ssdnm
środowiskowe
studia doktoranckie
z nauk matematycznych

Marcin Witkowski

Uniwersytet A. Mickiewicza w Poznaniu

Nonrepetitive sequences on arithmetic progressions

Praca semestralna nr 2

(semestr zimowy 2010/11)

Opiekun pracy: Jarosław Grytczuk

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

NONREPETITIVE SEQUENCES ON ARITHMETIC PROGRESSIONS

JAROSŁAW GRYTCZUK, JAKUB KOZIK, AND MARCIN WITKOWSKI

ABSTRACT. A sequence $S = s_1 s_2 \dots s_n$ is *nonrepetitive* if no two adjacent blocks of S are identical. In 1906 Thue proved that there exist arbitrarily long nonrepetitive sequences over 3-element set of symbols. We study a generalization of nonrepetitive sequences involving arithmetic progressions. We prove that for every $k \geqslant 1$ there exist arbitrarily long sequences over at most $2k + 10\sqrt{k}$ symbols whose subsequences determind by arithmetic progressions with common differences from the set $\{1, 2, \dots, k\}$ are nonrepetitive. This improves a previous bound obtained in [8]. Our approach is based on a technique introduced recently in [11] which was originally inspired by algorithmic proof of the Lovász Local Lemma due to Moser and Tardos [13]. We also discuss some related problems that can be succesfully attacked by this method.

1. INTRODUCTION

When for a sequence $S = s_1 s_2 \dots s_n$ there exists a nonempty block of symbols that repeats next to itself, i.e. $XX = x_1 \dots x_h x_1 \dots x_h$, then we call such a block a *repetition* of size h. A sequence is *nonrepetitive* if it does not contain a repetition of any size $h \geqslant 1$. For example, the sequence 1231312 contains a repetition 3131 of size two, while 123132123 is nonrepetitive.

It is easy to see that if we can use only two symbols, then the longest nonrepetitive sequence has length three. However in 1906 Thue [14] proved, by a remarkable inductive construction, that there exist arbitrarily long nonrepetitive sequences over just three different symbols (see also [5], [4]). This discovery found many unexpected applications in diverse areas, inspiring a stream of research and leading to the emergence of new branches of mathematics with a variety of challenging open problems (see [1], [3], [7], [10], [12]).

One particular variant, proposed in [3], asks for *nonrepetitive tilings*, i.e., assignments of symbols to lattice points of the plane so that all lines in prescribed directions are nonrepetitive. This idea led Currie and Simpson [6] to consider sequences with stronger property that all subsequences taken over arithmetic progressions of bounded common differences are nonrepetitive. Let $k \geqslant 1$ be a fixed positive integer and let $S(k)$ be the family of subsequences of S of the form $s_i s_{i+d} s_{i+2d} \dots s_{i+td}$ with $d \in \{1, 2, \dots, k\}$, $1 \leqslant i \leqslant k-1$, $t = \lfloor n/d \rfloor$. If every element of $S(k)$ is a nonrepetitive sequence, then S is called *nonrepetitive up to* mod k (see [6]). Let $M(k)$

Key words and phrases. nonrepetitive sequence, randomized algorithm, the Lovász Local Lemma.

This paper was written as a semester paper under the supervision of dr hab. Jarosław Grytczuk in the framework of joint PhD programme SSDNM.

denote the minimal number of symbols needed to create arbitrarily long sequences nonrepetitive up to mod k. So, Thue's theorem says simply that $M(1) = 3$. It is easy to see that $M(k) \geqslant k+2$ for every $k \geqslant 1$, and one may suspect that equality always holds.

Conjecture 1. *$M(k) = k + 2$ for every $k \geqslant 1$.*

This conjecture has been confirmed so far only for $k = 2$, 3, and 5 (it is not known for $k = 4$, curiously) by providing Thue type constructions of the desired sequences. However, using the Lovász Local Lemma (see [2]) it was proved in [8] that $M(k) \leqslant e^{33}k$ holds for any k. In this paper we improve the last bound substantially by proving that $M(k) \leqslant 2k+o(k)$. Our method is inspired by the recent constructive proof of the the Lovász Local Lemma due to Moser and Tardos [13]. As in the related paper [11] (as well as in the original nonconstructive approach), a stronger statement is obtained where symbols are chosen from prescribed lists assigned to the positions in a sequence. Also restriction to $\{1, 2, \ldots, k\}$ as the set of common differences is not crucial; we get the same result for any k-element set of positive integers.

2. The algorithm

We present an algorithm that generates consecutive terms of a sequence S by choosing symbols at random (uniformly and independently), and every time a repetition occurs on any element of $S(k)$, it erases the repeated block and continues from the smallest non-occupied position. We always erase this block which contains the last chosen element in order to ensure that after this removal the remaining sequence stays nonrepetitive.

We show that for a given lists of symbols L_i, each of size at least $2k+10\sqrt{k}$ and any given number n, Algorithm 1 computes a sequence of length n which is nonrepetitive up to mod k. As mentioned before random elements in Algorithm 1 are chosen independently with uniform distribution. The general idea is to prove that Algorithm 1 cannot work long enough for all possible evaluations of the random experiments. Otherwise we would get that we can compress a random string better than it is possible.

Theorem 1. *For every $n \geqslant 1$ and for every sequence of sets $L_1, \ldots, L_n$, each of size at least $2k+10\sqrt{k}$, there is a sequence $S = s_1 \ldots s_n$ nonrepetitive up to* mod *k such that $s_i \in L_i$ for every $i = 1, 2, \ldots, n$.*

Proof. Our goal is to show that there exists a value M such that after M iterations there would exist an evaluation of $r_1, \ldots, r_M$ for which Algorithm 1 produces a sequence of length n which is nonrepetitive up to mod k.

We are going to encode efficiently a random sequence of bits used in line (3) of the algorithm. As one can notice in step (3) we forbid to use as for s_i any of the symbols from k previous and k subsequent positions. This implies that we need at least $2k$ symbols on each list L_i for our algorithm to work. What we want to show is that additional $10\sqrt{k}$ symbols on each list suffice for the algorithm to obtain a proper sequence.

Let $r_j (1 \leqslant j \leqslant M)$ be a sequence of random variables, each of them taking values from 1 up to $2k + 10\sqrt{k}$. We pick elements from the lists with uniform distribution. Since there are at most $2k$ forbidden elements when

Algorithm 1 Choosing a sequence which is non-repetitive up to mod k

1: $i \leftarrow 1$
2: **while** $i \leq n$ **do**
3: $s_i \leftarrow$ random element of $L_i \backslash \{s_{i-k}, s_{i-k+1}, ..., s_{i+k-1}, s_{i+k}\}$
4: **if** $s_1, ..., s_n$ is non-repetitive in respect to non-zero elements **then**
5: $i \leftarrow$ smallest index j for which $s_j = 0$
6: **else**
7: from the set of the longest repetitions in S choose a

$$s_{j-2h\cdot d+d}, ..., s_{j-h\cdot d}, s_{j-h\cdot d+d}, ..., s_j$$

with the largest index of the first element $j - 2h \cdot d + d$
8: **if** $i \leq j - hd$ **then**
9: $c \leftarrow j - 2h \cdot d + d$
10: **else**
11: $c \leftarrow j - h \cdot d + d$
12: **end if**
13: **for** $j = 1$ **to** d **do**
14: $s_c \leftarrow 0$
15: $c \leftarrow c + d$
16: **end for**
17: $i \leftarrow$ smallest index j for which $s_j = 0$
18: **end if**
19: **end while**

the algorithm evaluates step (3), then there are at least $(10\sqrt{k})^M$ possible evaluations for the sequence (r_j). We endow each fixed evaluation of a random sequence with some additional structure. Consider the following five elements:

- Route R on the upper right quadrant of a grid from coordinate $(0, 0)$ to coordinate $(2M, 0)$ on $2M$ steps with possible moves $(1, 1)$ and $(1, -1)$ which never goes below the axis $y = 0$.
- A sequence D of numbers between 1 and k corresponding to the peaks on route R. Where by peak we mean a move $(1, 1)$ followed immediately by the move $(1, -1)$.
- A sequence O of numbers -1 or 1 corresponding to the peaks on route R.
- A sequence P of integer numbers, one for every peak, which sum is equal at most M.
- A sequence S produced by an Algorithm 1 at the end of computations.

A pentad (R, D, O, P, S) will be called a *log*. The way we encode steps of Algorithm 1 into log is the following:

Each time the algorithm executes line (3) we put a move $(1, 1)$ on route R and for every execution of line (10) or (14) we put a move $(1, -1)$ on route R. Notice that in lines (10) and (14) the algorithm can set zero only to s_c which are non-zero, therefore the number of down-steps on route R never excess the number of up-steps, and it never goes below axis $y = 0$. At the end of computations we add to the route R a one down-step for each element of S

which is non-zero. This brings us to the point $(2M, 0)$. Whenever Algorithm 1 executes line (7) we add into the sequence D a difference d of a chosen longest repetition. Then if (8) is true then we add 1 to the sequence O, else we add -1. For the sequence P we add a j for which c equals i in the loop between lines (12) and (15). S is just a sequence of s_i produced by Algorithm 1 after M executions of line (3).

Claim 1. *Every log corresponds to unique evaluation of* $r_1, \ldots, r_M$.

Proof. With given log (R, D, O, P, S) we are going to decode $r_1, \ldots, r_M$. At first we use information from route R and sequences D and P to determine which s_i were non-zero at each step of the algorithm and to find coordinations of elements which were zeroed at (10) and (14) steps of Algorithm 1. Notice that each operation of setting a non-zero value to some s_i corresponds to the up-step $(1, 1)$ on the route R and each zeroing of s_i corresponds to some down-step $(1, -1)$ on route R. We examine route R in order from the point $(0, 0)$ to the point $(2M, 0)$. Assume the first peak occurs after jth step. Since it is first time we erase some elements s_i we knew that $s_1, \ldots, s_j$ are the only non-zero elements at this point. Now we use information encoded in D and P. We look at the number of consecutive down-steps on R (which in this case is equal to p_1) and remember that for this peak we zeroed s_j, $s_j - d_1$, $s_j - 2d_1, \ldots, s_j - p_1 d_1$. Then again each up-step on R denote setting some non-zero value to the zeroed position with the smallest index i. Remember this index i for each up-step on R. For the second and all forthcoming peaks the analysis will be the same. Assume we are considering the lth peak. Remember the index of the last set element of S and the number of consecutive down-steps on R after this peak (denote the first value by i and the second one by t). We remember that zeroed values were those with indices $i - p_l d_l, i - (p_l - 1)d_l, \ldots, i, i + d_l, \ldots, i + (t - p_l)d_l$. We repeat those operations until we get to the end of R.

After this preparatory step we move on to decode $r_1, \ldots, r_M$. We proceed by considering elements of R, but now in reverse order, from the last point $(2M, 0)$ to the first $(0, 0)$. This time we use information encoded in S and O and the knowledge about S determined in a preparatory step. As we said before, each up-step $(1, 1)$ on the route R corresponds to some r_i. In the preparatory analysis we already determined the index of elements r_i on S for such steps. So going backward on R there is some number of down-steps corresponding to non-zero elements of S. We skip them and move on. Then each time there is an up-step on R we assign to r_j value from appropriate s_i (i was determined in the preparatory step) and set s_i to 0. We must then remember about excluded forbidden symbols from k preceding and k following places on S. Let z be the number of s_j, $(k - i \leqslant j \leqslant k + i)$ such that $s_j < s_i$. Then $r_i = s_i - z$. On the other hand, if there is a consecutive sequence of t down-steps on R we assign to $s_{i-p_l d_l}, s_{i-(p_l-1)d_l}, \ldots, s_i, s_{i+d_l}, \ldots, s_{i+(t-p_l)d_l}$ corresponding values $s_{i-p_l d_l+o_l kt}, s_{i-(p_l-1)d_l+o_l kt}, \ldots, s_{i+o_l kt}, s_{i+d_l+o_l kt}, \ldots, s_{i+(t-p_l)d_l+o_l kt}$. This are the values from the repetitions erased at (10) or (14) step of Algorithm 1. □

We showed that each sequence of random symbols corresponds to some log, and that this mapping is injective. This implies that the number of

different logs is always greater than or equal to the number of feasible evaluations of $r_1, \ldots, r_M$. Let L be the size of the set of all possible logs. To calculate L we have to determine the number of different structures for each element in a log. The number of all possible routes on the upper right quadrant of a grid of length $2M$ with possible moves $(1,1)$ and $(1,-1)$ is the well known figure and equals the Mth Catalan number. Notice that since a route cannot go below axis $y = 0$ and each peak is followed by at least one down move $(1,-1)$, the number of peaks on such a route cannot exceed $M/2$. This implies that there are at most $k^{M/2}$ possible evaluations for a sequence D and $2^{M/2}$ for the sequence O. The sequence S consists of n elements of value between 0 and $2k + 10\sqrt{k}$ which gives us $(2k + 10\sqrt{k} + 1)^n$ possible evaluations for this sequence. For the sequence $P = (p_1, \ldots, p_n)$ we have to determine maximum value of the product $p_1 p_2 \ldots p_n$ with $p_1 + \ldots + p_n = M$. The inequality between the arithmetic and geometric means implies that the maximum is obtained when all p_i are the same. Denote their common value by x. Then we must determine $\max\left(x^{\frac{M}{x}}\right)$. Since

$$\left(x^{\frac{M}{x}}\right)' = x^{\frac{M}{x}}\left(\frac{M}{x^2} - \frac{M\log(x)}{x^2}\right),$$

we get that the maximum value is obtained with $x = e$ and equals $\approx 1.44467^M < 1.5^M$.
Finally that brings us to the conclusion that the number of possible logs equals

$$(2k + 10\sqrt{k} + 1)^n \frac{1}{M+1}\binom{2M}{M} k^{M/2} 2^{M/2} (1.5)^M.$$

Comparing with the number of evaluations of a sequence (r_j) we get inequality

$$(10\sqrt{k})^M \leqslant (2k + 10\sqrt{k} + 1)^n \frac{1}{M+1}\binom{2M}{M} k^{M/2} 2^{M/2} (1.5)^M.$$

Asymptotically, Catalan numbers grow as $C_n \approx \frac{4^n}{n^{3/2}\sqrt{\pi}}$ which implies

$$(10\sqrt{k})^M \leqslant \frac{(10\sqrt{k})^M}{M\sqrt{\pi M}}(2k + 10\sqrt{k} + 1)^n.$$

Therefore there exists an evaluation of $r_1, \ldots, r_m$ such that Algorithm 1 finds a sequence which is nonrepetitive up to mod k in at most M steps, with $M\sqrt{\pi M} \geqslant (2k + 10\sqrt{k} + 1)^n$ (otherwise we could compress a sequence of random bits). □

We can notice that the above proof gives a stronger result.

Remark 1. *Theorem 1 holds for any set K of size k for which we want to avoid repetitions in subsequences mod p, with $p \in K$.*

Proof. Notice that in the proof of the Theorem 1 we concentrate only on the number of forbidden substructures, not their values. Given an arbitrary set of forbidden differences of size k we can just order and numerate them from 1 up to k and repeat the previous proof avoiding repetitions in arithmetic progressions with differences from this set. □

3. A related problem

As was stated in the introduction the problem of finding sequences nonrepetitive up to mod k has its origin in some geometric problem. We can apply our proof to a problem from this setting. The following problem concerning nonrepetitive coloring of discrete sets of points in $\mathbb{R}^n$ was considered in [8]. Let P be a discrete set of points and let L be a fixed set of lines in $\mathbb{R}^n$. A coloring of P is *nonrepetitive* (with respect to L) if no sequence of consecutive points on any $l \in L$ is colored repetitively. For a point $p \in P$ let $i(p)$ denote the number of lines from L incident with p and let $I = I(P, L) = \max\{i(p) : p \in P\}$ be the maximum incidence of the configuration (P, L). Using the Lovász Local Lemma it was proved that $Ie^{(8I^2+8I-4)/(I-1)^2}$ colors is sufficient to get such coloring. Adopting the proof of Theorem 1 we can get a better bound.

Theorem 2. *Let (P, L) be a configuration of points and lines in $\mathbb{R}^n$ with finite maximum incidence $I > 2$. If $C \geqslant 2I + 10\sqrt{I}$, then there is a nonrepetitive C-coloring of P with respect to L.*

Proof. The argument is pretty much the same as in the proof of Theorem 1. We provide an algorithm for which each point is colored at random by one of $2I + 10\sqrt{I}$ colors. Fix any linear ordering of all points in P. We color them in this order using Algorithm 1, where arithmetic progressions are changed into lines in $\mathbb{R}^n$. Again for given p and every L such that $p \in L$ we forbid to use colors already assigned to points preceding and following p on L. This gives us at most $2I$ forbidden colors for each vertex. So, by analogy to the previous proof one can show that additional $10\sqrt{I}$ suffice to get a nonrepetitive coloring of P with respect to L. For a log (R, D, O, P, S) we take the same sets as in last case with exceptions that now D keeps the information about the line for which we get a repetition (values between 1 and I) and S is a sequence of numbers between 0 and I. Then all calculations run similarly as before. □

References

[1] J-P. Allouche, J. Shallit, Automatic sequences. Theory, applications, generalizations, Cambridge University Press, Cambridge, 2003.

[2] N. Alon, J.H. Spencer, The probabilistic method, Second Edition, John Wiley & Sons, Inc., New York, 2000.

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

[4] J. Berstel, Axel Thue's work on repetitions in words; in P. Leroux, C. Reutenauer (eds.), Séries formelles et combinatoire algébrique Publications du LaCIM,, Université du Québec a Montréal, p 65-80, 1992.

[5] J. Berstel, Axel Thue's papers on repetitions in words: a translation, Publications du LaCIM, vol 20, Université du Québec a Montréal, 1995.

[6] J. Currie and J. Simpson, Non-repetitive Tilings, The Electron. J. Comb., 9 (2002), 2–8.

[7] J. Currie, Pattern avoidance; themes and variations, Theor. Comput. Sci., 339 (2005), 7–18.

[8] J.Grytczuk, Thue-Like Sequences and Rainbow Arithmetic Progressions, The Electr. J. Comb., 9(1) (2002), Research Paper 44, 10.

[9] J. Grytczuk, Nonrepetitive colorings of graphs - a survey, Int. J. Math. Math. Sci. (2007), Art. ID 74639, 10 pp.

[10] J. Grytczuk, Thue type problems for graphs, points, and numbers, Discrete Math. 308 (2008) 4419–4429.
[11] J. Grytczuk, J. Kozik, P. Micek, A new approach to nonrepetitive sequences (submitted).
[12] M. Lothaire, Combinatorics on Words, Addison-Wesley, Reading MA, 1983.
[13] R. Moser, G. Tardos, A constructive proof of the general Lovász local lemma, J. ACM, 57 (2010), Art. 11, 15.
[14] A. Thue, Über unendliche Zeichenreichen, Norske Vid. Selsk. Skr., I Mat. Nat. Kl., Christiania, 7 (1906), 1-22.

Department of Theoretical Computer Science, Jagiellonian University, Kraków, Poland and Faculty of Mathematics and Information Science, Warsaw University of Technology, Warszawa, Poland
E-mail address: grytczuk@tcs.uj.edu.pl

Department of Theoretical Computer Science, Jagiellonian University, Kraków, Poland
E-mail address: jkozik@tcs.uj.edu.pl

Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poznań, Poland
E-mail address: mw@amu.edu.pl