Habilitation à diriger des recherches

présentée devant

L'UNIVERSITÉ DE MONTPELLIER

École Doctorale I2S

Pascal Ochem

Évitement de motif

Contents

1 Résumé en français

Mes deux grandes thématiques de recherche sont la théorie des graphes et la combinatoire des mots. Dans une moindre mesure, j'ai aussi travaillé en théorie des nombres à l'amélioration de bornes inférieures sur certains paramètres d'un éventuel nombre parfait impair.

En théorie des graphes, je me suis surtout intéressé à plusieurs problèmes de coloration, notamment l'homomorphisme, et à des problèmes de représentation géométrique de graphes. Mes travaux en combinatoire des mots concernent l'évitement de régularités dans les mots infinis, et c'est le sujet principal de cette introduction. L'avantage de cette double casquette est que, dans certains cas, j'ai pu transférer des techniques de preuves d'un domaine à l'autre (voir section 3). Cela facilite aussi l'étude de problématiques à l'interface des graphes et des mots, comme les seuils de répétitions dans les graphes [J26,J54].

Ce document traite plus particulièrement de l'évitement de motifs, qui fait donc partie du domaine de l'évitement de régularités dans les mots infinis et qui est une composante importante de la combinatoire des mots depuis son origine. Pour cette habilitation, je ne voulais pas simplement citer ou reprouver des résultats de mes articles. Je choisi donc de parler uniquement d'évitement de motifs car c'est un domaine très cohérent et celui pour lequel il y a le plus de petites choses intéresantes qui n'ont pas eu l'opportunité d'apparaître dans un de mes articles. Cette thèse est ainsi l'occasion de donner des idées, des motivations ou des résultats non publiés et de présenter génériquement les techniques de preuves (avec leurs champs d'applicabilité) qui ont été utilisées dans différents articles.

la recherche est gratifiante pour ses résultats et aussi en tant qu'activité sociale. Je remercie tous mes co-auteurs, j'ai eu notamment le plaisir de travailler récemment, intensément et en personne avec Alexandre Pinlou, Michaël Rao, Mickael Montassier, Francesca Fiorenzi, Guillaume Guégan, Matthieu Rosenfeld, Golnaz Badkobeh, Élise Vaslet, Louis Esperet, Daniel Gonçalves, Stéphane Bessy et Dieter Rautenbach.

Une version électronique et au moins aussi récente de ce mémoire est disponible :
http://www.lirmm.fr/~ochem/hdr/

L'évitement en combinatoire des mots

Une introduction à l'évitement de régularités dans les mots infinis commence inévitablement par les travaux de Thue sur les mots sans carrés.

Un facteur f d'un mot m est un mot formé des lettres consécutives de m. Par exemple, *is*, *issis*, *mississipi*, *mis* et le mot vide ε sont tous des facteurs du mot *mississipi*. Un préfixe non-vide de m est un facteur contenant la première lettre de m. Un suffixe non-vide de m est un facteur contenant la dernière lettre de m.

Un carré est un mot de longueur $2n$ tel que le préfixe et le suffixe de longueur n sont égaux. Par exemple, *tonton*, *coco*, *uu*, *yyyyyy* et *abcbabcb* sont des carrés. On dit qu'un mot contient un carré si un de ses facteurs est un carré. Par exemple, *banane* contient le carré $anan = (an)^2$ et *chocolat* est sans carré.

Sur un alphabet de deux lettres, disons $\{a, b\}$, les plus longs mots sans carré sont *aba* et *bab*. Thue [50] a montré qu'il existe des mots sans carré arbitrairement grands sur un alphabet de trois lettres.

Ce résultat, par ses améliorations, ses généralisations successives et ses variations, est le point de départ d'une littérature foisonante dans le domaine de l'évitabilité, qui est central en combinatoire des mots.

En premier lieu, comme l'explique très bien Berstel [7], les résultats précis de Thue lui-même nous apporte bien plus que la simple existence d'un mot infini sans carré sur trois lettres. Pour tout $k \geqslant 2$, on note par $\Sigma_k = \{\mathtt{0}, \mathtt{1}, \ldots, \mathtt{k-1}\}$ l'alphabet à k lettres et par Σ_k^* l'ensemble des mots sur Σ_k. Un morphisme σ est une application de Σ_d^* dans Σ_f^* telle que pour tout $uv \in \Sigma_d$, on a $\sigma(uv) = \sigma(u)\sigma(v)$. Ainsi, on a $\sigma(\varepsilon) = \varepsilon$ et on décrit σ en donnant l'image $\sigma(x)$ de chaque lettre $x \in \Sigma_d$. Si $d = f$, et si $\mathtt{0}$ est un préfixe de $\sigma(\mathtt{0})$, on peut définir le point fixe $w = \sigma^\omega(\mathtt{0})$ obtenu en applicant une infinité de fois (ω) le morphisme σ à $\mathtt{0}$. Dans les cas non-bizarres qui nous intéressent, w est un mot infini et il est nommé "point fixe" en tant que solution de $w = \sigma(w)$.

On peut maintenant énoncer un des résultats de Thue:
Soit $\sigma : \Sigma_3 \mapsto \Sigma_3$ le morphisme défini par

- $\sigma(\mathtt{0}) = \mathtt{012}$,
- $\sigma(\mathtt{1}) = \mathtt{02}$,
- $\sigma(\mathtt{2}) = \mathtt{1}$.

Le point fixe $b_3 = \sigma^\omega(\mathtt{0})$ est un mot ternaire infini sans carré.

Complexité en facteur Plusieurs autres constructions de mots sans carré ont ensuite utilisé des morphismes, ce qui a donné lieu à une étude approfondie des morphismes sans carré (i.e. tels que l'image de tout mot sans carré est sans carré) [4, 6, 14].

Devant l'abondance des mots ternaires sans carré connus, on a voulu montrer qu'il existe *beaucoup* de mots ternaires sans carrés en évaluant le *taux de croissance exponentiel* du langage des mots ternaires sans carré, que l'on définit comme suit. La complexité d'un langage $L \subset \Sigma_k^*$ est le nombre $c_L(n) = |L \cap \Sigma_k^n|$ de mots de taille n de L. Un langage L est factoriel si pour tout mot m appartenant à L, les facteurs de m appartiennent aussi à L. Le taux de croissance exponentiel de L est ainsi $gr(L) = \lim_{n\to\infty} c_L(n)^{\frac{1}{n}}$. Par un argument standard de sous-multiplicativité, cette limite existe effectivement si L est factoriel.

Le taux de croissance exponentiel du langage factoriel $L \subset \Sigma_k^*$ est tel que

- $gr(L) = 0$ si L ne contient pas de mot infini,
- $gr(L) = k$ si $L = \Sigma_k^*$,
- $1 \leqslant gr(L) < k$ sinon.

Si $gr(L) > 1$, on dit que L est exponentiel. Une première série de bornes inférieures sur le taux de croissance exponentiel des mots ternaires sans carrés, à savoir $2^{\frac{1}{17}}$ [21] en 1998, $65^{\frac{1}{40}}$ [24] en 2001 et $110^{\frac{1}{42}}$ [49] en 2003, a utilisé des constructions à base de morphismes. J'ai participé à l'effort pour déterminer cette constante en abaissant la borne supérieure de 1.30193813 [46] en 2004 à 1.30178859 [C2] en 2006 avec une technique de matrice de transition.

Une autre série de bornes inférieures et supérieures par Kolpakov et Rao [31] et Shur [48] utilisant d'autres méthodes a abouti à une évaluation remarquablement précise de cette constante: entre 1.301759 et 1.301762.

Langage de faible complexité On a vu que Thue a montré que b_3 ne contient pas de carrés. Plus exactement, il a montré que b_3 évite les carrés ainsi que les facteurs 010 et 212, et que b_3 est le seul mot infini ternaire qui évite les carrés, 010 et 212. C'est-à-dire que pour tout facteur f de b_3, tous les mots ternaires évitant les carrés, 010, 212 et f sont finis.

Un autre résultat de Thue de ce type concerne l'évitement des chevauchements (overlaps) dans les mots binaires. Un overlap est un mot de la forme $awawa$ avec $a \in \Sigma_2$ et $w \in \Sigma_2^*$. Thue a montré que le point fixe b_2 du morphisme $\texttt{0} \to \texttt{01}$, $\texttt{1} \to \texttt{10}$ est le seul mot infini binaire sans overlap. Avec Golnaz Badkobeh [3] (voir Section 6.7), nous avons identifié d'autres langages binaires ne contenant qu'un seul mot infini. Ces langages sont définis en interdisant de grands carrés et un ensemble fini de facteurs.

Mots de Dejean Les carrés et les overlaps considérés par Thue ont étés interprétés comme des répétitions par Françoise Dejean [18]. Une répétition dans un mot w est une paire de mots non-vides p et e telle que pe est facteur de w et e est préfixe de pe. Si pe est une répétition, alors sa période est $|p|$ et son exposant est $\frac{|pe|}{|p|}$. Pour définir une notion d'évitement de répétition, on dit qu'un mot est α^+-free (resp. α-free) s'il ne contient pas de répétition d'exposant β tel que $\beta > \alpha$ (resp. $\beta \geqslant \alpha$). Ainsi, un mot sans overlap tel que b_2 est 2^+-free et un mot sans carré tel que b_3 est 2-free. Dejean a posé la question naturelle de la meilleure propriété qu'on puisse obtenir avec un mot à k lettres et elle définit pour cela le seuil de répétition $RT(k)$ comme le plus petit α tel qu'il existe un mot α^+-free sur k lettres. On sait que $RT(2) = 2$ et Dejean a montré que $RT(3) = \frac{7}{4}$ grâce à un point fixe de morphisme uniforme. Elle a aussi conjecturé les autres valeurs de $RT(k)$:

- $RT(4) = \frac{7}{5}$, montré par Pansiot [40].
- $RT(k) = \frac{k}{k-1}$ pour tout $k \geqslant 5$, montré par divers auteurs [10, 33, 42].

On verra à la Section 4 de nombreuses constructions consistant en l'image par un morphisme d'un mot de Dejean, c'est-à-dire d'un mot infini $RT(k)^+$-free sur k-lettres.

Motifs Il est temps d'en venir à la généralisation des carrés de Thue qui nous intéresse: le motif. Un motif p est un mot fini sur un alphabet $\Delta = \{A, B, C, \ldots\}$ de variables. Une occurrence de p dans un mot w est un morphisme $h : \Delta^* \to \Sigma^*$ non-effaçant tel que $h(p)$ est un facteur de w. Par exemple, le mot $w = \texttt{010000101001}$ contient plusieurs occurrences du motif $p = AABB$: w contient le facteur 000101 qui correspond à l'occurrence $A \to \texttt{0}$, $B \to \texttt{01}$, ainsi que les occurrences $A \to \texttt{0}$, $B \to \texttt{0}$; $A \to \texttt{0}$, $B \to \texttt{10}$; et $A \to \texttt{01}$, $B \to \texttt{0}$.

Ainsi, les occurrences du motif AA sont les carrés et donc les mots sans carrés évitent AA. De même, les mots sans overlap évitent à la fois AAA et $ABABA$. L'indice d'évitabilité $\lambda(p)$ du motif p est la taille du plus petit alphabet Σ tel qu'il existe un mot infini sur Σ qui ne contienne aucune occurrence de p. L'existence de mots infinis ternaires sans carrés et la finitude des mots binaires sans carrés signifient donc que $\lambda(AA) = 3$. Un motif est évitable si il a un indice d'évitabilité fini.

La majeure partie de ce document concerne la détermination de l'indice d'évitabilité d'un maximum de motifs. Et avant de présenter encore d'autres variantes de l'évitement de carré ou de motifs, auxquels j'ai par ailleurs pu contribuer, voici quelques propriétés qui font de l'indice d'évitabilité une notion particulièrement élégante.

1. Sans perte de généralité, les mots évitants sont uniformément récurrents (i.e., pour tout facteur f, il existe n tel que tout facteur de longueur n du mot évitant contient f). Cela facilite toutes sortes de raisonnements.

2. Les mots évitants sont apériodiques. Ce qui est bien car sinon, puisqu'ils sont uniformément récurrents, ils seraient périodiques, ce qui serait assez monotone.

3. Si un mot w contient un motif, alors l'image de w par un morphisme non-effaçant contient aussi le motif.

4. On connait une caractérisation des motifs évitables.

Évitement au sens abélien et additif Ces variantes considérent de nouvelles notions d'égalité entre facteurs. Deux mots x et y sont égaux au sens abélien si pour toute lettre a de l'alphabet, x et y contiennent le même nombre d'occurrences de a. Si l'alphabet est constitué d'entiers, deux mots x et y sont égaux au sens additif si ils ont la même longueur et la même somme de lettres. Ce sont des notions d'évitement plus fortes puisque deux mots égaux au sens classique sont égaux au sens abélien et deux mots égaux au sens abélien sont égaux au sens additif. Par exemple, `1432` est un carré au sens additif mais pas au sens abélien, et `210120` est un carré au sens abélien mais pas au sens classique.

Ces variantes sont beaucoup plus difficiles, notamment parce que des contraintes bien plus fortes pèsent sur les grands facteurs. Elles ne conservent pas la propriété 4 et la question principale du domaine est de savoir si les carrés sont évitables au sens additif. J'ai un argument heuristique qui suggère une réponse négative : http://www.lirmm.fr/~ochem/additive_square.htm
En revanche, on sait que les cubes (motif AAA) sont 3-évitables au sens additif [12, 43] et que les carrés sont 4-évitables au sens abélien [30]. Rosenfeld [47] a aussi montré la 2-évitabilité au sens abélien de nombreux motifs à deux variables.

Évitement par des mots partiels Un mot partiel est un mot sur l'alphabet $\Sigma_k \cup \{\diamond\}$. Un mot partiel représente l'ensemble des mots sur Σ_k qu'on peut obtenir en remplaçant indépendamment tous les occurrences du symbole $\diamond$ par une lettre de Σ_k. Blanchet-Sadri et ses co-auteurs se sont spécialisés dans les mots partiels, et notamment l'évitement de motifs par des mots partiels (voir [8] pour un survey de leurs résultats). Par exemple, avec le motif $AABB$ sur Σ_3, le mot $\mathtt{12}\diamond\mathtt{20}\diamond\mathtt{20}$ contient l'occurrence $A \to \mathtt{12}$, $B \to \mathtt{0}$ car `121200` est un facteur de `12120020`, qui est compatible avec $\mathtt{12}\diamond\mathtt{20}\diamond\mathtt{20}$.

Étant donné une taille d'alphabet k et un motif P, on peut ainsi chercher à maximiser la fréquence du symbole $\diamond$ dans un mot partiel infini sur $\Sigma_k \cup \{\diamond\}$ évitant P. Dans le même contexte, on peut aussi chercher à minimiser la distance maximale entre deux occurrences consécutives du symbole $\diamond$. En particulier, montrer qu'un motif est évité par un mot partiel avec une densité non nulle de $\diamond$ implique que le langage des mots évitant le motif est exponentiel.

Évitement de motifs avec reverse Currie, Lafrance, Mol et Rampersad [15, 16, 17] ont considéré la possibilité de faire correspondre une occurrence d'une variable avec son image miroir. Par exemple, `210120` contient l'occurrence $A \to \mathtt{21}$, $B \to \mathtt{0}$ du motif ABA^RB et du motif ABA^RB^R. C'est-à-dire que puisque $A \to \mathtt{21}$, on $A^R \to (\mathtt{21})^R = \mathtt{12}$.

Avec cette extension de la notion de motif, on perd la propriété 2, puisque par exemple le mot périodique $(\mathtt{01})^\infty$ évite AA^R. On perd aussi la propriété 3 puisque le mot $\mathtt{0}^\infty$ contient AA^R alors que son image par le morphisme $\mathtt{0} \to \mathtt{01}$ évite AA^R. Une question importante de l'évitement de motifs est l'existence de motifs évitables d'indice au moins 6 (voir Question 4). Puisqu'il semble moins difficile d'obtenir des motifs avec reverse d'indice 5 que des motifs sans reverse d'indice 5, on peut espérer attaquer la question 4 en trouvant d'abord un motif avec reverse d'indice 6.

Évitement dans des graphes La coloration non-répétitive d'un graphe simple G (pas d'orientation, pas de boucle, pas d'arête multiple) est une coloration des sommets de G telle que pour tout chemin de G (qui ne s'intersecte pas lui-même), la sequence de couleurs sur ce chemin est un mot sans carré. Comme pour la coloration classique, l'objectif est de minimiser le nombre de couleurs et le nombre chromatique associé est appelé *Thue number* du graphe. Le Thue number d'une classe de graphes est le maximum des Thue numbers des graphes de la classe.

Le Thue number des arbres est au plus 4, le Thue number des graphes de degré maximum Δ est $O(\Delta^2)$, et le Thue number des graphes de treewidth au plus t est $O\left(4^t\right)$. On sait que le Thue number des graphes planaires est au moins 11 et, récemment, Dujmović et al. [20] ont montré que le Thue number des graphes planaires est au plus 768.

On dit qu'un motif P est évitable dans les graphes si le nombre de couleurs nécessaire à l'éviter est borné en fonction de $\lambda(P)$ et du degré maximum du graphe. Grytczuk [25] a montré que tous les motifs doubled, i.e. tels que chaque variable apparait au moins deux fois, sont évitables dans les graphes. Il a aussi conjecturé que tous les motifs évitables sont évitables dans les graphes. Matthieu Rosenfeld et moi avons donné un motif contre-exemple qui est 2-évitable mais inévitable dans les graphes de degré maximum 3 [39] (voir Section 6.1).

Mes travaux sur l'évitement de motifs

J'ai commencé à étudier l'évitement de motifs en thèse où j'ai montré avec Lucian Ilie et Jeffrey Shallit que $ABCBABC$ est 2-évitable [29] et où j'ai terminé la détermination de l'indice d'évitabilité des motifs à 3 variables [34].

Les articles inclus dans ce document (Section 6) concernent ma recherche post-thèse en évitement de motifs, auxquels j'ai ajouté mon article avec Golnaz Badkobeh sur l'évitement de grands carrés et d'un ensemble fini de facteurs [3] (voir Section 6.7) parce qu'il a été précurseur de la notion de motif "essentiellement évité" par un nombre fini de mots morphiques, discutée en Section 5.

Par ordre chronologique, j'ai d'abord étudié plus en détail le motif $AABBCABBA$ (ou la formule équivalente $AABB.ABBA$) [35] (voir Section 6.5), notamment pour mettre en évidence la notion de types de mots différents évitant une formule, discutée en Section 5. Puis, avec Alexandre Pinlou en 2014 [37] (voir Section 6.6), nous avons utilisé la méthode dite de compression d'entropie pour démontrer cette conjecture de Julien Cassaigne : tout motif à n variables et de longueur au moins $3 \times 2^{n-1}$ est 2-évitable. J'ai ensuite simplifié cette méthode pour l'appliquer aux motifs doubled (i.e., tels que toute variable apparait au moins deux fois) et montré qu'ils sont tous 3-évitables [36] (voir Section 6.3). Cette nouvelle méthode reste non-constructive et est une généralisation de la "power series method" de la littérature. Elle est décrite en Section 3. Avec Guillaume Guégan, nous avons aussi utilisé la nouvelle méthode pour donner une preuve relativement courte que les shuffle squares sont 7-évitables [27].

Enfin, Les trois derniers articles traitent des formules à au plus trois variables. Avec Guilhem Gamard, Gwenaël Richomme et Patrice Séébold [22], nous avons considéré les formules ternaires les plus dures à éviter, dans un sens défini par Ronald Clark [13] et expliqué en Section 2. Nous avons fini de déterminer l'indice de ces formules ternaires en montrant que $\lambda(ABCA.BCAB.CABC) = 3$, $\lambda(ABCBA.CBABC) = 2$ et $\lambda(ABCA.CABC.BCB) = 3$. La méthode utilisé pour $ABCBA.CBABC$, $ABCA.CABC.BCB$ et la plupart des autres formules dites "nice" est décrite en Section 4. Avec

Matthieu Rosenfeld, nous avons déterminé l'indice de toutes les formules binaires [38] (voir Section 6.2) et certaines d'entres elles, telles que $AA.ABA.ABBA$, se sont avérées être essentiellement évitées par 1, 2 ou 4 mots morphiques binaires, comme on l'explique en Section 5. Nous avons aussi déterminé l'indice de certaines formules ternaires intéressantes [39] (voir Section 6.1), notamment $ABACA.ABCA$ qui est évitée par un nombre polynomial de mots binaires sans pour autant être essentiellement évitée par un nombre fini de mots morphiques (voir Section 5).

Plan du manuscrit

Ce manuscrit est stucturé par rapport aux méthodes de preuve utilisées dans les articles mentionés précédemment. C'est un point de vue transversal à la chronologie des articles, puisqu'une même méthode apparait dans plusieurs articles et un même article peut utiliser plusieurs méthodes.

On commence par présenter formellement le domaine en Section 2 avec les définitions nécessaires et l'état de l'art.

La première méthode, présentée en Section 3, est un argument de comptage permettant de donner une borne inférieure sur le taux de croissance exponentiel du langage des mots évitant un motif. C'est une méthode non-constructive et elle s'applique aux motifs doubled.

En Section 4, on donne une méthode qui borne l'indice d'une formule en construisant un mot évitant qui est l'image morphique d'un mot de Dejean. En s'intéressant à l'applicabilité de cette méthode, on est amené à définir et étudier les formules nice, qui sont l'objet principal de la section.

Les deux méthodes précédentes produisent un nombre exponentiel de mots évitants. Elles ne fonctionnent donc pas pour les formules dont le nombre de mots évitants est polynomial. La Section 5 est dédiée à ces formules ayant un nombre polynomial de mots évitants.

2 Pattern avoidance

A pattern p is a non-empty word over an alphabet $\Delta = \{A, B, C, \dots\}$ of capital letters called *variables*. An *occurrence* of p in a word w is a non-erasing morphism $h : \Delta^* \to \Sigma^*$ such that $h(p)$ is a factor of w. For example, the word $w = \texttt{010000101001}$ contains various occurrences of the pattern $p = AABB$. The factor $\texttt{000101}$ of w corresponds to the occurrence $A \to \texttt{0}$, $B \to \texttt{01}$. Also, w contains the occurrences $A \to \texttt{0}$, $B \to \texttt{0}$; $A \to \texttt{0}$, $B \to \texttt{10}$; and $A \to \texttt{01}$, $B \to \texttt{0}$.

The *avoidability index* $\lambda(p)$ of a pattern p is the size of the smallest alphabet Σ such that there exists an infinite word w over Σ containing no occurrence of p. Since there is no other index defined in this document, the avoidability index is often refered to as simply the index. We say that a pattern p is k-avoidable if $\lambda(p) \leqslant k$. Keeping our example, we can check by hand or computer that every sufficiently long word over the 2-letter alphabet contains $AABB$. Also, there exists an infinite word over the 3-letter alphabet that avoids squares [50], that is, occurrences of AA. Since no infinite binary word avoids $AABB$ whereas some infinite ternary words avoid $AABB$, $\lambda(AABB) = 3$.

Question 1. *Is $\lambda(p)$ computable?*

Let us review the results and conjectures (remotely) related to this interesting question. We need a few definitions. A finite factor f is recurrent in an infinite word w if w contains infinitely many occurrences of f. An infinite word w is recurrent if all of its finite factors are recurrent in w. A finite factor f is uniformly recurrent in an infinite word w if there exists an integer ℓ such that f occurs in every factor of length ℓ of w. An infinite recurrent word w is uniformly recurrent if all of its finite factors are uniformly recurrent in w, that is, there exists a function ℓ such that every finite factor f of w occurs in every factor of length $\ell(|f|)$ of w. Let $v(p)$ denote the number of distinct variables in the pattern p. Let us introduce the Zimin operator. If p does not contain the variable X, then $Z(p) = pXp$. The Zimin operator can iterated, so that $Z^0(p) = p$, $Z^1(p) = Z(p) = pXp$, $Z^2(p) = Z(Z(p)) = pXpYpXp$, $\dots$, $Z^{i+1}(p) = Z^i(p)WZ^i(p)$. We define the partial order $\preceq$ on patterns as follows: $p_1 \preceq p_2$ if and only if $Z^{v(p_1)}(p_2)$, considered as a word, contains an occurrence of p_1. If $p_1 \preceq p_2$, then we say that p_1 divides p_2.

Lemma 2.

(1) If L is a factorial language and is infinite, then L contains a uniformly recurrent word.

(2) If $p_1 \preceq p_2$, then every recurrent word avoiding p_1 also avoids p_2.

(3) If $p_1 \preceq p_2$, then $\lambda(p_1) \geqslant \lambda(p_2)$.

Proof.

1. This is a classical result in dynamic symbolic [41], which can be proved as follows. We define the total order $<$ on finite words as follows: if $|f_1| < |f_2|$ or if $|f_1| = |f_2|$ and f_1 is lexicographically smaller than f_2, then $f_1 < f_2$. Suppose for contradiction that there exists an infinite factorial language L such that every infinite word w in L is not uniformly recurrent.

   To every infinite word w in L, we denote by f_w the smallest factor of w with respect to $<$ that is not uniformly recurrent. Now, we consider an infinite word w in L that maximizes f_w with respect to $<$. Since f_w is not uniformly recurrent, there exist arbitrarily long factors of w avoiding f_w. By compacity, there also exists an infinite word w' with the following properties:

   - For every word $v < f_w$, either v is not a factor of w' or v is uniformly recurrent in w'.

- f_w is not a factor of w'.

  Thus, w' is an infinite word in L such that $f_w < f_{w'}$. This contradicts the maximality of f_w.

2. If a recurrent word w contains an occurrence of p_2, then w also contains an occurrence of $Z^{v(p_1)}(p_2)$. Since $p_1 \preceq p_2$, $Z^{v(p_1)}(p_2)$ contains an occurrence of p_1. Thus, if w contains an occurrence of p_2, then w also contains an occurrence p_1. This is the contrapositive of what we need to prove.

3. This is a consequence of (2).

□

We are ready to consider *unavoidable* patterns. A pattern p is unavoidable if $\lambda(p) = \infty$, that is, for every fixed k, no infinite word over Σ_k avoids p. Every word of length one is an occurrence of the pattern A, so that $\lambda(A) = \infty$. So, if a pattern p is such that $p \preceq A$, then $\lambda(p) \geqslant \lambda(A) = \infty$ by Lemma 2.3. Bean, Ehrenfeucht, and McNulty [4] and Zimin [51] proved that the necessary condition "p occurs in $Z^{v(p)}(A)$" is actually a characterization of unavoidable patterns. This characterization implies an algorithm to decide whether a pattern is unavoidable and Zimin also obtained a faster algorithm.

This implies that Question 1 can be restricted to avoidable patterns, that is, patterns with finite index. Moreover, the index of an avoidable pattern is not too large.

Theorem 3. *[32] If p is avoidable, then $\lambda(p) \leqslant v(p) + 4$.*

Thus, the computability of $\lambda(p)$ reduces to the decidability of $\lambda(p) \leqslant k$, given p and k as input. Actually, avoidable patterns with index up to 5 only have been exhibited [13].

Question 4. *Does there exist an avoidable pattern with index at least 6?*

A negative answer to Question 4 would further simplify Question 1. However, I conjecture that there exist available patterns with arbitrarily large index.

One last conjecture about the structure of avoiding words relates to Question 1. A word is *pure morphic* if it is the fixed point $m^\omega(0)$ of some morphism m. A word is *morphic* if it is the morphic image of a pure morphic word.

Conjecture 5. [Cassaigne's conjecture] For every avoidable pattern p, there exists a morphic word over $\lambda(p)$ letters avoiding p.

The pattern $ABCABC$ shows that we cannot require the morphic word in Cassaigne's conjecture to be pure morphic. Indeed, $\lambda(ABCABC) = 2$ whereas every pure morphic binary word contains arbitrarily long squares. Cassaigne [11] also gave an algorithm that decides, under a mild assumption, whether a morphic word avoids a pattern. Thus, proving the conjecture would reduce Question 1 from the existence of an infinite word to the existence of two finite morphisms.

Furthermore, assuming the stronger version of Cassaigne's conjecture that takes this mild assumption into account, Question 1 would have a positive answer. Given an avoidable pattern p and an integer k, both answers to whether $\lambda(p) \leqslant k$ would be semi-decidable: the answer YES is provable by looking for two finite morphisms generating an avoiding morphic word and the answer NO is provable by checking that the set of words avoiding p over Σ_k is finite.

Since Question 4 is difficult, we consider first patterns with low index. Intuitively, patterns p that are long enough with respect to $v(p)$ should have low index. The following families of patterns show what does "long enough" mean.

- if $p_t = Z^{t-1}(A)$, then $v(p_t) = t$, $|p_t| = 2^t - 1$, and $\lambda(p_t) = \lambda(A) = \infty$.
- if $p_t = Z^{t-1}(AA)$, then $v(p_t) = t$, $|p_t| = 3 \times 2^{t-1} - 1$, and $\lambda(p_t) = \lambda(AA) = 3$.
- if $p_t = Z^{t-2}(AABAB)$, then $v(p_t) = t$, $|p_t| = 3 \times 2^{t-1} - 1$, and $\lambda(p_t) = \lambda(AABAB) = 3$.

Cassaigne [32] asked to prove that these constructions are extremal. This has been settled independently by Blanchet-Sadri and Woodhouse [9] and Pinlou and me [37] (see Section 6.6).

Theorem 6 ([9, 37])**.** *Let p be a pattern.*

(a) If $|p| \geqslant 3 \times 2^{v(p)-1}$, then $\lambda(p) \leqslant 2$.

(b) If $|p| \geqslant 2^{v(p)}$, then $\lambda(p) \leqslant 3$.

The first step towards this result is to notice that every pattern p such that $|p| \geqslant k \times 2^{v(p)-1}$ (with $k = 2$ or $k = 3$ and k is fixed) contains as a factor a pattern p' such that $|p'| \geqslant k \times 2^{v(p')-1}$ and p' is *doubled*. A pattern is doubled if it contains no *isolated* variable, that is, a variable that appears exactly once.

Theorem 7 ([36], see Section 6.3)**.** *Every doubled pattern is 3-avoidable.*

The proofs of Theorems 6.a and 7 both use the construction of infinite words via morphisms when $v(p)$ is small and a non-constructive method when $v(p)$ is large. The non-constructive method is described in Section 3 and the constructions are described in Section 4.

An introduction to formulas. By Theorem 7, a pattern with index at least 4 contains at least one isolated variable, such as the variable C in $AABBCABBA$. Recall that by Lemma 2.(1), we can focus on recurrent words. A recurrent word w avoiding $AABBCABBA$ cannot contain the factor `0000`, since `0000x0000` would be an occurrence of $AABBCABBA$. Similarly, w cannot contain both the factors `0011` and `0110`, since this implies an occurrence of $AABBCABBA$ such that $A \to$ `0` and $B \to$ `1`. More generally, w cannot contain the same occurrence of both the patterns $AABB$ and $ABBA$.

We thus define the formula f corresponding to a pattern p as follows: f is obtained from p by replacing every isolated variable of p by a dot. Thus, $AABB.ABBA$ corresponds to $AABBCABBA$. In a formula, a maximal block of consecutive variables with no dot is a *fragment*. So $AABB$ and $ABBA$ are the fragments of the formula $AABB.ABBA$. Then an occurrence of a formula f in a word w is a non-erasing morphism $h : \Delta^* \to \Sigma^*$ such that the h-image of every fragment of f is a factor of w. The avoidability index of a formula is defined accordingly. From previous discussion, if f corresponds to p, then $f \preceq p$ and $p \preceq f$. Thus $\lambda(p) = \lambda(f)$. Without loss of generality, a formula does not contain a fragment that is a factor of another fragment. Also, every variable appears at least twice in a formula. We also define the number $v(f)$ of variables in the formula f. So, if f corresponds to p, then $v(f) \leqslant v(p)$.

Formulas appear to be easier to manipulate than patterns. For example, the order of the fragments in a formula does not matter.

Cassaigne [11] began and I [34] finished the determination of the index of every pattern with at most 3 variables. Their index is either infinite or at most 3.

This also holds for formulas with at most 2 variables: for every avoidable binary formula f, either

- $AA \preceq f$ and thus $\lambda(f) \leqslant 3$, or
- $f = ABA.BAB$ and thus $\lambda(f) = 3$

The precise distinction between index 2 and index 3 is as follows:

Theorem 8 ([38], see Section 6.2)**.** *Let f be an avoidable binary formula. If f or the reverse of f divides a formula in the following set, then $\lambda(f) = 3$. Otherwise, $\lambda(f) = 2$.*

- $AAB.ABA.ABB.BBA.BAB.BAA$
- $AAB.ABBA$
- $AAB.BBAB$
- $AAB.BBAA$
- $AAB.BABB$
- $AAB.BABAA$
- $ABA.ABBA$
- $AABA.BAAB$

Proving that a formula f in the list above is not 2-avoidable is easy by backtracking, that is, depth-first exploration of every binary word avoiding f. More generally, to obtain a lower bound on the index of a formula, we use the relation $\preceq$ and backtracking, possibly with some standard optimisations.

To prove the 2-avoidability of most of the other avoidable binary formulas, we use the method described in Section 4. This approach allows to show that exponentially many binary words avoid the formula. Interestingly, every 2-avoidable binary formula is avoided by a (binary) morphic image of the word b_3 defined below.

- b_2 is the fixed point of $\mathtt{0} \mapsto \mathtt{01}$, $\mathtt{1} \mapsto \mathtt{10}$.
- b_3 is the fixed point of $\mathtt{0} \mapsto \mathtt{012}$, $\mathtt{1} \mapsto \mathtt{02}$, $\mathtt{2} \mapsto \mathtt{1}$.
- b_4 is the fixed point of $\mathtt{0} \mapsto \mathtt{01}$, $\mathtt{1} \mapsto \mathtt{03}$, $\mathtt{2} \mapsto \mathtt{21}$, $\mathtt{3} \mapsto \mathtt{23}$.
- b_5 is the fixed point of $\mathtt{0} \mapsto \mathtt{01}$, $\mathtt{1} \mapsto \mathtt{23}$, $\mathtt{2} \mapsto \mathtt{4}$, $\mathtt{3} \mapsto \mathtt{21}$, $\mathtt{4} \mapsto \mathtt{0}$.

There remain some binary formulas that do not admit exponentially many binary avoiding words. Let w and w' be infinite (right infinite or bi-infinite) words. We say that w and w' are equivalent if they have the same set of recurrent factors. We write $w \sim w'$ if w and w' are equivalent. We say that a finite set of infinite words $\mathcal{S}$ *essentially avoids* a formula f on a given alphabet (or any other description of forbidden factors) if every infinite word avoiding f is equivalent to a word in $\mathcal{S}$. For example, we discuss in Section 5 that if g_x and g_t are the morphisms defined below, then $\{g_x(b_3), g_t(b_3)\}$ essentially avoids $ABA.AABB$.

$$\begin{array}{ll} g_x(\mathtt{0}) = \mathtt{01110}, & g_t(\mathtt{0}) = \mathtt{01011011010}, \\ g_x(\mathtt{1}) = \mathtt{0110}, & g_t(\mathtt{1}) = \mathtt{01011010}, \\ g_x(\mathtt{2}) = \mathtt{0}. & g_t(\mathtt{2}) = \mathtt{010}. \end{array}$$

As previously mentioned, patterns with index up to 5 are known:

- $\lambda(AB.BA.AC.CA.BC) = 4$. [2]
- $\lambda(AB.BA.AC.BC.CDA.DCD) = 5$. [13]

Let us focus on the lower bounds. Since an occurrence of AA is an occurrence of $AB.BA.AC.CA.BC$, every word avoiding $AB.BA.AC.CA.BC$ is square-free. This means that $AB.BA.AC.CA.BC \preceq AA$. Then it suffices to check that no infinite ternary square-free word avoids $AB.BA.AC.CA.BC$. Clark has a rather tedious argument that only square-free words should be considered among words avoiding $AB.BA.AC.BC.CDA.DCD$. However, using Clark's observation that $Z^1(AA) = AABAA$ contains the occurrence $A \to A$, $B \to A$, $D \to A$, and $C \to B$ of the formula $AB.BA.AC.BC.CDA.DCD$, we obtain that $AB.BA.AC.BC.CDA.DCD \preceq AA$. Then Lemma 2 tells that the relevant words avoiding $AB.BA.AC.BC.CDA.DCD$ are recurrent and square-free. We also notice that $Z^1(ABACBAB) = ABACBABDABACBAB$ contains the occurrence $A \to C$, $B \to BA$, $C \to B$, and $D \to A$ of the formula $AB.BA.AC.BC.CDA.DCD$. Then again, recurrent words avoiding $AB.BA.AC.BC.CDA.DCD$ must avoid $ABA.BAB$.

Noticeably, $AB.BA.AC.CA.BC$ is an earlier example of formula with few avoiding words (over Σ_4). Let b_4, b'_4, and b''_4 be the three non-equivalent words obtained from b_4 by permutations of Σ_4.

Theorem 9. *[2] $\{b_4, b'_4, b''_4\}$ essentially avoids $AB.BA.AC.CA.BC$.*

Thue [50] found similar examples, but they require to forbid more than just one formula:

Theorem 10. *[50]*

- *b_2 essentially avoids $ABABA$, 000, and 111 (or $ABABA$ and AAA).*
- *b_3 essentially avoids AA, 010, and 212.*
- *$m_1(b_5)$ essentially avoids AA, 010, and 020.*
- *$m_2(b_5)$ essentially avoids AA, 121, and 212.*

Concerning the last two results of Theorem 10, the morphisms m_1 and m_2 given below, as well as the morphism defining b_5 given above, appear in one of our paper [3] (see Section 6.7) and are smaller than the morphisms used by Thue. To be fair with Thue's work, our simpler morphisms were found by computer and computers were less efficient in the beginning of the twentieth century.

$$\begin{array}{ll}
m_1(0) = 012, & m_2(0) = 02, \\
m_1(1) = 1, & m_2(1) = 1, \\
m_1(2) = 02, & m_2(2) = 0, \\
m_1(3) = 12, & m_2(3) = 12, \\
m_1(4) = \varepsilon. & m_2(4) = \varepsilon.
\end{array}$$

As a digression, we mention that the paper [3] (See Section 6.7) also contains many results of the form "the binary morphic word w essentially avoids squares with period at least t and the set F of forbidden factors", such that w is a morphic image of either b_3 or b_5. Together with Theorems 9 and 10, this shows the ubiquity of the pure morphic words b_2, b_3, b_4, and b_5 in pattern avoidance. To study formulas with high index, Clark [13] has introduced the notion of *n-avoidance basis*, which is the smallest set of formulas with the following property: for every $i \leqslant n$ and for every avoidable formula f with i variables, there exists at least one formula f' with at most i variables in the n-avoidance basis such that $f' \preceq f$.

From the definition, it is not hard to obtain that the 1-avoidance basis is $\{AA\}$ and the 2-avoidance basis is $\{AA, ABA.BAB\}$. Clark obtained that the 3-avoidance basis is:

- AA ($\lambda = 3$ [50])
- $ABA.BAB$ ($\lambda = 3$ [11])
- $ABCA.BCAB.CABC$ ($\lambda = 3$ [22])
- $ABCBA.CBABC$ ($\lambda = 2$ [22])
- $ABCA.CABC.BCB$ ($\lambda = 3$ [22])
- $ABCA.BCAB.CBC$ ($\lambda = 3$, reverse of $ABCA.CABC.BCB$)
- $AB.AC.BA.CA.CB$ ($\lambda = 4$ [2])

The set of all formulas that belong to the n-avoidance basis (for some n) is simply called the ∞-avoidance basis. The ∞-avoidance basis gathers the avoidable formulas that are the hardest to avoid, in some sense. We say that a formula in the ∞-avoidance basis is *minimally avoidable*.

Let us describe a practical way to determine whether a formula is minimally avoidable. The operation of *splitting* a formula f on a fragment ϕ consists in replacing ϕ by two fragments, namely the prefix and the suffix of length $|\phi| - 1$ of ϕ. The obtained formula is not necessarily avoidable, and even if it is, it might need to be put in standard form, that is, if one of the new fragments is a factor of another fragment, then it must be deleted. For example, if we split the formula $ABCA.BCAB.BCB.CBA$ on the fragment CBA then we obtain $ABCA.BCAB.BCB.CB.BA$, which gives $ABCA.BCAB.BCB.BA$ since CB is a factor of BCB. Similarly, if we split $ABCA.BCAB.BCB.CBA$ on $BCAB$, then we eventually obtain $ABCA.CAB.BCB.CBA$, and if we split $ABCA.BCAB.BCB.CBA$ on BCB, then we eventually obtain $ABCA.BCAB.CBA$. Notice that if we split a formula f on some fragment to obtain a formula f', then $f' \prec f$. Thus, splitting can be seen as a way of obtaining formulas that are slightly harder to avoid than a given formula. Then, a formula is minimally avoidable if and only if splitting on any of its fragments leads to an unavoidable formula.

Clark identified in the ∞-avoidance basis the family of *circular* formulas consisting of AA, $ABA.BAB$, $ABCA.BCAB.CABC$, $ABCDA.BCDAB.CDABC.DABCD$, ... Checking that any circular formula is minimally avoidable is easy since all the fragments play symmetric roles. So we split $AB \cdots XA$ into $AB \cdots X.B \cdots XA$ and we obtain a formula that divides the pattern $B \cdots XAB \cdots X = Z(B \cdots X)$. This pattern is equivalent to $B \cdots X$, which is unavoidable. Thus, every circular formula is indeed minimally avoidable. We use morphic images of b_4 to obtain the index of every circular formula with at least 3 variables. Since this proof technique involving conjugacy classes seems specific to circular formulas, we do not detail it here and refer to the paper [22] (see Section 6.4). The proofs of $\lambda(ABCBA.CBABC) = 2$ and $\lambda(ABCA.CABC.BCB) = 3$ follow the framework in Section 4.

3 The non-constructive method

One of the first application of the entropy compression method was to prove a generalization of the fact that AA is 4-avoidable in the context of list-coloring [26]. We are not interested in list-coloring, but we present an easy proof, without morphisms, that there exist at least 2^i square-free words of length i over 4 letters. Let n_i be the number of such words of length i. We show by induction on i that $\frac{n_i}{n_{i-1}} \geqslant 2$. We consider the $4n_{i-1}$ words of length i with a square-free prefix of length $i-1$. Some these words contain a square as a suffix. The number of such words of the form uvv with $|v| = j$ is at most the number of square-free prefixes uv of length $i-j$, that is, n_{i-j}.

So we have $n_i \geqslant 4n_{i-1} - \sum_{1\leqslant j\leqslant i/2} n_{i-j}$, which gives $\frac{n_i}{n_{i-1}} \geqslant 4 - \sum_{1\leqslant j\leqslant i/2} \frac{n_{i-j}}{n_{i-1}}$

$$\begin{array}{rcll}\frac{n_i}{n_{i-1}} & \geqslant & 4 - \sum_{1\leqslant j\leqslant i/2} \frac{n_i - j}{n_{i-1}} & \\ & \geqslant & 4 - \sum_{1\leqslant j\leqslant i/2} \frac{1}{2^{j-1}} & \text{by induction} \\ & \geqslant & 4 - \sum_{j\geqslant 0} \frac{1}{2^j} & \\ & = & 2 & \end{array}$$

The full generality of the method is described and used in [36, 27] (see [36] in Section 6.3). We omit the proof of correction of the method, which is a generalization of the arguments in our example. We only describe how to use the method.

Let $L \subset \Sigma_m^*$ be a factorial language defined by a set F of forbidden factors of length at least 2. We define L' as the set of words w such that w is not in L and the prefix of length $|w|-1$ of w is in L. For every forbidden factor $f \in F$, we choose a number $1 \leqslant s_f \leqslant |f|$. Then, for every $i \geqslant 1$, we define an integer a_i such that

$$a_i \geqslant \max_{u\in L} \left|\left\{v \in \Sigma_m^i \mid uv \in L',\ uv = bf,\ f \in F,\ s_f = i\right\}\right|.$$

We consider the formal power series $P(x) = 1 - mx + \sum_{i\geqslant 1} a_i x^i$. If $P(x)$ has a positive real root x_0, then $n_i \geqslant x_0^{-i}$ for every $i \geqslant 0$.

In our example, we use $m = 4$, $s_f = \frac{|f|}{2}$. Given a prefix w, there exists at most one way to extend w with i letters in order to create a suffix square of length $2i$. So we take $a_i = 1$ for every $i \geqslant 1$. We obtain $P(x) = 1 - 4x + \sum_{i\geqslant 1} x^i = \frac{1}{1-x} - 4x = \frac{(1-2x)^2}{1-x}$ with root $x_0 = \frac{1}{2}$. We can conclude that $n_i \geqslant 2^i$. In particular, $\lambda(AA) \leqslant 4$.

The novelty of this method compared to the generating function method is the parameter s_f, which allows more flexibility to estimate the number of words with some forbidden suffix. Indeed, the generating function method corresponds to the case $s_f = |f|$. As noted by Rampersad, doing so leads to the weaker result that $\lambda(AA) \leqslant 7$.

Another example is about binary words avoiding sufficiently large squares. Let SQ_t denote the pattern corresponding to squares with period at least t, that is, $SQ_1 = AA$, $SQ_2 = ABAB$, $SQ_3 = ABCABC$, and so on.

We know that $\lambda(SQ_2) = 3$ and $\lambda(SQ_3) = 2$. Our method easily shows that $\lambda(SQ_5) = 2$. We use $m = 2$ and $s_f = \frac{|f|}{2}$, so that $a_i = 1$ for every $i \geqslant 5$. We obtain $P(x) = 1 - 2x + \sum_{i\geqslant 5} x^i = 1 - 2x + \frac{x^5}{1-x} = \frac{1-3x+2x^2+x^5}{1-x}$, which has the positive root $x_0 = 0.56984\ldots$.

If we use $s_f = |f|$ and try to avoid SQ_t on the binary alphabet, then the default bound is that there at most 2^i squares with period i. This gives $P(x) = 1 - 2x + \sum_{i\geqslant t} 2^i x^{2i} = 1 - 2x + \frac{(2x^2)^t}{1-2x^2}$, which admits a root for $t \geqslant 7$. To summarize, using $s_f = |f|$ we can prove $\lambda(SQ_7) = 2$, using $s_f = \frac{|f|}{2}$ we can prove $\lambda(SQ_5) = 2$, and we obtain $\lambda(SQ_3) = 2$ with morphisms.

These examples show that the ability to go beyond the case $s_f = |f|$ may allow to get a tighter result but often does not reach an optimal result. Within our framework of pattern avoidance, this method has been successfully applied to doubled patterns only, see Theorem 7. Again, tighter bounds are conjectured:

Conjecture 11. There exist finitely many doubled patterns with index 3, and thus the other doubled patterns are 2-avoidable.

4 Nice formulas

In this section, we consider another useful tool in pattern avoidance that has been defined in [36] (See Section 6.3) and already used implicitly in [34]. The *avoidability exponent* $AE(p)$ of a pattern p is the largest real α such that every α-free word avoids p. We extend this definition to formulas.

Let us show that $AE(ABCBA.CBABC) = \frac{4}{3}$. Suppose for contradiction that a $\frac{4}{3}$-free word contains an occurrence h of $ABCBA.CBABC$. In the rest of this section, we write $y = |h(Y)|$ for every variable Y. The factor $h(ABCBA)$ is a repetition with period $|h(ABCB)|$. So we have $\frac{a+b+c+b+a}{a+b+c+b} < \frac{4}{3}$. This simplifies to $2a < 2b + c$. Similarly, $CBABC$ gives $2c < a + 2b$, BAB gives $2b < a$, and BCB gives $2b < c$. Summing up these four inequalities gives $2a+4b+2c < 2a+4b+2c$, which is a contradiction. On the other hand, the word 0123420156786587683420123 4 is $\left(\frac{4}{3}^+\right)$-free and contains an occurrence of $ABCBA.CBABC$ with $A = 01$, $B = 2$, and $C = 34$. As an exercice, prove that $AE(ABCA.CABC.BCB) = \frac{5-\sqrt{5}}{2}$.

Of course, the avoidability exponent is related to divisibility.

Lemma 12. *If $f \preceq g$, then $AE(f) \leqslant AE(g)$.*

The avoidability exponent depends on the repetitions induced by f. We have $AE(f) = 1$ for formulas such as $f = AB.BA.AC.CA.BC$ or $f = AB.BA.AC.BC.CDA.DCD$ that do not have enough repetitions. That is, for every $\varepsilon > 0$, there exists a $(1 + \varepsilon)$-free word that contains an occurrence of f.

Let us investigate formulas with non-trivial avoidability exponent, that is, $AE(f) > 1$. A formula f is *nice* if for every variable X of f there exists a fragment of f that contains X at least twice. To show that a nice formula has a non-trivial avoidability exponent (see Lemma 13), we first introduce a notion of minimality for nice formulas similar to the avoidance basis for all formulas. A nice formula f is *minimally nice* if there exists no nice formula g such that $v(g) \leqslant v(f)$ and $g \prec f$. Alternatively, splitting a minimally nice formula on any of its fragments leads to a non-nice formula. The following property of every minimally nice formula is easy to derive. If a variable V appears as a prefix of a fragment ϕ, then

- V is also a suffix of ϕ (since otherwise we can split on ϕ and obtain a nice formula),
- ϕ contains exactly two occurrences of V (since otherwise we can remove the prefix letter V from ϕ and obtain a nice formula),
- V is neither a prefix nor a suffix of any fragment other than ϕ (since otherwise we can remove this prefix/suffix letter V from the other fragment and obtain a nice formula),
- Every fragment other than ϕ contains at most one occurrence of V (since otherwise we can remove the prefix letter V from ϕ and obtain a nice formula).

Lemma 13. *If f is a nice formula, then $AE(f) \geqslant 1 + 2^{1-v(f)}$.*

Proof. The lemma is obvious if $v(f) = 1$. Suppose that f contradicts the lemma. Since $1 + 2^{1-v(f)}$ is decreasing with $v(f)$, we can assume that f is a minimally nice formula by Lemma 12.

Then there exists a $\left(1 + 2^{1-v(f)}\right)$-free word w containing an occurrence h of f. Let X be a variable of f such that $|h(X)| \geqslant |h(Y)|$ for every variable Y. Thus, for every sequence s of variables, $|h(s)| \leqslant |s| \times |h(X)|$. Since f is nice, f contains a factor of the form XzX. Without loss of generality, z does not contain X, so that $v(z) \leqslant v(f) - 1$.

If $|z| \geqslant 2^{v(z)}$, then z contains a doubled pattern d with at most $v(z)$ variables. This contradicts the fact that f is minimally nice.

If $|z| \leqslant 2^{v(z)}-1$, then the exponent of $h(XzX)$ in w is $\frac{|h(XzX)|}{|h(Xz)|} = 1+\frac{|h(X)|}{|h(Xz)|} \geqslant 1+\frac{|h(X)|}{|Xz|\times|h(X)|} = 1+\frac{1}{|Xz|} \geqslant 1+\frac{1}{2^{v(z)}} \geqslant 1+\frac{1}{2^{v(f)-1}} = 1+2^{1-v(f)}$. This contradicts that w is $\left(1+2^{1-v(f)}\right)$-free. □

We will describe below a method to construct infinite words avoiding a formula. This method can be applied if and only if the formula f satisfies $AE(f) > 1$. So we are interested in characterizing the formulas f such that $AE(f) > 1$. By Lemmas 12 and 13, if f is a formula such that there exists a nice formula g satisfying $g \preceq f$, then $AE(f) > 1$. Now we prove that the converse also holds, which gives the following characterization.

Lemma 14. *A formula f satisfies $AE(f) > 1$ if and only if there exists a nice formula g such that $g \preceq f$.*

Proof. What remains to prove is that for every formula f that is not divisible by a nice formula and for every $\varepsilon > 0$, there exists an infinite $(1+\varepsilon)$-free word w containing an occurrence of f, such that the size of the alphabet of w only depends on f and ε.

First, we consider the equivalent pattern p obtained from f by replacing every dot by a distinct variable that does not appear in f. We will actually construct an occurrence of p. Then we construct a family f_i of pseudo-formulas as follows. We start with $f_0 = p$. To obtain f_{i+1} from f_i, we choose a variable that appears at most once in every fragment of f_i. This variable is given the alias name V_i and every occurrence of V_i is replaced by a dot. We say that f_i is a pseudo-formula since we do not try to normalize f_i, that is, f_i can contain consecutive dots and f_i can contain fragments that are factors of other fragments. However, we still have a notion of fragment for a pseudo-formula. Since f is not divisible by a nice formula, this process ends with the pseudo-formula $f_{v(p)}$ with no variable and $|p|$ consecutive dots. The goal of this process is to obtain the ordering V_0, V_1, …, $V_{v(p)-1}$ on the variables of p.

The image of every V_i is a finite factor w_i of a Dejean word over an alphabet of $\lfloor\varepsilon^{-1}\rfloor+2$ letters, so that w_i is $(1+\varepsilon)$-free. The alphabets are disjoint: if $i \neq j$, then w_i and w_j have no common letter. Finally, we define the length of w_i as follows: $\left|w_{v(p)-1}\right| = 1$ and $|w_i| = \lfloor\varepsilon^{-1}\rfloor \times |p| \times |w_{i+1}|$ for every i such that $0 \leqslant i \leqslant v(p)-2$. Let us show by contradiction that the constructed occurrence h of p is $(1+\varepsilon)$-free. Consider a repetition xyx of exponent at least $1+\varepsilon$ that is maximal, that is, which cannot be extended to a repetition with the same period and larger exponent. Since every w_i is $(1+\varepsilon)$-free and since two matching letters must come from distinct occurrences of the same variable, then $x = h(x')$ and $y = h(y')$ where x' and y' are factors of p. Our ordering of the variables of p implies that y' contains a variable V_i such that $i < j$ for every variable V_j in x'. Thus, $|y| \geqslant |w_i| = \lfloor\varepsilon^{-1}\rfloor \times |p| \times |w_{i+1}| \geqslant \lfloor\varepsilon^{-1}\rfloor \times |x|$, which contradicts the fact that the exponent of xyx is at least $1+\varepsilon$.

To obtain the infinite word w, we can insert our occurrence of p into a bi-infinite $(1+\varepsilon)$-free word over an alphabet of $\lfloor\varepsilon^{-1}\rfloor+2$ new letters. So w is an infinite $(1+\varepsilon)$-free word over an alphabet of $v(p)\left(\lfloor\varepsilon^{-1}\rfloor+2\right)+1$ letters which contains an occurrence of f. □

The bound in Lemma 13 is probably very far from optimal. The circular formulas show that $AE(f)$ can be as low as $1+(v(f))^{-1}$. However, lower avoidability exponents exist among nice formulas with at least 4 variables, such as $AE(ABCDBACBD) = 1.246266172\ldots$.

Let us detail how this value can be obtained. When we consider a repetition uvu in an α-free word, we derive that $\frac{|uvu|}{|uv|} < \alpha$, which gives $\beta|u| < |v|$ with $\alpha = 1+\frac{1}{\beta+1}$. We consider an occurrence

h of the pattern. The maximal repetitions in $ABCDBACBD$ are $ABCDBA$, $BCDB$, $BACB$, $CDBAC$, and $DBACBD$. They imply the following inequalities.

$$\begin{cases} \beta a \leqslant 2b + c + d \\ \beta b \leqslant c + d \\ \beta b \leqslant a + c \\ \beta c \leqslant a + b + d \\ \beta d \leqslant a + 2b + c \end{cases}$$

We look for the smallest β such that this system has no solution. Notice that a and d play symmetric roles. Thus, we can set $a = d$ and simplify the system.

$$\begin{cases} \beta a \leqslant a + 2b + c \\ \beta b \leqslant a + c \\ \beta c \leqslant 2a + b \end{cases}$$

Then β is the largest eigenvalue of the matrix $\left[\begin{smallmatrix} 1 & 1 & 2 \\ 2 & 0 & 1 \\ 1 & 1 & 0 \end{smallmatrix}\right]$ that corresponds to the latter system. So $\beta = 3.060647027\ldots$ is the largest root of the characteristic polynomial $x^3 - x^2 - 5x - 4$. Then $\alpha = 1 + \frac{1}{\beta+1} = 1.246266172\ldots$

This matrix approach is a convenient trick to use when possible. It was used in particular for some doubled patterns such that every variable occurs exactly twice [36] (See Section 6.3). It may fail if the number of inequalities is strictly greater than the number of variables or if the formula contains a repetition uvu such that $|u| \geqslant 2$. In any case, we can fix a rational value to β and ask a computer algebra system whether the system of inequalities is solvable. Then we can get arbitrarily good approximations of β (and thus α) by a dichotomy method.

Recall that the repetition threshold $RT(n)$ is the smallest real number α such that there exists an infinite a^+-free word over Σ_n. The proof of Dejean's conjecture established that $RT(2) = 2$, $RT(3) = \frac{7}{5}$, $RT(4) = \frac{7}{4}$, and $RT(n) = \frac{n}{n-1}$ for every $n \geqslant 5$. An infinite $RT(n)^+$-free word over Σ_n is called a Dejean word. By Lemma 13, every nice formula is avoidable since it is avoided by a Dejean word over a sufficiently large alphabet.

Thus, if a formula is nice and minimally avoidable, then it is minimally nice. This is the case for every formula in the 3-avoidance basis, except $AB.AC.BA.CA.CB$. However, a minimally nice formula is not necessarily minimally avoidable. Indeed, we have shown [39] (see Section 6.1) that the set of minimally nice ternary formulas consists of the nice formulas in the 3-avoidance basis, together with the minimally nice formulas in Table 1 that are divisible by $AB.AC.BA.CA.CB$.

Avoiding a nice formula Recall that a nice formula f is such that $AE(f) > 1$. We consider the smallest integer n such that $RT(n) < AE(f)$. Thus, every Dejean word over Σ_n avoids f, which already gives $\lambda(f) \leqslant n$. Recall that a morphism is q-uniform if the image of every letter has length q. Also, a uniform morphism $h : \Sigma_s^* \to \Sigma_e^*$ is *synchronizing* if for any $a, b, c \in \Sigma_s$ and $v, w \in \Sigma_e^*$, if $h(ab) = vh(c)w$, then either $v = \varepsilon$ and $a = c$ or $w = \varepsilon$ and $b = c$. For increasing values of q, we look for a q-uniform morphism $m : \Sigma_n^* \to \Sigma_k^*$ such that $m(w)$ avoids f for every $RT(n)^+$-free word $w \in \Sigma_n^\ell$, where ℓ is given by Lemma 15 below.

Lemma 15. *[34] Let $\alpha, \beta \in \mathbb{Q}$, $1 < \alpha < \beta < 2$ and $n \in \mathbb{N}^*$. Let $h : \Sigma_s^* \to \Sigma_e^*$ be a synchronizing q-uniform morphism (with $q \geqslant 1$). If $h(w)$ is (β^+, n)-free for every α^+-free word w such that $|w| < \max\left(\frac{2\beta}{\beta-\alpha}, \frac{2(q-1)(2\beta-1)}{q(\beta-1)}\right)$, then $h(t)$ is (β^+, n)-free for every (finite or infinite) α^+-free word t.*

- $ABA.BCB.CAC$
- $ABCA.BCAB.CBAC$ and its reverse
- $ABCA.BAB.CAC$
- $ABCA.BAB.CBC$ and its reverse
- $ABCA.BAB.CBAC$ and its reverse
- $ABCBA.CABC$ and its reverse
- $ABCBA.CAC$

Table 1: The minimally nice ternary formulas that are not minimally avoidable.

Given such a candidate morphism m, we use Lemma 15 to show that for every $RT(n)^+$-free word $w \in \Sigma_n^*$, the image $m(w)$ is (β^+, t)-free. The pair (β, t) is chosen such that $RT(n) < \beta < AE(f)$ and t is the smallest possible for the corresponding β. If $\beta < AE(f)$, then every occurrence h of f in a (β^+, t)-free word is such that the length of the h-image of every variable of f is bounded from above by a function of t and f only. Thus, the h-image of every fragment of f has bounded length and we can check that f is avoided by inspecting a finite set of factors of words of the form $m(w)$.

The index of a nice formula So what can be said about the index of a nice formula? So far, all the nice formulas that have been considered are 3-avoidable. This includes doubled patterns [36] (See Section 6.3), circular formulas [22] (see Section 6.4), the nice formulas in the 3-avoidance basis [22] (see Section 6.4), and the minimally nice ternary formulas in Table 1 [39] (see Section 6.1). As mentioned earlier, the last two results cover all the minimally nice ternary formulas and thus all the nice ternary formulas.

Theorem 16 ([22, 39]). *Every nice formula with at most* 3 *variables is* 3*-avoidable.*

We have a risky conjecture that would generalize both Theorems 7 and 16.

Conjecture 17. Every nice formula is 3-avoidable.

Let us show that Conjecture 17 would be best possible, in the sense that there exists infinitely many nice formulas with index 3. This also implies that Conjecture 11 cannot be generalized to nice formulas. For every $i \geqslant 2$, let T_i be the formula such that $T_2 = ABA.BAB$, $T_3 = ABA.BCB.CAC$, $T_4 = ABA.BCB.CDC.DAD$, and so on. More formally, T_i is the formula with i variables $A_0, \ldots, A_{i-1}$ which contains the i fragments of length three of the form $A_jA_{j+1}A_j$ such that the indices are taken modulo i.

Theorem 18. *For every $i \geqslant 2$, $\lambda(T_i) = 3$*

Proof. We have applied the method in this section to prove that the image of every $(7/4^+)$-free word over Σ_4 by the following 26-uniform morphism avoids $ABA.BAB$ [38] (see Section 6.2).

$$\begin{aligned}
0 &\mapsto \mathtt{0012110220011202110022012 1} \\
1 &\mapsto \mathtt{001122002110012022101220 21} \\
2 &\mapsto \mathtt{001120221100122002101202 21} \\
3 &\mapsto \mathtt{001120021100122021012002 21}
\end{aligned}$$

The useful property of these ternary words is that they are $(3/2, 3)$-free. In these words, for every occurrence m of ABA with $|m(B)| \leqslant |m(A)|$, we thus have $|m(A)| = |m(B)| = 1$. It is easy to check that the set of these occurrences is $\{101, 121, 202\}$. This implies that these ternary words also avoid T_i for every $i \geqslant 2$. This proves $\lambda(T_i) \leqslant 3$.

To show that T_i is not 2-avoidable, we consider the formula $H = ABA.BAB.ACA.CAC.BCB.CBC$. Standard backtracking shows that $\lambda(H) > 2$. The following mapping m shows that $T_i \preceq H$:

- If j is odd, then $m(A_j) = B$.
- If i is odd, then $m(A_{i-1}) = C$.
- Otherwise, $m(A_j) = A$.

Let us take $i = 7$ as an example. So $m(A_1) = m(A_3) = m(A_5) = B$, $m(A_6) = C$, and $m(A_0) = m(A_2) = m(A_4) = B$, which gives

$$\begin{aligned}
m(T_7) &= m(A_0A_1A_0.A_1A_2A_1.A_2A_3A_2.A_3A_4A_3.A_4A_5A_6.A_6A_0A_6) \\
&= ABA.BAB.ABA.BAB.ABA.BCB.CAC \\
&= ABA.BAB.BCB.CAC \\
&\preceq H.
\end{aligned}$$

By Lemma 2.3, $\lambda(T_i) \geqslant \lambda(H) \geqslant 3$. □

The number of fragments of a minimally avoidable formula Interestingly, the notion of (minimally) nice formula is helpful in proving the following.

Theorem 19. *The only formula with exactly one fragment in the ∞-avoidance basis is AA.*

Proof. A formula with one fragment is a doubled pattern. Since it is in the ∞-avoidance basis, it is a minimally nice formula. By the properties of minimally nice formulas discussed above, the unique fragment of the formula is either AA or is of the form ApA such that p does not contain the variable A. Thus, p is a doubled pattern such that $p \prec ApA$, which contradicts that ApA is minimally avoidable. □

By contrast, the ∞-avoidance basis contains infinitely many formulas with exactly two fragments. In particular, it contains the family of *two-birds* formulas, which consists of $ABA.BAB$, $ABCBA.CBABC$, $ABCDCBA.DCBABCD$, and so on. Every two-birds formula is nice. Let us check that every two-birds formula $AB\cdots X\cdots BA.X\cdots A\cdots X$ is minimally avoidable. Since the two fragments play symmetric roles, it is sufficient to split on the first fragment. We obtain the formula $AB\cdots X\cdots B.B\cdots X\cdots BA.X\cdots A\cdots X$ which divides the pattern $B\cdots X\cdots BAB\cdots X\cdots B = Z(B\cdots X\cdots B)$. This pattern is equivalent to $B\cdots X\cdots B$, which is unavoidable. Thus, every two-birds formula is indeed minimally avoidable.

Concerning the index of two-birds formulas, we have seen that $\lambda(ABA.BAB) = 3$ and $\lambda(ABCBA.CBABC) = 2$. Computer experiments suggest that larger two-birds formulas are easier to avoid.

Conjecture 20. Every two-birds formula with at least 3 variables is 2-avoidable.

5 Few avoiding words

A first remark is that the topic of "few infinite avoiding words" is limited to the alphabet $\Sigma_{\lambda(f)}$:

Remark 21. For every avoidable formula f, there exists at least $2^{\frac{n}{\lambda(f)}}$ words avoiding f of length n over $\Sigma_{\lambda(f)+1}$.

We have seen in Section 2 that:

1. Cassaigne conjectures that for every avoidable formula f, there exists a morphic word avoiding f over $\Sigma_{\lambda(f)}$.
2. Cassaigne's algorithm can usually test whether a given morphic word avoids a given formula.
3. For some formulas f, there exists a finite set S of morphic words that essentially avoids f.

In this Section, we focus on proof techniques for the third item. In the known examples, $|S|$ is either 1, 2, 3 or 4. So we are in the following situation: we consider a formula f and we have identified a candidate set S. Using Cassaigne's algorithm, we have also proved that every word in S actually avoids f. There remains to show, by contrapositive, that every recurrent word over $\Sigma_{\lambda(f)}$ that is not equivalent to a word in S contains an occurrence of f.

We will see that we can consider every word in S separately. As seen in Section 2, $\{g_x(b_3), g_t(b_3)\}$ essentially avoids $f = ABA.AABB$.

$$\begin{aligned} g_x(\mathtt{0}) &= \mathtt{01110}, & g_t(\mathtt{0}) &= \mathtt{01011011010},\\ g_x(\mathtt{1}) &= \mathtt{0110}, & g_t(\mathtt{1}) &= \mathtt{01011010},\\ g_x(\mathtt{2}) &= \mathtt{0}. & g_t(\mathtt{2}) &= \mathtt{010}. \end{aligned}$$

We check by backtracking that no infinite binary word avoids f, 010, and 0011. Obviously, no word avoiding f contains both 010 and 0011. Thus, there are a priori at most two kinds of infinite binary words avoiding f, depending on whether they avoid 010 or 0011. Then it suffices to show separately that

1. $g_x(b_3)$ essentially avoids f and 010,
2. $g_t(b_3)$ essentially avoids f and 0011.

Let us show that $g_x(b_3)$ essentially avoids f and 010. We see that $g_x(b_3)$ avoids 010 and we use Cassaigne's algorithm to show that $g_x(b_3)$ avoids f. Actually, Cassaigne did it himself in his Ph.D. [11] since $g_x(b_3)$ is said to avoid the pattern $ABACAABB$. Then, we check by computer that no infinite binary word avoids f, 010, and the prefix of length 500 of $g_x(b_3)$. This shows that every infinite binary word as the same set of recurrent factors of length 50 as $g_x(b_3)$. In particular, if w_2 is a recurrent binary word starting with $g_x(\mathtt{0})$, then $w_2 = g_x(w_3)$ for some ternary word w_3. Moreover, w_3 as the same set of factors of length 10 as b_3. So, w_3 avoids 010, 212, and squares XX with $|X| \leqslant 3$. Now w_3 must also avoid factors of the form $\mathtt{2}YY$, since for every $p, s \in \Sigma_3$, the word $g_x(p\mathtt{2}YYs) = V\mathtt{000}U\mathtt{0}U\mathtt{0}W$ contains the occurrence $A \to \mathtt{0}$, $B \to \mathtt{0}U$ of $f = ABA.AABB$. Then we use the following result.

Theorem 22 ([38], see Section 6.2)**.** *b_3 essentially avoids 010, 212, XX with $1 \leqslant |X| \leqslant 3$, and $\mathtt{2}YY$ with $|Y| \geqslant 4$.*

Thus, w_3 is equivalent to b_3. This shows that $g_x(b_3)$ essentially avoids f and 010.

Classifying all the ternary formulas, as we did for binary formulas, does not seem feasible due to the large number of ternary formulas that must be considered. Thus, we decided to focus on finding ternary formulas with an interesting set of avoiding words.

First, we wanted to characterize b_3 with only one forbidden formula and no forbidden factor. This is not possible since the set of factors of b_3 is not closed by letter permutation. Let b'_3 (resp. b''_3) be the word obtained from b_3 by exchanging 1 and 0 (resp. 1 and 2). Thus, if b_3 avoids a formula, then so do b'_3 and b''_3.

We consider the following formulas.

- $f_b = ABCAB.ABCBA.ACB.BAC$
- $f_1 = ABCA.BCAB.BCB.CBA$
- $f_2 = ABCAB.BCB.AC$
- $f_3 = ABCA.BCAB.ACB.BCB$
- $f_4 = ABCA.BCAB.BCB.AC.BA$

Theorem 23 ([39], see Section 6.1)**.** *Let $f \in \{f_b, f_1, f_2, f_3, f_4\}$. $\{b_3, b'_3, b''_3\}$ essentially avoids f.*

By considering the fomulas obtained by divisibility and reverse, we obtain 144 non-equivalent formulas f such that $\{b_3, b'_3, b''_3\}$ essentially avoids f.

We have no similar result for b_2 or b_5.

Question 24. *Does there exist a formula f such that b_2 essentially avoids f?*

So far, for every formula f that is avoided by polynomially many words, there actually exists a finite set of morphic words that essentially avoids f. Our next result shows that there exists another kind of formula that is avoided by polynomially many words.

We consider the morphisms $m_a : 0 \mapsto 001,\ 1 \mapsto 101$ and $m_b : 0 \mapsto 010,\ 1 \mapsto 110$. That is, $m_a(x) = x01$ and $m_b(x) = x10$ for every $x \in \Sigma_2$. We construct the set W of binary words as follows:

- $0 \in W$.
- If $v \in W$, then $m_a(v) \in W$ and $m_b(v) \in W$.
- If $v \in W$ and v' is a factor of v, then $v' \in W$.

Theorem 25 ([39], see Section 6.1)**.** *Let $f \in \{ABACA.ABCA, ABAC.BACA.ABCA\}$. The set of words u such that u is recurrent in an infinite binary word avoiding f is W.*

From the recursive definition of W, we obtain that the factor complexity of W is $\Theta(n^\alpha)$ where $\alpha = \frac{\ln 6}{\ln 3} = 1 + \frac{\ln 2}{\ln 3} \approx 1.6309$. Devyatov [19] has recently shown that the factor complexity of a morphic word is either $O(n \ln(n))$ or $\Theta\left(n^{1+1/k}\right)$ for some integer $k \geqslant 1$. Thus, the two formulas in Theorem 25 are avoided by polynomially many words. However, no finite set of morphic words essentially avoids them.

Let us finish this section with another risky conjecture. We have seen that the nice formula $ABA.AABB$ is essentially avoided by two binary morphic words. This implies that its index cannot be determined with the method in Section 4 using images of arbitrary Dejean words, since it implies an exponential complexity. However, notice that $ABA.AABB$ and the other similar binary formulas are not minimally nice.

Conjecture 26. Every doubled pattern and every minimally nice formula is avoided by exponentially many words.

References

[1] J.P. Allouche, J. Cassaigne, J. Shallit, and L.Q. Zamboni. A taxonomy of morphic sequences. arXiv:1711.10807

[2] K.A. Baker, G.F. McNulty, and W. Taylor. Growth problems for avoidable words. *Theor. Comput. Sci.* **69(3)** (1989), 319–345.

[3] G. Badkobeh and P. Ochem. Characterization of some binary words with few squares. *Theor. Comput. Sci.* **588** (2015), 73–80.

[4] D.R. Bean, A. Ehrenfeucht, and G.F. McNulty. Avoidable patterns in strings of symbols. *Pacific J. of Math.* **85** (1979), 261–294.

[5] J. Bell, T.L. Goh. Exponential lower bounds for the number of words of uniform length avoiding a pattern. *Inform. and Comput.* **205** (2007), 1295-1306.

[6] J. Berstel. Mots sans carré et morphismes itérés. *Discrete Mathematics* **29(3)** (1980), 235–244.

[7] J. Berstel. Axel Thue's papers on repetitions in words: a translation. *Départements de mathématiques et d'informatique, Université du Québec à Montréal* **20** (1995).

[8] K. Black, F. Blanchet-Sadri, I. Coley, B. Woodhouse, and A. Zemke. Pattern avoidance in partial words dense with holes. *Journal of Automata, Languages and Combinatorics* **22(4)** (2017), 209–241

[9] F. Blanchet-Sadri, B. Woodhouse. Strict bounds for pattern avoidance. *Theor. Comput. Sci.* **506** (2013), 17–27.

[10] A. Carpi. On Dejean's conjecture over large alphabets. *Theor. Comput. Sci.* **385(1–3)** (2007), 137–151.

[11] J. Cassaigne. *Motifs évitables et régularité dans les mots.* PhD thesis, Université Paris VI, 1994.

[12] J. Cassaigne, J.D. Currie, L. Schaeffer, and J. Shallit: Avoiding three consecutive blocks of the same size and same sum. *J. ACM* **61(2)** (2014) 10:1–10:17

[13] R.J. Clark. *Avoidable formulas in combinatorics on words.* PhD thesis, University of California, Los Angeles, 2001. Available at
http://www.lirmm.fr/~ochem/morphisms/clark_thesis.pdf

[14] M. Crochemore. Sharp characterizations of squarefree morphisms. emphTheor. Comput. Sci. **18(2)** (1982), 221–226.

[15] J.D. Currie and Ph. Lafrance. Avoidability index for binary patterns with reversal. *Electron. J. Comb.* **23(1)** (2016), #P1.36.

[16] J.D. Currie, L. Mol, and N. Rampersad: Avoidance bases for formulas with reversal emphTheor. Comput. Sci. **738** (2018), 25–41.

[17] J.D. Currie and N. Rampersad. Growth rate of binary words avoiding xxx^R. emphTheor. Comput. Sci. **609** (2016), 456–468.

[18] F. Dejean. Sur un théorème de Thue. *J. Combin. Theory. Ser. A* **13** (1972), 90–99.

[19] R. Devyatov. On Subword Complexity of Morphic Sequences. arXiv:1502.02310

[20] V. Dujmović, L. Esperet, G. Joret, B. Walczak, and D. Wood. Planar Graphs Have Bounded Nonrepetitive Chromatic Number. arXiv:1904.05269

[21] S.B. Ekhad and D. Zeilberger. There are more than $2^{n/17}$ n-letter ternary square-free words. *Journal of Integer Sequences* **1** (1998), Article 98.1.9

[22] G. Gamard, P. Ochem, G. Richomme, and P. Séébold. Avoidability of circular formulas. *Theor. Comput. Sci.* **726** (2018), 1–4.

[23] D. Gonçalves, M. Montassier, and A. Pinlou. Entropy compression method applied to graph colorings. arXiv:1406.4380

[24] U. Grimm. Improved Bounds on the Number of Ternary Square-Free Words. *Journal of Integer Sequences* **4** (2001), Article 01.2.7

[25] J. Grytczuk. Pattern avoidance on graphs. *Discrete Math.* **307(1112)** (2007), 1341-1346.

[26] J. Grytczuk, J. Kozik, P. Micek. New approach to nonrepetitive sequences. *Random Structures & Algorithms* **42** (2013), 214–225.

[27] G. Guégan and P. Ochem. A short proof that shuffle squares are 7-avoidable. *RAIRO - Theoret. Informatics Appl.* **50(1)** (2016), 101–103.

[28] G. Guégan, K. Knauer, J. Rollin, and T. Ueckerdt. The interval number of a planar graph is at most three. arXiv:11805.02947

[29] L. Ilie, P. Ochem, and J.O. Shallit. A generalization of repetition threshold. *Theor. Comput. Sci.* **92(2)** (2004), 71–76.

[30] V. Keränen Abelian squares are avoidable on 4 letters. ICALP 1992. LNCS **623** (1992).

[31] R. Kolpakov and M. Rao. On the number of Dejean words over alphabets of 5, 6, 7, 8, 9 and 10 letters. *Theor. Comput. Sci.* **412(46)** (2011), 6507–6516.

[32] M. Lothaire. *Algebraic Combinatorics on Words.* Cambridge Univ. Press, 2002.

[33] J. Moulin Ollagnier. Proof of Dejean's conjecture for alphabets with 5, 6, 7, 8, 9, 10 and 11 letters. *Theor. Comput. Sci.* **95(2)** (1992), 187–205.

[34] P. Ochem. A generator of morphisms for infinite words. *RAIRO - Theor. Inform. Appl.* **40** (2006), 427–441.

[35] P. Ochem. Binary words avoiding the pattern AABBCABBA. *RAIRO - Theoret. Informatics Appl.* **44(1)** (2010), 151–158.

[36] P. Ochem. Doubled patterns are 3-avoidable. *Electron. J. Comb.* **23(1)** (2016), #P1.19.

[37] P. Ochem and A. Pinlou. Application of entropy compression in pattern avoidance. *Electron. J. Comb.* **21(2)** (2014), #RP2.7.

[38] P. Ochem and M. Rosenfeld. Avoidability of formulas with two variables. Proceedings of the 20th international Conference, *DLT* 2016, Montréal, Lect. Notes Comput. Sci. 9840:344-354, S. Brlek and C. Reutenauer eds., 2016. *Electron. J. Comb.* **24(4)** (2017), #P4.30.

[39] P. Ochem and M. Rosenfeld. On some interesting ternary formulas. In *6th International Conference on Words (Words 2017)*, Montreal, Canada, September 11-15 2017. *Electron. J. Comb.* **26(1)** (2019), #RP1.12.

[40] J.-J. Pansiot. A propos d'une conjecture de F. Dejean sur les répétitions dans les mots. *Discrete Appl. Math.* **7(3)** (1984), 297–311.

[41] Pytheas Fogg. Substitutions in Dynamics, Arithmetics and Combinatorics. Springer Science & Business Media, 2002.

[42] M. Rao. Last cases of Dejean's conjecture. *Theor. Comput. Sci.* **412(27)** (2011), 3010–3018.

[43] M. Rao. On some generalizations of abelian power avoidability. *Theor. Comput. Sci.* **601** (2015), 39–46.

[44] N. Rampersad. Further applications of a power series method for pattern avoidance. *Electron. J. Comb.* **18(1)** (2011), #P134.

[45] P. Roth. Every binary pattern of length six is avoidable on the two-letter alphabet. *Acta Inform.* **29** (1992), 95–107.

[46] C. Richard and U. Grimm. On the entropy and letter frequencies of ternary square-free words, *Electron. J. Comb.* **11** (2004), #R14

[47] M. Rosenfeld. Every binary pattern of length greater than 14 is abelian-2-avoidable. MFCS 2016, 81:1-81:11

[48] A. Shur. Growth rates of complexity of power-free languages. *Theor. Comput. Sci.* **411(34-36)** (2010), 3209–3223.

[49] X. Sun. New Lower Bound On The Number of Ternary Square-Free Words. *Journal of Integer Sequences* **6** (2003), Article 03.3.2

[50] A. Thue. Über unendliche Zeichenreihen. *Norske Vid. Selsk. Skr. I. Mat. Nat. Kl. Christiania* **7** (1906), 1–22.

[51] A.I. Zimin. Blocking sets of terms. *Math. USSR Sbornik* **47(2)** (1984), 353–364. English translation.

6 Materials

6.1 Ternary formulas

On some interesting ternary formulas

Pascal Ochem[*]
LIRMM, CNRS
Université de Montpellier
France
ochem@lirmm.fr

Matthieu Rosenfeld
LIP, ENS de Lyon, CNRS, UCBL
Université de Lyon
France
matthieu.rosenfeld@ens-lyon.fr

Submitted: TBD; Accepted: TBD; Published: TBD

© The authors. Released under the CC BY-ND license (International 4.0).

Abstract

We obtain the following results about the avoidance of ternary formulas. Up to renaming of the letters, the only infinite ternary words avoiding the formula $ABCAB.ABCBA.ACB.BAC$ (resp. $ABCA.BCAB.BCB.CBA$) are the ones that have the same set of recurrent factors as the fixed point of $\texttt{0} \mapsto \texttt{012}$, $\texttt{1} \mapsto \texttt{02}$, $\texttt{2} \mapsto \texttt{1}$. The formula $ABAC.BACA.ABCA$ is avoided by polynomially many binary words (w.r.t. to their lengths) and there exist arbitrarily many infinite binary words with different sets of recurrent factors that avoid it. If every variable of a ternary formula appears at least twice in the same fragment, then the formula is 3-avoidable. The pattern $ABACADABCA$ is unavoidable for the class of C_4-minor-free graphs with maximum degree 3. This disproves a conjecture of Grytczuk. The formula $ABCA.ACBA$, or equivalently the palindromic pattern $ABCADACBA$, has avoidability index 4.

Mathematics Subject Classifications: 68R15

1 Introduction

A *pattern* p is a non-empty finite word over an alphabet $\Delta = \{A, B, C, \ldots\}$ of capital letters called *variables*. An *occurrence* of p in a word w is a non-erasing morphism $h : \Delta^* \to \Sigma^*$ such that $h(p)$ is a factor of w. The *avoidability index* $\lambda(p)$ of a pattern p is the size of the smallest alphabet Σ such that there exists an infinite word over Σ containing no occurrence of p.

A variable that appears only once in a pattern is said to be *isolated*. Following Cassaigne [4], we associate a pattern p with the *formula* f obtained by replacing every isolated variable in p by a dot. For example, the pattern $AABCABBDBBAA$ gives the formula

[*] The authors were partially supported by the ANR project CoCoGro (ANR-16-CE40-0005).

$AAB.ABB.BBAA$. The factors that are separated by dots are called *fragments*. So AAB, ABB, and $BBAA$ are the fragments of $AAB.ABB.BBAA$.

An *occurrence* of a formula f in a word w is a non-erasing morphism $h : \Delta^* \to \Sigma^*$ such that the h-image of every fragment of f is a factor of w. As for patterns, the avoidability index $\lambda(f)$ of a formula f is the size of the smallest alphabet allowing the existence of an infinite word containing no occurrence of f. Clearly, if a formula f is associated with a pattern p, every word avoiding f also avoids p, so $\lambda(p) \leqslant \lambda(f)$. Recall that an infinite word is *recurrent* if every finite factor appears infinitely many times and that any infinite factorial language contains a recurrent word (see Proposition 5.1.13 of [8] for instance). Thus, if there exists an infinite word over Σ avoiding p, then there exists an infinite recurrent word over Σ avoiding p. This recurrent word avoiding p also avoids f, so that $\lambda(p) = \lambda(f)$. Without loss of generality, a formula is such that no variable is isolated and no fragment is a factor of another fragment. We say that a formula f is *divisible* by a formula f' if f does not avoid f', that is, there is a non-erasing morphism h such that the image of any fragment of f' under h is a factor of a fragment of f. If f is divisible by f', then every word avoiding f' also avoids f. Let $\Sigma_k = \{0, 1, \ldots, k-1\}$ denote the k-letter alphabet. We denote by Σ_k^n the k^n words of length n over Σ_k.

A formula is *binary* if it has at most 2 variables. We have recently determined the avoidability index of every binary formula [14]. This exhaustive study led to the discovery of some binary formulas that are avoided by only a few binary words. Determining the avoidability index of every ternary formula would be a huge task. However, we have identified some interesting ternary formulas and this paper describes their properties.

We say that two infinite words are equivalent if they have the same set of factors. Let b_3 be the fixed point of $\texttt{0} \mapsto \texttt{012}$, $\texttt{1} \mapsto \texttt{02}$, $\texttt{2} \mapsto \texttt{1}$. A famous result of Thue [2, 15, 16] can be stated as follows:

Theorem 1. *[2, 15, 16] Every recurrent ternary word avoiding AA, $\texttt{010}$, and $\texttt{212}$ is equivalent to b_3.*

In Section 2, we obtain a similar result for b_3 by forbidding one ternary formula but without forbidding explicit factors in Σ_3^*. In Section 3, we describe the set of binary words avoiding $ABACA.ABCA$ and $ABAC.BACA.ABCA$. We show that these formulas are avoided by polynomially many binary words (w.r.t. to their lengths) and that there exist infinitely many recurrent binary words with different sets of recurrent factors that avoid them. In the terminology of [14], these formulas are not essentially avoided by a finite set of morphic words. In Section 4, we consider *nice* formulas. A formula f is nice if for every variable X of f, there exists a fragment of f that contains X at least twice. This notion generalizes to formulas the notion of a *doubled* pattern (that is, a pattern that contains every variable at least twice). Every doubled pattern is 3-avoidable [13]. We show that every ternary nice formula is 3-avoidable. In Section 5, we show that $ABACADABCA$ is a 2-avoidable pattern that is unavoidable on graphs with maximum degree 3. In Section 6, we show that there exists a palindromic pattern with index 4.

A preliminary version of this paper, without the results in Sections 4 and 6, has been presented at WORDS 2017.

2 Formulas closely related to b_3

For every letter $c \in \Sigma_3$, $\sigma_c : \Sigma_3^* \mapsto \Sigma_3^*$ is the morphism such that $\sigma_c(a) = b$, $\sigma_c(b) = a$, and $\sigma_c(c) = c$ with $\{a, b, c\} = \Sigma_3$. So σ_c is the morphism that fixes c and exchanges the two other letters.

We consider the following formulas.

- $f_b = ABCAB.ABCBA.ACB.BAC$
- $f_1 = ABCA.BCAB.BCB.CBA$
- $f_2 = ABCAB.BCB.AC$
- $f_3 = ABCA.BCAB.ACB.BCB$
- $f_4 = ABCA.BCAB.BCB.AC.BA$

Notice that f_b is divisible by f_1, f_2, f_3, f_4.

Theorem 2. *Let $f \in \{f_b, f_1, f_2, f_3, f_4\}$. Every ternary recurrent word avoiding f is equivalent to b_3, $\sigma_0(b_3)$, or $\sigma_2(b_3)$.*

By considering divisibility, we can deduce that Theorem 2 holds for 72 ternary formulas. Since b_3, $\sigma_0(b_3)$, and $\sigma_2(b_3)$ are equivalent to their reverses, Theorem 2 also holds for the 72 reverse ternary formulas.

Proof. Using Cassaigne's algorithm [3], we have checked that b_3 avoids f_i, for $1 \leqslant i \leqslant 4$. By symmetry, $\sigma_0(b_3)$ and $\sigma_2(b_3)$ also avoid f_i.

Let w be a ternary recurrent word w avoiding f_b. Assume towards a contradiction that w contains a square uu. Then there exists a non-empty word v such that $uuvuu$ is a factor of w. Thus, w contains an occurrence of f_b given by the morphism $A \mapsto u, B \mapsto u, C \mapsto v$. This contradiction shows that w is square-free.

An occurrence h of a ternary formula over Σ_3 is said to be *basic* if $\{h(A), h(B), h(C)\} = \Sigma_3$. As already noticed by Thue [2], no infinite ternary word avoids squares and `012`. So, every infinite ternary square-free word contains the 6 factors obtained by letter permutation of `012`. Thus, an infinite ternary square-free word contains a basic occurrence of f_b if and only if it contains the same basic occurrence of $ABCAB.ABCBA$. Therefore, w contains no basic occurrence of $ABCAB.ABCBA$.

A computer check shows that the longest ternary words avoiding f_b, squares, `021020120`, `102101201`, and `210212012` have length 159. So we assume without loss of generality that w contains `021020120`.

Assume towards a contradiction that w contains `010`. Since w is square-free, w contains `20102`. Moreover, w contains the factor `20120` of `021020120`. So w contains the basic occurrence $A \mapsto$ `2`, $B \mapsto$ `0`, $C \mapsto$ `1` of $ABCAB.ABCBA$. This contradiction shows that w avoids `010`.

Assume towards a contradiction that w contains `212`. Since w is square-free, w contains `02120`. Moreover, w contains the factor `02102` of `021020120`. So w contains the basic occurrence $A \mapsto$ 0, $B \mapsto$ 2, $C \mapsto$ 1 of $ABCAB.ABCBA$. This contradiction shows that w avoids 212.

Since w avoids squares, 010, and 212, Theorem 1 implies that w is equivalent to b_3. By symmetry, every ternary recurrent word avoiding f_b is equivalent to b_3, $\sigma_0(b_3)$, or $\sigma_2(b_3)$. □

3 Avoidability of *ABACA.ABCA* and *ABAC.BACA.ABCA*

Following the terminology in [14], we say that a finite set of infinite words $\mathcal{M}$ *essentially avoids* a formula f if every infinite word over $\Sigma_{\lambda(f)}$ avoiding f has the same set of recurrent factors as a word in $\mathcal{M}$. Let us list all the formulas (up to symetries) from the literature that are known to be essentially avoided by a finite set of words.

- Five binary formulas are known to be essentially avoided by a finite set of binary morphic words [14].
- $\{b_3, \sigma_0(b_3), \sigma_2(b_3)\}$ essentially avoids the ternary formulas in Section 2.
- $\{b_4, b'_4, b''_4\}$ essentially avoids $AB.AC.BA.CA.CB$ [1], where b_4 is the fixed point of 0 $\mapsto$ 01, 1 $\mapsto$ 21, 2 $\mapsto$ 03, 3 $\mapsto$ 23, b'_4 is obtained from b_4 by exchanging 0 and 1, and b''_4 is obtained from b_4 by exchanging 0 and 3.

The formulas listed above are also the only ones known to be avoided by polynomially many words (w.r.t. to their lengths). In this section, we show that the formulas $ABACA.ABCA$ and $ABAC.BACA.ABCA$ behave differently: they are avoided by polynomially many binary words but they are not essentially avoided by a finite set of morphic words.

We consider the morphisms m_a : 0 $\mapsto$ 001, 1 $\mapsto$ 101 and m_b : 0 $\mapsto$ 010, 1 $\mapsto$ 110. That is, $m_a(x) = x$01 and $m_b(x) = x$10 for every $x \in \Sigma_2$. We construct the set S of binary words as follows:

- 0 $\in S$.
- If $v \in S$, then $m_a(v) \in S$ and $m_b(v) \in S$.
- If $v \in S$ and v' is a factor of v, then $v' \in S$.

Theorem 3. *Let $f \in \{ABACA.ABCA, ABAC.BACA.ABCA\}$. The set of words u such that u is recurrent in an infinite binary word avoiding f is S.*

Proof. Let R be the set of words u such that u is recurrent in an infinite binary word avoiding $ABACA.ABCA$. Let R' be the set of words u such that u is recurrent in an infinite binary word avoiding $ABAC.BACA.ABCA$. An occurrence of $ABACA.ABCA$ is also an occurrence of $ABAC.BACA.ABCA$, so that $R' \subseteq R$.

Let us show that $R \subseteq S$. We study the small factors of a recurrent binary word w avoiding $ABACA.ABCA$. Notice that w avoids the pattern $ABAAA$ since it contains the occurrence $A \mapsto A$, $B \mapsto B$, $C \mapsto A$ of $ABACA.ABCA$. Since w contains recurrent factors only, w also avoids AAA.

A computer check shows that the longest binary words avoiding $ABACA.ABCA$, AAA, $\texttt{1001101001}$, and $\texttt{0110010110}$ have length 53. So we assume without loss of generality that w contains $\texttt{1001101001}$.

Assume towards a contradiction that w contains $\texttt{1100}$. Since w avoids AAA, w contains $\texttt{011001}$. Then w contains the occurrence $A \mapsto \texttt{01}, B \mapsto \texttt{1}, C \mapsto \texttt{0}$ of $ABACA.ABCA$. This contradiction shows that w avoids $\texttt{1100}$.

Since w contains $\texttt{0110}$, the occurrence $A \mapsto \texttt{0}, B \mapsto \texttt{1}, C \mapsto \texttt{1}$ of $ABACA.ABCA$ shows that w avoids $\texttt{01010}$. Similarly, w contains $\texttt{1001}$ and avoids $\texttt{10101}$.

Assume towards a contradiction that w contains $\texttt{0101}$. Since w avoids $\texttt{01010}$ and $\texttt{10101}$, w contains $\texttt{001011}$. Moreover, w avoids AAA, so w contains $\texttt{10010110}$. Then w contains the occurrence $A \mapsto \texttt{10}, B \mapsto \texttt{0}, C \mapsto \texttt{1}$ of $ABACA.ABCA$. This contradiction shows that w avoids $\texttt{0101}$.

So w avoids every factor in $\{\texttt{000}, \texttt{111}, \texttt{0101}, \texttt{1100}\}$. Thus, it is not difficult to check that if we extend any factor $\texttt{01}$ in w to three letters to the right, we get either $\texttt{01001}$ or $\texttt{01101}$, that is, $\texttt{01}x\texttt{01}$ with $x \in \Sigma_2$. This implies that w is the m_a-image of some binary word.

Obviously, the image by a non-erasing morphism of a word containing a formula also contains the formula. Thus, the pre-image of w by m_a also avoids $ABACA.ABCA$. This shows that $R \subseteq S$.

Let us show that $S \subseteq R'$, that is, every word in S avoids $ABAC.BACA.ABCA$. Assume towards a contradiction that a finite word $w \in S$ avoids $ABAC.BACA.ABCA$ and that $m_a(w)$ contains an occurrence h of $ABAC.BACA.ABCA$.

If we write $w = w_0w_1w_2w_3\ldots$, then the word $m_a(w) = w_0\texttt{01}w_1\texttt{01}w_2\texttt{01}w_3\texttt{01}\ldots$ is such that:

- Every factor $\texttt{00}$ occurs at position $0 \pmod 3$.
- Every factor $\texttt{01}$ occurs at position $1 \pmod 3$.
- Every factor $\texttt{11}$ occurs at position $2 \pmod 3$.
- Every factor $\texttt{10}$ occurs at position 0 or $2 \pmod 3$, depending on whether the factor $\texttt{1}w_i\texttt{0}$ is $\texttt{100}$ or $\texttt{110}$.

We say that a factor s is *gentle* if either $|s| \geqslant 3$ or $s \in \{\texttt{00}, \texttt{01}, \texttt{11}\}$. By the previous remarks, all the occurrences of the same gentle factor have the same position modulo 3.

First, we consider the case when $h(A)$ is gentle. This implies that the distance between two occurrences of $h(A)$ is $0 \pmod 3$. Since $m_a(w)$ contains the factors $h(ABA)$, $h(ACA)$, and $h(ABCA)$, we deduce that

- $|h(AB)| = |h(A)| + |h(B)| \equiv 0 \pmod 3$.
- $|h(AC)| = |h(A)| + |h(C)| \equiv 0 \pmod 3$.

- $|h(ABC)| = |h(A)| + |h(B)| + |h(C)| \equiv 0 \pmod 3$.

This gives $|h(A)| \equiv |h(B)| \equiv |h(C)| \equiv 0 \pmod 3$. Clearly, such an occurrence of the formula in $m_a(w)$ implies an occurrence of the formula in w, which is a contradiction.

Now we consider the case when $h(B)$ is gentle. If $h(CA)$ is also gentle, then the factors $h(BACA)$ and $h(BCA)$ imply that $|h(A)| \equiv 0 \pmod 3$. Thus, $h(A)$ is gentle and the first case applies. If $h(CA)$ is not gentle, then $h(CA) = \mathtt{10}$, that is, $h(C) = \mathtt{1}$ and $h(A) = \mathtt{0}$. Thus, $m_a(w)$ contains both $h(BAC) = h(B)\mathtt{01}$ and $h(BCA) = h(B)\mathtt{10}$. Since $h(B)$ is gentle, this implies that 01 and 10 have the same position modulo 3, which is impossible.

The case when $h(C)$ is gentle is symmetrical. If $h(AB)$ is gentle, then $h(ABAC)$ and $h(ABC)$ imply that $|h(A)| \equiv 0 \pmod 3$. If $h(AB)$ is not gentle, then $h(A) = \mathtt{1}$ and $h(B) = \mathtt{0}$. Thus, $m_a(w)$ contains both $h(ABC) = \mathtt{10}h(C)$ and $h(BAC) = \mathtt{01}h(C)$. Since $h(C)$ is gentle, this implies that 10 and 01 have the same position modulo 3, which is impossible.

Finally, if $h(A)$, $h(B)$, and $h(C)$ are not gentle, then the length of the three fragments of the formula is $2|h(A)| + |h(B)| + |h(C)| \leqslant 8$. So it suffices to consider the factors of length at most 8 in S to check that no such occurrence exists.

This shows that $S \subseteq R'$. Since $R' \subseteq R \subseteq S \subseteq R'$, we obtain $R' = R = S$, which proves Theorem 3. □

Corollary 4. *Neither $ABACA.ABCA$ nor $ABAC.BACA.ABCA$ is essentially avoided by a finite set of morphic words.*

Proof. Let $c(n) = |S \cap \Sigma_2^n|$ denote the number of words of length n in S. By construction of S,

$$c(n) = 2\sum_{0\leqslant i\leqslant 2} c\left(\left\lceil \tfrac{n-i}{3}\right\rceil\right) \text{ for every } n \geqslant 8.$$

Thus $c(n) = \Theta\left(n^{\ln 6/\ln 3}\right) = \Theta\left(n^{1+\ln 2/\ln 3}\right)$. Devyatov [7] has recently shown that the factor complexity (i.e. the number of factors of length n) of a morphic word is either $O\left(n\ln(n)\right)$ or $\Theta\left(n^{1+1/k}\right)$ for some integer $k \geqslant 1$. Thus, S cannot be the union of the factors of a finite number of morphic words. □

4 Ternary nice formulas

Clark [5] introduced the notion of *n-avoidance basis* for formulas, which is the smallest set of formulas with the following property: for every $i \leqslant n$, every avoidable formula with i variables is divisible by at least one formula with at most i variables in the n-avoidance basis. See [5, 9] for more discussions about the n-avoidance basis. The avoidability index of every formula in the 3-avoidance basis has been determined:

- AA ($\lambda = 3$ [15])
- $ABA.BAB$ ($\lambda = 3$ [4])
- $ABCA.BCAB.CABC$ ($\lambda = 3$ [9])

- $ABCBA.CBABC$ ($\lambda = 2$ [9])
- $ABCA.CABC.BCB$ ($\lambda = 3$ [9])
- $ABCA.BCAB.CBC$ ($\lambda = 3$, reverse of $ABCA.CABC.BCB$)
- $AB.AC.BA.CA.CB$ ($\lambda = 4$ [1])

Recall that a formula f is *nice* if for every variable X of f, there exists a fragment of f that contains X at least twice. Every formula in the 3-avoidance basis except $AB.AC.BA.CA.CB$ is both nice and 3-avoidable. This raised the question in [9] whether every nice formula is 3-avoidable, which would generalize the 3-avoidability of doubled patterns. In this section, we answer this question positively for ternary formulas.

Theorem 5. *Every nice formula with at most* 3 *variables is* 3*-avoidable.*

We say that a nice formula is minimal if it is not divisible by another nice formula with at most the same number of variables. The following property of every minimal nice formula is easy to derive. If a variable V appears as a prefix of a fragment ϕ, then

- V is also a suffix of ϕ,
- ϕ contains exactly two occurrences of V,
- V is neither a prefix nor a suffix of any fragment other than ϕ,
- Every fragment other than ϕ contains at most one occurrence of V.

Thus, if f is a minimal nice formula with $n \geqslant 2$ variables, then f has at most n fragments. Moreover, every fragment has length at most $2 + 2^{n-1} - 1 = 2^{n-1} + 1$, since otherwise it would contain a doubled pattern as a factor.

This implies an algorithm to list the minimal nice formulas with at most n variables. The table below lists the formulas that need to be shown 3-avoidable, that is, the minimal nice formulas with at most 3 variables that do not belong to the 3-avoidance basis. Also, if two distinct formulas are the reverse of each other, then only one of them appears in the table and the given avoiding word avoids both formulas. Some of these formulas are avoided by b_3 and the proof uses Cassaigne's algorithm [3] as in Section 2. The other formulas are each avoided by the image by a uniform morphism of either any infinite $\left(\frac{5}{4}^+\right)$-free word w_5 over Σ_5 or any infinite $\left(\frac{7}{5}^+\right)$-free word w_4 over Σ_4. We refer to [12, 13] for details about the technique to prove avoidance with morphic images of (α^+)-free words.

Formula	Closed under reversal?	Avoidability exponent	Avoiding word
$ABA.BCB.CAC$	yes	1.5	b_3
$ABCA.BCAB.CBAC$	no	1.333333333	b_3
$ABCA.BAB.CAC$	yes	1.414213562	$g_v(w_4)$
$ABCA.BAB.CBC$	no	1.430159709	$g_w(w_4)$
$ABCA.BAB.CBAC$	no	1.381966011	$g_x(w_5)$
$ABCBA.CABC$	no	1.361103081	$g_y(w_5)$
$ABCBA.CAC$	yes	1.396608253	$g_z(w_5)$

g_v	g_w	g_x	g_y	g_z
$0 \to 01220,$	$0 \to 02111,$	$0 \to 021110,$	$0 \to 022,$	$0 \to 120201,$
$1 \to 01110,$	$1 \to 01121,$	$1 \to 012221,$	$1 \to 021,$	$1 \to 100002,$
$2 \to 00212,$	$2 \to 00222,$	$2 \to 011120,$	$2 \to 012,$	$2 \to 022221,$
$3 \to 00112.$	$3 \to 00122.$	$3 \to 002211,$	$3 \to 011,$	$3 \to 011112,$
		$4 \to 001122.$	$4 \to 000.$	$4 \to 001122.$

5 A counter-example to a conjecture of Grytczuk

Grytczuk [10] considered the notion of pattern avoidance on graphs. This generalizes the definition of nonrepetitive coloring, which corresponds to the pattern AA. Given a pattern p and a graph G, the avoidability index $\lambda(p, G)$ is the smallest number of colors needed to color the vertices of G such that every path in G induces a word avoiding p.

We think that the natural framework is that of directed graphs with no loops and no multiple arcs, but such that opposite arcs (i.e., digons) are allowed. An oriented path in a directed graph $\overrightarrow{G}$ is a sequence of distinct vertices $v_1, v_2, \ldots, v_k$ such that $\overrightarrow{G}$ contains all the arcs $\overrightarrow{v_iv_{i+1}}$ such that $1 \leqslant i \leqslant k-1$.

A pattern occurs in a vertex-colored directed graph $\overrightarrow{G}$ if the sequence of colors on a directed path of $\overrightarrow{G}$ induces an occurrence of the pattern. Informally, the orientation of the path corresponds to the reading direction. We define $\lambda\left(p, \overrightarrow{G}\right)$ as the smallest number of colors such that there exists a vertex coloring avoiding p. This way, $\lambda(p) = \lambda\left(p, \overrightarrow{P}\right)$, where $\overrightarrow{P}$ is the infinite oriented path with vertices v_i and arcs $\overrightarrow{v_iv_{i+1}}$, for every $i \geqslant 0$.

Thus, an undirected graph corresponds to a symmetric directed graph: for every pair of distinct vertices u and v, either there exists no arc between u and v, or there exist both the arcs $\overrightarrow{uv}$ and $\overrightarrow{vu}$. Let P denote the infinite undirected path. We prefer the framework of directed graphs because, even though $\lambda\left(AA, \overrightarrow{P}\right) = \lambda(AA, P) = 3$, there exist patterns such that $\lambda\left(p, \overrightarrow{P}\right) < \lambda(p, P)$. For example, $\lambda(ABCACB) = \lambda\left(ABCACB, \overrightarrow{P}\right) = 2$ [12], whereas $\lambda(ABCACB, P) = 3$ since a computer check shows that the longest binary words avoiding both $ABCACB$ and its reverse $ABCBAC$ have length 23. The equivalence between avoiding a pattern and its corresponding formula holds for $\overrightarrow{P}$ but does not generalize to other directed graphs. So we do not try to define a notion of avoidance for formulas on graphs or directed graphs.

A conjecture of Grytczuk [10] says that for every avoidable pattern p, there exists a function g such that $\lambda(p,G) \leqslant g(\Delta(G))$, where G is an undirected graph and $\Delta(G)$ denotes its maximum degree. Grytczuk [10] obtained that his conjecture holds for doubled patterns.

As a counterexample, we consider the pattern $ABACADABCA$ which is 2-avoidable by the result in Section 3. Of course, $ABACADABCA$ is not doubled because of the isolated variable D. Let us show that $ABACADABCA$ is unavoidable on the infinite oriented graph $\overrightarrow{G}$ with vertices v_i and arcs $\overrightarrow{v_iv_{i+1}}$ and $\overrightarrow{v_{100i}v_{100i+2}}$, for every $i \geqslant 0$. Notice that $\overrightarrow{G}$ is obtained from $\overrightarrow{P}$ by adding the arcs $\overrightarrow{v_{100i}v_{100i+2}}$. The constant 100 in the construction is arbitrary and can be replaced by any constant.

Suppose that $\overrightarrow{G}$ is colored with k colors. Consider the factors in the subgraph $\overrightarrow{P}$ induced by the paths from $v_{300ik+1}$ to $v_{300ik+200k+1}$, for every $i \geqslant 0$. Since these factors have bounded length, the same factor appears on two disjoint such paths p_l and p_r (such that p_l is on the left of p_r). Notice that p_l contains $2k+1$ vertices with index $\equiv 1 \pmod{100}$. By the pigeon-hole principle, p_l contains three such vertices with the same color a. Thus, p_l contains an occurrence of $ABACA$ such that $A \mapsto a$ on vertices with index $\equiv 1 \pmod{100}$. The same is true for p_r. In $\overrightarrow{G}$, the occurrences of $ABACA$ in p_l and p_r imply an occurrence of $ABACADABCA$ since we can skip an occurrence of the variable A in p_l thanks to some arc of the form $\overrightarrow{v_{100j}v_{100j+2}}$.

This shows that $ABACADABCA$ is unavoidable on $\overrightarrow{G}$. So Grytczuk's conjecture is disproved since $\overrightarrow{G}$ has maximum degree 3. It is also a counterexample to Conjecture 6 in [6] which states that every avoidable pattern is avoidable on the infinite graph with vertices $\{v_0, v_1, \ldots\}$ and the arcs $\overrightarrow{v_iv_{i+1}}$ and $\overrightarrow{v_iv_{i+2}}$ for every $i \geqslant 0$.

6 A palindrome with index 4

Mikhailova [11] considered the largest avoidability index $\mathcal{P}$ of an avoidable pattern that is a palindrome. She proved that $\mathcal{P} \leqslant 16$. An obvious lower bound is $\mathcal{P} \geqslant \lambda(AA) = 3$. For a better lower bound, we consider the palindromic pattern $ABCADACBA$ or, equivalently, the ternary formula $f = ABCA.ACBA$. Since it is a ternary formula, f is 4-avoidable. More precisely, f is not nice because of the variable C, so the only formula in the 3-avoidance basis that divides f is $AB.AC.BA.CA.CB$, which is avoided by b_4.

Let us show that f is not 3-avoidable. Let w be a ternary recurrent word avoiding f. Assume towards a contradiction that w contains a square uu. Then there exists a non-empty word v such that $uuvuu$ is a factor of w. Thus, w contains an occurrence of f given by the morphism $A \mapsto u, B \mapsto u, C \mapsto v$. This contradiction shows that w is square-free. A computer check shows that no infinite ternary square-free word avoids f. This holds even if we forbid only squares and every occurrence h of f such that $|h(A)| = 1$ and $|h(B)| + |h(C)| \leqslant 5$. Thus, $\mathcal{P} \geqslant \lambda(ABCADACBA) = \lambda(ABCA.ACBA) = 4$.

References

[1] K. A. Baker, G. F. McNulty, and W. Taylor. Growth problems for avoidable words. *Theoret. Comput. Sci.*, 69(3):319–345, 1989.

[2] J. Berstel. *Axel Thue's papers on repetitions in words: a translation*, volume 20 of *Publications du LACIM*. Université du Québec à Montréal, 1994.

[3] J. Cassaigne. *An Algorithm to Test if a Given Circular HDOL-Language Avoids a Pattern. IFIP Congress*, pages 459–464, 1994.

[4] J. Cassaigne. *Motifs évitables et régularité dans les mots.* PhD thesis, Université Paris VI, 1994.

[5] R. J. Clark. *Avoidable formulas in combinatorics on words.* PhD thesis, University of California, Los Angeles, 2001. Available at
`http://www.lirmm.fr/~ochem/morphisms/clark_thesis.pdf`

[6] M. Debski, U. Pastwa, and K. Wesek. Grasshopper avoidance of patterns. *Electron. J. Combinatorics.*, 23(4):#P4.17, 2016.

[7] R. Devyatov. On subword complexity of morphic sequences. *Math. USSR Sbornik*, 2015. `arXiv:1502.02310`

[8] Pytheas Fogg. Substitutions in Dynamics, Arithmetics and Combinatorics. Springer Science & Business Media, 2002.

[9] G. Gamard, P. Ochem, G. Richomme, and P. Séébold. Avoidability of circular formulas. *Theor. Comput. Sci.*, 726:1–4, 2018.

[10] J. Grytczuk. Pattern avoidance on graphs. *Discrete Math.*, 307(11-12):1341–1346, 2007.

[11] I. Mikhailova. On the avoidability index of palindromes. *Matematicheskie Zametki.*, 93(4):634–636, 2013.

[12] P. Ochem. A generator of morphisms for infinite words. *RAIRO - Theoret. Informatics Appl.*, 40:427–441, 2006.

[13] P. Ochem. Doubled patterns are 3-avoidable. *Electron. J. Combinatorics.*, 23(1):#P1.19, 2016.

[14] P. Ochem and M. Rosenfeld. Avoidability of formulas with two variables. *Electron. J. Combin.*, 24(4):#P4.30, 2017.

[15] A. Thue. Über unendliche Zeichenreihen. *Norske Vid. Selsk. Skr. I. Mat. Nat. Kl. Christiania*, 7:1–22, 1906.

[16] A. Thue. Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen. *Norske Vid. Selsk. Skr. I. Mat. Nat. Kl. Christiania,*, 10:1–67, 1912.

6.2 Binary formulas

Avoidability of formulas with two variables

Pascal Ochem*and Matthieu Rosenfeld†

October 13, 2016

Abstract

In combinatorics on words, a word w over an alphabet Σ is said to avoid a pattern p over an alphabet Δ of variables if there is no factor f of w such that $f = h(p)$ where $h : \Delta^* \to \Sigma^*$ is a non-erasing morphism. A pattern p is said to be k-avoidable if there exists an infinite word over a k-letter alphabet that avoids p. We consider the patterns such that at most two variables appear at least twice, or equivalently, the formulas with at most two variables. For each such formula, we determine whether it is 2-avoidable, and if it is 2-avoidable, we determine whether it is avoided by exponentially many binary words.

Keywords: Word; Pattern avoidance.

1 Introduction

A *pattern* p is a non-empty finite word over an alphabet $\Delta = \{A, B, C, \ldots\}$ of capital letters called *variables*. An *occurrence* of p in a word w is a non-erasing morphism $h : \Delta^* \to \Sigma^*$ such that $h(p)$ is a factor of w. The *avoidability index* $\lambda(p)$ of a pattern p is the size of the smallest alphabet Σ such that there exists an infinite word over Σ containing no occurrence of p. Bean, Ehrenfeucht, and McNulty [3] and Zimin [11] characterized unavoidable patterns, i.e., such that $\lambda(p) = \infty$. We say that a pattern p is *t-avoidable* if $\lambda(p) \leqslant t$. For more

*LIRMM, CNRS, Université de Montpellier, France. ochem@lirmm.fr

†LIP, ENS de Lyon, CNRS, UCBL, Université de Lyon, France. matthieu.rosenfeld@ens-lyon.fr

informations on pattern avoidability, we refer to Chapter 3 of Lothaire's book [6].

A variable that appears only once in a pattern is said to be *isolated*. Following Cassaigne [4], we associate to a pattern p the *formula* f obtained by replacing every isolated variable in p by a dot. The factors between the dots are called *fragments*.

An *occurrence* of f in a word w is a non-erasing morphism $h : \Delta^* \to \Sigma^*$ such that the h-image of every fragment of f is a factor of w. As for patterns, the avoidability index $\lambda(f)$ of a formula f is the size of the smallest alphabet allowing an infinite word containing no occurrence of p. Clearly, every word avoiding f also avoids p, so $\lambda(p) \leqslant \lambda(f)$. Recall that an infinite word is *recurrent* if every finite factor appears infinitely many times. If there exists an infinite word over Σ avoiding p, then there there exists an infinite recurrent word over Σ avoiding p. This recurrent word also avoids f, so that $\lambda(p) = \lambda(f)$. Without loss of generality, a formula is such that no variable is isolated and no fragment is a factor of another fragment.

Cassaigne [4] began and Ochem [7] finished the determination of the avoidability index of every pattern with at most 3 variables. A *doubled* pattern contains every variable at least twice. Thus, a doubled pattern is a formula with exactly one fragment. Every doubled pattern is 3-avoidable [9]. A formula is said to be *binary* if it has at most 2 variables. In this paper, we determine the avoidability index of every binary formula.

We say that a formula f is *divisible* by a formula f' if f does not avoid f', that is, there is a non-erasing morphism such that the image of any fragment of f' by h is a factor of a fragment of f. If f is divisible by f', then every word avoiding f' also avoids f and thus $\lambda(f) \leqslant \lambda(f')$. Moreover, the reverse f^R of a formula f satisfies $\lambda(f^R) = \lambda(f)$. For example, the fact that $ABA.AABB$ is 2-avoidable implies that $ABAABB$ and $BAB.AABB$ are 2-avoidable. See Cassaigne [4] and Clark [5] for more information on formulas and divisibility. For convenience, we say that an avoidable formula f is *exponential* (resp. *polynomial*) if the number of words in $\Sigma^n_{\lambda(f)}$ avoiding f is exponential (resp. polynomial) in n.

First, we check that every avoidable binary formula is 3-avoidable. Since $\lambda(AA) = 3$, every formula containing a square is 3-avoidable. Then, the only square free avoidable binary formula is $ABA.BAB$ with avoidability index 3 [4]. Thus, we have to distinguish between avoidable binary formulas with avoidability index 2 and 3. A binary formula is minimally 2-avoidable if it is

2-avoidable and is not divisible by any other 2-avoidable binary formula. A binary formula f is maximally 2-unavoidable if it is 2-unavoidable and every other binary formula that is divisible by f is 2-avoidable.

Theorem 1.

Up to symmetry, the maximally 2-unavoidable binary formulas are:

- $AAB.ABA.ABB.BBA.BAB.BAA$
- $AAB.ABBA$
- $AAB.BBAB$
- $AAB.BBAA$
- $AAB.BABB$
- $AAB.BABAA$
- $ABA.ABBA$
- $AABA.BAAB$

Up to symmetry, the minimally 2-avoidable binary formulas are:

- $AA.ABA.ABBA$ *(polynomial)*
- $ABA.AABB$ *(polynomial)*
- $AABA.ABB.BBA$ *(polynomial)*
- $AA.ABA.BABB$ *(exponential)*
- $AA.ABB.BBAB$ *(exponential)*
- $AA.ABAB.BB$ *(exponential)*
- $AA.ABBA.BAB$ *(exponential)*
- $AAB.ABB.BBAA$ *(exponential)*
- $AAB.ABBA.BAA$ *(exponential)*
- $AABB.ABBA$ *(exponential)*
- $ABAB.BABA$ *(exponential)*

- *$AABA.BABA$ (exponential)*
- *AAA (exponential)*
- *$ABA.BAAB.BAB$ (exponential)*
- *$AABA.ABAA.BAB$ (exponential)*
- *$AABA.ABAA.BAAB$ (exponential)*
- *$ABAAB$ (exponential)*

Given a binary formula f, we can use Theorem 1 to find $\lambda(f)$. Now, we also consider the problem whether an avoidable binary formula is polynomial or exponential. If $\lambda(f) = 3$, then either f contains a square or $f = ABA.BAB$, so that f is exponential. Thus, we consider only the case $\lambda(f) = 2$. If f is divisible by an exponential 2-avoidable formula given in Theorem 1, then f is known to be exponential. This leaves open the case such that f is only divisible by polynomial 2-avoidable formulas. The next result settles every open case.

Theorem 2.

The following formulas are polynomial:

- $BBA.ABA.AABB$
- $AABA.AABB$

The following formulas are exponential:

- $BAB.ABA.AABB$
- $AAB.ABA.ABBA$
- $BAA.ABA.AABB$
- $BBA.AABA.AABB$

To obtain the 2-unavoidability of the formulas in the first part of Theorem 1, we use a standard backtracking algorithm. Figure 1 gives the maximal length and number of binary words avoiding each maximally 2-unavoidable formula.

In Section 3, we consider the polynomial formulas in Theorems 1 and 2. The proof uses a technical lemma given in Section 2. Then we consider in Section 4 the exponential formulas in Theorems 1 and 2.

A preliminary version of this paper, without Theorem 2, has been presented at DLT 2016.

Formula	Maximal length of a binary word avoiding this formula	Number of binary words avoiding this formula
$AAB.BBAA$	22	1428
$AAB.ABA.ABB.BBA.BAB.BAA$	23	810
$AAB.BBAB$	23	1662
$AABA.BAAB$	26	2124
$AAB.ABBA$	30	1684
$AAB.BABAA$	42	71002
$AAB.BABB$	69	9252
$ABA.ABBA$	90	31572

Figure 1: The number and maximal length of binary words avoiding the maximally 2-unavoidable formulas.

2 The useful lemma

Let us define the following words:

- b_2 is the fixed point of $\texttt{0} \mapsto \texttt{01}$, $\texttt{1} \mapsto \texttt{10}$.
- b_3 is the fixed point of $\texttt{0} \mapsto \texttt{012}$, $\texttt{1} \mapsto \texttt{02}$, $\texttt{2} \mapsto \texttt{1}$.
- b_4 is the fixed point of $\texttt{0} \mapsto \texttt{01}$, $\texttt{1} \mapsto \texttt{03}$, $\texttt{2} \mapsto \texttt{21}$, $\texttt{3} \mapsto \texttt{23}$.
- b_5 is the fixed point of $\texttt{0} \mapsto \texttt{01}$, $\texttt{1} \mapsto \texttt{23}$, $\texttt{2} \mapsto \texttt{4}$, $\texttt{3} \mapsto \texttt{21}$, $\texttt{4} \mapsto \texttt{0}$.

Let w and w' be infinite (right infinite or bi-infinite) words. We say that w and w' are equivalent if they have the same set of finite factors. We write $w \sim w'$ if w and w' are equivalent. A famous result of Thue [10] can be stated as follows:

Theorem 3. *[10] Every bi-infinite ternary word avoiding 010, 212, and squares is equivalent to b_3.*

Given an alphabet Σ and forbidden structures S, we say that a finite set W of infinite words over Σ *essentially avoids* S if every word in W avoids S and every bi-infinite words over Σ avoiding S is equivalent to one of the words in S. If W contains only one word w, we denote the set W by w instead of $\{w\}$. Then we can restate Theorem 3: b_3 essentially avoids 010, 212, and squares

The results in the next section involve b_3. We have tried without success to prove them by using Theorem 3. We need the following stronger property of b_3:

Lemma 4. *b_3 essentially avoids* $\mathtt{010}$*,* $\mathtt{212}$*, XX with $1 \leqslant |X| \leqslant 3$, and $\mathtt{2}YY$ with $|Y| \geqslant 4$.*

Proof. We start by checking by computer that b_3 has the same set of factors of length 100 as every bi-infinite ternary word avoiding $\mathtt{010}$, $\mathtt{212}$, XX with $1 \leqslant |X| \leqslant 3$, and $\mathtt{2}YY$ with $|Y| \geqslant 4$. The set of the forbidden factors of b_3 of length at most 4 is $F = \{\mathtt{00}, \mathtt{11}, \mathtt{22}, \mathtt{010}, \mathtt{212}, \mathtt{0202}, \mathtt{2020}, \mathtt{1021}, \mathtt{1201}\}$. To finish the proof, we use Theorem 3 and we suppose for contradiction that w is a bi-infinite ternary word that contains a large square MM and avoids both F and large factors of the form $\mathtt{2}YY$.

- Case $M = \mathtt{0}N$. Then w contains $MM = \mathtt{0}N\mathtt{0}N$. Since $\mathtt{00} \in F$ and $\mathtt{2}YY$ is forbidden, w contains $\mathtt{10}N\mathtt{0}N$. Since $\{\mathtt{11}, \mathtt{010}\} \subset F$, w contains $210N0N$. If $N = P\mathtt{1}$, then w contains $\mathtt{210}P\mathtt{10}P\mathtt{1}$, which contains $\mathtt{2}YY$ with $Y = 10P$. So $N = P2$ and w contains $\mathtt{210}P\mathtt{20}P\mathtt{2}$. If $P = Q\mathtt{1}$, then w contains $\mathtt{210}Q\mathtt{120}Q\mathtt{12}$. Since $\{\mathtt{11}, \mathtt{212}\} \subset F$, the factor $Q\mathtt{12}$ implies that $Q = R\mathtt{0}$ and w contains $\mathtt{210}R\mathtt{0120}R\mathtt{012}$. Moreover, since $\{\mathtt{00}, \mathtt{1201}\} \subset F$, the factor $\mathtt{120}R$ implies that $R = \mathtt{2}S$ and w contains $\mathtt{2102}S\mathtt{01202}S\mathtt{012}$. Then there is no possible prefix letter for S: $\mathtt{0}$ gives $\mathtt{2020}$, $\mathtt{1}$ gives $\mathtt{1021}$, and $\mathtt{2}$ gives $\mathtt{22}$. This rules out the case $P = Q\mathtt{1}$. So $P = Q\mathtt{0}$ and w contains $\mathtt{210}Q\mathtt{020}Q\mathtt{02}$. The factor $Q\mathtt{020}Q$ implies that $Q = \mathtt{1}R\mathtt{1}$, so that w contains $\mathtt{2101}R\mathtt{10201}R\mathtt{102}$. Since $\{\mathtt{11}, \mathtt{010}\} \subset F$, the factor $\mathtt{01}R$ implies that $R = \mathtt{2}S$, so that w contains $\mathtt{21012}S\mathtt{102012}S\mathtt{102}$. The only possible right extension with respect to F of $\mathtt{102}$ is $\mathtt{102012}$. So w contains $\mathtt{21012}S\mathtt{102012}S\mathtt{102012}$, which contains $\mathtt{2}YY$ with $Y = S\mathtt{102012}$.

- Case $M = \mathtt{1}N$. Then w contains $MM = \mathtt{1}N\mathtt{1}N$. In order to avoid $\mathtt{11}$ and $\mathtt{2}YY$, w must contain $\mathtt{01}N\mathtt{1}N$. If $N = P\mathtt{0}$, then w contains $\mathtt{01}P\mathtt{01}P\mathtt{0}$. So w contains the large square $\mathtt{01}P\mathtt{01}P$ and this case is covered by the previous item. So $N = P2$ and w contains $\mathtt{01}P\mathtt{21}P\mathtt{2}$. Then there is no possible prefix letter for P: $\mathtt{0}$ gives $\mathtt{010}$, $\mathtt{1}$ gives $\mathtt{11}$, and 2 gives $\mathtt{212}$.

- Case $M = \mathtt{2}N$. Then w contains $MM = 2N2N$. If $N = P\mathtt{1}$, then w contains $\mathtt{2}P\mathtt{12}P\mathtt{1}$. This factor cannot extend to $\mathtt{2}P\mathtt{12}P\mathtt{12}$, since this is $2YY$ with $Y = P12$. So w contains $2P12P10$. Then there is no possible suffix letter for P: 0 gives 010, 1 gives 11, and 2 gives 212. This rules out the case $N = P1$. So $N = P0$ and w contains $2P02P0$. This factor cannot extend to $02P02P0$, since this contains the large square $02P02P$ and this case is covered by the first item. Thus w contains $12P02P0$. If $P = Q1$, then w contains $12Q102Q10$. Since $\{22, 1021\} \subset F$, the factor $102Q$ implies that $Q = 0R$, so that w contains $120R1020R10$. Then there is no possible prefix letter for R: 0 gives 00, 1 gives 1201, and 2 gives 0202. This rules out the case $P = Q1$. So $P = Q2$ and w contains $12Q202Q20$. The factor $Q202$ implies that $Q = R1$ and w contains $12R1202R120$. Since $\{00, 1201\} \subset F$, w contains $12R1202R1202$, which contains $2YY$ with $Y = R1202$.

□

3 Polynomial formulas

Let us detail the binary words avoiding the polynomial formulas in Theorems 1 and 2.

Lemma 5.

- $\{g_x(b_3), g_y(b_3), g_z(b_3), g_{\overline{z}}(b_3)\}$ *essentially avoids* $AA.ABA.ABBA$.
- $g_x(b_3)$ *essentially avoids* $AABA.ABB.BBA$.
- *Let* f *be either* $ABA.AABB$, $BBA.ABA.AABB$, *or* $AABA.AABB$. *Then* $\{g_x(b_3), g_t(b_3)\}$ *essentially avoids* f.

The words avoiding these formulas are morphic images of b_3 by the morphisms given below. Let $\overline{w}$ denote the word obtained from the (finite or bi-infinite) binary word w by exchanging 0 and 1. Obviously, if w avoids a given formula, then so does $\overline{w}$. A (bi-infinite) binary word w is *self-complementary* if $w \sim \overline{w}$. The words $g_x(b_3)$, $g_y(b_3)$, and $g_t(b_3)$ are self-complementary. Since the frequency of 0 in $g_z(b_3)$ is $\frac{5}{9}$, $g_z(b_3)$ is not self-complementary. Then $g_{\overline{z}}$ is obtained from g_z by exchanging 0 and 1, so that $g_{\overline{z}}(b_3) = \overline{g_z(b_3)}$.

$$\begin{array}{llll}
g_x(0) = 01110, & g_y(0) = 0111, & g_z(0) = 0001, & g_t(0) = 01011011010, \\
g_x(1) = 0110, & g_y(1) = 01, & g_z(1) = 001, & g_t(1) = 01011010, \\
g_x(2) = 0. & g_y(2) = 00. & g_z(2) = 11. & g_t(2) = 010.
\end{array}$$

Let us first state interesting properties of the morphisms and the formulas in Lemma 5.

Lemma 6. *For every $p, s \in \Sigma_3$, $Y \in \Sigma_3^*$ with $|Y| \geqslant 4$, and $g \in \{g_x, g_y, g_z, g_{\overline{z}}, g_t\}$, the word $g(p2YYs)$ contains an occurrence of $AABA.AABBA$.*

Proof.

- Since $\mathtt{0}$ is a prefix and a suffix of the g_x-image of every letter, $g_x(p2YYs) = V\mathtt{000}U\mathtt{00}U\mathtt{00}W$ contains an occurrence of $AABA.AABBA$ with $A = \mathtt{0}$ and $B = \mathtt{0}U\mathtt{0}$.

- Since $\mathtt{0}$ is a prefix of the g_y-image of every letter, $g_y(2YYs) = \mathtt{000}U\mathtt{0}U\mathtt{0}V$ with $U, V \in \Sigma_3^+$, which contains an occurrence of $AABA.AABBA$ with $A = \mathtt{0}$ and $B = \mathtt{0}U$.

- Since $\mathtt{1}$ is a suffix of the g_z-image of every letter, $g_z(p2YY) = \mathtt{111}U\mathtt{1}U\mathtt{1}$ contains an occurrence of $AABA.AABBA$ with $A = \mathtt{1}$ and $B = \mathtt{1}U$.

- Since $g_{\overline{z}}(p2YY) = \overline{g_z(p2YY)}$, $g_{\overline{z}}(s2YY)$ contains an occurrence of $AABA.AABBA$.

- Since $\mathtt{010}$ is a prefix and a suffix of the g_t-image of every letter, $g_t(p2YYs) = V\mathtt{010010010}U\mathtt{010010}U\mathtt{010010}W$ contains an occurrence of $AABA.AABBA$ with $A = \mathtt{010}$ and $B = \mathtt{010}U\mathtt{010}$.

□

Lemma 7. *$AABA.AABBA$ is divisible by every formula in Lemma 5.*

We are now ready to prove Lemma 5. To prove the avoidability, we have implemented Cassaigne's algorithm that decides, under mild assumptions, whether a morphic word avoids a formula [4]. We have to explain how the long enough binary words avoiding a formula can be split into 4 or 2 distinct incompatible types. A similar phenomenon has been described for $AABB.ABBA$ [8].

First, consider any infinite binary word w avoiding $AA.ABA.ABBA$. A computer check shows by backtracking that w must contain the factor $\mathtt{01110001110}$. In particular, w contains $\mathtt{00}$. Thus, w cannot contain both $\mathtt{010}$ and $\mathtt{0110}$, since it would produce an occurrence of $AA.ABA.ABBA$. Moreover, a computer check shows by backtracking that w cannot avoid both

010 and 0110. So, w must contain either 010 or 0110 (this is an exclusive or). By symmetry, w must contain either 101 or 1001. There are thus at most 4 possibilities for w, depending on which subset of $\{010, 0110, 101, 1001\}$ appears among the factors of w, see Figure 2.

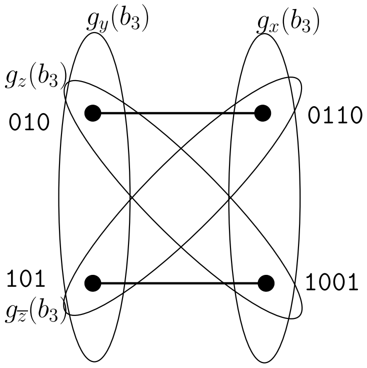

Figure 2: The four infinite binary words avoiding $AA.ABA.ABBA$.

Also, consider any infinite binary word w avoiding f, where f is either $ABA.AABB$, $BBA.ABA.AABB$, or $AABA.AABB$. Notice that the formulas $BBA.ABA.AABB$ and $AABA.AABB$ are divisible by $ABA.AABB$. We check by backtracking that no infinite binary word avoids f, 0010, and 00110. A word containing both 0010 and 00110 contains an occurrence of $AABA.AABBA$, and thus an occurrence of f by Lemma 7. So w does not contain both 0010 and 00110. Thus, there are two possibilities for w depending on whether it contains 0010 or 00110.

Now, our tasks of the form "prove that a set of morphic words essentially avoids one formula" are reduced to (more) tasks of the form "prove that one morphic word essentially avoids one formula and a set of factors".

Since all the proofs of such reduced tasks are very similar, we only detail the proof that $g_y(b_3)$ essentially avoids $AA.ABA.ABBA$, 0110, and 1001. We check that the set of prolongable binary words of length 100 avoiding $AA.ABA.ABBA$, 0110, and 1001 is exactly the set of factors of length 100 of $g_y(b_3)$. Using Cassaigne's notion of circular morphism [4], this is sufficient to prove that every bi-infinite binary word of this type is the g_y-image of some bi-infinite ternary word w_3. It also ensures that w_3 and b_3 have the same set of small factors. Suppose for contradiction that $w_3 \not\sim b_3$. By Lemma 4, w_3 contains a factor $2YY$ with $|Y| \geqslant 4$. Since w_3 is bi-infinite, w_3 even contains a factor $p2YYs$ with $p, s \in \Sigma_3$. By Lemma 6, $g_y(w_3)$

contains an occurrence of $AABA.AABBA$ and by Lemma 7, $g_y(w_3)$ contains an occurrence of $AA.ABA.ABBA$. This contradiction shows that $w_3 \sim b_3$. So $g_y(b_3)$ essentially avoids $AA.ABA.ABBA$, `0110`, and `1001`.

4 Exponential formulas

Given a morphism $g : \Sigma_3^* \to \Sigma_2^*$, an sqf-$g$-image is the image by g of a (finite or infinite) ternary square free word. With an abuse of language, we say that g avoids a set of formulas if every sqf-g-image avoids every formula in the set. For every 2-avoidable exponential formula f in Theorems 1 and 2, we give below a uniform morphism g that avoids f. If possible, we simultaneously avoid the reverse formula f^R of f. We also avoid large squares. Let SQ_t denote the pattern corresponding to squares of period at least t, that is, $SQ_1 = AA$, $SQ_2 = ABAB$, $SQ_3 = ABCABC$, and so on. The morphism g avoids SQ_t with t as small as possible. Since $\lambda(SQ_2)$, a binary word avoiding SQ_3 is necessarily best possible in terms of length of avoided squares.

- $f = AA.ABA.BABB$. This 22-uniform morphism avoids $\left\{f, f^R, SQ_6\right\}$:

  $$\begin{aligned}\texttt{0} &\mapsto \texttt{0001101101110011100011}\\ \texttt{1} &\mapsto \texttt{0001101101110001100011}\\ \texttt{2} &\mapsto \texttt{0001101101100011100111}\end{aligned}$$

  This 44-uniform morphism avoids $\{f, SQ_5\}$:

  $$\begin{aligned}\texttt{0} &\mapsto \texttt{00010010011000111001001100010011100100100111}\\ \texttt{1} &\mapsto \texttt{00010010011000100111001001100011100100100111}\\ \texttt{2} &\mapsto \texttt{00010010011000100111001001001100011100100111}\end{aligned}$$

  Notice that $\left\{f, f^R, SQ_5\right\}$ is 2-unavoidable and $\{f, SQ_4\}$ is 2-unavoidable.

- $f = AA.ABB.BBAB$. This 60-uniform morphism avoids $\left\{f, f^R, SQ_{11}\right\}$:

  $$\begin{aligned}\texttt{0} &\mapsto \texttt{000110011100011001110011000111000110011100011100110001110011}\\ \texttt{1} &\mapsto \texttt{000110011100011001110001110011000111000110011100110001110011}\\ \texttt{2} &\mapsto \texttt{000110011100011001110001100111000111001100011100110001110011}\end{aligned}$$

  Notice that $\{f, SQ_{10}\}$ is 2-unavoidable.

- $f = AA.ABAB.BB$ is self-reverse. This 11-uniform morphism avoids $\{f, SQ_4\}$:

$$\begin{aligned}
0 &\mapsto \mathtt{00100110111}\\
1 &\mapsto \mathtt{00100110001}\\
2 &\mapsto \mathtt{00100011011}
\end{aligned}$$

  Notice that $\{f, SQ_3\}$ is 2-unavoidable.

- $f = AA.ABBA.BAB$ is self-reverse. This 30-uniform morphism avoids $\{f, SQ_6\}$:

$$\begin{aligned}
0 &\mapsto \mathtt{000110001110011000110011100111}\\
1 &\mapsto \mathtt{000110001100111001100011100111}\\
2 &\mapsto \mathtt{000110001100011001110011100111}
\end{aligned}$$

  Notice that $\{f, SQ_5\}$ is 2-unavoidable.

- $f = AAB.ABB.BBAA$ is self-reverse. This 30-uniform morphism avoids $\{f, SQ_5\}$:

$$\begin{aligned}
0 &\mapsto \mathtt{000100101110100010110111011101}\\
1 &\mapsto \mathtt{000100101101110100010111011101}\\
2 &\mapsto \mathtt{000100010001011101110111010001}
\end{aligned}$$

  Notice that $\{f, SQ_4\}$ is 2-unavoidable.

- $f = AAB.ABBA.BAA$ is self-reverse. This 38-uniform morphism avoids $\{f, SQ_5\}$:

$$\begin{aligned}
0 &\mapsto \mathtt{00010001000101110111010001011100011101}\\
1 &\mapsto \mathtt{00010001000101110100011100010111011101}\\
2 &\mapsto \mathtt{00010001000101110001110100010111011101}
\end{aligned}$$

  Notice that $\{f, SQ_4\}$ is 2-unavoidable.

- $f = AABB.ABBA$. This 193-uniform morphism avoids $\{f, SQ_{16}\}$:

$$\begin{aligned}
0 \mapsto\ &\mathtt{00010001011011101100010110111000101101110111000101100010001011}\\
&\mathtt{011101100010110111011100010110111011000101101110001011011101110001}\\
&\mathtt{0110001000101101110001011011101110001011011101100010110111 0001011}\\
1 \mapsto\ &\mathtt{00010001011011101100010110111000101101110111000101100010001011}\\
&\mathtt{011100010110111011100010110111011000101101110001011011101110001011}\\
&\mathtt{0001000101101110110001011011101110001011011101100010110111 0001011}\\
2 \mapsto\ &\mathtt{00010001011011100010110111011100010110001000101101110110001011}\\
&\mathtt{011101110001011011101100010110111000101101110111000101100010001011}\\
&\mathtt{0111011000101101110001011011101110001011011101100010110111 0001011}
\end{aligned}$$

Notice that $\{f, f^R\}$ is 2-unavoidable and $\{f, SQ_{15}\}$ is 2-unavoidable. Previous papers [7, 8] have considered a 102-uniform morphism to avoid $\{f, SQ_{27}\}$.

- $f = ABAB.BABA$ is self-reverse. This 50-uniform morphism avoids $\{f, SQ_3\}$, see [7]:

  0 ↦ 00011001011000111001011001110001011100101100010111
  1 ↦ 00011001011000101110010110011100010110001110010111
  2 ↦ 00011001011000101110010110001110010111000101100111

  Notice that a binary word avoiding $\{f, SQ_3\}$ contains only the squares 00, 11, and 0101 (or 00, 11, and 1010).

- $f = AABA.BABA$: A case analysis of the small factors shows that a recurrent binary word avoids $\{f, f^R, SQ_3\}$ if and only if it contains only the squares 00, 11, and 0101 (or 00, 11, and 1010). Thus, the previous 50-uniform morphism that avoids $\{ABAB.BABA, SQ_3\}$ also avoids $\{f, f^R, SQ_3\}$.

- $f = AAA$ is self-reverse. This 32-uniform morphism avoids $\{f, SQ_4\}$:

  0 ↦ 00101001101101001011001001101011
  1 ↦ 00101001101100101101001001101011
  2 ↦ 00100101101001001101101001011011

  Notice that $\{f, SQ_3\}$ is 2-unavoidable.

- $f = ABA.BAAB.BAB$ is self-reverse. This 10-uniform morphism avoids $\{f, SQ_3\}$:

  0 ↦ 0001110101
  1 ↦ 0001011101
  2 ↦ 0001010111

- $f = AABA.ABAA.BAB$ is self-reverse. This 57-uniform morphism avoids $\{f, SQ_6\}$:

  0 ↦ 000101011100010110010101100010111001011000101011100101011
  1 ↦ 000101011100010110010101100010101110010110001011100101011
  2 ↦ 000101011100010110010101100010101110010101100010111001011

  Notice that $\{f, SQ_5\}$ is 2-unavoidable.

- $f = AABA.ABAA.BAAB$ is self-reverse. This 30-uniform morphism avoids $\{f, SQ_3\}$:

  0 $\mapsto$ 000101110001110101000101011101

  1 $\mapsto$ 000101110001110100010101110101

  2 $\mapsto$ 000101110001010111010100011101

- $f = ABAAB$. This 10-uniform morphism avoids $\left\{f, f^R, SQ_3\right\}$, see [7]:

  0 $\mapsto$ 0001110101

  1 $\mapsto$ 0000111101

  2 $\mapsto$ 0000101111

- $f = BAB.ABA.AABB$ is self-reverse. This 16-uniform morphism avoids $\{f, SQ_5\}$:

  0 $\mapsto$ 0101110111011101

  1 $\mapsto$ 0100010111010001

  2 $\mapsto$ 0001010111010100

  Notice that $\{f, SQ_4\}$ is 2-unavoidable.

- $f = AAB.ABA.ABBA$ is avoided with its reverse. This 84-uniform morphism avoids $\left\{f, f^R, SQ_5\right\}$:

  0 $\mapsto$ 000100010111000111010001000101110111010001011100011101000101110111
  010001110001011101

  1 $\mapsto$ 000100010111000111010001000101110100011100010111011101000101110001
  110100010111011101

  2 $\mapsto$ 000100010111000111010001000101110100011100010111010001000101110001
  110100010111011101

  Notice that $\{f, SQ_4\}$ is 2-unavoidable.

- $f = BAA.ABA.AABB$. This 304-uniform morphism avoids $\{f, SQ_7\}$:

  0 ↦ 000110001100111000111001100011001110011100110001100011001110011000
  1110001100111001110011000110011100011100110001100011001110011000111000
  1100111001110011000110001100111000111001100011001110011100110001110001
  1001110011000110001100111001110011000110011100011100110001100011001110
  0111001100011100011001110011

  1 ↦ 000110001100111000111001100011001110011100110001100011001110011000
  1110001100111001110011000110011100011100110001100011001110011000111000
  1100111001110011000110001100111000111001100011001110011100110001100011
  0011100110001110001100111001110011000110011100011100110001100011001110
  0111001100011100011001110011

  2 ↦ 000110001100111000111001100011001110011100110001100011001110011000
  1110001100111001110011000110001100111000111001100011001110011100110001
  1100011001110011000110001100111001110011000110011100011100110001100011
  0011100110001110001100111001110011000110001100111000111001100011001110
  0111001100011100011001110011

  Using the morphism g_w below and the technique in [1], we can show that $g_w(b_3)$ essentially avoids $\{f, SQ_6\}$:

  $g_w(0) =$ 011100111001110001100111001100011000110
  $g_w(1) =$ 011100111001100011000110
  $g_w(2) =$ 001110011000110

  Notice that $\left\{f, f^R\right\}$ is 2-unavoidable and $\{f, SQ_5\}$ is 2-unavoidable.

- $f = BBA.AABA.AABB$. This 160-uniform morphism avoids $\left\{f, f^R, SQ_{21}\right\}$:

  0 ↦ 000101100101110001011100101100010111000101100101110010110001011100
  1011000101100101110010110001011100010110010111000101110010110001011001
  011100101100010111001011

  1 ↦ 000101100101110001011100101100010111000101100101110010110001011100
  1011000101100101110001011001011100101100010111000101110010110001011001
  011100101100010111001011

  2 ↦ 000101100101110001011001011100101100010111000101100101110001011100
  1011000101100101110010110001011100010111001011000101100101110001011001
  011100101100010111001011

This 202-uniform morphism avoids $\{f, SQ_5\}$:

0 $\mapsto$ 000110100111011010001101010001110110100110110101000111011010001101
0100011101101010001101001110110100110110101000110100111011010100011101
101000110101000111011010100011010011101101010001110110100110110101
1 $\mapsto$ 000110100111011010001101010001110110100110110101000110100111011010
1000111011010001101010001110110101000110100111011010100011101101001101
101010001101001110110100110110101000111011010001101010001110110101
2 $\mapsto$ 000110100111011010001101010001110110100110110101000110100111011010
1000111011010001101010001110110101000110100111011010011011010100011101
101000110101000111011010100011010011101101010001110110100110110101

Notice that $\left\{f, f^R, SQ_{20}\right\}$ is 2-unavoidable and $\{f, SQ_4\}$ is 2-unavoidable.

We start by checking that every morphism is synchronizing, that is, for every letters $a, b, c \in \Sigma_3$, the factor $g(a)$ only appears as a prefix or a suffix in $g(bc)$.

For every q-morphism g, the sqf-g-images are claimed to avoid SQ_t with $2t < q$. Let us prove that SQ_t is avoided. We check exhaustively that the sqf-g-images contain no square uu such that $t \leqslant |u| \leqslant 2q - 2$. Now suppose for contradiction that an sqf-g-image contains a square uu with $|u| \geqslant 2q - 1$. The condition $|u| \geqslant 2q - 1$ implies that u contains a factor $g(a)$ with $a \in \Sigma_3$. This factor $g(a)$ only appears as the g-image of the letter a because g is synchronizing. Thus the distance between any two factors u in an sqf-g-image is a multiple of q. Since uu is a factor of an sqf-g-image, we have $q \mid |u|$. Also, the center of the square uu cannot lie between the g-images of two consecutive letters, since otherwise there would be a square in the pre-image. The only remaining possibility is that the ternary square free word contains a factor $aXbXc$ with $a, b, c \in \Sigma_3$ and $X \in \Sigma_3^+$ such that $g(aXbXc) = bsYpsYpe$ contains the square $uu = sYpsYp$, where $g(X) = Y$, $g(a) = bs$, $g(b) = ps$, $g(c) = pe$. Then, we also have $a \neq b$ and $b \neq c$ since $aXbXc$ is square free. Then abc is square free and $g(abc) = bspspe$ contains a square with period $|s| + |p| = |g(a)| = q$. This is a contradiction since the sqf-g-images contain no square with period q.

Let us show that for every formula f above and corresponding morphism g, g avoids f. Notice that f is not square free, since the only avoidable square free binary formula is $ABA.BAB$, which is not 2-avoidable. We distinguish two kinds of formula.

A formula is *easy* if every appearing variable is contained in at least one square. Every potential occurrence of an easy formula then satisfies $|A| < t$

and $|B| < t$ since SQ_t is avoided. The longest fragment of every easy formula has length 4. So, to check that g avoids an easy formula, it is sufficient to consider the set of factors of the sqf-g-images with length at most $4(t-1)$.

A formula is *tough* if one of the variables is not contained in any square. The tough formulas have been named so that this variable is B. The tough formulas are $ABA.BAAB.BAB$, $ABAAB$, $AABA.ABAA.BAAB$, and $AABA.ABAA.BAB$. As before, every potential occurrence of a tough formula satisfies $|A| < t$ since SQ_t is avoided. Suppose for contradiction that $|B| \geqslant 2q-1$. By previous discussion, the distance between any two occurrences of B in an sqf-g-image is a multiple of q. The case of $ABA.BAAB.BAB$ can be settled as follows. The factor $BAAB$ implies that q divides $|BAA|$ and the factor BAB implies that q divides $|BA|$. This implies that q divides $|A|$, which contradicts $|A| < t$. For the other formulas, only one fragment contains B twice. This fragment is said to be *important*. Since $|A| < t$, the important fragment is a repetition which is "almost" a square. The important fragment is $\boldsymbol{B}A\boldsymbol{B}$ for $AABA.ABAA.BAB$, $\boldsymbol{B}AA\boldsymbol{B}$ for $AABA.ABAA.BAAB$, and $\boldsymbol{AB}A\boldsymbol{AB}$ for $ABAAB$. Informally, this almost square implies a factor $aXbXc$ in the ternary pre-image, such that $|a| = |c| = 1$ and $1 \leqslant |b| \leqslant 2$. If $|X|$ is small, then $|B|$ is small and we check exhaustively that there exists no small occurrence of f. If $|X|$ is large, there would exist a ternary square free factor $aYbYc$ with $|Y|$ small, such that $g(aYbYc)$ contains the important fragment of an occurrence of f if and only if $g(aXbXc)$ contains the important fragment of a smaller occurrence of f.

5 Concluding remarks

From our results, every minimally 2-avoidable binary formula, and thus every 2-avoidable binary formula, is avoided by some morphic image of b_3.

What can we forbid so that there exists only polynomially many avoiding words ? The known examples from the literature [1, 2, 10] are:

- one pattern and two factors:
  - b_3 essentially avoids AA, `010`, and `212`.
  - A morphic image of b_5 essentially avoids AA, `010`, and `020`.
  - A morphic image of b_5 essentially avoids AA, `121`, and `212`.
  - b_2 essentially avoids $ABABA$, `000`, and `111`.

- two patterns: b_2 essentially avoids $ABABA$ and AAA.
- one formula over three variables: b_4 and two words obtained from b_4 by letter permutation essentially avoid $AB.AC.BA.BC.CA$.

Now we can extend this list:

- one formula over two variables:
    - $g_x(b_3)$ essentially avoids $AAB.BAA.BBAB$.
    - $\{g_x(b_3), g_t(b_3)\}$ essentially avoids $ABA.AABB$ (or $BBA.ABA.AABB$, or $AABA.AABB$).
    - $\{g_x(b_3), g_y(b_3), g_z(b_3), g_{\overline{z}}(b_3)\}$ essentially avoids $AA.ABA.ABBA$.
- one pattern over three variables: $ABACAABB$ (same as $ABA.AABB$) or $AABACAABB$ (same as $AABA.AABB$).

References

[1] G. Badkobeh and P. Ochem. Characterization of some binary words with few squares. *Theoret. Comput. Sci.*, 588:73–80, 2015.

[2] K. A. Baker, G. F. McNulty, and W. Taylor. Growth problems for avoidable words. *Theoretical Computer Science*, 69(3):319 – 345, 1989.

[3] D. R. Bean, A. Ehrenfeucht, and G. F. McNulty. Avoidable patterns in strings of symbols. *Pacific J. of Math.*, 85:261–294, 1979.

[4] J. Cassaigne. *Motifs évitables et régularité dans les mots.* PhD thesis, Université Paris VI, 1994. URL: http://www.lirmm.fr/~ochem/morphisms/clark_thesis.pdf.

[5] R. J. Clark. *Avoidable formulas in combinatorics on words.* PhD thesis, University of California, Los Angeles, 2001.

[6] M. Lothaire. *Algebraic Combinatorics on Words.* Cambridge Univ. Press, 2002.

[7] P. Ochem. A generator of morphisms for infinite words. *RAIRO - Theoret. Informatics Appl.*, 40:427–441, 2006.

[8] P. Ochem. Binary words avoiding the pattern AABBCABBA. *RAIRO - Theoret. Informatics Appl.*, 44(1):151–158, 2010.

[9] P. Ochem. Doubled patterns are 3-avoidable. *Electron. J. Combinatorics.*, 23(1), 2016.

[10] A. Thue. Über unendliche Zeichenreihen. *Norske Vid. Selsk. Skr. I. Mat. Nat. Kl. Christiania*, 7:1–22, 1906.

[11] A. I. Zimin. Blocking sets of terms. *Math. USSR Sbornik*, 47(2):353–364, 1984.

6.3 Doubled patterns

Doubled patterns are 3-avoidable

Pascal Ochem
LIRMM, Université de Montpellier, CNRS
Montpellier, France
ochem@lirmm.fr

Submitted: June, 2015; Accepted: XX; Published: XX
Mathematics Subject Classifications: 68R15

Abstract

In combinatorics on words, a word w over an alphabet Σ is said to avoid a pattern p over an alphabet Δ if there is no factor f of w such that $f = h(p)$ where $h : \Delta^* \to \Sigma^*$ is a non-erasing morphism. A pattern p is said to be k-avoidable if there exists an infinite word over a k-letter alphabet that avoids p. A pattern is said to be doubled if no variable occurs only once. Doubled patterns with at most 3 variables and doubled patterns with at least 6 variables are 3-avoidable. We show that doubled patterns with 4 and 5 variables are also 3-avoidable.

Keywords: Word; Pattern avoidance.

1 Introduction

A pattern p is a non-empty word over an alphabet $\Delta = \{A, B, C, \dots\}$ of capital letters called *variables*. An *occurrence* of p in a word w is a non-erasing morphism $h : \Delta^* \to \Sigma^*$ such that $h(p)$ is a factor of w. The avoidability index $\lambda(p)$ of a pattern p is the size of the smallest alphabet Σ such that there exists an infinite word w over Σ containing no occurrence of p. Bean, Ehrenfeucht, and McNulty [2] and Zimin [14] characterized unavoidable patterns, i.e., such that $\lambda(p) = \infty$. We say that a pattern p is t-avoidable if $\lambda(p) \leqslant t$. For more informations on pattern avoidability, we refer to Chapter 3 of Lothaire's book [8].

It follows from their characterization that every unavoidable pattern contains a variable that occurs once. Equivalently, every doubled pattern is avoidable. Our result is that :

Theorem 1. *Every doubled pattern is* 3*-avoidable.*

Let $v(p)$ be the number of distinct variables of the pattern p. For $v(p) \leqslant 3$, Cassaigne [5] began and I [10] finished the determination of the avoidability index of every pattern with at most 3 variables. It implies in particular that every avoidable pattern with at most 3 variables is 3-avoidable. Moreover, Bell and Goh [3] obtained that every doubled pattern p such that $v(p) \geqslant 6$ is 3-avoidable.

Therefore, as noticed in the conclusion of [11], there remains to prove Theorem 1 for every pattern p such that $4 \leqslant v(p) \leqslant 5$. In this paper, we use both constructions of infinite words and a non-constructive method to settle the cases $4 \leqslant v(p) \leqslant 5$.

Recently, Blanchet-Sadri and Woodhouse [4] and Ochem and Pinlou [11] independently obtained the following.

Theorem 2 ([4, 11])**.** *Let p be a pattern.*

(a) If p has length at least $3 \times 2^{v(p)-1}$ then $\lambda(p) \leqslant 2$.

(b) If p has length at least $2^{v(p)}$ then $\lambda(p) \leqslant 3$.

As noticed in these papers, if p has length at least $2^{v(p)}$ then p contains a doubled pattern as a factor. Thus, Theorem 1 implies Theorem 2.(b).

2 Extending the power series method

In this section, we borrow an idea from the entropy compression method to extend the power series method as used by Bell and Goh [3], Rampersad [13], and Blanchet-Sadri and Woodhouse [4].

Let us describe the method. Let $L \subset \Sigma_m^*$ be a factorial language defined by a set F of forbidden factors of length at least 2. We denote the factor complexity of L by $n_i = |L \cap \Sigma_m^i|$. We define L' as the set of words w such that w is not in L and the prefix of length $|w|-1$ of w is in L. For every forbidden factor $f \in F$, we choose a number $1 \leqslant s_f \leqslant |f|$. Then, for every $i \geqslant 1$, we define an integer a_i such that

$$a_i \geqslant \max_{u \in L} \left| \left\{ v \in \Sigma_m^i \mid uv \in L',\ uv = bf,\ f \in F,\ s_f = i \right\} \right| .$$

We consider the formal power series $P(x) = 1 - mx + \sum_{i \geqslant 1} a_i x^i$. If $P(x)$ has a positive real root x_0, then $n_i \geqslant x_0^{-i}$ for every $i \geqslant 0$.

Let us rewrite that $P(x_0) = 1 - mx_0 + \sum_{i \geqslant 1} a_i x_0^i = 0$ as

$$m - \sum_{i \geqslant 1} a_i x_0^{i-1} = x_0^{-1} \tag{1}$$

Since $n_0 = 1$, we will prove by induction that $\frac{n_i}{n_{i-1}} \geqslant x_0^{-1}$ in order to obtain that $n_i \geqslant x_0^{-i}$ for every $i \geqslant 0$. By using (1), we obtain the base case: $\frac{n_1}{n_0} = n_1 = m \geqslant x_0^{-1}$. Now, for every length $i \geqslant 1$, there are:

- m^i words in Σ_m^i,
- n_i words in L,

- at most $\sum_{1\leqslant j\leqslant i} n_{i-j}a_j$ words in L',
- $m(m^{i-1}-n_{i-1})$ words in $\Sigma_m^i \setminus \{L \cup L'\}$.

This gives $n_i+\sum_{1\leqslant j\leqslant i} n_j a_{i-j}+m(m^{i-1}-n_{i-1}) \geqslant m^i$, that is, $n_i \geqslant mn_{i-1}-\sum_{1\leqslant j\leqslant i} n_{i-j}a_j$.

$$\begin{aligned}
\tfrac{n_i}{n_{i-1}} &\geqslant m-\textstyle\sum_{1\leqslant j\leqslant i} a_j \tfrac{n_{i-j}}{n_{i-1}} & \\
&\geqslant m-\textstyle\sum_{1\leqslant j\leqslant i} a_j x_0^{j-1} & \text{By induction}\\
&\geqslant m-\textstyle\sum_{j\geqslant 1} a_j x_0^{j-1} & \\
&= x_0^{-1} & \text{By (1)}
\end{aligned}$$

The power series method used in previous papers [3, 4, 13] corresponds to the special case such that $s_f=|f|$ for every forbidden factor. Our condition is that $P(x)=0$ for some $x>0$ whereas the condition in these papers is that every coefficient of the series expansion of $\frac{1}{P(x)}$ is positive. The two conditions are actually equivalent (Miller [9] uses a similar criterion). The result in [12] concerns series of the form $S(x)=1+a_1x+a_2x^2+a_3x^3+\ldots$ with real coefficients such that $a_1<0$ and $a_i\geqslant 0$ for every $i\geqslant 2$. It states that every coefficient of the series $1/S(x)=b_0+b_1x+b_2x^2+b_3x^3+\ldots$ is positive if and only if $S(x)$ has a positive real root x_0. Moreover, we have $b_i\geqslant x_0^{-i}$ for every $i\geqslant 0$.

The entropy compression method as developed by Gonçalves, Montassier, and Pinlou [6] uses a condition equivalent to $P(x)=0$. The benefit of the present method is that we get an exponential lower bound on the factor complexity. It is not clear whether it is possible to get such a lower bound when using entropy compression for graph coloring, since words have a simpler structure than graphs.

3 Applying the method

In this section, we show that some doubled patterns on 4 and 5 variables are 3-avoidable. Given a pattern p, every occurrence f of p is a forbidden factor. With an abuse of notation, we denote by $|A|$ the length of the image of the variable A of p in the occurrence f. This notation is used to define the length s_f.

Let us first consider doubled patterns with 4 variables. We begin with patterns of length 9, so that one variable, say A, appears 3 times. We set $s_f=|f|$. Using the obvious upper bound on the number of pattern occurrences, we obtain

$$\begin{aligned}
P(x) &= 1-3x+\textstyle\sum_{a,b,c,d\geqslant 1} 3^{a+b+c+d}x^{3a+2b+2c+2d}\\
&= 1-3x+\textstyle\sum_{a,b,c,d\geqslant 1} \left(3x^3\right)^a\left(3x^2\right)^b\left(3x^2\right)^c\left(3x^2\right)^d\\
&= 1-3x+\left(\textstyle\sum_{a\geqslant 1}\left(3x^3\right)^a\right)\left(\textstyle\sum_{b\geqslant 1}\left(3x^2\right)^b\right)\left(\textstyle\sum_{c\geqslant 1}\left(3x^2\right)^c\right)\left(\textstyle\sum_{d\geqslant 1}\left(3x^2\right)^d\right)\\
&= 1-3x+\left(\tfrac{1}{1-3x^3}-1\right)\left(\tfrac{1}{1-3x^2}-1\right)\left(\tfrac{1}{1-3x^2}-1\right)\left(\tfrac{1}{1-3x^2}-1\right)\\
&= 1-3x+\left(\tfrac{1}{1-3x^3}-1\right)\left(\tfrac{1}{1-3x^2}-1\right)^3\\
&= \tfrac{1-3x-9x^2+24x^3+36x^4-54x^5-108x^6+243x^8+162x^9-243x^{10}}{(1-3x^3)(1-3x^2)^3}.
\end{aligned}$$

Then $P(x)$ admits $x_0=0.3400\ldots$ as its smallest positive real root. So, every doubled pattern p with 4 variables and length 9 is 3-avoidable and there exist at least $x_0^{-n}>2.941^n$

ternary words avoiding p. Notice that for patterns with 4 variables and length at least 10, every term of $\sum_{a,b,c,d\geqslant 1} 3^{a+b+c+d}x^{3a+2b+2c+2d}$ in $P(x)$ gets multiplied by some positive power of x. Since $0 < x < 1$, every term is now smaller than in the previous case. So $P(x)$ admits a smallest positive real root that is smaller than $0.3400\ldots$ Thus, these patterns are also 3-avoidable.

Now, we consider patterns with length 8, so that every variable appears exactly twice. If such a pattern has $ABCD$ as a prefix, then we set $s_f = \frac{|f|}{2} = |A| + |B| + |C| + |D|$. So we obtain $P(x) = 1 - 3x + \sum_{a,b,c,d\geqslant 1} x^{a+b+c+d} = 1 - 3x + \left(\frac{1}{1-x} - 1\right)^4$. Then $P(x)$ admits $0.3819\ldots$ as its smallest positive real root, so that this pattern is 3-avoidable.

Among the remaining patterns, we rule out patterns containing an occurrence of a doubled pattern with at most 3 variables. Also, if one pattern is the reverse of another, then they have the same avoidability index and we consider only one of the two. Thus, there remain the following patterns: $ABACBDCD$, $ABACDBDC$, $ABACDCBD$, $ABCADBDC$, $ABCADCBD$, $ABCADCDB$, and $ABCBDADC$.

Now we consider doubled patterns with 5 variables. Similarly, we rule out every pattern of length at least 11 with the method by setting $s_f = |f|$. Then we check that $P(x) = 1 - 3x + \sum_{a,b,c,d,e\geqslant 1} 3^{a+b+c+d+e}x^{3a+2b+2c+2d+2e} = 1 - 3x + \left(\frac{1}{1-3x^3} - 1\right)\left(\frac{1}{1-3x^2} - 1\right)^4$ has a positive real root.

We also rule out every pattern of length 10 having ABC as a prefix. We set $s_f = |f| - |ABC| = |A| + |B| + |C| + 2|D| + 2|E|$. Then we check that $P(x) = 1 - 3x + \sum_{a,b,c,d,e\geqslant 1} 3^{d+e}x^{a+b+c+2d+2e} = 1 - 3x + \left(\frac{1}{1-x} - 1\right)^3\left(\frac{1}{1-3x^2} - 1\right)^2$ has a positive real root.

Again, we rule out patterns containing an occurrence of a doubled pattern with at most 4 variables and patterns whose reversed pattern is already considered. Thus, there remain the following patterns: $ABACBDCEDE$, $ABACDBCEDE$, and $ABACDBDECE$.

4 Sporadic doubled patterns

In this section, we consider the 10 doubled patterns on 4 and 5 variables whose 3-avoidability has not been obtained in the previous section.

We define the *avoidability exponent* $AE(p)$ of a pattern p as the largest real α such that every α-free word avoids p. This notion is not pertinent e.g. for the pattern $ABWBAXACYCAZBC$ studied by Baker, McNulty, and Taylor [1], since for every $\epsilon > 0$, there exists a $(1+\epsilon)$-free word containing an occurrence of that pattern. However, $AE(p) > 1$ for every doubled pattern. To see that, consider a factor $A\ldots A$ of p. If an α-free word contains an occurrence of p, then the image of this factor is a repetition such that the image of A cannot be too large compared to the images of the variables occurring between the As in p. We have similar constraints for every variable and this set of constraints becomes unsatisfiable as α decreases towards 1. We present one way of obtaining a lower bound on the avoidability exponent for a doubled pattern p of length $2v(p)$. We construct the $v(p) \times v(p)$ matrix M such that $M_{i,j}$ is the number of occurrences of the variable X_j between the two occurrences of the variable X_i. Let us show that $AE(p) \geqslant 1 + \frac{1}{\beta+1}$ where β is the largest eigenvalue of M. We consider an occurrence of p and we note $\ell_i = |A_i|$. In an α-free word, the image of the factor $X_i \dots X_i$ of p implies that $\frac{2\ell_i+\sum_{1\leqslant j\leqslant v(p)} M_{i,j}\ell_j}{\ell_i+\sum_{1\leqslant j\leqslant v(p)} M_{i,j}\ell_j} < \alpha$, that is, $l_i < \frac{\alpha-1}{2-\alpha}\sum_{1\leqslant j\leqslant v(p)} M_{i,j}\ell_j$. Thus, the vector $V = \begin{bmatrix} \ell_1 \\ \vdots \\ \ell_{v(p)} \end{bmatrix}$ must satisfy $V < \frac{\alpha-1}{2-\alpha}MV$. This implies that the largest eigenvalue β of M satisfies $\beta > \frac{2-\alpha}{\alpha-1}$, that is, $\alpha > 1 + \frac{1}{\beta+1}$. Hence, if $\alpha \leqslant 1 + \frac{1}{\beta+1}$, then every α-free word avoids p. So $AE(p) \geqslant 1 + \frac{1}{\beta+1}$.

For example if $p = ABACDCBD$, then we get $M = \begin{bmatrix} 0&1&0&0 \\ 1&0&0&1 \\ 0&2&0&1 \\ 0&1&1&0 \end{bmatrix}$, $\beta = 1.9403\dots$, and $AE(p) \geqslant 1 + \frac{1}{\beta+1} = 1.3400\dots$. The avoidability exponents of the 10 patterns considered in this section range from $AE(ABCADBDC) \geqslant 1.292893219$ to $AE(ABACBDCD) \geqslant 1.381966011$. For each pattern p among the 10, we give a uniform morphism $m : \Sigma_5^* \to \Sigma_2^*$ such that for every $\left(\frac{5}{4}^+\right)$-free word $w \in \Sigma_5^*$, we have that $m(w)$ avoids p. The proof that p is avoided follows the method in [10]. Since there exist exponentially many $\left(\frac{5}{4}^+\right)$-free words over Σ_5 [7], there exist exponentially many binary words avoiding p.

- $AE(ABACBDCD) \geqslant 1.381966011$, 17-uniform morphism

$$\begin{aligned}
0 &\mapsto 00000111101010110 \\
1 &\mapsto 00000110100100110 \\
2 &\mapsto 00000011100110111 \\
3 &\mapsto 00000011010101111 \\
4 &\mapsto 00000011001001011
\end{aligned}$$

- $AE(ABACDBDC) \geqslant \frac{4}{3} = 1.333333333$, 33-uniform morphism

$$\begin{aligned}
0 &\mapsto 000000101101000111111011001010111 \\
1 &\mapsto 000000100110100001111101001010111 \\
2 &\mapsto 000000010110100001111111010010111 \\
3 &\mapsto 000000010011010100011111010010111 \\
4 &\mapsto 000000010011001000001111010010111
\end{aligned}$$

- $AE(ABACDCBD) \geqslant 1.340090632$, 28-uniform morphism

$$\begin{aligned}
0 &\mapsto 0000101010001110010000111111 \\
1 &\mapsto 0000001111010001101001111111 \\
2 &\mapsto 0000001101000011110100100111 \\
3 &\mapsto 0000001011110000110100111111 \\
4 &\mapsto 0000001010111100100001111111
\end{aligned}$$

- $AE(ABCADBDC) \geqslant 1.292893219$, 21-uniform morphism

$$\begin{array}{l}
0 \mapsto 000011101101011111010 \\
1 \mapsto 000010110100100111101 \\
2 \mapsto 000001101110100101111 \\
3 \mapsto 000001101011001111111 \\
4 \mapsto 000000110111010111111
\end{array}$$

- $AE(ABCADCBD) \geqslant 1.295597743$, 22-uniform morphism

$$\begin{array}{l}
0 \mapsto 0000011011010100011111 \\
1 \mapsto 0000011010101001001111 \\
2 \mapsto 0000001101100100111111 \\
3 \mapsto 0000001010110000111111 \\
4 \mapsto 0000000110101001110111
\end{array}$$

- $AE(ABCADCDB) \geqslant 1.327621756$, 26-uniform morphism

$$\begin{array}{l}
0 \mapsto 00000011110010101011000111 \\
1 \mapsto 00000011010111111001011011 \\
2 \mapsto 00000010011111101001110111 \\
3 \mapsto 00000001001111110001010111 \\
4 \mapsto 00000001000111111001010111
\end{array}$$

- $AE(ABCBDADC) \geqslant 1.302775638$, 33-uniform morphism

$$\begin{array}{l}
0 \mapsto 000000101111110011000110011111101 \\
1 \mapsto 000000101111001000001100111111101 \\
2 \mapsto 000000011011111001100000100111101 \\
3 \mapsto 000000011010101011000001001111101 \\
4 \mapsto 000000010111111001010101001111011
\end{array}$$

- $AE(ABACBDCEDE) \geqslant 1.366025404$, 15-uniform morphism

$$\begin{array}{l}
0 \mapsto 001011011110000 \\
1 \mapsto 001010100111111 \\
2 \mapsto 000110010011000 \\
3 \mapsto 000011111111100 \\
4 \mapsto 000011010101110
\end{array}$$

- $AE(ABACDBCEDE) \geqslant 1.302775638$, 18-uniform morphism

$$\begin{array}{l}
0 \mapsto 000010110100100111 \\
1 \mapsto 000010100111111111 \\
2 \mapsto 000000110110011111 \\
3 \mapsto 000000101010101111 \\
4 \mapsto 000000000111100111
\end{array}$$

- $AE(ABACDBDECE) \geqslant 1.320416579$, 22-uniform morphism

$$\begin{aligned}
0 &\mapsto 0000001111110001011011\\
1 &\mapsto 0000001111100100110101\\
2 &\mapsto 0000001111100001101101\\
3 &\mapsto 0000001111001001011100\\
4 &\mapsto 0000001111000010101100
\end{aligned}$$

5 Simultaneous avoidance of doubled patterns

Bell and Goh [3] have also considered the avoidance of multiple patterns simultaneously and ask (question 3) whether there exist an infinite word over a finite alphabet that avoids every doubled pattern. We give a negative answer.

A word w is *n-splitted* if $|w| \equiv 0 \pmod{n}$ and every factor w_i such that $w = w_1 w_2 \dots w_n$ and $|w_i| = \frac{|w|}{n}$ for $1 \leqslant i \leqslant n$ contains every letter in w. An n-splitted pattern is defined similarly. Let us prove by induction on k that every word $w \in \Sigma_k^{n^k}$ contains an n-splitted factor. The assertion is true for $k = 1$. Now, if the word $w \in \Sigma_k^{n^k}$ is not itself n-splitted, then by definition it must contain a factor w_i that does not contain every letter of w. So we have $w_i \in \Sigma_{k-1}^{n^{k-1}}$. By induction, w_i contains an n-splitted factor, and so does w.

This implies that for every fixed n, every infinite word over a finite alphabet contains n-splitted factors. Moreover, an n-splitted word is an occurrence of an n-splitted pattern such that every variable has a distinct image of length 1. So, for every fixed n, the set of all n-splitted patterns is not avoidable by an infinite word over a finite alphabet.

Notice that if $n \geqslant 2$, then an n-splitted word (resp. pattern) contains a 2-splitted word (resp. pattern) and a 2-splitted word (resp. pattern) is doubled.

6 Conclusion

Our results answer to the first of two questions of our previous paper [11]. The second question is whether there exists a finite k such that every doubled pattern with at least k variables is 2-avoidable. As already noticed [11], such a k is at least 5 since, e.g., $ABCCBADD$ is not 2-avoidable.

Acknowledgments

I am grateful to Narad Rampersad for comments on a draft of the paper, to Vladimir Dotsenko for pointing out the result in [12], and to Andrei Romashchenko for translating this paper.

References

[1] K.A. Baker, G.F. McNulty, and W. Taylor. Growth problems for avoidable words, *Theoret. Comput. Sci.* **69** (1989), 319–345.

[2] D.R. Bean, A. Ehrenfeucht, and G.F. McNulty. Avoidable patterns in strings of symbols. *Pacific J. of Math.* **85** (1979), 261–294.

[3] J. Bell, T. L. Goh. Exponential lower bounds for the number of words of uniform length avoiding a pattern. *Inform. and Comput.* **205** (2007), 1295-1306.

[4] F. Blanchet-Sadri, B. Woodhouse. Strict bounds for pattern avoidance. *Theor. Comput. Sci.* **506** (2013), 17–27.

[5] J. Cassaigne. Motifs évitables et régularité dans les mots. Thèse de Doctorat, Université Paris VI, Juillet 1994.

[6] D. Gonçalves, M. Montassier, and A. Pinlou. Entropy compression method applied to graph colorings. arXiv:1406.4380

[7] R. Kolpakov and M. Rao: On the number of Dejan words over alphabets of 5, 6, 7, 8, 9 and 10 letters. *Theor. Comput. Sci.* **412(46)** (2011), 6507–6516.

[8] M. Lothaire. Algebraic Combinatorics on Words. *Cambridge Univ. Press* (2002).

[9] J. Miller. Two notes on subshifts. *Proc. Amer. Math. Soc.* **140** (2012), 1617–1622.

[10] P. Ochem. A generator of morphisms for infinite words. *RAIRO: Theoret. Informatics Appl.* **40** (2006), 427–441.

[11] P. Ochem and A. Pinlou. Application of entropy compression in pattern avoidance. *Electron. J. Combinatorics.* **21(2)** (2014), #RP2.7.

[12] D. I. Piotkovskii. On the growth of graded algebras with a small number of defining relations. *Uspekhi Mat. Nauk.* **48:3(291)** (1993), 199–200.

[13] N. Rampersad. Further applications of a power series method for pattern avoidance. *Electron. J. Combinatorics.* **18(1)** (2011), #P134.

[14] A.I. Zimin. Blocking sets of terms. *Math. USSR Sbornik* **47(2)** (1984), 353–364. English translation.

6.4 Circular formulas

Avoidability of circular formulas

Guilhem Gamard[a], Pascal Ochem[a,b], Gwenaël Richomme[a,c], Patrice Séébold[a,c]

[a] *LIRMM, Université de Montpellier and CNRS, France*
[b] *CNRS*
[c] *Université Paul-Valéry Montpellier 3*

Abstract

Clark has defined the notion of n-avoidance basis which contains the avoidable formulas with at most n variables that are closest to be unavoidable in some sense. The family C_i of circular formulas is such that $C_1 = AA$, $C_2 = ABA.BAB$, $C_3 = ABCA.BCAB.CABC$ and so on. For every $i \leqslant n$, the n-avoidance basis contains C_i. Clark showed that the avoidability index of every circular formula and of every formula in the 3-avoidance basis (and thus of every avoidable formula containing at most 3 variables) is at most 4. We determine exactly the avoidability index of these formulas.

1. Introduction

A *pattern* p is a non-empty finite word over an alphabet $\Delta = \{A, B, C, \dots\}$ of capital letters called *variables*. An *occurrence* of p in a word w is a non-erasing morphism $h : \Delta^* \to \Sigma^*$ such that $h(p)$ is a factor of w. The *avoidability index* $\lambda(p)$ of a pattern p is the size of the smallest alphabet Σ such that there exists an infinite word over Σ containing no occurrence of p. Bean, Ehrenfeucht, and McNulty [2] and Zimin [13] characterized unavoidable patterns, i.e., such that $\lambda(p) = \infty$. We say that a pattern p is *t-avoidable* if $\lambda(p) \leqslant t$. For more information on pattern avoidability, we refer to Chapter 3 of Lothaire's book [8]. See also this book for basic notions in Combinatorics on Words.

A variable that appears only once in a pattern is said to be *isolated*. Following Cassaigne [3], we associate to a pattern p the *formula* f obtained by replacing every isolated variable in p by a dot. The factors between the dots are called *fragments*.

An *occurrence* of a formula f in a word w is a non-erasing morphism $h : \Delta^* \to \Sigma^*$ such that the h-image of every fragment of f is a factor of w. As for patterns, the avoidability index $\lambda(f)$ of a formula f is the size of the smallest alphabet allowing the existence of an infinite word containing no occurrence of f. Clearly, if a formula f is associated to a pattern p, every word avoiding f also avoids p, so $\lambda(p) \leqslant \lambda(f)$. Recall that an infinite word is *recurrent* if every finite factor appears infinitely many times. If there exists an infinite word over Σ avoiding p, then there exists an infinite recurrent word over Σ avoiding p. This recurrent word also avoids f, so that $\lambda(p) = \lambda(f)$. Without loss of generality,

Preprint submitted to Elsevier *December 1, 2017*

a formula is such that no variable is isolated and no fragment is a factor of another fragment.

Cassaigne [3] began and Ochem [9] finished the determination of the avoidability index of every pattern with at most 3 variables. A *doubled* pattern contains every variable at least twice. Thus, a doubled pattern is a formula with exactly one fragment. Every doubled pattern is 3-avoidable [10]. A formula is said to be *binary* if it has at most 2 variables. The avoidability index of every binary formula has been recently determined [11]. We say that a formula f is *divisible* by a formula f' if f does not avoid f', that is, there is a non-erasing morphism h such that the image of every fragment of f' by h is a factor of a fragment of f. If f is divisible by f', then every word avoiding f' also avoids f and thus $\lambda(f) \leqslant \lambda(f')$. Moreover, the reverse f^R of a formula f satisfies $\lambda(f^R) = \lambda(f)$. For example, the fact that $ABA.AABB$ is 2-avoidable implies that $ABAABB$ and $BAB.AABB$ are 2-avoidable. See Cassaigne [3] and Clark [4] for more information on formulas and divisibility.

Clark [4] has introduced the notion of *n-avoidance basis* for formulas, which is the smallest set of formulas with the following property: for every $i \leqslant n$, every avoidable formula with i variables is divisible by at least one formula with at most i variables in the n-avoidance basis.

From the definition, it is not hard to obtain that the 1-avoidance basis is $\{AA\}$ and the 2-avoidance basis is $\{AA, ABA.BAB\}$. Clark obtained that the 3-avoidance basis is composed of the following formulas:

- AA
- $ABA.BAB$
- $ABCA.BCAB.CABC$
- $ABCBA.CBABC$
- $ABCA.CABC.BCB$
- $ABCA.BCAB.CBC$
- $AB.AC.BA.CA.CB$

The following properties of the avoidance basis are derived.

- The n-avoidance basis is a subset of the $(n+1)$-avoidance basis.
- The n-avoidance basis is closed under reverse. (In particular, $ABCA.BCAB.CBC$ is the reverse of $ABCA.CABC.BCB$.)
- Two formulas in the n-avoidance basis with the same number of variables are incomparable by divisibility. (However, AA is divisible $AB.AC.BA.CA.CB$.)
- The n-avoidance basis is computable.

The *circular formula* C_t is the formula over $t \geqslant 1$ variables $A_0, \dots, A_{t-1}$ containing the t fragments of the form $A_i A_{i+1} \dots A_{i+t}$ such that the indices are taken modulo t. Thus, the first three formulas in the 3-avoidance basis, namely $C_1 = AA$, $C_2 = ABA.BAB$, and $C_3 = ABCA.BCAB.CABC$, are also the first three circular formulas. More generally, for every $t \leqslant n$, the n-avoidance basis contains C_t.

It is known that $\lambda(AA) = 3$ [12], $\lambda(ABA.BAB) = 3$ [3], and $\lambda(AB.AC.BA.CA.CB) = 4$ [1]. Actually, $AB.AC.BA.CA.CB$ is avoided by the fixed point $b_4 = 0121032101230321\dots$ of the morphism given below.

$$\begin{array}{l} 0 \mapsto 01 \\ 1 \mapsto 21 \\ 2 \mapsto 03 \\ 3 \mapsto 23 \end{array}$$

Clark [4] obtained that b_4 also avoids C_i for every $i \geqslant 1$, so that $\lambda(C_i) \leqslant 4$ for every $i \geqslant 1$. He also showed that the avoidability index of the other formulas in the 3-avoidance basis is at most 4. Our main results finish the determination of the avoidability index of the circular formulas (Theorem 1) and the formulas in the 3-avoidance basis (Theorem 4).

2. Conjugacy classes and circular formulas

In this section, we determine the avoidability index of circular formulas.

Theorem 1. *$\lambda(C_3) = 3$. $\forall i \geqslant 4$, $\lambda(C_i) = 2$.*

We consider a notion that appears to be useful in the study of circular formulas. A *conjugacy class* is the set of all the conjugates of a given word, including the word itself. The length of a conjugacy class is the common length of the words in the conjugacy class. A word contains a conjugacy class if it contains every word in the conjugacy class as a factor. Consider the uniform morphisms given below.

$$\begin{array}{lll} g_2(0) = 0000101001110110100 & g_3(0) = 0010 & g_6(0) = 01230 \\ g_2(1) = 0011100010100111101 & g_3(1) = 1122 & g_6(1) = 24134 \\ g_2(2) = 0000111100010110100 & g_3(2) = 0200 & g_6(2) = 52340 \\ g_2(3) = 0011110110100111101 & g_3(3) = 1212 & g_6(3) = 24513 \end{array}$$

Lemma 2.

- *The word $g_2(b_4)$ avoids every conjugacy class of length at least 5.*
- *The word $g_3(b_4)$ avoids every conjugacy class of length at least 3.*
- *The word $g_6(b_4)$ avoids every conjugacy class of length at least 2.*

Proof. We only detail the proof for $g_2(b_4)$, since the proofs for $g_3(b_4)$ and $g_6(b_4)$ are similar. Notice that g_2 is 19-uniform. First, a computer check shows that $g_2(b_4)$ contains no conjugacy class of length i with $5 \leqslant i \leqslant 55$ (i.e., $2 \times 19 + 17$).

Suppose for contradiction that $g_2(b_4)$ contains a conjugacy class of length at least 56 (i.e., $2 \times 19 + 18$). Then every element of the conjugacy class contains a factor $g_2(ab)$ with $a, b \in \Sigma_4$. In particular, one of the elements of the conjugacy class can be written as $g_2(ab)s$. The word $g_2(b)sg_2(a)$ is also a factor of $g_2(b_4)$. A computer check shows that for every letters α, β, and γ in Σ_4 such that $g_2(\alpha)$ is a factor of $g_2(\beta\gamma)$, $g_2(\alpha)$ is either a prefix or a suffix of $g_2(\beta\gamma)$. This implies that s belongs to $g_2(\Sigma_4^+)$.

Thus, the conjugacy class contains a word $w = g_2(\ell_1\ell_2 \dots \ell_k) = x_1x_2...x_{19k}$. Consider the conjugate $\tilde{w} = x_7x_8 \dots x_{19k}x_1x_2x_3x_4x_5x_6$. Observe that the prefixes of length 6 of $g_2(0)$, $g_2(1)$, $g_2(2)$, and $g_2(3)$ are different. Also, the suffixes of length 12 of $g_2(0)$, $g_2(1)$, $g_2(2)$, and $g_2(3)$ are different. Then the prefix $x_7 \dots x_{19}$ and the suffix $x_1 \dots x_6$ of $\tilde{w}$ both force the letter ℓ_1 in the pre-image. That is, b_4 contains $\ell_1\ell_2 \dots \ell_k\ell_1$. Similarly, the conjugate of w that starts with the letter $x_{19(r-1)+7}$ implies that b_4 contains $\ell_r \dots \ell_k\ell_1 \dots \ell_r$. Thus, b_4 contains an occurrence of the formula C_k. This is a contradiction since Clark [4] has shown that b_4 avoids every circular formula C_i with $i \geqslant 1$. □

Notice that if a word contains an occurrence of C_i, then it contains a conjugacy class of length at least i. Thus, a word avoiding every conjugacy class of length at least i also avoids every circular formula C_t with $t \geqslant i$. Moreover, $g_2(b_4)$ contains no occurrence of C_4 such that the length of the image of every variable is 1. By Lemma 2, this gives the next result, which proves Theorem 1.

Corollary 3. *The word $g_3(b_4)$ avoids every circular formula C_i with $i \geqslant 3$. The word $g_2(b_4)$ avoids every circular formula C_i with $i \geqslant 4$.*

3. Remaining formulas in the 3-avoidance basis

In this section, we prove the following result which completes the determination of the avoidability index of the formulas in the 3-avoidance basis.

Theorem 4. $\lambda(ABCBA.CBABC) = 2$. $\lambda(ABCA.CABC.BCB) = 3$.

Notice that $\lambda(ABCBA.CBABC) = 2$ implies the well-known fact that $\lambda(ABABA) = 2$. It also implies that $\lambda(ABCBABC) = 2$, which was first obtained in [6].

For both formulas, we give a uniform morphism m such that for every $\left(\frac{5}{4}^+\right)$-free word $w \in \Sigma_5^*$, the word $m(w)$ avoids the formula. Since there exist exponentially many $\left(\frac{5}{4}^+\right)$-free words over Σ_5 [7], there exist exponentially many words avoiding the formula. The proof that the formula is avoided follows the method in [9].

To avoid $ABCBA.CBABC$, we use this 15-uniform morphism:

$$\begin{aligned}
m_{15}(0) &= \texttt{001111010010110}\\
m_{15}(1) &= \texttt{001110100101110}\\
m_{15}(2) &= \texttt{001101001011110}\\
m_{15}(3) &= \texttt{000111010001011}\\
m_{15}(4) &= \texttt{000110100001011}
\end{aligned}$$

First, we show that the m_{15}-image of every $\left(\frac{5}{4}^+\right)$-free word w is $\left(\frac{97}{75}^+, 61\right)$-free, that is, $m_{15}(w)$ contains no repetition with period at least 61 and exponent strictly greater than $\frac{97}{75}$. By Lemma 2.1 in [9], it is sufficient to check this property for every $\left(\frac{5}{4}^+\right)$-free word w such that $|w| < \frac{2\times\frac{97}{75}}{\frac{97}{75}-\frac{5}{4}} < 60$. Consider a potential occurrence h of $ABCBA.CBABC$ and write $a = |h(A)|$, $b = |h(B)|$, $c = |h(C)|$. Suppose that $a + b \geqslant 61$. The factor $h(BAB)$ is then a repetition with period $a + b \geqslant 61$, so that its exponent satisfies $\frac{a+2b}{a+b} \leqslant \frac{97}{75}$. This gives $53b \leqslant 22a$. Similarly, BCB implies $53b \leqslant 22c$, $ABCBA$ implies $53a \leqslant 22(2b+c)$, and $CBABC$ implies $53c \leqslant 22(a + 2b)$. Summing up these inequalities gives $53a + 106b + 53c \leqslant 44a + 88b + 44c$, which is a contradiction. Thus, we have $a + b \leqslant 60$. By symmetry, we also have $b + c \leqslant 60$. Using these inequalities, we check exhaustively that $h(w)$ contains no occurrence of $ABCBA.CBABC$.

To avoid $ABCA.CABC.BCB$ and its reverse $ABCA.BCAB.CBC$ simultaneously, we use this 6-uniform morphism:

$$\begin{aligned}
m_6(\mathtt{0}) &= \mathtt{021210}\\
m_6(\mathtt{1}) &= \mathtt{012220}\\
m_6(\mathtt{2}) &= \mathtt{012111}\\
m_6(\mathtt{3}) &= \mathtt{002221}\\
m_6(\mathtt{4}) &= \mathtt{001112}
\end{aligned}$$

We check that the m_6-image of every $\left(\frac{5}{4}^+\right)$-free word w is $\left(\frac{13}{10}^+, 25\right)$-free. By Lemma 2.1 in [9], it is sufficient to check this property for $\left(\frac{5}{4}^+\right)$-free word w such that $|w| < \frac{2\times\frac{13}{10}}{\frac{13}{10}-\frac{5}{4}} = 52$.

Let us consider the formula $ABCA.CABC.BCB$. Suppose that $b + c \geqslant 25$. Then $ABCA$ implies $7a \leqslant 3(b + c)$, $CABC$ implies $7c \leqslant 3(a + b)$, and BCB implies $7b \leqslant 3c$. Summing up these inequalities gives $7a+7b+7c \leqslant 3a+6b+6c$, which is a contradiction. Thus $b + c \leqslant 24$. Suppose that $a \geqslant 23$. Then $ABCA$ implies $a \leqslant \frac{3}{7}(b + c) \leqslant \frac{72}{7} < 23$, which is a contradiction. Thus $a \leqslant 22$. For the formula $ABCA.BCAB.CBC$, the same argument holds except that the roles of B and C are switched, so that we also obtain $b + c \leqslant 24$ and $a \leqslant 22$. Then we check exhaustively that $h(w)$ contains no occurrence of $ABCA.CABC.BCB$ and no occurrence of $ABCA.BCAB.CBC$.

It can be noticed that arguments using repetition to forbid patterns has also been used in [5]

4. Concluding remarks

A major open question is whether there exist avoidable formulas with arbitrarily large avoidability index. If such formulas exist, some of them necessarily belong to the n-avoidance basis for increasing values of n. With the example of circular formulas, Clark noticed that belonging to the n-avoidance basis and having many variables does not imply a large avoidability index. Our results strengthen this remark and show that the n-avoidance basis contains a 2-avoidable formula on t variables for every $3 \leqslant t \leqslant n$.

A formula f is *nice* if for every variable X of f there exists a fragment of f that contains X at least twice. This notion generalizes the notion of doubled pattern, which corresponds to a nice formula with one fragment. Notice that every formula in the 3-avoidance basis is nice except $AB.AC.BA.CA.CB$. Thus, our results imply that the nice formulas in the 3-avoidance basis are 3-avoidable. Is every nice formula 3-avoidable?

Concerning conjugacy classes, we propose the following conjecture:

Conjecture 5. *There exists an infinite word in Σ_5^* that avoids every conjugacy class of length at least 2.*

Associated to the results in Lemma 2, this would give the smallest alphabet that allows to avoid every conjugacy class of length at least i, for every i.

References

References

[1] K. A. Baker, G. F. McNulty, and W. Taylor. Growth problems for avoidable words. *Theoret. Comput. Sci.*, 69(3):319–345, 1989.

[2] D. R. Bean, A. Ehrenfeucht, and G. F. McNulty. Avoidable patterns in strings of symbols. *Pacific J. Math.*, 85:261–294, 1979.

[3] J. Cassaigne. *Motifs évitables et régularité dans les mots.* PhD thesis, Université Paris VI, 1994.

[4] R. J. Clark. *Avoidable formulas in combinatorics on words.* PhD thesis, University of California, Los Angeles, 2001. Available at
http://www.lirmm.fr/~ochem/morphisms/clark_thesis.pdf

[5] J.D. Currie and V. Linek. Avoiding Patterns in the Abelian Sense. *Canadian J. Math.*, 53:696–714, 2001.

[6] L. Ilie, P. Ochem, and J.O. Shallit. A generalization of repetition threshold. *Theoret. Comput. Sci.*, 92(2):71–76, 2004.

[7] R. Kolpakov and M. Rao. On the number of Dejean words over alphabets of 5, 6, 7, 8, 9 and 10 letters. *Theoret. Comput. Sci.*, 412(46):6507–6516, 2011.

[8] M. Lothaire. *Algebraic Combinatorics on Words.* Cambridge Univ. Press, 2002.

[9] P. Ochem. A generator of morphisms for infinite words. *RAIRO - Theor. Inform. Appl.*, 40:427–441, 2006.

[10] P. Ochem. Doubled patterns are 3-avoidable. *Electron. J. Combin.*, **23(1)** (2016), #P1.19.

[11] P. Ochem and M. Rosenfeld. Avoidability of formulas with two variables. *Electron. J. Combin.* **24(4)** (2017), #P4.30.

[12] A. Thue. Über unendliche Zeichenreihen. *Norske Vid. Selsk. Skr. I. Mat. Nat. Kl. Christiania*, 7:1–22, 1906.

[13] A. I. Zimin. Blocking sets of terms. *Math. USSR Sbornik*, 47(2):353–364, 1984.

6.5 The formula *AABB.ABBA*

Theoretical Informatics and Applications Will be set by the publisher
Informatique Théorique et Applications

BINARY WORDS AVOIDING THE PATTERN *AABBCABBA*

Pascal Ochem[1, 2]

Abstract. We show that there are three types of infinite words over the two-letter alphabet $\{0,1\}$ that avoid the pattern $AABBCABBA$. These types, P, E_0, and E_1, differ by the factor complexity and the asymptotic frequency of the letter 0. Type P has polynomial factor complexity and letter frequency $\frac{1}{2}$. Type E_0 has exponential factor complexity and the frequency of the letter 0 is at least 0.45622 and at most 0.48684. Type E_1 is obtained from type E_0 by exchanging 0 and 1.

1991 Mathematics Subject Classification. 68R15.

1. Introduction

This paper deals with pattern avoidability [4, 7]. Let Σ_s denote the s-letter alphabet $\{0,1,\ldots,s-1\}$. A *pattern* is a finite word over the alphabet of capital letters $\{A,B,\ldots\}$. An *occurrence* of a pattern is obtained by replacing each alphabet letter by a non-empty word. For example, the word 0111010011 is an occurrence of the pattern $ABBA$ where $A \mapsto 011$ and $B \mapsto 10$; it also contains another occurrence of this pattern (i.e. 1001) as a factor. A word *avoids* a pattern P if it contains no occurrence of P as a factor. The *avoidability index* $\lambda(P)$ of the pattern P is the smallest alphabet size over which an infinite word avoiding P exists. Patterns such as A, ABC, ABA, $ABACBA$ cannot be avoided with any finite alphabet. These patterns such that $\lambda(P) = \infty$ are said to be *unavoidable* and have been characterized by Zimin [11].

Let t_n be the number of words of length n in a language. If that language is closed under taking factors, which is the case for words avoiding a pattern, then t_n is sub-multiplicative and the *growth rate* $\lim_{n\to\infty}(t_n)^{\frac{1}{n}}$ is well-defined. See the survey of Berstel [3] for more information on the growth rate. For a given a pattern P, once its avoidability index is known, it is interesting to consider

[1] CNRS, Lab. J.V. Poncelet, Moscow; e-mail: `ochem@lri.fr`

[2] LRI, Bât 490 Université Paris-Sud 11, 91405 Orsay Cedex France

© EDP Sciences 1999

the factor complexity of words avoiding P over $\Sigma_{\lambda(P)}$, in order to know whether P is "barely" or "easily" avoided over $\Sigma_{\lambda(P)}$. For example, it is known that $\lambda(ABDACEBAFCAGCB) = 4$ and that there are only polynomialy many words over Σ_4 avoiding that pattern [1], so their growth rate is 1. On the other hand, $\lambda(AA) = 3$ and there are exponentially many ternary square-free words, since their growth rate is > 1.30125 [6].

In this paper, we show that binary words avoiding $AABBCABBA$ can be classified into three disjoint types P, E_0, and E_1. Type E_1 is obtained from type E_0 by exchanging 0 and 1. There are polynomialy many words of type P and the asymptotic frequency of the letter 0 in words of type P is $\frac{1}{2}$. There are exponentially many words of type E_0 but their growth rate is small. When it is defined, the frequency of the letter 0 in an infinite word of type E_0 is between 0.45622 and 0.48684. Type E_1 is obtained from type E_0 by exchanging 0 and 1.

2. Three types of words avoiding $AABBCABBA$

A finite word is *recurrent* in an infinite word w if it appears as a factor of w infinitely many times. An infinite word w is *recurrent* if all its finite factors are recurrent in w. We are interested in infinite binary recurrent words avoiding the pattern $AABBCABBA$. Such words equivalently avoid the formula $AABB.ABBA$ (see [4, 5] for more on formulas). This means that for every occurrence of $AABB$ (e.g., 000011) that appears, the corresponding occurrence of $ABBA$ (so, 001100) does not appear, or vice-versa. To see this, suppose that both an occurrence of $AABB$ and the corresponding occurrence of $ABBA$ appear in an infinite recurrent word w. Since these two occurrences are recurrent factors in w, then w must contain, from left to right, the mentioned occurrence of $AABB$, followed by one letter, and then an infinite suffix that has to contain the corresponding occurrence of $ABBA$. This creates an occurrence of $AABBCABBA$.

Remark 2.1. An infinite recurrent word avoiding $AABBCABBA$ also avoids the patterns $AABBA$ and $AAAA$.

This remark is a straigtforward consequence of the property on formulas mentioned above. An occurrence of $AABBA$ contains an occurrence of $AABB$ and the corresponding occurrence of $ABBA$. An occurrence of $AAAA$ is both an occurrence of $AABB$ such that $A = B$ and the corresponding occurrence of $ABBA$.

Figure 1 is a graph whose vertices are the occurrences of length 4 of $AABB$ or $ABBA$ that might be recurrent in an infinite binary word avoiding $AABBCABBA$. The factors 0000 and 1111 have been ruled out since they are occurrences of $AAAA$ (see Remark 2.1). An edge stands for an incompatibility between an occurrence of $AABB$ and the corresponding occurrence of $ABBA$: two factors associated to adjacent vertices cannot be recurrent in a same infinite word avoiding $AABBCABBA$. So, given an infinite binary recurrent word w avoiding $AABBCABBA$, we can associate the set of vertices of the graph that appear as factors in w. Moreover, this set is an independent set.

Let us check that neither an independent set of size at most one nor $\{0011, 1100\}$ can be associated to an infinite binary recurrent word avoiding $AABBCABBA$. By symmetry and maximality, we only need to consider the case of the sets $\{0110\}$ and $\{0011, 1100\}$. In the case of the set $\{0110\}$ (resp. $\{0011, 1100\}$), we can enumerate lexicographically all binary words avoiding the patterns $AABBCABBA$, $AABBA$, and $AAAA$, and the factors 0011, 1100, and 1001 (resp. the factors 0110 and 1001).

There remain three potential types for an infinite binary recurrent word avoiding $AABBCABBA$, that we call P, E_0, and E_1. These three types respectively contain factors in $\{0110, 1001\}$, $\{1100, 0110\}$, and $\{0011, 1001\}$. Notice that by exchanging 0 and 1, type P stays unchanged, type E_0 becomes type E_1, and type E_1 becomes type E_0.

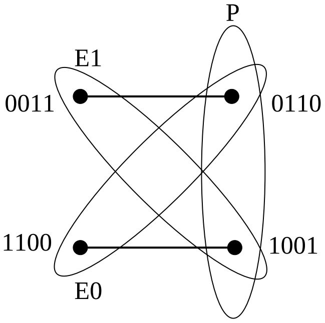

Figure 1. Graph of incompatibilities between factors of length 4

3. Type P has polynomial growth

Let t be the fixed point of the morphism $0 \mapsto 012$, $1 \mapsto 02$, $2 \mapsto 1$, and let h be the morphism defined by

$$\begin{aligned} 0 &\mapsto 0010110111011101001, \\ 1 &\mapsto 00101101101001, \\ 2 &\mapsto 00010. \end{aligned}$$

In this section, we give a characterization of words of type P:

Theorem 3.1. *The set of factors of type P is the set of factors of $h(t)$.*

The following lemma about t is needed in the proof of Theorem 3.1.

Lemma 3.2. *If w is an infinite recurrent ternary square-free word, the following assertions are equivalent:*

- *w has the same set of factors as t,*
- *w contains neither 010 nor 212 as a factor,*

- *w does not contain factors of the form $0v1v0$ with $v \in \Sigma_3^*$.*

Proof. The equivalence of the first and the second assertion is a well known result of Thue (see [2] for a translation). Let us prove the equivalence of the second and third assertion, which is that when considering recurrent languages of ternary square-free word, avoiding factors of the form $0v1v0$ with $v \in \Sigma_3^*$ is equivalent to avoiding the factors 010 and 212. Because of square-freeness, avoiding $0v1v0$ is equivalent to avoiding 010, 02120, and $02v'212v'20$ with $v' \in \Sigma_3^*$. Because it is a recurrent language, avoiding 02120 is equivalent to avoiding 212, since 02120 is the only possible extension of 212 that does not create a square. □

Let us prove one direction of Theorem 3.1, namely that $h(t)$ contains only factors of type P. Since t is recurrent, so is $h(t)$. Since $h(t)$ contains 0110 and 1001, it remains to check that $h(t)$ avoids $AABBCABBA$. First, we show that $h(t)$ contains no square xx with $|x| > 4$. It is easy to check that no such *large square* appears in the h-image of a factor of t of length at most two. Notice also that for every letter $i \in \Sigma_3$, the factor $h(i)$ appears only in $h(t)$ as the h-image of the letter i. This implies that any large square would be a factor of a word of the form $h(pvmvs)$ with $v \in \Sigma_3^*$, $p, m, s \in \Sigma_3$, $p \neq m$, and $m \neq s$. So there would be a large square also in $h(pms)$, which happens only in the case $pms = 010$. Since t contains no factors of the form $0v1v0$ by Lemma 3.2, $h(t)$ contains no square xx with $|x| > 4$. So we can list all the occurrences of the pattern $AABB$ in $h(t)$, because their length is at most 16. Then we can check that for every occurrence of the pattern $AABB$ in $h(t)$, the corresponding occurrence of $ABBA$ is not a factor of $h(t)$.

Now, we prove the other direction of Theorem 3.1, namely that every factor of type P is a factor of $h(t)$. First, we check that a factor of type P is a factor of the h-image of some ternary word. We consider the language P' of binary words avoiding 0011, 1100, $AAAA$, $AABBA$, and $AABBCABBA$. It contains P by Remark 2.1. We compute the set of factors in P' of length $|h(0)| + |h(1)| = 33$ and remove from this set factors that are not prolongable in P'. This can be done with the method described in Section 4, until this set becomes equal to the set of factors of $h(t)$ of length 33. In this set, every factor with prefix $h(i)$ for some $i \in \Sigma_3$ is such that the prefix $h(i)$ is followed by either $h((i+1) \pmod 3)$ or $h((i+2) \pmod 3)$. Thus, a factor of type P is a factor of the h-image of some ternary word.

Let $L \subset \Sigma_3^*$ denote the language of words whose h-image is of type P. Since factors of type P are reccurrent, words in L are bi-prolongable in L. Let $u \in \Sigma_3^+$. We suppose now that L contains a square occurrence uu. Because of the prolongability, this implies that L contains a factor $puus$ for some $p, s \in \Sigma_3$. Since 00 is a common proper prefix of $h(1)$, $h(2)$, and $h(3)$, we can write $h(u) = 00r$ for some $r \in \Sigma_2^+$. The following three cover every possible values of p and s. Each case is ruled out because it contains an occurrence of $AABBA$, which is forbidden by Remark 2.1.

- If $s = 2$, then $h(uu2) = 00r00r00010$ contains an occurrence of $AABBA$ with $A = 0$ and $B = r00$.

- If $p = 2$, then $h(2uus)$ has a prefix $0001000r00r00$ that contains an occurrence of $AABBA$ with $A = 0$ and $B = 0r0$.
- If $p, s \in \{0, 1\}$, then $h(puus)$ contains a factor $0100100r00r0010$ because 01001 is a common suffix of $h(0)$ and $h(1)$, and 0010 is a common prefix of $h(0)$ and $h(1)$. This factor is an occurrence of $AABBA$ with $A = 010$ and $B = 0r0$.

This shows that the language L contains square-free words only.

Factors of the form $0v1v0$ with $v \in \Sigma_3^*$ are not in L since their image by h contains the factor $1101001h(v)00101101101001h(v)0010110111$ which is an occurrence of $AABBA$ with $A = 1$ and $B = 01001h(v)001011011$.

To summarize, every factor of type P is a factor of the h-image of some recurrent ternary square-free word avoiding factors of the form $0v1v0$ with $v \in \Sigma_3^*$. By Lemma 3.2, every factor of type P is thus a factor of $h(t)$. This concludes the proof of Theorem 3.1.

As a corollary of Theorem 3.1, words of type P have polynomial growth.

4. Types E_0 and E_1 have exponential growth

Theorem 4.1. *The growth rate for words of type E_0 is between* 1.002584956 *and* 1.02930952*.*

Proof. For the lower bound, we extend the result [7] that the image of any ternary $\frac{7}{4}^+$-free word by the following 102-uniform morphism k avoids $AABBCABBA$.

$$0 \mapsto w0010110111011100010110001000101101110$$
$$1 \mapsto w1100010110111011100010110001000101101$$
$$2 \mapsto w1110001011000100010110111011000101101$$

with $w =$11000101101110111000101101110110001011011100010110001000101101110.

These words avoiding $AABBCABBA$ are actually of type E_0 since they are recurrent and contain the factors 1100 and 0110.

Kolpakov [6] has shown that the growth rate of ternary $\frac{7}{4}^+$-free (resp. square-free) words is at least 1.245 (resp. 1.30125).

Ternary $\frac{7}{4}^+$-free words were used [7] as pre-image for k in order to have simple and standardized proofs. To get the lower bound of Theorem 4.1, we need the stronger statement that the k-image of any ternary square-free word avoids $AABBCABBA$. We can prove this by checking that the k-image of any ternary square-free word of length 3 contains no square xx with $|x| > 26$. Then again, for each occurrence of $AABB$ in the k-image of some ternary square-free word, we can check that the corresponding occurrence of $ABBA$ does not appear. The growth rate of words of type E_0 is thus at least $1.30125^{1/102} > 1.0025849$.

For the upper bound, we basically use our method [9] that gave an upper bound on the growth rate of ternary square-free words. We have noticed that the notion of prolongability is much more important for words of type E_0 than for ternary square-free words (maybe because the growth rate is much lower). For example, in a ternary square-free word pws such that $|w| = 50$ and $|p| = |s| = 15$, the factor w is very probably a recurent factor in some infinite ternary square-free word. This is not the case for type E_0. We take this behavior into account by computing iteratively a set of words of some length avoiding $AABBCABBA$, 0011, and 1001 from another such set. These sets contain all words of type E_0 of the specified length but maybe also other words that are not prolongable. Let $f(n, e, S, k)$ be the function that computes the set of words w such that pws avoids $AABBCABBA$, 0011, and 1001, $|w| = n$, $|p| = |s| = e$, and every factor of length k of pws belongs to S, where S is a previously computed set of words of length k. For example (with fictional values), we can first compute a set of words of length 40 from scratch: $S_1 \leftarrow f(40, 5, \emptyset, 0)$. Then a set of words of length 50 from S_1: $S_2 \leftarrow f(50, 10, S_1, 40)$. Then another set of words of length 50 from S_2: $S_3 \leftarrow f(50, 10, S_2, 50)$. Of course, we have that $S_3 \subseteq S_2$ and hope that $S_3 \subset S_2$. Maybe even the set of prefixes of length 40 of words in S_3 is smaller than the initial set S_1. The user thus computes sets of words of increasing size and obtain a set of words that are prolongable by at least e letters, where e is the second parameter in the final call. Cassaigne [4] described a similar method using Rauzy graphs. We have obtained a set S of words of length 360 that are prolongable by 40 letters to the left and to the right.

The upper bound in Theorem 4.1 has been obtained by applying the transfert matrix method [9] with parameters $k = 359$ and $l = 101$. That is, we constructed a matrix M such that $M[i, j]$ is the number of factors of length $k + l = 460$ whose prefix (resp. suffix) is the i^{th} (resp. j^{th}) factor of length k. Then the upper bound is obtained by taking the l^{th} root of the largest eigenvalue of M. Compared to the calculation described in [9], we made the following modifications: we used an adjacency list representation, because the matrix here is much sparser, and we required that only the words w of length $k + l$ such that every factor of w of length 360 belongs to S are taken into account in the matrix. Shur [10] presented another method for upper bounds on the growth rate that gives a better result for ternary square-free words. It would be interesting to check if his method also gives a better bound for words of type E_0. □

5. Letter frequencies

Let $|v|_i$ denote the number of occurrences of the letter i in the finite word v.

Theorem 5.1. *Let w be an infinite recurrent word avoiding $AABBCABBA$. For all $\varepsilon > 0$, there exists an integer n_ε such that the frequency $\frac{|v|_0}{|v|}$ of the letter 0 in every finite factor v of w with length at least n_ε is in*

- $\left[\frac{1}{2} - \varepsilon, \frac{1}{2} + \varepsilon\right]$ *if w is of type P,*
- $\left[\frac{271}{594} - \varepsilon, \frac{37}{76} + \varepsilon\right]$ *if w is of type E_0,*
- $\left[\frac{39}{76} - \varepsilon, \frac{323}{594} + \varepsilon\right]$ *if w is of type E_1.*

Proof. Let us check that infinite words of type P have letter frequency $\frac{1}{2}$. It is well-known (and easy to check) that the letters of Σ_3 have equal frequencies in the fixed point t of the morphism $0 \mapsto 012$, $1 \mapsto 02$, $2 \mapsto 1$. Now, by Theorem 3.1, words of type P are factors of the image of t by a morphism h that satisfies $|h(0)|_0 + |h(1)|_0 + |h(2)|_0 = |h(0)|_1 + |h(1)|_1 + |h(2)|_1$.

For types E_0 and E_1, we only have to compute lower bounds, since if x is a lower bound on the frequency of the letter 0 for type E_i, then $(1-x)$ is an upper bound on the frequency of the letter 0 for type E_{1-i}. These lower bounds were obtained using our method [8] with a "suffix cover". A suffix cover C of a langage L is a set of factors such that every large enough and prolongable enough word in L has a suffix that belongs to C. We used the suffix cover $C_0 = \{00, 110111000101100010001 0, 1100010, 110, 1\}$ for type E_0, and the suffix cover $C_1 = \{0011101001000 1000, 01110111010010001000, 0100011101001000, 010010001001110100 1000, 0111011101001000, 0100, 010, 01110, 1\}$ for type E_1.

To check that C_0 is a suffix cover of E_0, it is sufficient to verify that every word in the set S computed in Section 4 has a suffix in C_0, because S contains every factor of type E_0 of length 360. We also check that the complement of every word in S has a suffix in C_1. Now, to prove for example that the asymptotic frequency of the letter 0 is at least $\frac{271}{594}$ in an infinite word of type E_0, we verify with backtracking that, for every $u \in C_0$, there exists no right infinite binary word w such that uw is of type E_0 and $\frac{|p|_0}{|p|} < \frac{271}{594}$ for every finite prefix p of w. □

It is noticeable that these three sets of potential frequencies are disjoint: if w is an infinite binary recurrent word avoiding $AABBCABBA$ with defined letter frequencies, then the frequency of 0 is in $\left[\frac{271}{594}, \frac{37}{76}\right] \cup \left\{\frac{1}{2}\right\} \cup \left[\frac{39}{76}, \frac{323}{594}\right] =$ $[0.45622\ldots, 0.48684\ldots] \cup \{0.5\} \cup [0.51315\ldots, 0.54377\ldots]$. The infinite words of type E_0 obtained by the construction in [7] and in Section 4 are of type E_0 and the frequency of the letter 0 is $\frac{48}{102} = \frac{8}{17} = 0.47058\ldots$.

6. Conclusion

Infinite binary recurrent words avoiding $AABBCABBA$ split into three types when considering the factors of length 4. Informally, such splittings happen because the letter C appears only once in the pattern, but is not necessarily related to the length of factors. Nothing prevents a priori from further sub-splittings into sub-types when considering larger factor lengths. Type P obviously cannot be split. Since types E_0 and E_1 are symmetrical, we can focus on type E_0 and consider the set S of words of type E_0 of length 360 discussed in Section 4. We have checked that for every two (distinct) words $w_1, w_2 \in S$, and for every occurrence of $AABB$ appearing in w_1, the corresponding occurrence of $ABBA$ does not appear in w_2. This means that no sub-splitting happens for length 360. We leave as an open question whether such a sub-splitting exists.

We do not know how to prove a negative answer. A positive answer could be obtained by constructing an infinite word of type E_0 containing a particular occurrence of $AABB$ (as a recurrent factor) and another one containing the corresponding occurrence of $ABBA$.

References

[1] K. A. Baker, G. F. McNulty, and W. Taylor. Growth problems for avoidable words, *Theoret. Comput. Sci.* **69** (1989) 319–345.

[2] J. Berstel. Axel Thue's papers on repetitions in words: a translation. *Publications du LaCIM*, Département de mathématiques et d'informatique **95**, Université du Québec Montréal (1995).
`http://www-igm.univ-mlv.fr/~berstel/Articles/1994ThueTranslation.pdf`

[3] J. Berstel. Growth of repetition-free words - a review. *Theoret. Comput. Sci.* **340(2)** (2005) 280–290.

[4] J. Cassaigne. Motifs évitables et régularité dans les mots, Thèse de Doctorat, Université Paris VI, Juillet 1994.

[5] R.J. Clark. Avoidable formulas in combinatorics on words, Ph.D. Thesis, University of California, Los Angeles (2001).

[6] R. Kolpakov. Efficient lower bounds on the number of repetition-free words *J. Integer Sequences* 10(3):Article 07.3.2 (2007).

[7] P. Ochem. A generator of morphisms for infinite words, *RAIRO: Theoret. Informatics Appl.* **40** (2006) 427–441.

[8] P. Ochem. Letter frequency in infinite repetition-free words, *Theoret. Comput. Sci.* **380** (2007) 388–392.

[9] P. Ochem and T. Reix. Upper bound on the number of ternary square-free words, *Proceedings of the Workshop on Words and Automata (WOWA'06)* (St Petersburg, June 2006).
`http://www.lri.fr/~ochem/morphisms/wowa.ps`

[10] A. M. Shur. Combinatorial Complexity of Regular Languages. CSR 2008. *LNCS* **5010** (2008) 289-301.

[11] A.I. Zimin. Blocking sets of terms, *Math. USSR Sbornik* **47(2)** (1984) 353–364. English translation.

Communicated by (The editor will be set by the publisher).

6.6 Long patterns

Application of entropy compression in pattern avoidance

Pascal Ochem   Alexandre Pinlou[*]
LIRMM, Université Montpellier 2, CNRS
Montpellier, France
{pascal.ochem,alexandre.pinlou}@lirmm.fr

Submitted: Jan 7, 2013; Accepted: XX; Published: XX
Mathematics Subject Classifications: 68R15

Abstract

In combinatorics on words, a word w over an alphabet Σ is said to avoid a pattern p over an alphabet Δ if there is no factor f of w such that $f = h(p)$ where $h : \Delta^* \to \Sigma^*$ is a non-erasing morphism. A pattern p is said to be k-avoidable if there exists an infinite word over a k-letter alphabet that avoids p. We give a positive answer to Problem 3.3.2 in Lothaire's book "Algebraic combinatorics on words", that is, every pattern with k variables of length at least 2^k (resp. $3 \times 2^{k-1}$) is 3-avoidable (resp. 2-avoidable). This conjecture was first stated by Cassaigne in his thesis in 1994. This improves previous bounds due to Bell and Goh, and Rampersad.

Keywords: Word; Pattern avoidance.

1 Introduction

A pattern p is a non-empty word over an alphabet $\Delta = \{A, B, C, \dots\}$ of capital letters called *variables*. An *occurrence* of p in a word w is a non-erasing morphism $h : \Delta^* \to \Sigma^*$ such that $h(p)$ is a factor of w. The avoidability index $\lambda(p)$ of a pattern p is the size of the smallest alphabet Σ such that there exists an infinite word w over Σ containing no occurrence of p. Bean, Ehrenfeucht, and McNulty [1] and Zimin [16] characterized unavoidable patterns, i.e., such that $\lambda(p) = \infty$. We say that a pattern p is t-avoidable if $\lambda(p) \leqslant t$. For more informations on pattern avoidability, we refer to Chapter 3 of Lothaire's book [8].

[*] Second affiliation: Département MIAp, Université Paul-Valéry, Montpellier 3, Route de Mende, 34199 Montpellier, France

In this paper, we consider upper bounds on the avoidability index of long enough patterns with k variables. Bell and Goh [2] and Rampersad [12] used a method based on power series and obtained the following bounds. Let $v(p)$ be the number of distinct variables of the pattern p.

Theorem 1 ([2, 12])**.** *Let p be a pattern.*

(a) If p has length at least $2^{v(p)}$ then $\lambda(p) \leqslant 4$. [2]

(b) If p has length at least $3^{v(p)}$ then $\lambda(p) \leqslant 3$. [12]

(c) If p has length at least $4^{v(p)}$ then $\lambda(p) = 2$. [12]

Our main result improves these bounds:

Theorem 2. *Let p be a pattern.*

(a) If p has length at least $2^{v(p)}$ then $\lambda(p) \leqslant 3$.

(b) If p has length at least $3 \times 2^{v(p)-1}$ then $\lambda(p) = 2$.

Theorem 2 gives a positive answer to Problem 3.3.2 of Lothaire's book [8]. As noticed by Cassaigne [5, 8], both bounds of Theorem 2 are tight. The bound $2^{v(p)}$ in Theorem 2.(a) is tight in the sense that the patterns p in the family $\{A, ABA, ABACABA, ABACABADABACABA, \ldots\}$ have length $2^{v(p)} - 1$ and are unavoidable. Similarly, the bound $3 \times 2^{v(p)-1}$ in Theorem 2.(b) is tight in the sense that the patterns in the family $\{AA, AABAA, AABAACAABAA, AABAACAABAADAABAACAABAA, \ldots\}$ have length $3 \times 2^{v(p)-1} - 1$ and are not 2-avoidable. Hence, this shows that the upper bound 3 of Theorem 2.(a) is best possible.

The avoidability index of every pattern with at most 3 variables is known, thanks to various results in the literature. In particular, Theorem 2 is proved for every pattern p with $v(p) \leqslant 3$:

- For $v(p) = 1$, the famous results of Thue [14, 15] give $\lambda(AA) = 3$ and $\lambda(AAA) = 2$.
- For $v(p) = 2$, every binary pattern of length at least 4 contains a square, and is thus 3-avoidable. Moreover, Roth [13] proved that every binary pattern of length at least 6 is 2-avoidable.
- For $v(p) = 3$, Cassaigne [5] began and the first author [10] finished the determination of the avoidability index of every pattern with at most 3 variables. Every ternary pattern of length at least 8 is 3-avoidable and every binary pattern of length at least 12 is 2-avoidable.

So, there remains to prove Theorem 2 for every pattern p with $v(p) \geqslant 4$.

Section 2 is devoted to some preliminary results. We prove Theorem 2.(a) in Section 3 as a corollary of a result of Bell and Goh [2]. In Section 4, we prove Theorem 2.(b) using the so-called *entropy compression method*.

Very recently, Blanchet-Sadri and Woodhouse [4] independently proved Theorem 2 using completely different methods.

2 Preliminary results

Let p be a pattern over $\Delta = \{A, B, C, \ldots\}$. An *occurrence* of p in a word w over the alphabet Σ is a non-erasing morphism $h : \Delta^* \to \Sigma^*$ such that $h(p)$ is a factor of w. Note that two distinct occurrences of p may form the same factor. For example, if $p = ABA$, then the occurrence $h = (A \to 00; B \to 1)$ of p forms the factor $h(p) = h(ABA) = h(A)h(B)h(A) = 00100$; on the other hand, $h' = (A \to 0; B \to 010)$ is a distinct occurrence of p which forms the same factor $h'(p) = h'(ABA) = h'(A)h'(B)h'(A) = 00100$.

A pattern p is *doubled* if every variable of p appears at least twice in p. A pattern p is *balanced* if it is doubled and every variable of p appears both in the prefix and the suffix of length $\left\lfloor \frac{|p|}{2} \right\rfloor$ of p. Note that if the pattern has odd length, then the variable X that appears in the middle of p (i.e. in position $\left\lfloor \frac{|p|}{2} \right\rfloor + 1$) must appear also in the prefix and in the suffix in order to make p balanced.

Claim 3. *For every integer $f \geqslant 2$, every pattern p with length at least $f \times 2^{v(p)-1}$ contains a balanced pattern p' with length at least $f \times 2^{v(p')-1}$ as a factor.*

Proof. We prove this claim by induction on $v(p)$. If $v(p) = 1$, then p has size at least $f \geqslant 2$ and is clearly balanced. Suppose this is true for some $v(p) = n$, i.e. p with n variables and length at least $f \times 2^{n-1}$ contains a balanced pattern p' as a factor with length at least $f \times 2^{v(p')-1}$. Let $v(p) = n + 1$ and let p_1 (resp. p_2) be the prefix (resp. the suffix) of p of size $\left\lfloor \frac{|p|}{2} \right\rfloor$. If p is not balanced, then there exists a variable X in p that does not occur in p_i for some $i \in \{1, 2\}$. Thus, we have $v(p_i) \leqslant v(p) - 1 = n$ and $|p_i| \geqslant f \times 2^{n-1}$. Therefore, by induction hypothesis, p contains a balanced pattern p' with length at least $f \times 2^{v(p')-1}$ as a factor. □

In the following, we will only use the fact that the pattern p' in Claim 3 is doubled instead of balanced.

3 3-avoidable long patterns

We prove Theorem 2.(a) as a corollary of the following result of Bell and Goh [2]:

Lemma 4 ([2]). *Every doubled pattern with at least 6 variables is 3-avoidable.*

Proof of Theorem 2.(a). We want to prove that every pattern p with length at least $2^{v(p)}$ is 3-avoidable, or equivalently, that every pattern p with $v(p) \leqslant k$ and length at least 2^k is 3-avoidable. By Claim 3, every such pattern contains a doubled pattern p' as a factor with length at least $2^{v(p')}$. So there remains to show that every doubled pattern p with $v(p) \leqslant k$ and length at least 2^k is 3-avoidable. As discussed in the introduction, the case of patterns with at most 3 variables has been settled. Now, it is sufficient to prove that doubled patterns of length at least $2^4 = 16$ are 3-avoidable.

Suppose that p_1 is a doubled pattern containing a variable X that appears at least 4 times. Replace 2 occurrences of X with a new variable to obtain a pattern p_2. Example:

We replace the first and third occurrence of B in $p_1 = ABBCDBCABDDCB$ by a new variable E to obtain $p_2 = AEBCDECABDDCB$. Then p_2 is a doubled pattern such that $|p_1| = |p_2|$ and $\lambda(p_1) \leqslant \lambda(p_2)$, since every occurrence of p_1 is also an occurrence of p_2.

Given a doubled pattern p of length at least 16, we make such replacements as long as we can. We thus obtain a doubled pattern p' of length at least 16 such that $\lambda(p) \leqslant \lambda(p')$. Moreover, every variable in p' appears either 2 or 3 times and therefore p' contains at least $\lceil 16/3 \rceil = 6$ variables. So p' is 3-avoidable by Lemma 4. Thus p is 3-avoidable, which finishes the proof. □

4 2-avoidable long patterns

We want to prove that every pattern p with length at least $3 \times 2^{v(p)-1}$ is 2-avoidable, or equivalently, that every pattern p with $v(p) \leqslant k$ variables and length at least $3 \times 2^{k-1}$ is 2-avoidable. By Claim 3, every such pattern contains a doubled pattern p' as a factor with length at least $3 \times 2^{v(p')-1}$. So there remains to show that every doubled pattern p with $v(p) \leqslant k$ and length at least $3 \times 2^{k-1}$ is 2-avoidable.

As discussed in the introduction, the case of patterns with at most 3 variables has been settled. Now, it is sufficient to prove Theorem 2.(b) for doubled patterns with at least 4 variables.

Let $\Sigma = \{0, 1\}$ be the alphabet. For the remaining of this section, let $k \geqslant 4$ and $q(k) = 3 \times 2^{k-1}$.

Suppose by contradiction that there exists a doubled pattern p on k variables and length at least $q(k)$ that is not 2-avoidable. Then there exists an integer n such that every word $w \in \Sigma^n$ contains p. We put an arbitrary order on the k variables of p and call A_j the j-th variable of p.

4.1 The algorithm AvoidP

Let $V \in \{0, 1\}^t$ be a vector of length t. The algorithm AvoidP takes the vector V as input and returns a word w avoiding p and a data structure R that is called a *record* in the remaining of the paper.

The way we encode information in R at lines 5 and 7 will be explained in Subsection 4.2.

In the algorithm AvoidP, let w_i be the word w after i steps. Clearly, w_i avoids p at each step. By contradiction hypothesis, the resulting word w of the algorithm (that is w_t) has length less than n. We will prove that each output of the algorithm allows to determine the input. Then we obtain a contradiction by showing that the number of possible outputs is strictly smaller than the number of possible inputs when t is chosen large enough compared to n. This implies that every pattern p with at most k variables and length at least $q(k)$ is 2-avoidable.

To analyze the algorithm, we borrow ideas from graph coloring problems [6, 7]. These

```
Algorithm 1: AvoidP
Input  : $V$.
Output: $w$ (a word avoiding $p$) and $R$ (a data structure).
$w \leftarrow \epsilon$
$R \leftarrow \emptyset$
for $i \leftarrow 1$ to $t$ do
    Append $V[i]$ (the $i$-th letter of $V$) to $w$
    Encode in $R$ that a letter has been appended to $w$
    if $w$ contains a factor of length $\ell$ corresponding to an occurrence of $p$ then
        Encode in $R$ the occurrence of $p$
        Erase the suffix of length $\ell$ of $w$
return $R$, $w$
```

results are based on the Moser-Tardos [9] entropy-compression method which is an algorithmic proof of the Lovász Local Lemma.

4.2 The record R

An important part of the algorithm is to update the record R at each step of the algorithm. Let R_i be the record after i steps of the algorithm AvoidP. On one hand, given V as input of the algorithm, this produces a pair (R_t, w_t). On the other hand, given a pair (R_t, w_t), we will show in Lemma 6 that we can recover the entire input vector V. So, each input vector V produces a distinct pair (R_t, w_t).

Let $\mathcal{V}$ be the set of input vectors V of size t, let $\mathcal{R}$ be the set of records R produced by the algorithm AvoidP and let $\mathscr{O}$ be the set of different outputs (R_t, w_t). After the execution of the algorithm (t steps), w_t avoids p by definition and therefore $|w_t| < n$ by contradiction hypothesis. Hence, the number of possible final words w_t is independent from t (it is at most 2^n). We then clearly have $|\mathscr{O}| \leqslant 2^n \times |\mathcal{R}|$. We will prove that $|\mathcal{V}| \leqslant |\mathscr{O}|$ and that $|\mathcal{R}| = o(2^t)$ to obtain the contradiction $2^t = |\mathcal{V}| \leqslant |\mathscr{O}| \leqslant 2^n \times |\mathcal{R}| = o(2^t)$.

The record R is a triplet $R = (D, L, X)$ where D is a binary word (each element is 0 or 1), L is a vector of $(k-1)$-sets of non-zero integers and X is a binary word. At the beginning, D, L and X are empty. At step i of the algorithm, we append $V[i]$ to w_{i-1} to get w'_i.

If w'_i contains no occurrence of p, then we append 0 to D to get R_i and we set $w_i = w'_i$. Otherwise, suppose that w'_i contains an occurrence h of p that forms a factor $h(p)$ of length ℓ, that is, the suffix of length ℓ of w'_i is $h(p)$. Recall that A_j is the j-th variable of p. For $1 \leqslant j \leqslant k-1$, let $z_j = |h(A_1 \dots A_j)|$. Let $L' = \{z_1, z_2, \dots, z_{k-1}\}$ be a $(k-1)$-set of non-zero integers. To get R_i, we append the factor 01^ℓ to D; we add L' as the last element of L; and we append the factor $h(A_1 A_2 \dots A_k)$ to X.

Example 5.

Let us give an example with $k = 3$, $p = ACBBCBBABCAB$ and $V = [0,0,1,0,0,1,1,0,0,1,1,1,0,0,1,1,0,1,1,1,0,0,0,1,1,0]$. The variables of p were initially ordered as (A, B, C). For the first 24 steps, no occurrence of p appeared, so at each step $i \leqslant 24$, we append $V[i]$ to w_{i-1} and we append one 0 to D. Hence, at step 24, we have:

- $w_{24} = 001001100111001101110001$
- $R_{24} = \begin{cases} D &= 000000000000000000000000 = 0^{24} \\ L &= [\,] \\ X &= \epsilon \end{cases}$

Now, at step 25, we first append $V[25] = 1$ to w_{24} to get w'_{25}. The word w'_{25} contains an occurrence $h = (A \to 01; B \to 1; C \to 100)$ of p which forms a factor of length 21 (the 21 last letters of w'_{25}). Then we set $L' = \{|h(A)|, |h(AB)|\} = \{2, 3\}$. We obtain w_{25} from w'_{25} by erasing its suffix of length 21. To get R_{25}, we append the factor 01^{21} to D, we add L' as the last element of L, and we append the factor $h(ABC) = 011100$ to X. This gives:

- $w_{25} = 0010$
- $R_{25} = \begin{cases} D &= 0000000000000000000000000111111111111111111111 = 0^{25}1^{21} \\ L &= [\{2, 3\}] \\ X &= 011100 \end{cases}$

Let V_i be the vector V restricted to its i first elements. We will show that the pair (R_i, w_i) at some step i allows to recover V_i.

Lemma 6. *After i steps of the algorithm* AvoidP*, the pair (R_i, w_i) permits to recover V_i.*

Proof. Before step 1, we have $w_0 = \epsilon$, $R_0 = (\epsilon, [\,], \epsilon)$, and $V_0 = \epsilon$. Let $R_i = (D, L, X)$ be the record after step i, with $1 \leqslant i \leqslant t$.

Suppose that 0 is a suffix of D. This means that at step i, no occurrence of p was found: the algorithm appended $V[i]$ to w_{i-1} to get w_i. Therefore $V[i]$ is the last letter of w_i, say x. Then the word w_{i-1} is obtained from w_i by erasing the last letter and the record R_{i-1} is obtained from R_i by removing the suffix 0 of D. We recover V_{i-1} from (R_{i-1}, w_{i-1}) by induction hypothesis and we obtain $V_i = V_{i-1} \cdot x$.

Suppose now that 01^ℓ is a suffix of D. This means that an occurrence h of p has been created during step i such that $|h(p)| = \ell$. Let L' be the last element of L which is a $(k-1)$-set $L' = \{z_1, z_2, \ldots, z_{k-1}\}$. By construction of L', we have $|h(A_1)| = z_1$ and $|h(A_s)| = z_s - z_{s-1}$ for $2 \leqslant s \leqslant k - 1$. We know the pattern p, the total length of the factor $h(p)$ (that is ℓ) and the lengths of the $k - 1$ first variables of p in $h(p)$, so we are able to compute $|h(A_k)|$. Now, we can parse the suffix of length $\sum_{1\leqslant j\leqslant k}|h(A_j)|$ of X, which is the factor $h(A_1 \dots A_k)$, to obtain the factors $h(A_1), \dots, h(A_k)$. Thus, we have recovered the occurrence h of p.

Now, w_{i-1} is obtained by removing the last letter x of $w_i \cdot h(p)$. This letter x is $V[i]$, the letter appended to w_{i-1} at step i to get w'_i. The record R_{i-1} is obtained from R_i as follows: remove the suffix 01^{ℓ} from D, remove the last element of L, and remove the suffix $h(A_1 \dots A_k)$ of X. We recover V_{i-1} from (R_{i-1}, w_{i-1}) by induction hypothesis and we obtain $V_i = V_{i-1} \cdot x$.

□

The previous lemma proves that distinct input vectors cannot correspond to the same pair (R_t, w_t). So we get $|\mathcal{V}| \leqslant |\mathscr{O}|$.

4.3 Analysis of $\mathcal{R}$

Now we compute $|\mathcal{R}|$. Let $R = R_t = (D, L, X)$ be a given record produced by an execution of AvoidP. Let $\mathcal{D}$ be the set of such binary words D. For a given $D \in \mathcal{D}$, let $\mathcal{L}_D$ be the set of such vectors of $(k-1)$-sets of non-zero integers L compatible with D. Let $\mathcal{X}$ be the set of such binary words X.

We thus have $|\mathcal{R}| \leqslant |\mathcal{D}| \times \max_{D\in\mathcal{D}} |\mathcal{L}_D| \times |\mathcal{X}|$.

Let us give some useful information in order to get upper bounds on $|\mathcal{D}|$, $|\mathcal{X}|$, and $|\mathcal{L}_D|$. The algorithm runs in t steps. At each step, one letter is appended to w, so t letters have been appended and therefore the number of erased letters during the execution of the algorithm is $t - |w_t|$. At some steps, an occurrence h of p appears and the factor $h(p)$ is immediately erased. Let m be the number of erased factors during the execution of the algorithm. Let $h_i(p)$, $1 \leqslant i \leqslant m$, be the m erased factors. We have $|h_i(p)| \geqslant q(k)$ since each variable of p is a non-empty word and p has length at least $q(k)$. Moreover, we have $\sum_{1\leqslant i\leqslant m} |h_i(p)| = t - |w_t| \leqslant t$. Each time a factor $h_i(p)$ is erased, we add an element to L, so $|L| = m$.

4.3.1 Analysis of $\mathcal{D}$

In the binary word D, each 0 corresponds to an appended letter during the execution of the algorithm and each 1 corresponds to an erased letter. Therefore, D has length $2t - |w_t|$. Observe that every prefix in D contains at least as many 0's as 1's. Indeed, since a 1 corresponds to an erased letter x, this letter x had to be added first and thus there is a 0 before that corresponds to this 1. The word D is therefore a partial Dyck word. Since any erased factor $h_i(p)$ has length at least $q(k)$, any maximal sequence of 1's (which is called a *descent* in the sequel) in D has length at least $q(k)$. So D is a partial Dyck words with t 0's such that each descent has length at least $q(k)$.

Let $C_{t,r,d}$ (resp. $C_{t,d}$) be the number of partial Dyck words with t 0's and $t-r$ 1's (resp. Dyck words of length $2t$) such that all descents have length at least d.

Lemma 7. $C_{t,r,d} \leqslant C_{t+d,d}$.

Proof. We map every partial Dyck word y with t 0's and $t-r$ 1's to the Dyck word $y0^d1^{d+r}$, which has $t+d$ 0's and $t+d$ 1's. Since d is fixed, this mapping is injective. This proves the lemma. □

If $q(k) \geqslant d$, then we have $|\mathcal{D}| \leqslant C_{t,|w_t|,q(k)} \leqslant C_{t,|w_t|,d} \leqslant C_{t+d,d}$ by Lemma 7. Let $\phi_d(x) = 1 + \sum_{j \geqslant d} x^j = 1 + \frac{x^d}{1-x}$. The radius of convergence of ϕ_d is 1. The following lemma comes from a more general statement of Esperet and Parreau [7] and gives an upper bound on $|\mathcal{D}|$.

Lemma 8. *[7] Let d be an integer such that the equation $\phi_d(x) - x\phi'_d(x) = 0$ has a solution τ with $0 < \tau < 1$. Then τ is the unique solution of the equation in the open interval $(0,1)$. Moreover, there exists a constant c_d such that $C_{t,d} \leqslant c_d \gamma_d^t t^{-\frac{3}{2}}$ where $\gamma_d = \phi'_d(\tau) = \frac{\phi_d(\tau)}{\tau}$.*

The equation $\phi_d(x) - x\phi'_d(x) = 0$ is equivalent to $P(x) = (1-x)^2 + (1-d)x^d + (d-2)x^{d+1} = 0$. Since $P(0) = 1$ and $P(1) = -1$, $P(x) = 0$ has a solution τ in the open interval $(0,1)$. By Lemma 8, this solution is unique and, for some constant c_d, we have $C_{t+d,d} \leqslant c_d \gamma_d^{t+d}(t+d)^{-\frac{3}{2}}$ with $\gamma_d = \phi'_d(\tau)$. We clearly have $C_{t+d,d} = o(\gamma_d^t)$. So, we can compute γ_d for d fixed. We will use the following bounds: $\gamma_{24} \leqslant 1.27575$ and $\gamma_{48} \leqslant 1.15685$.

So, by Lemmas 7 and 8, when t is large enough, we have $|\mathcal{D}| < 1.27575^t$ (resp. $|\mathcal{D}| < 1.15685^t$) if the length of any descent is at least 24 (resp. 48).

4.3.2 Analysis of $\mathcal{X}$

Each erased factor $h_i(p)$ adds $|h_i(A_1 \dots A_k)|$ letters to X. Since p is doubled, we have $|h_i(p)| \geqslant 2|h_i(A_1 \dots A_k)| + q(k) - 2k \geqslant 2|h_i(A_1 \dots A_k)| + 24 - 2 \times 4$. This gives $|h_i(A_1 \dots A_k)| \leqslant \frac{|h_i(p)|}{2} - 8$. Since $\sum_{1 \leqslant i \leqslant m} |h_i(p)| \leqslant t$, we have $|X| = \sum_{1 \leqslant i \leqslant m} |h_i(A_1 \dots A_k)| \leqslant \sum_{1 \leqslant i \leqslant m} \left(\frac{|h_i(p)|}{2} - 8 \right) \leqslant \frac{t}{2} - 8m$. Therefore $|\mathcal{X}| \leqslant 2^{\frac{t}{2} - 8m + 1} \leqslant (\sqrt{2})^t$.

4.3.3 Analysis of $\mathcal{L}_D$

For a given $R = (D, L, X)$, the vector L is dependent on the partial Dyck word D. Indeed, by construction, the i-th element of L is a $(k-1)$-set of integers smaller than $\frac{\ell}{2}$ where ℓ is the length of the i-th descent of D. In this subsection, we compute an upper bound on the number of vectors L compatible with D for a given $D \in \mathcal{D}$ and thus we give an upper bound on $|\mathcal{L}_D|$.

Each element $L_i = \{z_1, z_2, \dots, z_{k-1}\}$ of L corresponds to the erased factor $h_i(p)$ and by construction we have $|h_i(A_1 \dots A_j)| = z_j$. By construction of D, $|h_i(p)|$ is the length of the i-th descent of D. Since D is fixed, $|h_i(p)|$ is fixed for every $1 \leqslant i \leqslant m$.

Let $s_k(\ell)$ be the number of such $(k-1)$-sets L_i that correspond to factors of length ℓ. Recall that $|h_i(p)| \geqslant q(k)$, so $s_k(\ell)$ is defined for $k \geqslant 4$ and $\ell \geqslant q(k)$. Each of the m elements of L corresponds to an erased factor, so $|\mathcal{L}_D| \leqslant s_k(|h_1(p)|) \times s_k(|h_2(p)|) \times \dots \times s_k(|h_m(p)|)$. Let $g_k(\ell) = s_k(\ell)^{\frac{1}{\ell}}$ be defined for $k \geqslant 4$ and $\ell \geqslant q(k)$. Then $|\mathcal{L}_D| \leqslant$

$g_k(|h_1(p)|)^{|h_1(p)|} \times g_k(|h_2(p)|)^{|h_2(p)|} \times \ldots \times g_k(|h_m(p)|)^{|h_m(p)|}$. So, if we are able to upper-bound $g_k(\ell)$ by some constant c for all $\ell \geqslant q(k)$, then we get $|\mathcal{L}_D| \leqslant c^{|h_1(p)|} \times c^{|h_2(p)|} \times \ldots \times c^{|h_m(p)|} \leqslant c^t$.

Now we bound $g_k(\ell)$ using two different methods depending on the number k of variables in p and the length $q(k)$ of p.

4.3.3.1 Bound on $g_k(\ell)$ for $k = 4$, $\ell \geqslant 96$ or $k \geqslant 5$, $\ell \geqslant q(k)$

As shown in Section 4.3.2, we have $|h_i(A_1 \ldots A_k)| \leqslant \frac{|h_i(p)|}{2} - 8$. For a given $L_i = \{z_1, z_2, \ldots, z_{k-1}\}$ that corresponds to $h_i(p)$, we thus have $z_{k-1} = |h_i(A_1 \ldots A_{k-1})| \leqslant \frac{|h_i(p)|}{2} - 9$. Therefore, L_i is a set of $(k-1)$ distinct integers between 1 and $\frac{|h_i(p)|}{2} - 9$. So $s_k(\ell) \leqslant \binom{\lfloor \ell/2 \rfloor}{k-1}$ and $g_k(\ell) \leqslant \binom{\lfloor \ell/2 \rfloor}{k-1}^{\frac{1}{\ell}}$. We can upper-bound $g_k(\ell)$ by $\overline{g_k}(\ell) = \left(\frac{(\ell/2)^{k-1}}{(k-1)!}\right)^{\frac{1}{\ell}}$ for $\ell \geqslant q(k)$.

Let us show that when k is fixed, $\overline{g_k}(\ell)$ is a decreasing function of ℓ for $\ell \geqslant q(k)$. The derivative $(\overline{g_k}(\ell))' = \overline{g_k}(\ell) \times \frac{1}{\ell^2} \times \left(k - 1 - \ln\left(\frac{(\ell/2)^{k-1}}{(k-1)!}\right)\right)$ is negative if and only if $k - 1 < \ln\left(\frac{(\ell/2)^{k-1}}{(k-1)!}\right)$, that is, if and only if $(k-1)!e^{k-1} < (\ell/2)^{k-1}$. This inequality holds since $(k-1)!e^{k-1} < ((k-1)e)^{k-1} < \left(3 \times 2^{k-2}\right)^{k-1} \leqslant (\ell/2)^{k-1}$.

We also have that $\overline{g_k}(q(k))$ is a decreasing function of k for $k \geqslant 4$ since we have checked using Maple that the only zero of its derivative is at $k \approx 3.37$ and that its derivative is negative for $k \geqslant 3.38$.

Thus, we get $g_k(\ell) < \overline{g_k}(\ell) \leqslant \overline{g_k}(q(k)) \leqslant \overline{g_5}(48) < 1.21973$ for all $k \geqslant 5$ and $\ell \geqslant q(k)$, and we get $g_4(\ell) < \overline{g_4}(\ell) \leqslant \overline{g_4}(96) < 1.10773$ for all $\ell \geqslant 96$. We chose the value 96 to distinguish between the cases, because it is the smallest value such that the argument holds.

4.3.3.2 Bound on $g_4(\ell)$ for $24 \leqslant \ell \leqslant 95$

The second method to bound the size of $g_4(\ell)$ is based on ordinary generating functions (OGF). Here, $k = 4$, so let A_1, A_2, A_3, A_4 be the four variables of p and let a_i be the number of instances of A_i in p. Therefore, $a_1 + a_2 + a_3 + a_4 = |p|$. Recall that each variable appears at least twice in p since p is doubled, so $a_i \geqslant 2$. Moreover, a factor of length ℓ, with $24 \leqslant \ell \leqslant 95$, necessarily corresponds to an occurrence of a pattern of length between 24 and 95. So we just have to consider patterns p with $24 \leqslant |p| \leqslant 95$.

Given $L_i = \{z_1, z_2, z_3\}$ an element of L corresponding to $h_i(p)$, we have $|h_i(A_1)| = z_1$, $|h_i(A_2)| = z_2 - z_1$, $|h_i(A_3)| = z_3 - z_2$ and $|h_i(A_4)| = \frac{|h_i(p)| - (a_1|h_i(A_1)| + a_2|h_i(A_2)| + a_3|h_i(A_3)|)}{a_4}$. Let $\mathcal{A}_p = \sum_{j \geqslant |p|} b_j\ x^j$ be the OGF of such sets L', i.e. b_j is the number of 3-sets $\{z_1, z_2, z_3\}$ that corresponds to a factor of length j formed by an occurrence of p. In other words, b_j is the number of 4-tuples $(\ell_1, \ell_2, \ell_3, \ell_4)$ such that $a_1 \times \ell_1 + a_2 \times \ell_2 + a_3 \times \ell_3 + a_4 \times \ell_4 = j$ and with $\ell_i \geqslant 1$ (since each variable of p corresponds to a non-empty word). So by definition of h_4, we have $h_4(\ell) = b_\ell$ and thus $g_4(\ell) = b_\ell^{\frac{1}{\ell}}$.

This kind of OGF has been studied and is similar to the well-known problem of counting the number of ways you can change a dollar [11]: you have only five types of coins (pennies, nickels, dimes, quarters, and half dollars) and you want to count the number of ways you can change any amount of cents. So, let $\mathcal{C} = \sum_{j \geqslant 1} c_j \ x^j$ be the OGF of the problem and thus any c_j is the number of ways you can change j cents. Then, for example, c_{100} corresponds to the number of ways you can change a dollar. Here, $\mathcal{C} = \frac{1}{1-x} \times \frac{1}{1-x^5} \times \frac{1}{1-x^{10}} \times \frac{1}{1-x^{25}} \times \frac{1}{1-x^{50}}$.

In our case, we have four coins with value a_1, a_2, a_3, and a_4 respectively (so we can have different types of coins with the same value) and each type of coins appears at least once (since $\ell_i \geqslant 1$). Thus we get $\mathcal{A}_p = \sum_{j \geqslant |p|} b_j \ x^j = \frac{x^{a_1}}{1-x^{a_1}} \times \frac{x^{a_2}}{1-x^{a_2}} \times \frac{x^{a_3}}{1-x^{a_3}} \times \frac{x^{a_4}}{1-x^{a_4}}$. We use Maple for our computation. For each $24 \leqslant |p| \leqslant 95$, for each 4-tuple (a_1, a_2, a_3, a_4) such that $\sum a_i = |p|$, we consider the associated OGF $\mathcal{A}_p$ and we compute, using Maple, the truncated series expansion up to the order 95, that gives $\mathcal{A}_p = b_{24}x^{24} + b_{25}x^{25} + \ldots + b_{95}x^{95} + O(x^{96})$ with explicit values for the coefficients b_j. So, for any $24 \leqslant \ell \leqslant 95$, $g_4(\ell)$ is upper-bounded by the maximum of $b_\ell^{\frac{1}{\ell}}$ taken over all $\mathcal{A}_p$. Maple gives that $b_\ell^{\frac{1}{\ell}}$ is maximal for $|p| = 24$, $(a_1, a_2, a_3, a_4) = (2, 2, 2, 18)$, and $\ell = 46$: in this case, $b_{46} = 84$ (i.e. there exist 84 distinct 3-sets L_i that correspond to some factor of length 46 formed by an occurrence of a pattern of length 24 where three variables appear twice and one variable appears 18 times). So, $g_4(\ell) \leqslant 84^{\frac{1}{46}} < 1.10112$ for all $24 \leqslant \ell \leqslant 95$.

4.3.3.3 Bound on $g_k(\ell)$ for all $k \geqslant 4$

We can deduce from Paragraphs 4.3.3.1 and 4.3.3.2 the following.

If $k = 4$, then $g_4(\ell) < 1.10112$ for $24 \leqslant \ell \leqslant 95$ and $g_4(\ell) < 1.10773$ for $\ell \geqslant 96$. So for $k = 4$, we have $|\mathcal{L}_D| < (1.10773)^t$.

If $k \geqslant 5$, then $g_k(\ell) < 1.21973$ for $\ell \geqslant q(k)$. So for $k \geqslant 5$, we have $|\mathcal{L}_D| < (1.21973)^t$.

4.4 End of the proof

The bounds on $|\mathcal{L}_D|$ obtained in Subsection 4.3.3 hold for any fixed $D \in \mathcal{D}$. So they also hold for $\max_{D \in \mathcal{D}} |\mathcal{L}_D|$.

Aggregating the above analysis, we get the following. For $k \geqslant 5$, we have $q(k) \geqslant 48$: then $|\mathcal{R}| \leqslant |\mathcal{D}| \times \max_{D \in \mathcal{D}} |\mathcal{L}_D| \times |\mathcal{X}| \leqslant (1.15685 \times 1.21973 \times \sqrt{2})^t = o(2^t)$. For $k = 4$, we have $q(k) \geqslant 24$: then $|\mathcal{R}| \leqslant |\mathcal{D}| \times \max_{D \in \mathcal{D}} |\mathcal{L}_D| \times |\mathcal{X}| \leqslant (1.27575 \times 1.10773 \times \sqrt{2})^t = o(2^t)$.

Thus for all $k \geqslant 4$, $|\mathcal{R}| = o(2^t)$ and so we obtain the desired contradiction:

$$2^t = |\mathcal{V}| \leqslant |\mathscr{O}| \leqslant 2^n \times |\mathcal{R}| = 2^n \times o(2^t) = o(2^t).$$

5 Conclusion

In our results, we heavily use the fact that the patterns are doubled. The fact that the patterns are long is convenient for our proofs but does not seem so important. So we ask whether every doubled pattern is 3-avoidable. By the remarks in Section 1 and by Lemma 4, the only remaining cases are doubled patterns with 4 and 5 variables. Also, does there exist a finite k such that every doubled pattern with at least k variables is

2-avoidable ? Using the standard backtracking algorithm, we have checked by computer that ABCCBADD is not 2-avoidable. So we know that such a k is at least 5.

Acknowledgments

We are grateful to the anonymous referees for their helpful suggestions to improve the presentation of the paper.

References

[1] D.R. Bean, A. Ehrenfeucht, and G.F. McNulty, Avoidable Patterns in Strings of Symbols, *Pacific J. of Math.* **85** (1979) 261–294.

[2] J. Bell, T. L. Goh. Exponential lower bounds for the number of words of uniform length avoiding a pattern, *Inform. and Comput.* **205** (2007), 1295-1306.

[3] J. Berstel. Axel Thue's work on repetitions in words. Invited Lecture at the 4th Conference on Formal Power Series and Algebraic Combinatorics, Montreal, 1992, June 1992. Available at `http://www-igm.univ-mlv.fr/~berstel/index.html`.

[4] F. Blanchet-Sadri, B. Woodhouse. Strict Bounds for Pattern Avoidance. *Theor. Comput. Sci.* **506** (2013), 17–27.

[5] J. Cassaigne. Motifs évitables et régularité dans les mots, Thèse de Doctorat, Université Paris VI, Juillet 1994.

[6] V. Dujmović, G. Joret, J. Kozik, and D. R. Wood. Nonrepetitive Colouring via Entropy. *Combinatorica*, to appear, 2013+ (Also available on `arXiv:1112.5524`).

[7] L. Esperet and A. Parreau. Acyclic edge-coloring using entropy compression. *European Journal of Combinatorics* **36(4)** (2013), 1019–1027.

[8] M. Lothaire. Algebraic Combinatorics on Words. *Cambridge Univ. Press* (2002).

[9] R. A. Moser, G. Tardos. A constructive proof of the general Lovasz local lemma. *J. ACM*, **57(2)** (2010), p. 11:1-11:15.

[10] P. Ochem. A generator of morphisms for infinite words. *RAIRO: Theoret. Informatics Appl.* **40** (2006) 427–441.

[11] G. Pólya, R. E. Tarjan, D. R. Woods. Notes on Introductory Combinatorics. *Progress in Computer Science*, Birkhäuser (1983).

[12] N. Rampersad. Further applications of a power series method for pattern avoidance. *Electron. J. Combinatorics.* **18(1)** (2011), #P134.

[13] P. Roth. Every binary pattern of length six is avoidable on the two-letter alphabet. *Acta Inform.* **29** (1992), 95–107.

[14] A. Thue. Über unendliche Zeichenreihen. *Norske Vid. Selsk. Skr. I. Mat. Nat. Kl. Christiania*, 7:1–22, 1906.

[15] A. Thue. Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen. *Norske Vid. Selsk. Skr. I. Mat. Nat. Kl. Christiania*, 10:1–67, 1912.

[16] A.I. Zimin. Blocking sets of terms. *Math. USSR Sbornik* **47(2)** (1984) 353–364. English translation.

6.7 Binary words with few distinct squares

Characterization of some binary words with few squares

Golnaz Badkobeh[a], Pascal Ochem[b]

[a] *Department of Computer Science, University of Sheffield, UK*
[b] *CNRS - LIRMM, Montpellier, France*

Abstract

Thue proved that the factors occurring infinitely many times in square-free words over {0,1,2} avoiding the factors in {010,212} are the factors of the fixed point of the morphism 0 ↦ 012, 1 ↦ 02, 2 ↦ 1. He similarly characterized square-free words avoiding {010,020} and {121,212} as the factors of two morphic words. In this paper, we exhibit smaller morphisms to define these two square-free morphic words and we give such characterizations for six types of binary words containing few distinct squares.

1. Introduction

Let Σ_k denote the k-letter alphabet $\{0,1,\ldots,k\text{-}1\}$. Let ε denote the empty word. A finite word is *recurrent* in an infinite word w if it appears as a factor of w infinitely many times. An infinite word w is *recurrent* if all its finite factors are recurrent in w. If a morphism f is such that $f(0)$ starts with 0, then the *fixed point* of f is the unique word $w = f^\infty(0)$ starting with 0 and satisfying $w = f(w)$. An infinite word is *pure morphic* if it is the fixed point of a morphism. An infinite word is *morphic* if it is the image $g(f^\infty(0))$ by a morphism g of a pure morphic word $f^\infty(0)$. The *factor complexity* of an infinite word or a language is the number of factors of length n of the infinite word or the language. A pattern P is a finite word of variables over the alphabet $\{A, B, \ldots\}$. A word w (finite or infinite) *avoids* a pattern P if for every substitution ϕ of the variables of P with non-empty words, $\phi(P)$ is not a factor of w. Given a finite alphabet Σ_k, a finite set $\mathcal{P}$ of patterns, and a finite set $\mathcal{F}$ of factors over Σ_k, we say that $\mathcal{P} \cup \mathcal{F}$ *characterizes* a morphic word w over Σ_k if w avoids $\mathcal{P} \cup \mathcal{F}$ and every recurrent factor of an infinite word avoiding $\mathcal{P} \cup \mathcal{F}$ is a factor of w. In other words, $\mathcal{P} \cup \mathcal{F}$ characterizes w if and only if every recurrent word over Σ_k avoiding $\mathcal{P} \cup \mathcal{F}$ has the same set of factors as w. In our results, we do not specify the alphabet size k since Σ_k corresponds to the set of letters appearing in $\mathcal{F}$. A *repetition* is a factor of the form $r = u^n v$ where u is non-empty and v is a prefix of u. Then $|u|$ is the *period* of the repetition r and its *exponent* is $|r|/|u|$. A *square* is a repetition of exponent 2. Equivalently, it is an occurrence of the pattern AA. An overlap is a repetition with exponent strictly greater than 2.

Thue [3, 10, 11] gave the following characterization of overlap-free binary words: $\{ABABA\} \cup \{000,111\}$ characterizes the fixed point of the morphism

Preprint submitted to Elsevier *October 7, 2016*

$0 \mapsto 01$, $1 \mapsto 10$. Concerning ternary square-free words, he proved that

- $\{AA\} \cup \{010,212\}$ characterizes the fixed point of $f_3 : 0 \mapsto 012$, $1 \mapsto 02$, $2 \mapsto 1$,
- $\{AA\} \cup \{010,020\}$ characterizes the morphic word $T_1(f_T^\infty(0))$,
- $\{AA\} \cup \{121,212\}$ characterizes the morphic word $T_2(f_T^\infty(0))$,

where the morphisms f_T, T_1, and T_2 are given below.

$$\begin{array}{lll}
f_T(0) = 012, & T_1(0) = 01210212, & T_2(0) = 021012, \\
f_T(1) = 0432, & T_1(1) = 01210120212, & T_2(1) = 02102012, \\
f_T(2) = 0134, & T_1(2) = 01210212021, & T_2(2) = 02101201, \\
f_T(3) = 013432, & T_1(3) = 012102120210120212, & T_2(3) = 0210120102012, \\
f_T(4) = 0434. & T_1(4) = 0121012021. & T_2(4) = 0210201.
\end{array}$$

To obtain the last two results, Thue first proved that $f_T^\infty(0)$ is characterized by $\{AA\} \cup \{02,03,10,14,21,23,24,30,31,41,42,040,132,404,1201,2012\}$.

In this paper, we prove such characterizations mostly for the binary words considered by the first author [1]. We also obtain smaller morphisms for Thue's words avoiding $\{AA\} \cup \{010,020\}$ and $\{AA\} \cup \{121,212\}$ as well as a characterization for words avoiding the patterns $AABBCC$ (i.e., three consecutive squares), $ABCABC$ and a finite set of factors. The results are summarized in Table 1. The first column shows the description of the considered language given in the literature. It is either given by forbidden sets of patterns and factors, or by the notation (e, n, m), which means that we consider the binary words avoiding repetitions with exponent strictly greater than e, containing exactly n distinct repetitions with exponent e as a factor, and containing the minimum number m of distinct squares. We use the notation SQ_t for the pattern corresponding to squares with period at least t, that is, $SQ_1 = AA$, $SQ_2 = ABAB$, $SQ_3 = ABCABC$, and so on. These languages actually have an equivalent definition with one forbidden pattern SQ_t and a finite set of forbidden factors. This standardized definition, given in the second column, is more suited for proving the characterization. The third column gives the corresponding morphic word. The fourth column indicates the section containing the corresponding set F_{xx} and morphism g_{xx}.

To define a morphic word $g(f^\infty(0))$, we allow that g is an *erasing* morphism, i.e., that the g-image of a letter is empty. Notice that replacing g by $h_c = g \circ f^c$ defines the same morphic word, and that h_c is non-erasing for some small constant c.

The proofs are obtained by computer using the technique described in the next section. An example of proof by hand is given for Theorem 3. The morphic words in Table 1 are gathered according the pure morphic word they are built on. We introduce in Section 3 a pure morphic word $f_5^\infty(0)$ similar to Thue's word $f_T^\infty(0)$ and we characterize some of its morphic images. Section 4 is devoted to characterizations of some morphic images of Thue's ternary pure morphic word $f_3^\infty(0)$.

Original form	Standardized form	Morphic word	Section
$\{AA\} \cup \{\mathtt{010},\mathtt{020}\}$	$\{AA\} \cup \{\mathtt{010},\mathtt{020}\}$	$M_1(f_5^\infty(\mathtt{0}))$	3.1
$\{AA\} \cup \{\mathtt{121},\mathtt{212}\}$	$\{AA\} \cup \{\mathtt{121},\mathtt{212}\}$	$M_2(f_5^\infty(\mathtt{0}))$	3.1
$(5/2, 2, 8)$	$\{SQ_7\} \cup F_8$	$g_8(f_5^\infty(\mathtt{0}))$	3.2
$(7/3, 2, 12)$	$\{SQ_9\} \cup F_{12}$	$g_{12}(f_5^\infty(\mathtt{0}))$	3.3
$(7/3, 1, 14)$	$\{SQ_9\} \cup F_{14}$	$g_{14}(f_5^\infty(\mathtt{0}))$	3.4
$\{AABBCC, SQ_3\} \cup F'_{cs}$	$\{SQ_3\} \cup F_{cs}$	$g_{cs}(f_5^\infty(\mathtt{0}))$	3.5
$(5/2, 1, 11)$	$\{SQ_5\} \cup F_{11}$	$g_{11}(f_3^\infty(\mathtt{0}))$	4.1
$(3, 2, 3) \cup F'_3$	$\{SQ_3\} \cup F_3$	$g_3(f_3^\infty(\mathtt{0}))$	4.2
$\{AABBCABBA\} \cup \{\mathtt{0011},\mathtt{1100}\}$	$\{SQ_5\} \cup F_q$	$g_q(f_3^\infty(\mathtt{0}))$	4.3

Figure 1: Table of results

2. Characterizing a morphic word

A morphism $f : \Sigma_k^* \to \Sigma_k^*$ is *primitive* if there exists $n \in \mathbb{N}$ such that $f^n(a)$ contains b for every $a, b \in \Sigma_k$. We are given a primitive morphism $f : \Sigma_k^* \to \Sigma_k^*$, a morphism $g : \Sigma_k^* \to \Sigma_{k'}^*$, and a finite set of factors $\mathcal{F}_m \subset \Sigma_{k'}^*$. We want to prove that $g(f^\infty(\mathtt{0}))$ is characterized by $\{SQ_t\} \cup \mathcal{F}_m$.

We assume that $g(f^\infty(\mathtt{0}))$ avoids $\{SQ_t\} \cup \mathcal{F}_m$. This can be checked using Cassaigne's algorithm [5] that determines if a morphic word defined by circular morphisms avoids a given pattern with constants. We refer to Cassaigne [5] for the definitions of circular morphisms, synchronization point, and synchronization delay. We can use an online implementation [4] of this algorithm. We also assume that the pure morphic word $f^\infty(\mathtt{0})$ is characterized by $\{AA\} \cup \mathcal{F}_p$ for some finite set of factors $\mathcal{F}_p \subset \Sigma_k^*$.

We compute the smallest integer c such that $\min\{|g(f^c(a))|,\ a \in \Sigma_k\} \geqslant t$. This c exists because f is primitive. We can consider the morphism $g' = g \circ f^c$ instead of g since we have $g'(f^\infty(\mathtt{0})) = g(f^\infty(\mathtt{0}))$.

First, we check that g' is circular. Then, we compute the set S_l of words v such that there exists a word $pvs \in \Sigma_{k'}^*$ avoiding $\{SQ_t\} \cup \mathcal{F}_m$, where $l = \max\{|u|, u \in \mathcal{F}_p\} \times \max\{|g'(a)|, a \in \Sigma_k\}$, $|v| = l$, and $|p| = |s| = 4l$. To do this, we simply perform a depth-first exploration of the words of length $9l$ avoiding $\{SQ_t\} \cup \mathcal{F}_m$ and for each of them, we put the central factor of length l in S_l. The running time of this brute-force approach is not so prohibitive precisely because the characterization implies a polynomial factor complexity. Finally, we check that every word in S_l is a factor of $g'(f^\infty(\mathtt{0}))$.

This implies that an infinite word over $\Sigma_{k'}$ avoiding $\{SQ_t\} \cup \mathcal{F}_m$ is the g'-image of an infinite word $w \in \Sigma_k^*$. Now w is square-free, since otherwise $g'(w)$ would contain a square of period at least t. Also w does not contain a word $y \in \mathcal{F}_p$, because $g'(y)$ is a word of length at most l that is not a factor of any word in S_l. So w avoids $\{AA\} \cup \mathcal{F}_p$, and thus has the same set of factors as $f^\infty(\mathtt{0})$. Thus, every infinite recurrent word over $\Sigma_{k'}$ avoiding $\{SQ_t\} \cup \mathcal{F}_m$ has the same set of factors as $g'(f^\infty(\mathtt{0}))$.

The programs we used are available at
`http://www.lirmm.fr/~ochem/morphisms/characterization.htm` .

3. A pure morphic word over Σ_5

We define the morphism f_5 from Σ_5^* to Σ_5^* as follows:

$$\begin{aligned} f_5(\mathtt{0}) &= \mathtt{01},\\ f_5(\mathtt{1}) &= \mathtt{23},\\ f_5(\mathtt{2}) &= \mathtt{4},\\ f_5(\mathtt{3}) &= \mathtt{21},\\ f_5(\mathtt{4}) &= \mathtt{0}. \end{aligned}$$

We also define the set

$$F_5 = \{\mathtt{02}, \mathtt{03}, \mathtt{13}, \mathtt{14}, \mathtt{20}, \mathtt{24}, \mathtt{31}, \mathtt{32}, \mathtt{40}, \mathtt{41}, \mathtt{43}, \mathtt{121}, \mathtt{212}, \mathtt{304}, \mathtt{3423}, \mathtt{4234}\}.$$

Theorem 1. *$\{AA\} \cup F_5$ characterizes $f_5^\infty(\mathtt{0})$.*

Proof. We adapt the method of the previous section for morphic words to the pure morphic word $f_5^\infty(\mathtt{0})$ by setting $g = g' = f_5$ and $\mathcal{F}_m = \mathcal{F}_p = F_5$. We set $l = \max\{|u|, u \in F_5\} \times \max\{|f_5(a)|, a \in \Sigma_k\} = 8$. We compute the set S_l of words v such that there exists a word $pvs \in \Sigma_5^*$ avoiding squares and F_5 with $|v| = l$ and $|p| = |s| = 4l$. Then we check that every word in S_l is a factor of $f_5^\infty(\mathtt{0})$.

The morphism f_5 is circular with synchronization delay 1. Indeed, for every factor of length 1 of the f_5-image of some word, we can insert at least one synchronization point $|$ between letter images:

$$\begin{aligned} &\mathtt{0} \text{ implies } |\mathtt{0},\\ &\mathtt{1} \text{ implies } \mathtt{1}|,\\ &\mathtt{2} \text{ implies } |\mathtt{2},\\ &\mathtt{3} \text{ implies } \mathtt{3}|,\\ &\mathtt{4} \text{ implies } |\mathtt{4}|. \end{aligned}$$

This implies that every infinite recurrent word over Σ_5 avoiding $\{AA\} \cup F_5$ is the f_5-image of some infinite recurrent word w over Σ_5. Notice that w must be square-free, since otherwise $f_5(w)$ would not avoid squares. Now suppose that w contains a factor $y \in F_5$. Then $f_5(y)$ must appear as a factor in S_l since $|f_5(y)| \leqslant 8 = l$. Every word in S_l is a factor of $f_5^\infty(\mathtt{0})$, so $f_5(y)$ should also be a factor of $f_5^\infty(\mathtt{0})$, which is a contradiction. So w avoids squares and F_5, which implies by induction that it has the same set of factors as $f_5^\infty(\mathtt{0})$. Finally, we have that every infinite recurrent word over Σ_5 avoiding $\{AA\} \cup F_5$ is of the form $f_5(w)$ where w has the same set of factors as $f_5^\infty(\mathtt{0})$, so that $f_5(w)$ also has the same set of factors as $f_5^\infty(\mathtt{0})$.

□

Since many morphic words in this paper are obtained as the image of $f_5^\infty(\mathtt{0})$, let us state some of its properties. In $f_5^\infty(\mathtt{0})$, the letters $\mathtt{0}$, $\mathtt{1}$, and $\mathtt{2}$ have frequency $\sqrt{5} - 2$ and the letters $\mathtt{3}$ and $\mathtt{4}$ have frequency $\left(7 - 3\sqrt{5}\right)/2$. Notice that $\{AA\} \cup F_5$, and thus the set of factors of $f_5^\infty(\mathtt{0})$, is invariant by the operation consisting in reversing the word and exchanging 3 and 4. This is trivially true for squares. For a word in F_5, say 40, we obtain 04 by reversing the word and we obtain 03 by exchanging 3 and 4, then we have that F_5 contains indeed 03. The factor complexity of $f_5^\infty(\mathtt{0})$ seems to be $4n+1$ for every factor length $n \geqslant 0$.

3.1. Smaller morphisms for Thue's words

Let M_1 and M_2 be the morphisms from Σ_5^* to Σ_3^* defined by

$$\begin{array}{ll}
M_1(\mathtt{0}) = \mathtt{012}, & M_2(\mathtt{0}) = \mathtt{02}, \\
M_1(\mathtt{1}) = \mathtt{1}, & M_2(\mathtt{1}) = \mathtt{1}, \\
M_1(\mathtt{2}) = \mathtt{02}, & M_2(\mathtt{2}) = \mathtt{0}, \\
M_1(\mathtt{3}) = \mathtt{12}, & M_2(\mathtt{3}) = \mathtt{12}, \\
M_1(\mathtt{4}) = \varepsilon. & M_2(\mathtt{4}) = \varepsilon.
\end{array}$$

Theorem 2.

- $\{AA\} \cup \{\mathtt{010}, \mathtt{020}\}$ *characterizes the morphic word* $M_1(f_5^\infty(\mathtt{0}))$,
- $\{AA\} \cup \{\mathtt{121}, \mathtt{212}\}$ *characterizes the morphic word* $M_2(f_5^\infty(\mathtt{0}))$.

Thue noticed that every word avoiding $\{AA\} \cup \{\mathtt{121},\mathtt{212}\}$ can be obtained from a word avoiding $\{AA\}\cup\{\mathtt{010},\mathtt{020}\}$ by deleting the letter immediately after each occurrence of the letter 0. This property is easy to check by comparing M_2 to M_1 and it explains why the same pure morphic word is used for both types of words. The morphisms M_1 and M_2 are the smallest possible. However, the morphisms $M_1' = M_1 \circ f_5$ and $M_2' = M_2 \circ f_5$ given below provide additional insight.

$$\begin{array}{ll}
M_1'(\mathtt{0}) = \mathtt{0121}, & M_2'(\mathtt{0}) = \mathtt{021}, \\
M_1'(\mathtt{1}) = \mathtt{0212}, & M_2'(\mathtt{1}) = \mathtt{012}, \\
M_1'(\mathtt{2}) = \varepsilon, & M_2'(\mathtt{2}) = \varepsilon, \\
M_1'(\mathtt{3}) = \mathtt{021}, & M_2'(\mathtt{3}) = \mathtt{01}, \\
M_1'(\mathtt{4}) = \mathtt{012}. & M_2'(\mathtt{4}) = \mathtt{02}.
\end{array}$$

The morphism M_1' exhibits natural properties of words avoiding $\{AA\}\cup\{\mathtt{010},\mathtt{020}\}$ and of $M_1(f_5^\infty(\mathtt{0}))$:

- The set $\{\mathtt{0121},\mathtt{0212},\mathtt{012},\mathtt{021}\}$ is a code for words avoiding $\{AA\}\cup\{\mathtt{010},\mathtt{020}\}$.
- The asymptotic frequencies of the factors 121 and 212 are equal since the letters 1 and 2 are symmetrical for words avoiding $\{AA\} \cup \{\mathtt{010},\mathtt{020}\}$.
- Similarly, the asymptotic frequencies of 0120 and 0210 are equal.
- By applying the symmetry of the factors of $f_5^\infty(\mathtt{0})$ to M_1', that is, reversing the M_1'-images of every letter and exchanging 3 and 4, we obtain the conjugate morphism of M_1' such that the common prefix 0 becomes the common suffix.

Except for the last, similar remarks hold for M_2'. The factor complexity of $M_1(f_5^\infty(\mathtt{0}))$ and $M_2(f_5^\infty(\mathtt{0}))$ seems to be $4n-2$ for every factor length $n \geqslant 2$.

3.2. Words containing two 5/2-repetitions and 8 squares

If an infinite binary word contains the repetitions $\mathtt{01010}$ and $\mathtt{10101}$ of exponent $5/2$ and no other overlap, then it contains at least 8 distinct squares. Moreover, if it contains exactly 8 distinct squares, then these 8 squares are $\mathtt{0}^2$, $\mathtt{1}^2$, $(\mathtt{01})^2$, $(\mathtt{10})^2$, $(\mathtt{0110})^2$, $(\mathtt{1001})^2$, $(\mathtt{011001})^2$, $(\mathtt{100110})^2$. Equivalently, a recurrent binary word containing these overlaps and squares avoids SQ_7 and the set

$$\begin{aligned}F_8 = \{&\mathtt{000,111,00100,11011,010010,010101,101010,101101,00110011,}\\&\mathtt{11001100,1011001011,0100110100}\}.\end{aligned}$$

Let g_8 be the morphism from Σ_5^* to Σ_2^* defined by

$$\begin{aligned}g_8(\mathtt{0}) &= \mathtt{011},\\ g_8(\mathtt{1}) &= \mathtt{0},\\ g_8(\mathtt{2}) &= \mathtt{01},\\ g_8(\mathtt{3}) &= \varepsilon,\\ g_8(\mathtt{4}) &= \varepsilon.\end{aligned}$$

Theorem 3. *$\{SQ_7\} \cup F_8$ characterizes $g_8(f_5^\infty(\mathtt{0}))$.*

Proof. We assume that $g_8(f_5^\infty(\mathtt{0}))$ avoids SQ_7 and F_8 and we prove the other direction of Theorem 3. That is, we suppose that G_8 is an infinite recurrent word avoiding $\{SQ_7\} \cup F_8$ and we show that every factor of G_8 is a factor of $g_8(f_5^\infty(\mathtt{0}))$. We consider the morphism $g'_8 = g_8 \circ f_5^5$ given below instead of g_8 because we have $\min\{|g'_8(a)|,\ a \in \Sigma_5\} = 9 \geqslant 7 = t$, as specified in the method.

$$\begin{aligned}g'_8(\mathtt{0}) &= \mathtt{0110010100110101100110 10},\\ g'_8(\mathtt{1}) &= \mathtt{011001011001101},\\ g'_8(\mathtt{2}) &= \mathtt{011001010},\\ g'_8(\mathtt{3}) &= \mathtt{0110010110011010},\\ g'_8(\mathtt{4}) &= \mathtt{01100101001101}.\end{aligned}$$

Let $p = \mathtt{01100101}$ be the common prefix of the factors $g'_8(a)$ for $a \in \Sigma_5$. It is easy to check that every occurrence of p in the g'_8-image of a word is the prefix of g'_8-image of a letter. So g'_8 has bounded synchronization delay. Moreover, a computer check shows that the factors of G_8 are factors of the g'_8-image of a word. Let $L \subset \Sigma_5^*$ denote the language of words whose g'_8-image is a factor of G_8. We show that L is the set of factors of $f_5^\infty(\mathtt{0})$. Suppose that L contains a square uu for some $u \in \Sigma_5^+$. Then G_8 contains the square $g'_8(uu)$ with period $|g'_8(u)| \geqslant 9$. This is a contradiction since G_8 avoids SQ_7, so L is square-free.

Now, for every $w \in F_5$, we suppose that $w \in L$ and obtain a contradiction:

- $w \in \{\mathtt{02,32}\}$: $g'_8(\mathtt{02})p$ and $g'_8(\mathtt{32})p$ both contain the square $\mathtt{1}g'_8(\mathtt{2})p = (\mathtt{001100101})^2$ with period 9 as a suffix.

- $w = \mathtt{03}$: $g'_8(\mathtt{03})p$ contains the square $(\mathtt{1001101001100101})^2$ with period 16 as a suffix.

- $w \in \{\texttt{13,41,43}\}$: A common suffix of $g_8'(\texttt{1})$ and $g_8'(\texttt{4})$ is $\texttt{1}$. A common prefix of $g_8'(\texttt{1})$ and $g_8'(\texttt{3})$ is $\texttt{011001011}$. So, in every case, $g_8'(w)$ contains the factor $\texttt{1011001011} \in F_8$.

- $w = \texttt{14}$: $g_8'(\texttt{14})p$ contains the square $(\texttt{00110101100101})^2$ with period 14 as a suffix.

- $w \in \{\texttt{20,24}\}$: $g_8'(\texttt{20})$ and $g_8'(\texttt{24})$ both contain the square $g_8'(\texttt{22})$ with period 9 as a prefix.

- $w = \texttt{31}$: $g_8'(\texttt{31})p$ contains the square $g_8'(\texttt{33})$ with period 16 as a prefix.

- $w = \texttt{40}$: $g_8'(\texttt{40})$ contains the square $g_8'(\texttt{44})$ with period 14 as a prefix.

- $w = \texttt{304}$: $g_8'(\texttt{304}) = \texttt{0110}(\texttt{010110011010011001010011})^2\texttt{01}$ contains a square with period 24.

- $w = \texttt{121}$: Since L is square-free and avoids $\{\texttt{13,14}\}$, L must contain $\texttt{1210}$. However, $g_8'(\texttt{1210})$ contains the square $g_8'(\texttt{1212})$ with period 24 as a prefix.

- $w = \texttt{212}$: Since L is square-free and avoids $\{\texttt{20,24}\}$, L must contain $\texttt{2123}$. However, $g_8'(\texttt{2123})$ contains the square $g_8'(\texttt{2121})$ with period 24 as a prefix.

- $w = \texttt{3423}$: Since L is square-free and avoids $\{\texttt{03,13,43}\}$, L must contain $\texttt{23423}$. Since L is square-free and avoids $\{\texttt{31,32}\}$, L must contain $\texttt{234230}$. However, $g_8'(\texttt{234230})$ contains the square $g_8'(\texttt{234234})$ with period 39 as a prefix.

- $w = \texttt{4234}$: Since L is square-free and avoids $\{\texttt{40,41,43}\}$, L must contain $\texttt{42342}$. Since L is square-free and avoids $\{\texttt{20,24}\}$, L must contain $\texttt{423421}$. However, $g_8'(\texttt{423421})p$ contains the square $g_8'(\texttt{423423})$ with period 39 as a prefix.

Therefore L is square-free and does not contain a factor in F_5, thus L is the set of factors as $f_5^\infty(\texttt{0})$ by Theorem 1. □

Notice that the last part of the proof above (that every word in F_p is a forbidden factor in L) differs from the computer check described in Section 2. The proof by hand exhibits a forbidden factor in $\{SQ_t\} \cup F_m$ for every word in F_p. The computer check does the contrapositive: It lists all words avoiding $\{SQ_t\} \cup F_m$ of some sufficient length and checks that they are g'-images of some word avoiding $\{AA\} \cup F_p$.

The factor complexity of $g_8(f_5^\infty(\texttt{0}))$ seems to be $4n-6$ for every factor length $n \geqslant 3$.

3.3. Words containing two 7/3-repetitions and 12 squares

If an infinite binary word contains the repetitions $\mathtt{0110110}$ and $\mathtt{1001001}$ of exponent $7/3$ and no other overlap, then it contains at least 12 distinct squares. Moreover, if it contains exactly 12 distinct squares, then these 12 squares are $\mathtt{0}^2$, $\mathtt{1}^2$, $(\mathtt{01})^2$, $(\mathtt{10})^2$, $(\mathtt{001})^2$, $(\mathtt{010})^2$, $(\mathtt{011})^2$, $(\mathtt{100})^2$, $(\mathtt{101})^2$, $(\mathtt{110})^2$, $(\mathtt{01101001})^2$, $(\mathtt{10010110})^2$. Equivalently, a recurrent binary word containing these overlaps and squares avoids SQ_9 and the set

$$\begin{aligned}F_{12} = \{&\mathtt{000,111,01010,10101,001100,110011,0010010,0100100,1011011,}\\ &\mathtt{1101101,0011010011,0101100101,1010011010,1100101100,}\\ &\mathtt{01001011010010}\}.\end{aligned}$$

Let g_{12} be the morphism from Σ_5^* to Σ_2^* defined by

$$\begin{aligned}g_{12}(\mathtt{0}) &= \mathtt{01},\\ g_{12}(\mathtt{1}) &= \mathtt{0},\\ g_{12}(\mathtt{2}) &= \mathtt{011},\\ g_{12}(\mathtt{3}) &= \varepsilon,\\ g_{12}(\mathtt{4}) &= \varepsilon.\end{aligned}$$

Theorem 4. *$\{SQ_9\} \cup F_{12}$ characterizes $g_{12}(f_5^\infty(\mathtt{0}))$.*

The factor complexity of $g_{12}(f_5^\infty(\mathtt{0}))$ seems to be $4n-6$ for every factor length $n \geqslant 3$.

3.4. Words containing one 7/3-repetition and 14 squares

If an infinite binary word contains the repetition $\mathtt{1001001}$ of exponent $7/3$ and no other overlap, then it contains at least 14 distinct squares. Moreover, if it contains exactly 14 distinct squares, then these 14 squares are $\mathtt{0}^2$, $\mathtt{1}^2$, $(\mathtt{01})^2$, $(\mathtt{10})^2$, $(\mathtt{001})^2$, $(\mathtt{010})^2$, $(\mathtt{100})^2$, $(\mathtt{101})^2$, $(\mathtt{0110})^2$, $(\mathtt{1001})^2$, $(\mathtt{100110})^2$, $(\mathtt{0100110})^2$, $(\mathtt{0110010})^2$, and $(\mathtt{10010110})^2$. Equivalently, a recurrent binary word containing these overlaps and squares avoids SQ_9 and the set

$$\begin{aligned}F_{14} = \{&\mathtt{000,111,11011,010101,101010,0010010,0100100,00110011,}\\ &\mathtt{11001100,101001101,101100101,0100101101,1100101100,}\\ &\mathtt{001001100100,010011010011,0011001001100,1011010010110011}\}.\end{aligned}$$

Let g_{14} be the morphism from Σ_5^* to Σ_2^* defined by

$$\begin{aligned}g_{14}(\mathtt{0}) &= \mathtt{01},\\ g_{14}(\mathtt{1}) &= \mathtt{00110},\\ g_{14}(\mathtt{2}) &= \mathtt{1},\\ g_{14}(\mathtt{3}) &= \mathtt{0010110},\\ g_{14}(\mathtt{4}) &= \mathtt{0110}.\end{aligned}$$

Theorem 5. *$\{SQ_9\} \cup F_{14}$ characterizes $g_{14}(f_5^\infty(\mathtt{0}))$.*

The factor complexity of $g_{14}(f_5^\infty(\mathtt{0}))$ seems to be $4n-1$ for every factor length $n \geqslant 11$.

3.5. Words avoiding AABBCC

The second author proved that the pattern $AABBCC$, i.e., three consecutive squares, can be avoided over the binary alphabet [8]. More precisely, there exist exponentially many binary words avoiding both $AABBCC$ and SQ_3. However, if we forbid also the factors in

$$F'_{cs} = \{\mathtt{0001110010110},\mathtt{0110100111000},\mathtt{1001011000111},\mathtt{1110001101001}\},$$

we obtain a characterization of the morphic word $g_{cs}(f_5^\infty(\mathtt{0}))$, where g_{cs} is the morphism from Σ_5^* to Σ_2^* defined by

$$\begin{aligned}
g_{cs}(\mathtt{0}) &= \mathtt{00101100011010},\\
g_{cs}(\mathtt{1}) &= \mathtt{0111},\\
g_{cs}(\mathtt{2}) &= \mathtt{0010111010},\\
g_{cs}(\mathtt{3}) &= \mathtt{011100011010},\\
g_{cs}(\mathtt{4}) &= \mathtt{001011000111}.
\end{aligned}$$

The word $g_{cs}(f_5^\infty(\mathtt{0}))$ avoids SQ_3 and the set

$$\begin{aligned}
F_{cs} = \{&\mathtt{0000},\mathtt{1111},\mathtt{01010},\mathtt{10101},\mathtt{011001},\mathtt{100110},\mathtt{0011101},\mathtt{1011100},\\
&\mathtt{1100010},\mathtt{00010111},\mathtt{11101000},\mathtt{0001110010110},\mathtt{0110100111000},\\
&\mathtt{1001011000111},\mathtt{1110001101001}\}
\end{aligned}$$

Theorem 6. *$\{AABBCC, SQ_3\} \cup F'_{cs}$ and $\{SQ_3\} \cup F_{cs}$ both characterize $g_{cs}(f_5^\infty(\mathtt{0}))$.*

The factor complexity of $g_{cs}(f_5^\infty(\mathtt{0}))$ seems to be $4n+4$ for every factor length $n \geqslant 6$.

4. Thue's ternary pure morphic word

Thue [3, 10, 11] proved that $\{AA\} \cup \{\mathtt{010},\mathtt{212}\}$ characterizes the fixed point of f_3. In this section, we give characterizations of three words that are morphic images of $f_3^\infty(\mathtt{0})$. It is not surprising that $f_3^\infty(\mathtt{0})$ appears in the context of characterizations: as soon as a morphism m is such that $m(\mathtt{0}) = axb$ and $m(\mathtt{1}) = ab$, the m-image of words of the form $0u1u0$, $u \in \Sigma_3^*$, contains a large square: $m(0u1u0) = axbm(u)abm(u)axb$ contains $(bm(u)a)^2$. Moreover, a ternary square-free word avoids factors of the form $0u1u0$ with $u \in \Sigma_3^*$ if and only if it avoids $\{010, 212\}$ [9]. So, the set of factors of a factorial langage containing only square-free factors in $\{m(0), m(1), m(2)\}^*$ such that $m(\mathtt{0}) = axb$ and $m(1) = ab$ is the set of factors of $m(f_3^\infty(\mathtt{0}))$. It is also easy to check that $\{AA\} \cup \{\mathtt{010},\mathtt{212}\}$ characterizes the same ternary word as $\{AA\} \cup \{\mathtt{1021},\mathtt{1201}\}$.

4.1. Words containing one 5/2-repetition and 11 squares

If an infinite binary word contains the repetition $\mathtt{10101}$ of exponent $5/2$ and no other overlap, then it contains at least 11 distinct squares. Moreover, if it contains exactly 11 distinct squares, then these 11 squares are $\mathtt{0}^2$, $\mathtt{1}^2$, $(\mathtt{01})^2$,

$(\mathtt{10})^2$, $(\mathtt{001})^2$, $(\mathtt{010})^2$, $(\mathtt{011})^2$, $(\mathtt{100})^2$, $(\mathtt{101})^2$, $(\mathtt{110})^2$, $(\mathtt{01100110})^2$. Equivalently, a recurrent binary word containing these overlaps and squares avoids SQ_7 and the set

$$F_{11} = \{\mathtt{000,111,01010,001100,0010010,0100100,1011011,1101101}\}.$$

Let g_{11} be the morphism from Σ_3^* to Σ_2^* defined by

$$\begin{aligned} g_{11}(\mathtt{0}) &= \mathtt{1001001101011001101001011001001101100} \\ &\ \mathtt{1011010011011001001101001011001101 01}, \\ g_{11}(\mathtt{1}) &= \mathtt{100100110100101}, \\ g_{11}(\mathtt{2}) &= \mathtt{1001001101100101101001101}. \end{aligned}$$

Theorem 7. *$\{SQ_5\} \cup F_{11}$ characterizes $g_{11}(f_3^\infty(\mathtt{0}))$.*

4.2. Words containing 3 squares

It is known that there exist exponentially many binary words containing only 3 distinct squares [7, 8]. Without loss of generality, we assume that these 3 squares are 00, 11, and 1010. To obtain a characterization, we forbid also the factors in $F_3' = \{\mathtt{01000110,10011101,1001101000,1110100110}\}$. If w is a recurrent binary word avoiding F_3' and squares distinct from 00, 11, and 1010, then w avoids SQ_3 and the set

$$F_3 = \{\mathtt{0000,0101,1111,01000110,10011101,1001101000,1110100110}\}.$$

Let g_3 be the morphism from Σ_3^* to Σ_2^* defined by

$$\begin{aligned} g_3(\mathtt{0}) &= \mathtt{000111}, \\ g_3(\mathtt{1}) &= \mathtt{0011}, \\ g_3(\mathtt{2}) &= \mathtt{01001110001101}. \end{aligned}$$

Theorem 8. *$\{SQ_3\} \cup F_3$ characterizes $g_3(f_3^\infty(\mathtt{0}))$.*

4.3. Words avoiding AABBCABBA

Another characterization has been obtained by the second author [9]: $\{AABBCABBA\} \cup \{\mathtt{0011,1100}\}$ characterizes $g_q(f_3^\infty(\mathtt{0}))$, where g_q is given below.

$$\begin{aligned} g_q(\mathtt{0}) &= \mathtt{0010110111011101001}, \\ g_q(\mathtt{1}) &= \mathtt{00101101101001}, \\ g_q(\mathtt{2}) &= \mathtt{00010}. \end{aligned}$$

Equivalently, $g_q(f_3^\infty(\mathtt{0}))$ is characterized by $\{SQ_5\} \cup F_q$ where

$$F_q = \{\mathtt{0000,0011,1100,1111,01010,10101,010111,101000,0001001,}$$
$$\mathtt{1110110,00100100,01011010,10100101,11011011,0110111010,1001000101}\}$$

5. Concluding remarks

We have seen in Section 4 why $f_3^\infty(\mathtt{0})$ appears often in the context of characterization. Also, we have seen in Section 3.1 why Thue's words avoiding $\{AA\} \cup \{\mathtt{010,020}\}$ and $\{AA\} \cup \{\mathtt{121,212}\}$ use the same pure morphic word $f_5^\infty(\mathtt{0})$. However, we do not see why $f_5^\infty(\mathtt{0})$ is used in other "natural" languages. It would be interesting to investigate its properties, in particular to prove that its factor complexity is $4n+1$ and that its critical exponent is $(5+\sqrt{5})/4$.

The fixed point of $\mathtt{0} \mapsto \mathtt{01}$, $\mathtt{1} \mapsto \mathtt{0}$, known as the Fibonacci word, seems to have the same set of factors as $g_{\mathtt{fib}}(f_5^\infty(\mathtt{0}))$, where $g_{\mathtt{fib}}$ is given below. Moreover, the Rote-Fibonacci word studied in [6] seems to have the same set of factors as $g_{\mathtt{rf}}(f_5^\infty(\mathtt{0}))$, where $g_{\mathtt{rf}}$ is given below.

$$\begin{array}{ll}
g_{\mathtt{fib}}(\mathtt{0}) = \mathtt{01}, & g_{\mathtt{rf}}(\mathtt{0}) = \mathtt{01}, \\
g_{\mathtt{fib}}(\mathtt{1}) = \mathtt{0}, & g_{\mathtt{rf}}(\mathtt{1}) = \mathtt{10}, \\
g_{\mathtt{fib}}(\mathtt{2}) = \mathtt{1}, & g_{\mathtt{rf}}(\mathtt{2}) = \varepsilon, \\
g_{\mathtt{fib}}(\mathtt{3}) = \mathtt{0}, & g_{\mathtt{rf}}(\mathtt{3}) = \mathtt{11}, \\
g_{\mathtt{fib}}(\mathtt{4}) = \mathtt{0}. & g_{\mathtt{rf}}(\mathtt{4}) = \mathtt{00}.
\end{array}$$

The method discussed in this paper is not able to prove such equivalences because the languages are not defined by avoiding large squares and a finite set of factors. Maybe it can be proven by the method used in [6] to recover many known results about the Fibonacci word.

Baker, McNulty, and Taylor [2] obtained that $ABXBAYACZCAWBC \cup \{02\}$ characterizes the fixed point of $\mathtt{0} \mapsto \mathtt{01}$, $\mathtt{1} \mapsto \mathtt{21}$, $\mathtt{2} \mapsto \mathtt{03}$, $\mathtt{3} \mapsto \mathtt{23}$ over Σ_4. Notice that the forbidden factor 02 is not crucial here, its only role is to distinguish one out of three symmetric versions obtained by permutation of the alphabet letters. So, characterizations are known for the patterns AA, $ABABA$, $ABCABC$, $AABBCC$, $AABBCABBA$, and $ABXBAYACZCAWBC$. An interesting open question is the following: Suppose that P is an avoidable pattern with avoidability index $\lambda(P) = k$. Is it possible to find a finite set $\mathcal{P}$ of patterns and a finite set $\mathcal{F}$ of factors such that $P \in \mathcal{P}$ and $\mathcal{P} \cup \mathcal{F}$ characterizes a morphic word over Σ_k ? This would be a strengthening of Cassaigne's conjecture stating that there exists a morphic word avoiding P over Σ_k.

References

[1] G. Badkobeh. Fewest repetitions vs maximal-exponent powers in infinite binary words, *Theoret. Comput. Sci.* **412** (2011), 6625–6633.

[2] K.A. Baker, G.F. McNulty, and W. Taylor. Growth problems for avoidable words, *Theoret. Comput. Sci* **69** (1989), 319–345.

[3] J. Berstel. Axel Thue's Papers on Repetitions in Words: a Translation. *Publications du Laboratoire de Combinatoire et d'Informatique Mathématique. Université du Québec à Montréal*, Number 20, February 1995.

[4] F. Blanchet-Sadri, K. Black, and A. Zemke. Avoidable patterns in partial words. http://www.uncg.edu/cmp/research/patterns/implementation.html

[5] J. Cassaigne. An algorithm to test if a given circular HD0L-language avoids a pattern. Information processing '94, Vol. I (Hamburg, 1994), 459–464, IFIP Trans. A Comput. Sci. Tech., A-51, North-Holland, Amsterdam, 1994

[6] C. F. Du, H. Mousavi, L. Schaeffer, and J. Shallit. Decision algorithms for Fibonacci-automatic words, with applications to pattern avoidance. arXiv:1406.0670

[7] T. Harju and D. Nowotka. Binary words with few squares. *Bull. EATCS* **89** (2006), 164–166.

[8] P. Ochem. A generator of morphisms for infinite words. *RAIRO - Theoret. Informatics Appl.* **40** (2006), 427–441.

[9] P. Ochem. Binary words avoiding the pattern AABBCABBA. *RAIRO - Theoret. Informatics Appl.* **44(1)** (2010), 151–158.

[10] A. Thue. Über unendliche Zeichenreihen. *Norske vid. Selsk. Skr. Mat. Nat. Kl.* **7** (1906), 1–22. Reprinted in *Selected Mathematical Papers of Axel Thue*, T. Nagell, editor, Universitetsforlaget, Oslo, 1977, pp. 139–158.

[11] A. Thue. Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen. *Norske vid. Selsk. Skr. Mat. Nat. Kl.* **1** (1912), 1–67. Reprinted in *Selected Mathematical Papers of Axel Thue*, T. Nagell, editor, Universitetsforlaget, Oslo, 1977, pp. 413–478.

Appendices

A Curriculum Vitae

État civil

Nom, prénom : OCHEM, Pascal
Date et lieu de naissance : 13 janvier 1978, Creutzwald (Moselle)
Nationalité : Française
E-mail : ochem@lirmm.fr

Parcours

- CR CNRS section 6
  - au LIRMM (UMR 5506) à Montpellier depuis juillet 2011.
  - au laboratoire Poncelet (UMI 5141) à Moscou de septembre 2008 à aout 2009.
  - au LRI (UMR 8623) à Orsay d'octobre 2007 à aout 2008 et de sept. 2009 à juin 2011.
- A.T.E.R temps plein à Bordeaux 4 de septembre 2006 à aout 2007.
- A.T.E.R mi-temps à Bordeaux 1 de septembre 2005 à aout 2006.
- Vacataire TD et TP à Bordeaux 1 de janvier 2003 à juin 2005.
- novembre 2005 : Doctorat d'informatique de l'Université Bordeaux 1.
  mention très honorable.
  LaBRI - Université Bordeaux 1.
  Sujet : Graph coloring and combinatorics on words.
  Directeur de thèse : Éric Sopena.
  Rapporteurs : Jean Berstel, Jan Kratochvíl.
  Membres du jury : Jean Berstel, Robert Cori, Jan Kratochvíl,
  André Raspaud, Gwénaël Richomme, Éric Sopena.

B Recherche scientifique post-thèse

En théorie des graphes, je me suis beaucoup intéressé à des questions de NP-complétude. En combinatoire des mots, j'ai continué d'étudier l'évitement de motifs, par l'utilisation classique de morphismes et par une méthode non-constructive que j'ai développée et qui est simple d'utilisation. En 2009, j'ai commencé un nouveau domaine de recherche : l'obtention de bornes sur divers paramètres d'un éventuel nombre parfait impair.

Théorie des graphes - NP-complétude

Beaucoup de problèmes de coloration ou de partition de sommets ou d'arêtes sont NP-complet pour les graphes en général, et il est devenu routinier de montrer, dès son introduction, qu'un nouveau problème est NP-complet. J'ai obtenu trois types de résultats qui vont au dela de cette routine.

Le premier de ces types est de montrer la NP-complétude pour une ou plusieurs classes de graphes les plus restreintes possibles. Nous avons ainsi montré [J41] qu'il est NP-complet de déterminer le interval number des graphes planaires 2-dégénérés de degré maximum 5. Cela répond à une question de West et Shmoys datant de 1984. Ce résultat a été obtenu en choisissant judicieusement le problème à réduire et en trouvant des gadgets qui vérifient les contraintes de la classe de graphe. Nous avons obtenu [J37] que le très classique problème de la clique maximum est APX-complete pour les graphes 2-interval. Cette fois, le choix du problème à réduire (stable maximum) et la réduction sont très simples, et la difficulté réside dans ce lemme purement structurel : si on prend un graphe, que l'on subdivise 4 fois chaque arête et que l'on prend le complémentaire, on obtient un graphe 2-interval.

Le deuxième type est de montrer un seul résultat assez fort pour s'appliquer à plusieurs problèmes comparables. Pour deux colorations A et B telles que tout graphe A-coloriable est aussi B-coloriable, on considère uniquement les instances qui soient soit A-coloriables, soit non-B-coloriables. La question est bien sur de discriminer ces cas. Une seule preuve suffit donc à montrer qu'a la fois A, B, ainsi que toute coloration intermédiaire, est NP-complète. C'est d'autant plus intéressant que la différence entre A et B est importante. Dans [J44], A est la coloration des sommets en deux couleurs tels que les composantes connexes monochromatiques sont de taille au plus 2 et B est la partition des sommets en deux couleurs tels que les composantes connexes monochromatiques sont de taille au plus k, pour tout $k \geqslant 2$ fixé. Dans [J41], pour tout $d \geqslant 2$ fixé, A représente les graphes de track number au plus d et B représente les graphes de interval number au plus d. Dans ce dernier cas, il existe un paramètre, le *local track number*, qui est toujours compris entre le track number et le interval number. Notre preuve implique donc que décider si le local track number est au plus d est aussi un problème NP-complet.

Le troisième type est de montrer des résultats *spéculatifs*. Considérons une classe de graphe C et une propriété P tels que savoir si tous les graphes de C satisfont P est une question ouverte. On veut alors montrer que ou bien tous les graphes de C satisfont P, ou alors décider si un graphe de C satisfait P est NP-complet. Pour cela, on doit construire les gadgets de la réduction de NP-complétude à partir d'un (hypothétique) contre-exemple à la question ouverte. Si la question ouverte est un jour résolue, on est gagnant dans tous les cas : on a un résultat structurel si la réponse est positive et on a la NP-complétude si la réponse est négative. Nous avions notamment considéré la conjecture de Steinberg [J28], qui correspond au cas où C est la classe des graphes planaires sans cycles de taille 4 ou 5 et P est la propriété d'être 3-coloriable. Récemment, un contre-exemple a été construit, ce qui implique qu'il est NP-complet de décicer si un graphe planaire sans cycles de taille 4 ou 5 est 3-coloriable. Dans [J39], C est la classe des graphes planaires orientés de maille g et P est l'existence d'un homomorphisme vers un graphe cible T. Nous avons aussi considéré les problèmes de partition de sommets d'un graphe planaire (simple) de maille g en graphes de degré borné [J38]. Cette compléxité spéculative a d'autres vertues que son concept amusant :

- Elle écarte le cas où certains graphes de C ne satisfont pas P mais P est testable en temps polynômial sur C.
- Elle permet d'obtenir des preuves classiques de NP-complétude sans avoir à construire explicitement le gadget de "forçage". Notamment, elle s'applique même si l'on ne dispose que de preuves non-constructives de graphes dans C ne satisfaisant pas P.
- Elle motive encore plus la recherche de graphes ne satisfaisant pas P dans des petites classes de graphes.

Combinatoire des mots

J'ai beaucoup progressé en matière d'évitement de motifs, et je dispose maintenant de 3 méthodes bien abouties. Le choix de la méthode à utiliser dépend de la forme du motif et de la complexité en facteur des mots évitants.

Quelques définitions sont indispensables. La *formule* associée à un motif s'obtient en remplaçant par un point toutes les variables n'apparaissant qu'une seule fois. Ainsi, le motif $ABCADCABCEBCB$ est associé à la formule $ABCA.CABC.BCB$. Un mot contient une formule si elle contient la même occurrence de tous les fragments. Par exemple, $w = \texttt{00011001101}$ contient $ABCA.CABC.BCB$ car w contient l'occurrence $\{A \mapsto \texttt{0}, B \mapsto \texttt{0}, C \mapsto \texttt{11}\}$ des fragments $ABCA$, $CABC$, BCB, soit respectivement 00110, 110011, 0110. L'*indice d'évitabilité* d'une formule ou d'un motif est la taille du plus petit alphabet permettant d'écrire un mot infini qui évite la formule ou le motif. Ainsi, quand on parle de mots évitants, il s'agit seulement de mots sur ce plus petit alphabet. Un motif et sa formule associée ont le même indice d'évitabilité. C'est pourquoi on ne parle plus que de formules dans la suite. Enfin, l'*avoidability exponent* d'une formule est le plus grand α tel que tout mot α-free évite la formule.

Je décris maintenant les 3 méthodes pour construire des mots infinis évitants.

1. Pour une formule dont l'avoidability exponent α est strictement supérieur à 1 et telle que la complexité en facteur des mots évitants est exponentielle, on utilise une construction par morphisme uniforme. Cette méthode est utilisé dans [J42] et [J51]. Les mots dont on prend l'image morphique sont des mots de Dejean, qui utilise un alphabet plus gros et sont β-free pour un certain $\beta < \alpha$. L'avantage de cette méthode est qu'on prouve aussi la complexité en facteur exponentielle.

2. Pour une formule telle que la complexité en facteur des mots évitants est polynômiale, on décrit exactement tous les mots morphiques évitants. Cette méthode est utilisé dans [J51]. Pour l'instant, tous les mots morphiques rencontrés étaient image du point fixe de $\texttt{0} \mapsto \texttt{012}$, $\texttt{1} \mapsto \texttt{02}$, $\texttt{2} \mapsto \texttt{1}$, et il serait intéressant de trouver d'autres types de mots morphiques. Cette méthode permet ensuite d'affirmer que les mots morphiques trouvés sont essentiellement les seuls à éviter la formule. Ainsi, on prouve que la complexité en facteur est polynômiale. On avait aussi utilisé une version plus primitive de cette méthode pour des mots binaires évitant de grands carrés et un ensemble fini de facteurs [J40].

3. Pour une formule avec un seul fragment (on parle de motif *doubled*), on utilise la méthode non-constructive que j'ai développé. Elle provient d'autres méthodes non-constructives sur les mots de la littérature et ainsi que d'une idée venant de la récente méthode de "compression d'entropie", qui lui donne toute son efficacité. Elle permet d'obtenir simplement une borne inférieure sur le taux de croissance exponentielle de la complexité en facteur. A priori, son champs d'application aux formules est compris dans celui de la méthode 1, mais elle s'applique facilement aux formules ayant beaucoup de variables. Aussi, elle permet de traiter en une seule fois un grand nombre de formules, par exemple si elles ont un préfix commun et le même nombre de chaque variables. Cette méthode est utilisée dans [J42], ainsi que dans [J43] où on l'applique aux "shuffle squares", qui peuvent être vus comme un ensemble de motifs doubled.

Nombres parfaits impairs

Un nombre entier est parfait si il est égal à la somme de ses diviseurs propres. Par exemple, $6 = 1+2+3$ et $28 = 1+2+4+7+14$. Une conjecture ancienne et difficile dit qu'il n'existe pas de nombre parfait impair N. Avec Michaël Rao, nous avons perfectionné une méthode algorithmique pour obtenir des bornes sur un éventuel nombre parfait impair. On note $\Omega(N)$ le nombre de facteurs premiers de N et $\omega(N)$ le nombre de facteurs premiers distincts de N. Nous avons montré que $N \geqslant 10^{1500}$ [J25] qui améliore la précédente borne, i.e., 10^{300}. Nous avons aussi montré que $\Omega(N) \geqslant \max(101,\ (18\omega(N)-31)/7,\ 2\omega(N)+51)$ [J25, J31].

C Formation par la recherche

Encadrement

Stagiaires : J'ai co-encadré les stages suivants au niveau licence et maîtrise.

- Thomas Picchetti (L3 ENS Lyon, 2010) : Le Thue choice number des chemins (avec Francesca Fiorenzi).
- Amal Mejdoub (M2 Montpellier 2, 2013) : La sommet-arboricité dans les graphes planaires (avec Daniel Gonçalves).
- Marc de Visme (L3 ENS Ulm, 2014) : Graphes d'intersection de courbes et pair-crossing number (avec Daniel Gonçalves).
- Nathanaël Gross–Humbert (L3 ENS Cachan, 2017) : Évitement de motifs et de formules.
- Thibaud Gamard (L3 Orsay, 2018) : Inclusions de classes de graphes.

Guillaume Guégan : J'ai été l'encadrant scientifique des travaux de thèse de Guillaume Guégan (L'encadrant administratif étant Stéphane Bessy) à partir d'octobre 2012. Des problèmes de santé ont contraint Guillaume à renoncer à terminer son manuscrit. Nos travaux en commun ont abouti à deux publications. L'un traite de la complexité de problèmes d'homomorphisme de graphes orientés [J39] et l'autre est une preuve non-constructive de l'existence d'un mot infini sur 7 lettres évitant les "shuffle squares" [J43]. Guillaume a aussi collaboré avec d'autres chercheurs pour obtenir une preuve plus simple (et plus correcte !) que le interval number d'un graphe planaire est au plus 3 [28].

Autres doctorants : J'ai aussi co-écrit avec d'autres doctorants sans leurs encadrants.

- Élise Vaslet [J22, J26] : j'ai rencontré Élise à WORDS 2009 alors qu'elle était en début de sa thèse avec Julien Cassaigne à Marseille. L'article d'Élise et notre article (avec Francesca Fiorenzi) portaient sur le même sujet, les "generalized repetition threshold". Nous avons décidé d'associer nos résultats complémentaires dans un article commun pour la version journal [J22]. Au cours des mois suivants, nous nous sommes revus 3 fois pour étudier encore les seuils de répétition, sur des graphes cette fois, ce qui a abouti à notre deuxième article commun [J26].
- Hervé Hocquard et Petru Valicov [J29] : j'ai rencontré Hervé et Petru au cours de plusieurs séjours en 2009 et 2010 au LABRI à Bordeaux alors qu'ils étaient en thèse avec André Raspaud et Mickael Montassier. Après mon séminaire sur des réductions de NP-complétude, ils sont venus me parler de "strong edge coloring". On a obtenu que cette coloration est NP-complete pour 4, 5, et 6 couleurs sur des classes de graphes très restreintes [J29].
- Golnaz Badkobeh [J40] : j'ai rencontré Golnaz aux journées Montoises 2010. Elle était en thèse à Londres avec Maxime Crochemore. Son article portait sur différents types de mots binaires infinis contenant peu de carrés et de chevauchements distincts. Elle avait dit conjecturer que certaines contraintes de cette forme forçait essentiellement un seul mot binaire infini. Les preuves de ses conjectures ont été les résultats principaux de ma présentation invitée à WORDS 2011. Ensuite je suis allé la voir à Londres en novembre 2011, et nous avons trouvé

d'autres cas intéressants de mots évitants des grands carrés et un ensemble fini de facteurs. Ce travail [J40] a aussi permis de trouver des constructions plus simples pour 2 mots infinis classiques étudiés par Thue en 1912.

- Valentin Garnero [J41] : Valentin a été un doctorant chez AlGCo, encadré par Dimitrios Thillikos et Ignasi Sau. Valentin a assisté à une discussion entre un prof invité, Aquiles Braga De Queiroz, et moi sur le interval number. À trois, nous avons obtenu des résultats de NP-complétude et Valentin a contribué à trouver et simplifier un graphe 2-track qui n'est pas une union arête-disjointe de 2 graphes d'intervalles [J41].

- Nazanin Movarraei [J49] : j'ai rencontré Nazanin au meeting HOSIGRA 2013 alors qu'elle était thésarde en Inde et où elle était très peu encadrée. André Raspaud m'a demandé de lui proposer des sujets de recherches. On a travaillé sur des problèmes d'homomorphisme et j'ai présenté un sous-ensemble des résultats de [J49] à BGW2014. Après cette conf, Nazanin a passé un mois à Montpellier et nous avons fini les preuves de [J49].

- Matthieu Rosenfeld [J51, C5] : J'ai rencontré Matthieu aux journées Montoises 2014 alors qu'il était en début de thèse à l'ENS Lyon avec Michaël Rao. Au cours de plusieurs séjours qui ont suivis (lui à Montpellier ou moi à Lyon), nous nous sommes partagé la charge d'écriture de code pour étudier l'évitement de formules. Nous avons d'abord traité exhaustivement les formules à 2 variables et trouvé des comportements intéressants [J51]. Les formules à 3 variables sont trop nombreuses pour une étude exhaustive, mais nous avons encore trouvé d'autres comportements intéressants pour certaines formules à 3 variables [C5].

Enseignement

J'ai enseigné aux M2 Math-info à Montpellier

- 2011-2012 : 9 heures de Théorie des graphes - initiation aux méthodes de preuve par déchargement.

- 2014-2015 : 15 heures de Combinatoire des mots - Languages factoriels, taux de croissance exponentiel et indice d'évitabilité.

J'ai également co-encadré avec Stéphane Bessy le TER d'un groupe de 4 étudiants de L3 math-info, à propos de la conjecture d'Entringer.

Animation scientifique

- Octobre 2012 : Organisation des 12èmes JCALM (Journées Combinatoire et Algorithmes du Littoral Méditerranéen) au LIRMM.

- Mars 2015 : Deux interventions au Lycée français Jean Giono de Turin dans le cadre de la fête de la science.

- Depuis juillet 2014 : Co-responsable du séminaire du Pôle Algo-Calcul du LIRMM.

Tâches collectives

- Arbitrage d'articles pour 14 journaux, principalement Disc. Appl. Math. (19), Discrete Math. (13), Inform. Process. Lett. (6), Theor. Comput. Sci. (5), Graphs and Combinatorics (4).
- Arbitrage d'une vingtaine d'articles en conférence, membre du comité de programme d'IWOCA 2013.
- Membre extérieur du comité de sélection pour le poste 27-MCF-695 à Montpellier 2 en 2011 (Grégory Lafitte).
- Membre du conseil de laboratoire du LRI de septembre 2010 à juin 2011.

Bonus

- Titulaire de la PEDR 2014-2017.

D Liste de publications

Revues d'audience internationale avec comité de rédaction

[J56] P. Ochem and M. Rosenfeld. On some interesting ternary formulas. In *11th International Conference on Words (Words 2017)*, Montreal, Canada, September 11-15 2017. *Electron. J. Comb.* **26(1)** (2019), #P1.12.

[J55] S. Bessy, P. Ochem, and D. Rautenbach. On the Kőnig-Egerváry theorem for k-paths. *J. Graph Theory* **91(1)** (2019), 73–87.

[J54] B. Lužar, P. Ochem, and A. Pinlou. On repetition thresholds of caterpillars and trees of bounded degree. *Electron. J. Comb.* **25(1)** (2018), #P1.61.

[J53] G. Gamard, P. Ochem, G. Richomme, and P. Séébold. Avoidability of circular formulas. *Theor. Comput. Sci.* **726** (2018), 1–4.

[J52] C. Duffy, G. MacGillivray, P. Ochem, and A. Raspaud. Oriented incidence colourings of digraphs. *Discussiones Mathematicae Graph Theory* **39** (2019), 191–210.

[J51] P. Ochem and M. Rosenfeld. Avoidability of formulas with two variables. In *Developments in Language Theory*, Montreal, Canada, July 25-28 2016. *Electron. J. Comb.* **24(4)** (2017), #P4.30.

[J50] P. Ochem. 2-subcoloring is NP-complete for planar comparability graphs. *Inform. Process. Lett.* **128** (2017), 46–48.

[J49] N. Movarraei and P. Ochem. Oriented, 2-edge-colored, and 2-vertex-colored homomorphisms. *Inform. Process. Lett.* **123** (2017), 42–46.

[J48] M. Bougeret and P. Ochem. The complexity of partitioning into disjoint cliques and a triangle-free graph. *Disc. Appl. Math.* **217(3)** (2017), 438–445.

[J47] S. Bessy, P. Ochem, and D. Rautenbach. Bounds on the exponential domination number. *Discrete Math.* **340(3)** (2017), 494–503.

[J46] P. Ochem, A. Pinlou, and S. Sen. Homomorphisms of 2-edge-colored triangle-free planar graphs. *J. Graph Theory* **85(1)** (2017), 258–277.

[J45] S. Bessy, P. Ochem, and D. Rautenbach. Exponential domination in subcubic graphs. *Electron. J. Comb.* **23(4)** (2016), #P4.42.

[J44] L. Esperet and P. Ochem. Islands in graphs on surfaces. *SIAM Journal on Discrete Mathematics* **30(1)** (2016), 206–219.

[J43] G. Guégan and P. Ochem. A short proof that shuffle squares are 7-avoidable. *Theor. Informatics Appl.* **50(1)** (2016), 101–103.

[J42] P. Ochem. Doubled patterns are 3-avoidable. *Electron. J. Comb.* **23(1)** (2016), #P1.19.

[J41] A. Braga De Queiroz, V. Garnero, and P. Ochem. On interval representations of graphs. *Disc. Appl. Math.* **202** (2016), 30–36.

[J40] G. Badkobeh and P. Ochem. Characterization of some binary words with few squares. *Theor. Comput. Sci.* **588** (2015), 73–80.

[J39] G. Guégan and P. Ochem. Complexity dichotomy for oriented homomorphism of planar graphs with large girth. *Theor. Comput. Sci.* **22(1)** (2015), 142–148.

[J38] M. Montassier and P. Ochem. Near-colorings: non-colorable graphs and NP-completeness. *Electron. J. Comb.* **22(1)** (2015), #P1.57.

[J37] M. Francis, D. Gonçalves, and P. Ochem. The maximum clique problem in multiple interval graphs. In *WG 2012*, Jerusalem, Israel, June 26-27 2012. *Algorithmica* **71(4)** (2015), 812–836.

[J36] P. Ochem and A. Pinlou. Application of entropy compression in pattern avoidance. *Electron. J. Comb.* **21(2)** (2014), #RP2.7.

[J35] P. Ochem and M. Rao. Another remark on the radical of an odd perfect number. *The Fibonacci Quarterly* **52(3)** (2014), 215–217.

[J34] R. Mercas, P. Ochem, A. Samsonov, and A.M. Shur. Binary patterns in binary cube-free words: avoidability and growth. *Theor. Informatics Appl.* **48(4)** (2014), 369–389.

[J33] P. Ochem. More on square-free words obtained from prefixes by permutations. *Fundamenta Informaticae* **132** (2014), 1–4.

[J32] P. Dorbec, M. Montassier, and P. Ochem. Vertex-partitions of graphs into cographs and stars. *J. Graph Theory* **75(1)** (2014), 75–90.

[J31] P. Ochem and M. Rao. On the number of prime factors of an odd perfect number. *Math. Comp.* **83** (2014), 2435–2439.

[J30] P. Ochem and A. Pinlou. Oriented coloring of triangle-free planar graphs and 2-outerplanar graphs. In *LAGOS 2011*, Bariloche, Argentina, March 28 - April 1st 2011. *Graphs and Combinatorics* **30(2)** (2014), 439–453.

[J29] H. Hocquard, P. Ochem, and P. Valicov. Strong edge-colouring and induced matchings. *Inform. Process. Lett.* **113(19-21)** (2013), 836–843.

[J28] L. Esperet, M. Montassier, P. Ochem, and A. Pinlou. A complexity dichotomy for the coloring of sparse graphs. *J. Graph Theory* **73(1)** (2013), 85–102.

[J27] L. Esperet, S. Gravier, M. Montassier, P. Ochem, and A. Parreau. Locally identifying coloring of graphs. *Electron. J. Comb.* **19(2)** (2012), #P40.

[J26] P. Ochem and É. Vaslet. Repetition thresholds for subdivided graphs and trees. In *Mons Days of Theoretical Computer Science*, Amiens, France, September 6-10 2010. *Theoret. Informatics Appl.* **46(1)** (2012), 123–130.

[J25] P. Ochem and M. Rao. Odd perfect numbers are greater than 10^{1500}. *Math. Comp.* **81** (2012), 1869–1877.

[J24] S.A. Fletcher, P.P. Nielsen, and P. Ochem. Sieve methods for odd perfect numbers. *Math. Comp.* **81** (2012), 1753–1776.

[J23] F. Fiorenzi, P. Ochem, P. Ossona de Mendez, and X. Zhu. Thue choosability of trees. *Disc. App. Math.* **159(17)** (2011), 2045–2049.

[J22] F. Fiorenzi, P. Ochem, and É. Vaslet. Bounds for the generalized repetition threshold. In *7th International Conference on Words (Words 2009)*, Salerno, Italy, September 14-18 2009. *Theor. Comput. Sci.* **412** (2011), 2955–2963.

[J21] A. Montejano, P. Ochem, A. Pinlou, A. Raspaud, É. Sopena. Homomorphisms of 2-edge-colored graphs. *Disc. App. Math.* **158(12)** (2010), 1365–1379.

[J20] J. Chalopin, D. Gonçalves, and P. Ochem. Planar graphs have 1-string representations. In *Proceedings of SODA 2007* 609–617. *Discrete and Computational Geometry* **43(3)** (2010), 626–647.

[J19] O.V. Borodin, A.O. Ivanova, M. Montassier, P. Ochem and A. Raspaud. Vertex decompositions of sparse graphs into an edgeless subgraph and a subgraph of maximum degree at most k. *J. Graph Theory* **65(2)** (2010), 83–93.

[J18] P. Ochem. Binary words avoiding the pattern AABBCABBA. *Theoret. Informatics Appl.* **44(1)** (2010), 151–158.

[J17] R. Kolpakov, G. Kucherov, and P. Ochem. On maximal repetitions of arbitrary exponent. *Inform. Process. Lett.* **110(7)** (2010), 252–256.

[J16] D. Gonçalves and P. Ochem. On star and caterpillar arboricity. *Discrete Math.* **309(11)** (2009), 3694–3702.

[J15] L. Esperet and P. Ochem. On circle graphs of girth at least five. In *EuroComb 2007*, Seville, September 11-15 2007, *ENDM* **29** (2007), 129–133. *Discrete Math.* **309(8)** (2009), 2217–2222.

[J14] J. Chalopin and P. Ochem. Dejean's conjecture and letter frequency. In *Mons Days of Theoretical Computer Science*, Rennes, August 30 - September 2 2006. *Theoret. Informatics Appl.* **42(3)** (2008), 477–480.

[J13] M. Montassier, P. Ochem, and A. Pinlou. Strong oriented chromatic number of planar graphs without short cycles. *DMTCS* **10(1)** (2008), 1–24.

[J12] L. Esperet, A. Labourel, and P. Ochem. On induced-universal graphs for the class of bounded-degree graphs. *Inform. Process. Lett.* **108(5)** (2008), 255–260.

[J11] P. Ochem and A. Pinlou. Oriented colorings of partial 2-trees. In *EuroComb 2007*, Seville, September 11-15 2007, *ENDM* **29** (2007), 195–199. *Inform. Process. Lett.* **108(2)** (2008), 82–86.

[J10] P. Ochem, A. Pinlou, and É. Sopena. On the oriented chromatic index of oriented graphs. *J. Graph Theory* **57(4)** (2008), 313–332.

[J9] P. Ochem, N. Rampersad, and J. Shallit. Avoiding approximate squares. In *Developments in language theory (DLT 2007)*, Turku, Finland, July 3-6 2007. *IJFCS* **19(3)** (2008), 633–648.

[J8] L. Esperet and P. Ochem. Oriented colorings of 2-outerplanar graphs. *Inform. Process. Lett.* **101(5)** (2007), 215–219.

[J7] P. Ochem. Letter frequency in infinite repetition-free words. In *5th International Conference on Words (Words 2005)*, Montreal, Canada, September 13-17 2005. *Theor. Comput. Sci.* **380** (2007), 388–392.

[J6] M. Montassier, P. Ochem, and A. Raspaud. On the acyclic choosability of graphs. In *GT04 Graph Theory 2004: a conference in memory of Claude Berge*, Paris, France, July 5-9 2004. *J. Graph Theory* **51(4)** (2006), 281–300.

[J5] P. Ochem. A generator of morphisms for infinite words. In *Proceedings of the Workshop on Word Avoidability, Complexity, and Morphisms*, Turku, Finland, July 17 2004. LaRIA Technical Report 2004-07, 9–14. *Theoret. Informatics and Appl.* **40** (2006), 427–441.

[J4] L. Ilie, P. Ochem, and J.O. Shallit. A generalization of Repetition Threshold. In *Proceedings of MFCS 2004: 29th International Symposium on Mathematical Foundations of Computer Science*, Prague, Czech Republic, August 22-27 2004, Lecture Notes in Computer Science, Springer, Vol. 3153, 818-826. *Theor. Comput. Sci.* **345** (2005), 359–369.

[J3] P. Ochem. Oriented colorings of triangle-free planar graphs. *Inform. Process. Lett.* **92(2)** (2004), 71–76.

[J2] J. Balogh, P. Ochem, and A. Pluhàr. On the interval number of special graphs. *J. Graph Theory* **46(4)** (2004) 241–253.

[J1] G. Kucherov, P. Ochem, and M. Rao. How many square occurrences must a binary sequence contain? *Electron. J. Comb.* **10(1)** (2003), #R12.

Actes de conférence d'audience internationale avec comité de sélection

[C4] P. Ochem and M. Rao. Minimum frequencies of occurrences of squares and letters in infinite words. In *Mons Days of Theoretical Computer Science (JM 2008)*, Mons, Belgium, August 27-30 2008.

[C3] P. Ochem. Unequal letter frequencies in ternary square-free words. In *6th International Conference on Words (Words 2007)*, Marseille, France, September 17-21 2007.

[C2] P. Ochem and T. Reix. Upper bound on the number of ternary square-free words. In *Workshop on Words and Automata (WOWA'06)*, St Petersburg, Russia, June 7 2006.

[C1] P. Ochem. Negative results on acyclic improper colorings. In *Proceedings of the 2005 European Conference on Combinatorics, Graph Theory and Applications* (EuroComb '05), Berlin, Germany, September 5-9 2005, *DMTCS Proceedings* 357–362.

Conférences invités

[I2] P. Ochem. Pattern avoidance. *Combinatorics and algorithmics on words*, special session of *Computability in Europe (CiE 2017)*, Turku, Finland, June 12-16 2017.

[I1] P. Ochem. Pattern avoidance and HDOL words. In *8th International Conference on Words (WORDS 2011)*, Prague, Czech republic, September 12-16 2011.

Chapitre de livre

[B1] P. Ochem, M. Rao, and M. Rosenfeld. Avoiding or limiting regularities in words. In Valérie Berthé and Michel Rigo, editors, *Sequences, Groups and Number Theory.* Chapter 5. Trends in Mathematics. Birkhuser, Cham.

E Projet de recherche

Je me suis intéressé à une large variété de domaines et de problèmes : homorphismes, mots sans carrés, complexité, motifs, partitions d'arêtes, nombres parfaits impairs, mots sans carrés, coloration non-répétitive... Bien sur, je vais continuer à écrire et co-écrire avec des collègues de tous horizons sur certains de ces sujets et sans doute de nouveaux. Bien sur, des problèmes ouverts intéressants, importants, et difficiles ont attirés mon attention et attendent toujours une solution. Pour qu'ils puissent faire partie d'un projet de recherche raisonnable, ils doivent aussi admettre un ou plusieurs angles d'attaque crédibles. Cela n'est pas le cas pour, par exemple, la conjecture de Cassaigne, la 3-coloration acyclique des graphes planaires de maille 5, l'amélioration des bornes sur le nombre chromatique des graphes planaires, la conjecture d'Hadwiger.

Dans les prochaines années, je vais entreprendre deux tâches ambitieuses présentées dans les deux prochaines sections. En plus de satisfaire aux critères mentionnés d'intéret, d'importance, de difficulté et d'abordabilité, elles ont le bon gout de faire partie chacune d'un projet ANR démmaré récemment. Cela me donnera les moyens de rencontrer mes collaborateurs.

ANR HOSIGRA : PLANAR H-COLORING

Le projet ANR HOSIGRA (HOmomorphisms of SIgned GRAphs), Porté par Reza Naserasr, commence en janvier 2018.

Dans le cadre du projet ANR HOSIGRA (Porté par Reza Naserasr, janvier 2018), je vais surtout m'intéresser à de l'homomorphisme de graphe. Le H-COLORING est le problème de décider si un graphe G admet un homomorphisme vers un graphe cible fixe H. Hell et Nešetřil ont montré que si H est un graphe simple, alors H-COLORING est polynômial si H est biparti et NP-complet sinon. Ensuite de nombreux travaux ont porté sur la généralisation de ce résultat au cas où H est un graphe dirigé et ont abouti à la récente preuve de la fameuse "CSP dichotomy conjecture" de Feder et Vardi. En restant dans le cadre des graphes simples, je veux maintenant étudier le cas où G est contraint, et la contrainte la plus naturelle et la plus intéressante est que G soit planaire. Le but ultime est la classification des graphes simples H selon la complexité du problème PLANAR H-COLORING.

Le résultat lui-même n'est pas une simple dichotomie entre biparti et non-biparti. En effet, en plus de K_4, il existe une infinité de graphes (incomparables selon l'ordre "homomorphique") tels que le PLANAR H-COLORING revient à tester la présence d'un nombre fini de sous-graphes, et donc pour lesquels PLANAR H-COLORING est polynômial.

Cependant, la majeure partie du travail consiste toujours en de nombreuses réductions de NP-complétude. Or, la planarité engendre deux types de difficultés par rapport à la preuve de Hell et Nešetřil. Premièrement, certaines opérations très utilisées ne conserve pas la planarité, comme par exemple connecter tous les sommets d'un graphe à une unique copie d'un gadget. Deuxièmement, il faut modifier l'ordre sur les graphes cibles utilisé par Hell et Nešetřil. En effet, ils considérent qu'un graphe cible est plus petit qu'un autre s'il a autant de sommets et plus d'arêtes. Or, faire des réductions en rajoutant des arêtes aux graphes cibles n'est pas très adapté à nos graphes cibles qui sont K_4-free.

Pour contourner la première difficulté, la stratégie va dépendre des symétries de H. Si H est peu symétrique, on peut faire une réduction assez standard de H vers un graphe cible H' à moins de sommets en ajoutant un gadget par sommet de G. Si H est très symétrique, où même sommettransitif comme l'icosahèdre, il faut faire une réduction ad-hoc. Cela implique de considérer les graphes cibles isolément, ce qui évite aussi la deuxième difficulté. On est donc très loin de l'efficacité de la preuve de Hell et Nešetřil qui traite le cas de tous les graphes cibles sans triangle en une seule réduction.

L'enjeu est donc de généraliser à de larges familles de graphes cibles certaines des réductions que j'ai déjà obtenues pour des graphes cibles isolés.

ANR COCOGRO : motifs d'indice au moins 6 et formules nice

L'évitement de motifs est un domaine vivant. La partie scientifique de cette HDR présente de nombreux phénomèmes intéressants que nous avons découvert récemment, comme par exemple des nouvelles formules polynomiales, dont l'ensemble des mots évitants peut ou ne peut pas être décrit par un nombre fini de mots morphiques. Aussi le développement de techniques puissantes pour borner l'indice d'évitabilité des formules a considérablement accru notre connaissance de ces indices.

Je suis maintenant prêt à attaquer une des grandes questions de l'évitement de motifs : existe-t'il des motifs dont l'indice d'évitabilité est 6 ? Ou même, existe-t'il des motifs dont l'indice d'évitabilité est fini mais arbitrairement grand ? Dans le cadre du projet ANR COCOGRO (Porté par Nathalie Aubrun, janvier 2017), je vais tenter de répondre positivement à cette deuxième question en identifiant des familles infinies de formules évitables qui sont "extrémales au sens de la divisibilité".

Un autre axe de recherche est de prouver la conjecture 17, c'est-à-dire montrer que toutes les formules *nice* sont 3-évitables, ou au moins obtenir une borne absolue sur leur indice d'évitabilité. Une formule est nice si pour toute variable V, il existe un fragment contenant au moins deux occurrences de V. Les formules nice généralisent donc les motifs doubled qui sont 3-évitables [J42]. Elles partagent aussi avec les motifs doubled la propriété sympatique que pour toute formule nice, il existe un mot de Dejean évitant sur un alphabet dont la taille dépend seulement du nombre de variables de la formule. Cependant, la méthode non-constructive que j'utilise pour les motifs doubled ne s'applique pas aux autres formules nice. Il faut donc développer un nouvel outil pour ces formules.

Pour une formule nice donnée, la méthode de la Section 4 permet généralement de montrer la 3-évitabilité. La difficulté est de traiter des familles infinies de formules nice en une fois, comme on l'a fait pour les formules circulaires [22] (voir Section 6.4) et les formules T_i mentionnées à la Section 4 comme exemples de formules nice d'indice exactement 3.

Une première étape est de montrer la 3-évitabilité pour d'autres familles de formules nice, notamment les formules two-birds $ABA.BAB$, $ABCBA.CBABC$, $ABCDCBA.DCBABCD$, ... qu'on a déjà évoquées. Ensuite, on essayera d'identifier les ingrédients intéressants dans les preuves d'évitabilité de ces "petites" familles de formules pour montrer l'évitabilité de familles de formules nice de plus en plus générales.

Plus tard : facteur de croissance des motifs

La conjecture de Cassaigne 5 implique que tout motif ou formule évitable est évité par un mot morphique. Sans perte de généralité, c'est un mot morphique uniformément récurrent et non-périodique. Par les théorèmes 12 et 17 de [1], la complexité en facteur de ce mot est linéaire.

On peut donc définir le facteur de croissance $g(P)$ du motif P comme le plus petit α tel qu'il existe un mot infini w et une constante β tel que la complexité de w (i.e. le nombre $c_w(n)$ de facteurs de longueur n) est au plus $\alpha n+\beta$ pour tout $n \geqslant 0$. Remarquons qu'il n'y a pas de contraintes sur la taille de l'alphabet de w. D'ailleurs, cette nouvelle notion est le pendant asymptotique de l'indice d'évitabilité : l'indice d'évitabilité minimise $c_w(1)$ alors que le facteur de croissance minimise $c_w(n)$ pour n grand. On conserve ainsi les arguments de divisibilité des motifs, et notamment que si $P_1 \preceq P_2$, alors $g(P_1) \geqslant g(P_2)$. L'exemple de la formule $F = AABA.ABB.BBA$ montre que ces deux objectifs de minimisation peuvent être incompatibles : Le seul mot binaire évitant F est $g_x(b_3)$ [J51], ce qui montre que $g(F) \leqslant \frac{10}{3}$, alors que le point fixe (ternaire) du morphisme $\mathtt{0} \rightarrow \mathtt{01}$, $\mathtt{1} \rightarrow \mathtt{0201}$, $\mathtt{2} \rightarrow \mathtt{21}$, qui évite F et même $AAB.ABA.ABB$, montre que $g(F) \leqslant 3$.

Le facteur de croissance ouvre un champs d'investigation immense. La première remarque triviale est $g(P) \geqslant 1$ puisque tout mot apériodique w vérifie $c_w(n) \geqslant n+1$ pour tout n. Cette borne est optimale pour $AAAA$. En effet, $g(AAAA) = 1$ car le mot de Fibonacci (i.e. le point fixe de $\mathtt{0} \rightarrow \mathtt{01}$, $\mathtt{1} \rightarrow \mathtt{0}$), est de complexité $n+1$ et évite $AAAA$. Pour le motif AA, on sait seulement que $g(AA) \leqslant \frac{10}{3}$ car b_3 est sans carré. Une première tâche est donc de développer des techniques pour les bornes inférieures afin de prouver que $g(AA) = \frac{10}{3}$. Un des outils prometteurs pour cela est l'analyse des graphes de Rauzy des mots évitants, qui permettrait de montrer que $g(P) \geqslant 2$ sous certaines conditions.

La conjecture de Cassaigne suggère que les mots morphiques sont les objets centraux de l'étude de l'indice d'évitabilité. Pour le facteur de croissance, il semble que les objets centraux soient les mots purement morphiques, i.e., les points fixes de morphismes. En effet, il semble si un mot évitant purement morphique minimise $c(n)$, alors lui appliquer un morphisme n'abaisse la complexité que d'une constante additive. Il est important que j'essaye de vérifier ou d'infirmer cette "hypothèse des points fixes" chaque fois que l'occasion se présente, car elle a de nombreuses conséquences. Par exemple, elle implique que $g(AA) = g(ABAB) = g(SQ_t)$, où SQ_t désigne les carrés de période au moins t. Aussi, elle implique que $g(ABABA) = g(ABABACDCDC)$ ou encore $g(ABABA) = g(AABAABAA)$.

Ces équivalences entre motifs ont deux conséquences intéressantes. D'abord elles diminuent l'effet "catalogue" qu'on peut avoir dans [J5,J51] quand on étudie exhaustivement toute une famille de motifs. Et surtout, elles facilitent la recherche de nouveaux points fixes utiles en suggérant de regarder des motifs non-équivalents aux motifs déja considérés.

Pour résumer, l'étude du facteur de croissance des motifs promet d'approfondir nos connaissances sur les points fixes de morphismes. Cela est complémentaire de l'étude de l'indice d'évitabilité qui met l'accent sur le morphisme externe des mots morphiques. Ce domaine, par son coté asymptotique, est aussi beaucoup plus proche des systèmes dynamiques. Mais surtout, et c'est ce qui me motive, il faudra comprendre le phénomème suivant : Le domaine connu de l'indice d'évitabilité des motifs évitable, $\{2, 3, 4, 5\}$, est restreint aux petits entiers alors que le domaine du facteur de croissance ne l'est pas. On s'attend donc à ce que le facteur de croissance soit une mesure plus discriminante de l'évitabilité. Or, ce n'est pas du tout flagrant dans les premières expérimentations que j'ai menées.