SUITES DE STURM, COMPLEXITÉ ET DIMENSION DE VAPNIK-CHERVONENKIS

MARIE-PAULE MULLER

RÉSUMÉ. La complexité d'une suite (à valeurs dans un alphabet fini) dénombre tous les facteurs de longueur donnée qui figurent dans cette suite. Pour une suite de Sturm, nous donnons ici une estimation du nombre de mots vus par une fenêtre qui est un ensemble fini quelconque, donc plus générale qu'un intervalle d'entiers consécutifs. Nous en déduisons que la dimension de Vapnik-Chervonenkis d'une suite de Sturm est égale à deux.

ABSTRACT.The complexity of a sequence (with values in a finite alphabet) counts the factors of given length in the sequence. For a Sturmian sequence, we give here an estimation of the number of words which are seen through a window which is an arbitrary finite set, therefore more general than an interval of consecutive integers. As a corollary, we prove that the VC-dimension of a Sturmian sequence is 2.

1. INTRODUCTION.

Parmi les suites à valeurs dans un alphabet fini de symboles, si l'on excepte les suites ultimement périodiques, les suites de Sturm sont celles qui sont "les plus régulières" possible. Elles se caractérisent par des propriétés remarquables qui se formulent dans des cadres divers : les systèmes dynamiques, comme M. Morse et G.A. Hedlund l'ont établi dès 1940 [7] ; la combinatoire ; l'arithmétique ; la théorie des graphes ; le calcul différentiel : des suites de Sturm apparaissent par exemple dans la séparation des zéros des solutions d'équations différentielles (de la forme y"+f(x).y=0 où f est périodique [7] p.40 et [6] p.862), dans la description des géodésiques du tore plat [7] et donc (via une transformation simple) des trajectoires de billard, dans les minimaux des systèmes dynamiques discrets, dans la représentation pixellisée des droites en informatique. Certaines suites de Sturm sont aussi des points

Key words and phrases. Mathematics Subject Classification (2000) : 68R15 Key Words : Sturmian sequence, mechanical sequence, complexity, VC-dimension, combinatorics on words.

fixes de morphismes sturmiens, que l'on peut voir comme des automorphismes positifs du groupe libre à deux générateurs ; des applications récentes concernent le groupe de tresses [5]. Les différentes caractérisations usuelles des suites de Sturm sont exposées par J. Berstel et P. Séébold dans [2], accompagnées des références bibliographiques de nombreux auteurs et résultats ; des exemples de suites binaires sont décrits dans [1]. D'autre part, P. Kurka [4] présente une étude détaillée des systèmes dynamiques symboliques.

La fonction complexité P d'une suite $x = (x_n)_{n\geq 0}$ à valeurs dans un alphabet fini dénombre les facteurs de longueur n de la suite : $P(n) = |\{x_i x_{i+1}...x_{i+n-1} \ : \ i \geq 0\}|$. Cette fonction, particulièrement intéressante pour les suites récurrentes, est liée aux propriétés arithmétiques de la densité de chacun des symboles dans la suite et est étudiée par de nombreux auteurs [8], [3]. Les suites de Sturm sont les suites pour lesquelles la complexité est $P(n) = n + 1$.

Une notion différente de complexité est celle définie par la dimension de Vapnik-Chervonenkis [9]. Elle est un indicateur de la complexité d'une famille de parties d'un ensemble donné : on recherche un plus grand sous-ensemble sur lequel cette famille induit tous les sous-ensembles possibles.

Nous déterminons la dimension de Vapnik-Chervonenkis pour les suites de Sturm, en abordant l'étude du nombre de séquences différentes que forme une suite de Sturm défilant par translation derrière une fenêtre qui n'est plus nécessairement un intervalle.

2. Définitions et rappels

2.1. **Dimension de Vapnik-Chervonenkis.** Soit X un ensemble, et $\mathcal{F}$ une famille de parties de X . Pour tout sous-ensemble fini $S \subseteq X$, on considère la famille de parties de S : $S \cap \mathcal{F} = \{S \cap F \ : \ F \in \mathcal{F}\}$. La *dimension de Vapnik-Chervonenkis* de $\mathcal{F}$ est la taille maximale d'un ensemble S *éparpillé* par $\mathcal{F}$, c'est-à-dire tel que toutes ses parties soient obtenues dans la famille induite :

$$VCdim(\mathcal{F}) = sup\{|S| \ : \ |S \cap \mathcal{F}| = 2^{|S|}\}$$

Une construction récursive permet de prouver que

$$|S \cap \mathcal{F}| \leq |\{T \subseteq S \ : \ |T \cap \mathcal{F}| = 2^{|T|}\}|$$

et donc que $|S \cap \mathcal{F}| \leq |\{T \subseteq S \ : \ |T| \leq VCdim(\mathcal{F})\}|$.

Une suite binaire $x : \mathbb{N} \to \{0, 1\}$ est l'indicatrice d'un sous-ensemble. La famille de parties de $\mathbb{N}$ que nous considérons correspond à la famille des sections finissantes de x , c'est-à-dire des suites translatées $T_k x$ $(k \in \mathbb{N})$, où $(T_k x)_n = x_{n+k}$. Pour un sous-ensemble du type particulier intervalle $S = \{0, \ldots, s-1\}$, on a donc $|S \cap \mathcal{F}| = P(s)$. Dans le cas

général, S est une famille finie d'entiers $n_1 < n_2 < \ldots < n_s$ et $S \cap \mathcal{F}$ est l'ensemble des mots de la forme $S \cap T_k x = x_{k+n_1} x_{k+n_2} \ldots x_{k+n_s}$ $(k \in \mathbb{N})$.

2.2. **Les suites de Sturm et les suites mécaniques.** La complexité P d'une suite apériodique (c'est-à-dire non ultimement périodique) est strictement croissante, et donc $P(n) \geq n+1$. Une *suite de Sturm* est une suite apériodique dont la complexité est minimale : $P(n) = n+1$.

Une suite de Sturm $x = (x_n)_{n \geq 0}$ est donc écrite avec deux symboles, disons 0 et 1 , et elle est nécessairement récurrente. Par convention, la *densité* de la suite sera celle du symbole 1 :

$$\alpha = lim_{n \to \infty} \frac{1}{n}(x_0 + x_1 + \ldots + x_{n-1})$$

Dans le cas d'une suite de Sturm, cette limite existe et est irrationnelle.

Parmi les nombreuses caractérisations connues, nous rappelons que les suites de Sturm s'identifient aux suites mécaniques à densité irrationnelle, que nous présentons brièvement.

On considère deux paramètres réels $0 < \alpha < 1$ et $0 \leq \beta < 1$. Pour $n \in \mathbb{N}$, le nombre $\alpha n + \beta$ est encadré par les deux entiers les plus proches $\lfloor \alpha n + \beta \rfloor$ et $\lceil \alpha n + \beta \rceil$. La *suite mécanique inférieure* (associée aux deux paramètres) est définie par

$$x_n = \lfloor \alpha(n+1) + \beta \rfloor - \lfloor \alpha n + \beta \rfloor \ \ (n \in \mathbb{N})$$

et la *suite mécanique supérieure* est définie par

$$x'_n = \lceil \alpha(n+1) + \beta \rceil - \lceil \alpha n + \beta \rceil \ \ (n \in \mathbb{N})$$

Ces deux suites, binaires, sont telles que

$$\lfloor \alpha n + \beta \rfloor = \lfloor \beta \rfloor + x_0 + x_1 + \ldots + x_{n-1} \leq \alpha n + \beta \leq$$

$$\leq \lceil \beta \rceil + x'_0 + x'_1 + \ldots + x'_{n-1} = \lceil \alpha n + \beta \rceil$$

pour $n \geq 1$.

Nous supposons dorénavant que α est irrationnel. Les deux suites mécaniques sont alors de Sturm. Si $\alpha n + \beta$ n'est jamais entier, elles sont identiques. S'il existe un entier $m \geq 0$ tel que $\alpha m + \beta \in \mathbb{N}$, elles ne diffèrent qu'aux indices $m-1$ et m (on a $x_{m-1} x_m = 10$ et $x'_{m-1} x'_m = 01$) et ont donc des sections finissantes identiques. Réciproquement, toute suite de Sturm est une suite mécanique, inférieure ou supérieure. De plus, toutes les suites mécaniques ayant même densité α ont les mêmes facteurs.

3. VC-dimension des suites de Sturm

Soit x une suite de Sturm. Une famille S d'entiers $n_1 < n_2 < \ldots < n_s$ étant donnée, on étudie le nombre de mots de la forme $S \cap T_k x = x_{k+n_1} x_{k+n_2} \ldots x_{k+n_s}$ $(k \in \mathbb{N})$.

Afin d'éviter une discussion selon le cas, on peut supposer que les 1 sont isolés dans x , c'est-à-dire que sa densité est telle que $0 < \alpha < \frac{1}{2}$ (quitte à échanger les symboles 1 et 0). Comme x est récurrente et que nous nous intéressons à la trace des $T_k x$ sur l'ensemble fini S , nous pouvons supposer aussi que x est une suite mécanique inférieure (quitte à remplacer x par une section finissante adéquate). En fait, comme toutes les suites mécaniques de même densité α ont les mêmes facteurs, en particulier les mêmes facteurs de longueur n_s, il suffit d'étudier le cas où x est la suite $(\lfloor (n+1)\alpha \rfloor - \lfloor n\alpha \rfloor)_{n \geq 0}$, (pour laquelle $\beta = 0$).

On désigne par $\mathbb{S}^1$ le cercle $\mathbb{R}/\mathbb{Z} = [0,1]/0 \sim 1$. Afin de lever l'ambiguité, un intervalle sur $\mathbb{S}^1$ sera décrit via un intervalle de $\mathbb{R}$, et l'intervalle $]t,\, t+\alpha[$, de longueur α, sera noté I_t. On rappelle que α est irrationnel.

Remarquons que $x_n = 1$ si et seulement si $(n+1)\alpha \in]0, \alpha[\subset \mathbb{S}^1$. Sur le cercle $\mathbb{S}^1$, soit $U = \{(n_i+1)\alpha \,:\, 1 \leq i \leq s\}$. Le i^{me} symbole du mot $S \cap T_k x = x_{k+n_1} x_{k+n_2} \ldots x_{k+n_s}$ est 1 si et seulement si $(n_i+1)\alpha \in]-k\alpha, (-k+1)\alpha[\subset \mathbb{S}^1$.

L'ensemble des mots $S \cap T_k x = x_{k+n_1} x_{k+n_2} \ldots x_{k+n_s}$ $(k \in \mathbb{N})$ est en bijection avec l'ensemble des parties de U de la forme $U \cap I_t$ $(-t \in \alpha\mathbb{N})$. Comme α est irrationnel, il suffit d'astreindre I_t à avoir ses extrémités dans $\mathbb{S}^1 \setminus U$ pour obtenir ces mêmes parties de U .

Les points de U sont réordonnés circulairement, $U = \{u_1,\, u_2, \ldots, u_s\}$ avec $0 < u_1 < u_2 < \ldots < u_s < 1$; parmi les composantes connexes de $\mathbb{S}^1 \setminus U$, celles qui sont de longueur strictement supérieure à α seront appelées "grands intervalles". Leur nombre p est tel que $0 \leq p < \frac{1}{\alpha}$. Les points de U sont ainsi assemblés en p sous-familles, séparées par ces grands intervalles.

L'intervalle I_t est donc déplacé par rotation sur le cercle, en étant astreint à avoir ses extrémités dans $\mathbb{S}^1 \setminus U$. Lorsque I_t est contenu dans un "grand intervalle", la configuration correspond toujours au mot nul $0 \ldots 0$. A part ce cas, chaque passage de l'une des extrémités de I_t en un point de U produit un nouveau mot. Lorsque les extrémités de I_t passent devant deux points de U distants de α, deux symboles sont échangés simultanément dans le mot ($\ldots 0 \ldots 1 \ldots$ devient $\ldots 1 \ldots 0 \ldots$ par exemple) ; deux tels points dans U correspondent à deux entiers consécutifs dans S .

Si S est formé de q intervalles maximaux (autrement dit, non contigus) d'entiers consécutifs, de longueurs respectives $s_1, \ldots, s_q$ (donc $s_1 + \ldots + s_q = s$), alors le nombre de paires d'entiers consécutifs est égal à $(s_1 - 1) + \ldots + (s_q - 1) = s - q$. On obtient la

Proposition 3.1. *Soit x une suite de Sturm, et S un sous-ensemble fini de $\mathbb{N}$. Alors*

$$|\{S \cap T_k x \,:\, k \in \mathbb{N}\}| = s + q - max\{p - 1,\, 0\}$$

où q est le nombre d'intervalles maximaux d'entiers consécutifs de S , α est la densité de x , et p est le nombre de composantes connexes de longueur strictement supérieure à α dans $\mathbb{S}^1 \setminus (\alpha S)$.

En particulier, $|\{S \cap T_k x \,:\, k \in \mathbb{N}\}| \leq s + q \leq 2s$.

Evidemment, $0 \leq p < \frac{1}{\alpha}$. Remarquons que si S est un intervalle ($q = 1$), alors on a nécessairement $p \leq 1$ et on retrouve la complexité $P(s) = s + 1$. On a aussi $p \leq 1$ dès que S contient un intervalle suffisamment long : s'il existe $i \leq q$ tel que $s_i \geq \left\lfloor \frac{1}{\alpha} \right\rfloor$. Signalons aussi qu'une transformation simple (morphisme sturmien) permet d'obtenir, à partir d'une suite de Sturm x donnée, une suite de Sturm x' dont la densité α' est telle $\frac{1}{3} < \alpha' < \frac{1}{2}$: pour une telle suite, il y a au plus deux "grands intervalles" et donc $|\{S \cap T_k x \,:\, k \in \mathbb{N}\}|$ est égal à $s + q$ ou à $s + q - 1$.

Dans la preuve de la proposition, on voit que la famille $\{T_k x \,:\, k \in \mathbb{N}\}$ induit sur S les mêmes parties que n'importe quelle sous-famille $\{T_k x \,:\, k \in K\}$ (où $K \subseteq \mathbb{N}$) pourvu que αK soit dense dans le cercle ; par exemple, $K = a\mathbb{N}$ (a>0 entier). Incidemment, on peut voir ainsi que tous les facteurs (de longueur a) sont obtenus en découpant la suite de Sturm en séquences consécutives $x_{ak} x_{ak+1} \ldots x_{ak+a-1}$ ($k \in \mathbb{N}$).

Corollaire 3.2. *Soit x une suite de Sturm à 1 isolés (densité $\alpha < \frac{1}{2}$). Si $\{S \cap T_k x \,:\, k \in \mathbb{N}\}$ contient le mot 1...1, alors $|\{S \cap T_k x \,:\, k \in \mathbb{N}\}| = 2s$.*

Démonstration. Tous les points de U sont rassemblés dans un intervalle ouvert de longueur α. Il y a donc un unique grand intervalle ($p = 1$) et de plus, S ne peut pas contenir deux entiers consécutifs ($q = s$). □

On peut voir aisément que pour tout mot $m \in \{0, 1\}^s$, il existe un ensemble S tel que $\{S \cap T_k x \,:\, k \in \mathbb{N}\}$ ait 2s éléments et contienne le mot m.

Corollaire 3.3. *La dimension de Vapnik-Chervonenkis d'une suite de Sturm est égale à 2.*

Démonstration. On peut supposer les 1 isolés dans x. La suite contient deux facteurs u et v, de même longueur, de la forme $u = 10\ldots010$ et $v = 10\ldots001$, d'où l'existence d'un ensemble S à deux éléments tel que $|\{S \cap T_k x \,:\, k \in \mathbb{N}\}| = 4 = 2^{|S|}$. □

Quelques remarques. En complément au corollaire 1, on peut remarquer qu'il existe une progression arithmétique arbitrairement longue $S = \{0,\, a,\, 2a, \ldots,\, (s-1)a\}$ pour laquelle le mot 1...1 appartient à $\{S \cap T_k x \,:\, k \in \mathbb{N}\}$ (il est évidemment inutile d'invoquer le théorème de Szemeredi ici : l'irrationalité de la densité α des suites de Sturm suffit et une valeur de a s'obtient en fonction de la longueur s voulue et du développement en fraction continue de α) ; dans cette situation, les $2s$ mots obtenus sont précisément les 0^j1^l et 1^j0^l $(j+l=s)$. On peut remarquer aussi que si $S = \{0,\, a,\, 2a, \ldots,\, (s-1)a\}$ est une progression arithmétique, alors $\{S \cap T_k x \,:\, k \in \mathbb{N}\}$ est l'ensemble des facteurs de longueur s de l'une quelconque des suites extraites $y^i = (y^i_n)_{n\geq 0} = (x_{i+an})_{n\geq 0}$ qui ont toutes les mêmes facteurs pour a fixé ; lorsque $a \geq 2$, la complexité de ces suites extraites est $P(s) = 2s$ lorsque s est assez grand.

Références

[1] J.P. Allouche, *Sur la complexité des suites infinies,* Bull. Belg. Math. Soc. 1(1994) 133-143.

[2] J. Berstel, P. Séébold, *Sturmian words,* Chap. 2 in Algebraic Combinatorics on Words, Lothaire 2001.

[3] J. Cassaigne, *Complexité et facteurs spéciaux,* Bull. Belg. Math. Soc. 4(1997) 67-88

[4] P. Kurka, *Topological and symbolic dynamics,* Société Mathématique de France 2003

[5] C. Kassel, C. Reutenauer, *Sturmian morphisms, the braid group B_4, Christoffel words and bases of F_2*, Prépublication IRMA 2004.

[6] M. Morse, G.A. Hedlund, *Symbolic dynamics,* Amer. J. Math. 60(1938) 815-866.

[7] M. Morse, G.A. Hedlund, *Symbolic dynamics II. Sturmian trajectories,* Amer. J. Math. 62(1940) 1-42.

[8] R. Tijdeman, *On the minimal complexity of infinite words,* Indag. Math. 10(1999) 123-129.

[9] V.N. Vapnik, A.Y. Chervonenkis, *The uniform convergence of frequencies of the appearance of events to their probabilities.* (Russian. English summary) Teor. Verojatnost. i Primenen. 16 (1971) 264-279, english translation Theor. Probability Appl. 16 (1971), 264 280.

Institut de Recherche Mathématique Avancée -UMR 7501 du CNRS
7 rue Descartes, F-67084 Strasbourg Cedex, France.
http ://www-irma.u-strasbg.fr *e-mail :* mpmuller@math.u-strasbg.fr
et : IUT Robert-Schuman, BP 10315, F-67411 Illkirch Cedex