, <http://www-igm.univ-mlv.fr/~berstel/>.
- [3] V. Berthé, A. de Luca, C. Reutenauer, *On an involution of Christoffel words and Sturmian morphisms*, <http://www.lirmm.fr/~berthe/prepublis.html>.
- [4] R. Graham, D. Knuth, O. Patashnik, *Concrete mathematics: a foundation for computer science*, Addison Wesley, 1994.
- [5] M. Lothaire, *Combinatorics on words*, Cambridge University Press, Cambridge, (2002), <http://www-igm.univ-mlv.fr/~berstel/Lothaire>.

- [6] M. Lothaire, *Algebraic Combinatorics on words*, Cambridge University Press, Cambridge, (2002), <http://www-igm.univ-mlv.fr/~berstel/Lothaire>.
- [7] M. Lothaire, *Applied Combinatorics on words*, Cambridge University Press, Cambridge, (2002), <http://www-igm.univ-mlv.fr/~berstel/Lothaire>.
- [8] N. Pytheas Fogg, *Substitutions in Dynamics, Arithmetics and Combinatorics*, Lect. Notes in Math. **1794**, Springer-Verlag, Berlin, (2002), <http://www.lirmm.fr/~berthe/Fogg.html>.