Discrete Applied Mathematics 217 (2017) 488–494

Contents lists available at ScienceDirect

Discrete Applied Mathematics

journal homepage: www.elsevier.com/locate/dam

Bannai et al. method proves the *d*-step conjecture for strings

Antoine Deza [*], Frantisek Franek

Advanced Optimization Laboratory, Department of Computing and Software, McMaster University, Hamilton, Ontario, Canada

ARTICLE INFO

Article history:
Received 30 April 2015
Received in revised form 19 September 2016
Accepted 27 September 2016
Available online 20 October 2016

Keywords:
Strings
Runs
Lyndon root
d-step approach

ABSTRACT

Inspired by the d-step approach used for investigating the diameter of polytopes, Deza and Franek introduced the d-step conjecture for runs stating that the number of runs in a string of length n with exactly d distinct symbols is at most $n-d$. Bannai et al. showed that the number of runs in a string is at most $n-3$ for $n \geq 5$ by mapping each run to a set of starting positions of Lyndon roots. We show that Bannai et al. method proves that the d-step conjecture for runs holds, and stress the structural properties of run-maximal strings. In particular, we show that, up to relabelling, there is a unique run-maximal string of length $2d$ with d distinct symbols. The number of runs in a string of length n is shown to be at most $n-4$ for $n \geq 9$.

1. Introduction

A *run* in a string $x[1..n]$ is a succinct notion of a maximal repetition. A run is usually encoded by in a triple (s, e, p) such that the substring $x[s..e]$ has a minimal period of p, $x[s..s+p-1]$ is primitive, $s+2p-1 \leq e$ and so $x[s..s+p-1]$ repeats at least twice, and either $s=1$ or $x[s-1] \neq x[s+p-2]$ and either $e=n$ or $x[e-p] \neq x[e+1]$, i.e. the periodicity can be extended neither to the left nor to the right. Thus, s encodes the start of the run, e the end of the run, and p its period. The substring $x[s..s+p-1]$ is the *root* of the run. For example, in the string $a\underline{abababa}a$, the underlined run is encoded by $(2, 8, 2)$, and its root ab is repeated 4 times, with the last repeat being incomplete. Runs, equal up to translation, may occur more than once in a string. For example, in the string $a\underline{abababa}aaaa\underline{abababa}a$, the underlined runs encoded by $(2, 8, 2)$ and $(13, 19, 2)$ are both counted.

Crochemore [4] showed in 1981 that the order of the number of maximal repetitions in a string of length n is $\mathcal{O}(n \log n)$. In 1999, Kolpakov and Kucherov [18] showed that the order of the largest number $\rho(n)$ of runs over all strings of length n is $\mathcal{O}(n)$, without exhibiting an explicit constant, and conjectured that $\rho(n) \leq n$. Rytter [23,24] determined such a constant in 2006, and the following years witnessed a tightening of the lower and upper bounds for the limit of $\rho(n)/n$, see [5,6,14–16,19,21,20,22]. In 2015, the conjecture was proven by Bannai et al. [3] who showed that $\rho(n) \leq n-1$, and $\rho(n) \leq n-3$ for $n \geq 5$, by using starts of specific Lyndon roots of each run; that is by mapping all runs to mutually disjoint subsets of the string indices.

Deza and Franek investigated the largest number $\rho_d(n)$ of runs over all strings of length n with exactly d distinct symbols. Similarities between $\rho_d(n)$ and the largest diameter $\Delta(d, n)$ over all polytopes of dimension d having n facets triggered the formulation of the d-step conjecture for strings stating that $\rho_d(n) \leq n-d$, see [8]. The proposed d-step approach proved that the following statements are equivalent $\{\rho_d(n) \leq n-d$ for all d and $n\}$, $\{\rho_d(2d) \leq d$ for all $d\}$, and $\{\rho_d(2d)$ is achieved for all d by a, up to relabelling, unique string $\}$. Considering binary strings, Fischer et al. [12] showed that $\rho_2(n) \leq \lceil 22n/23 \rceil$. While it is widely believed that $\rho_{d+1}(n) \leq \rho_d(n)$, and thus that $\rho(n) = \rho_2(n)$, no such results are known.

[*] Corresponding author.
E-mail addresses: deza@mcmaster.ca (A. Deza), franek@mcmaster.ca (F. Franek).

http://dx.doi.org/10.1016/j.dam.2016.09.036
0166-218X/© 2016 Elsevier B.V. All rights reserved.

Some properties concerning maximal strings are rather counterintuitive. For example, consider the largest number $\sigma_d(n)$ of distinct primitively rooted squares over all strings of length n with exactly d distinct symbols. It was similarly believed that the binary case is the key one; i.e. that $\sigma_{d+1}(n) \leq \sigma_d(n)$, and thus that $\sigma(n) = \sigma_2(n)$, till a counterexample was provided for $n = 33$ with $\sigma_3(33) > \sigma_2(33)$, see [9].

This paper aims at combining the Bannai et al. and d-step approaches in order to highlight the structural properties of run-maximal strings. Besides strengthening by one the upper bound to $\rho(n) \leq n - 4$ for $n \geq 9$, these structural properties may provide preliminary substantiation for the hypothesis that $\rho(n) \leq n - \lceil \log_2 n \rceil$. For more details and additional results concerning runs in strings we refer to [3] and references therein. Before presenting the main results in Section 2, we briefly recall the Bannai et al. and d-step approaches in the remainder of this section.

1.1. Preliminaries

Strings are indexed starting with 1, i.e. a string x of length n can be written either as $x[1..n]$ or $x[1]x[2]\ldots x[n]$. The *alphabet* of a string x is the set of all symbols occurring in x. A (d, n)-string refers to a string of length n with exactly d distinct symbols. A string x is a *rotation* of a string y if there are u and v such that $x = uv$ and $y = vu$, and the rotation is *trivial* when either u or v is the empty string. Let $\prec$ be a total order over the alphabet of a string x. The string x is *Lyndon with respect to* $\prec$ if x is lexicographically strictly smaller than any of its non-trivial rotations or, equivalently, if x is lexicographically strictly smaller than any of its suffixes. The lexicographic order of strings is induced in the usual manner by the order of the alphabet. Note that $\rho_1(1) = 0$ and $\rho_1(n) = 1$ for $n \geq 2$. Thus, we can assume that both d and n are at least 2 in the remainder of the paper.

1.2. A d-step approach for polytopes and its continuous analogue

We briefly recall the d-step approach used to investigate the Hirsch bound for the diameter of polytopes, and its continuous analogue, and provide some basic references.

A d-step approach for diameter-maximal polytopes

A polyhedron is the intersection of finitely many closed half-spaces, and a polytope is a bounded polyhedron. A (d, n)-polytope is a polytope of dimension d with n facets. The diameter $\delta(P)$ of a polytope P is the smallest integer such that any pair of vertices of P can be connected by an edge-path of length at most $\delta(P)$. Let $\Delta(d, n)$ denote the largest diameter over all (d, n)-polytopes. The Hirsch conjecture, posed in 1957, states that $\Delta(d, n) \leq n - d$. The values for $\Delta(d, n)$ are usually listed in a $(d, n - d)$ table where d is the index for the rows and $n - d$ the index for the columns. The following properties can be checked: $\Delta(d, n) \leq \Delta(d, n + 1)$, $\Delta(d, n) < \Delta(d + 1, n + 2)$, $\Delta(d, n) \leq \Delta(d + 1, n + 1)$ for $n \geq d$; and $\Delta(d, n) = \Delta(d + 1, n + 1)$ for $2d \geq n \geq d$. In other words, the maximum of $\Delta(d, n)$ within a column is achieved on the main diagonal and all values below a value on the main diagonal are equal to that value. The role played by the main diagonal of the $(d, n - d)$ table was underlined by Klee and Walkup [17] who showed the equivalency between the Hirsch conjecture and the d-step conjecture stating that $\Delta(d, 2d) \leq d$ for all d. The Hirsch conjecture was disproved by Santos [25] by exhibiting a violation on the main diagonal with $(d, n) = (43, 86)$; that is, Santos constructed a polytope in dimension 43 with 86 facets and a diameter of at least 44. Note that the d-cube is a $(d, 2d)$-polytope having diameter d and therefore $\Delta(d, 2d) \geq d$ for all d. The string $a_1a_1a_2a_2\ldots a_da_d$ is a $(d, 2d)$-string with d runs and therefore $\rho(d, 2d) \geq d$ for all d. While there is no obvious way to map the n facets of a (d, n)-polytope and the n characters of a (d, n)-string in general, one may map the d squares a_ia_i of the string $a_1a_1a_2a_2\ldots a_da_d$ and the d pairs of disjoint facets of the d-cube.

A d-step approach for curvature-maximal polytopes

Considering links between the currently most computationally successful algorithms for linear optimization; i.e., the simplex and central-path following primal–dual interior point methods, Deza et al. [11] proposed a continuous analogue of the Hirsch conjecture. The value of $\Delta(d, n)$ provides a lower bound for the number of iterations of simplex methods for the worst case behaviour. The curvature of a polytope, defined as the largest total curvature of the associated central path, can be regarded as the continuous analogue of its diameter. Considering the largest curvature $\Lambda(d, n)$ over all (d, n)-polytopes, Deza et al. [11] proved the following continuous analogue of the equivalence between the Hirsch conjecture and the d-step conjecture: if $\Lambda(d, 2d) = \mathcal{O}(d)$ for all d, then $\Lambda(d, n) = \mathcal{O}(n)$. Using a tropical linear optimization setting, Allamigeon et al. [1] constructed an exponential counterexample to the continuous analogue of the polynomial Hirsch conjecture by exhibiting a $(d, 3d/2)$-polytope with a curvature of at least $2^{d/2}$.

1.3. A d-step approach for strings

A d-step formulation for strings was proposed in [8] where it was shown that $\rho_d(n)$ and $\Delta(d, n)$ exhibit similarities and, in particular, that $\rho_d(n) \leq \rho_d(n + 1)$, $\rho_d(n) < \rho_{d+1}(n + 2)$, $\rho_d(n) \leq \rho_{d+1}(n + 1)$ for $n \geq d$; and $\rho_d(n) = \rho_{d+1}(n + 1)$ for $2d \geq n \geq d$. Consequently, the value of $\rho_d(n)$ is presented in a $(d, n - d)$ table where d is the index for the rows and $n - d$ the index for the columns, see Table 1 for a 20 × 20 portion of the $(d, n - d)$ table for $\rho_d(n)$. These properties noted in [8] show that the maximum of $\rho_d(n)$ within a column is achieved on the main diagonal and all values below a value on

Table 1
$(d, n-d)$ table for $\rho_d(n)$ with $2 \le d \le 20$ and $2 \le n-d \le 20$ with the various fonts illustrating Proposition 5.

		$n-d$																		
		2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	2	**2**	*2*	*3*	*4*	*5*	***5***	***6***	***7***	***8***	***8***	***10***	10	11	12	13	14	15	15	16
	3	2	**3**	*3*	*4*	*5*	*6*	***6***	***7***	***8***	***9***	***10***	***11***	11	12	13	14	15	16	16
	4	2	3	**4**	*4*	*5*	*6*	*7*	***7***	***8***	***9***	***10***	***11***	***12***	12	13	14	15	16	17
	5	2	3	4	**5**	*5*	*6*	*7*	*8*	***8***	***9***	***10***	***11***	***12***	***13***	13	14	15	16	17
	6	2	3	4	5	**6**	*6*	*7*	*8*	*9*	***9***	***10***	***11***	***12***	***13***	***14***	14	15	16	17
	7	2	3	4	5	6	**7**	*7*	*8*	*9*	*10*	***10***	***11***	***12***	***13***	***14***	***15***	15	16	17
	8	2	3	4	5	6	7	**8**	*8*	*9*	*10*	*11*	***11***	***12***	***13***	***14***	***15***	***16***	16	17
	9	2	3	4	5	6	7	8	**9**	*9*	*10*	*11*	*12*	***12***	***13***	***14***	***15***	***16***	***17***	17
	10	2	3	4	5	6	7	8	9	**10**	*10*	*11*	*12*	*13*	***13***	***14***	***15***	***16***	***17***	***18***
d	11	2	3	4	5	6	7	8	9	10	**11**	*11*	*12*	*13*	*14*	***14***	***15***	***16***	***17***	***18***
	12	2	3	4	5	6	7	8	9	10	11	**12**	*12*	*13*	*14*	*15*	***15***	***16***	***17***	***18***
	13	2	3	4	5	6	7	8	9	10	11	12	**13**	*13*	*14*	*15*	*16*	***16***	***17***	***18***
	14	2	3	4	5	6	7	8	9	10	11	12	13	**14**	*14*	*15*	*16*	*17*	***17***	***18***
	15	2	3	4	5	6	7	8	9	10	11	12	13	14	**15**	*15*	*16*	*17*	*18*	***18***
	16	2	3	4	5	6	7	8	9	10	11	12	13	14	15	**16**	*16*	*17*	*18*	*19*
	17	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	**17**	*17*	*18*	*19*
	18	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	**18**	*18*	*19*
	19	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	**19**	*19*
	20	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	**20**

the main diagonal are equal to that value. The main results and conjectures yielded by the d-step approach for strings are given in Proposition 1 and Conjecture 2.

Proposition 1 ([8]). *Let $\rho_d(n)$ be the largest number of runs over all strings of length n with exactly d distinct symbols, then*

(i) $\rho_d(n) \le n-d$ *for all d and n* $\Longleftrightarrow$ $\rho_d(2d) \le d$ *for all d,*
(ii) $\rho_d(2d) = \rho_d(2d+1) \Longrightarrow$ *the string* $a_1a_1a_2a_2\dots a_da_d$ *is, up to relabelling, the unique run-maximal string of length* $2d$ *with exactly d distinct symbols,*
(iii) $\rho_d(2d+1) = \rho_d(2d+2) = \rho_d(2d+3)$.

Conjecture 2 ([8]). *A string of length n with exactly d distinct symbols has at most* $n-d$ *runs; that is,* $\rho_d(n) \le n-d$.

Note that the d-step formulation was used in [2] to determine $\rho_d(n)$ for previously intractable values of d and n. In particular, the largest number of runs has been determined for binary strings of length up to 74.

1.4. Bannai et al. method for strings

A key idea of Bannai et al. method is to map the runs of a string to mutually disjoint subsets of its indices. Given a total order $\prec$ of the alphabet of a string, let $\prec^{-1}$ denote the reverse order. Consider a run $t = (i, j, p)$ in a string x. For $i \le k \le j-p$, all the substrings $x[k..k+p]$ are primitive, and at least one of them is Lyndon with respect to $\prec$, and at least one of them is Lyndon with respect to $\prec^{-1}$. This observation motivated the notion of L-roots for a run $t = (i, j, p)$:

Case 1: $j = n$ or $x[j-p+1] \succ x[j+1]$; then every substring $x[k..k+p]$, $i < k \le j-p$, that is Lyndon with respect to $\prec$, is necessarily a maximal Lyndon substring and is referred to as an L-root of t.
Case 2: $j < n$ and $x[j-p+1] \prec x[j+1]$; then every substring $x[k..k+p]$, $i < k \le j-p$, that is Lyndon with respect to $\prec^{-1}$, is necessarily a maximal Lyndon substring and is referred to as an L-root of t.

Note that $x[j-p+1] \ne x[j+1]$ as otherwise t could be extended by one position to the right, contradicting the maximality condition for a run. Thus, exactly one of the two cases holds. If the considered order is clear from the context, we simply use the term L-root. Note a slight modification of Bannai et al. terminology: our definition of L-roots excludes Lyndon substrings, of the length of the period of t, starting at the beginning of the run t. A run t is mapped to the set $\mathrm{Beg}(t)$ of the starting positions of all its L-roots. Bannai et al. [3] showed that $\mathrm{Beg}(t_1) \cap \mathrm{Beg}(t_2) = \emptyset$ for distinct runs t_1 and t_2; that is, the L-roots of two distinct runs never start at the same position—recall that L-roots of a run never start at the beginning of a run. This mapping implies that the number of runs of a string is at most its length. In addition, since no L-root starts at position 1, the number of runs of a string is strictly less than its length.

2. Bannai et al. method and the *d*-step conjecture for strings

The authors contacted Hideo Bannai in the summer of 2014 to point out the d-step conjecture for runs, and a proof of $\rho_d(n) \le n-d$ was subsequently added to [3] in January 2015, along with $\rho_d(n) \le n-d-1$ for $n \ge 2d+1$ which implies $\rho(n) \le n-3$ for $n \ge 5$. Besides providing alternative proofs for Lemmas 9 and 10, we show additional properties and strengthen by one the bound to $\rho_d(n) \leq n - d - 2$ for $n \geq 2d + 5$ which implies $\rho(n) \leq n - 4$ for $n \geq 9$. We wish to point out related results by Crochemore and Mercaş [7] and Fischer et al. [12] that both build up on Bannai et al. method to bound the number of runs. Note the overlap between the notions of *multiplicities of Lyndon roots for cubic runs* in [7], of *overloaded* in [12] and what we call *redundant*.

2.1. Main results

The following propositions are obtained by combining the d-step and Bannai et al. approaches. Proposition 3 strengthens by one the upper bound for the number of run in a string of length n, Proposition 4 illustrates that, in contrast with polytopes, the d-step conjecture holds for strings and the uniqueness of run-maximal stings whose length is twice its number of symbols, and Proposition 5 deals with strings whose length is at most twice its number of symbols plus 10.

Proposition 3. *Let $\rho(n)$ be the largest number of runs over all strings of length n, then $\rho(n) \leq n - 4$ for $n \geq 9$.*

Proof. Proposition 3 is a direct corollary of Lemmas 9, 10, and 11. □

Proposition 4. *The string $a_1a_1a_2a_2\ldots a_da_d$ is, up to relabelling, the unique run-maximal string of length 2d with exactly d distinct symbols.*

Proof. Proposition 4 is a direct corollary of items (ii) of Propositions 1 and 5. □

Proposition 5. *Let $\rho_d(n)$ be the largest number of runs over all strings of length n with exactly d distinct symbols, then*

(i) $\rho_d(n) = n - d$ *for* $2d \geq n$,
(ii) $\rho_d(2d+1) = \rho_d(2d+2) = \rho_d(2d+3) = \rho_d(2d+4) = n - d - 1$,
(iii) $\rho_d(2d+5) = \rho_d(2d+6) = \rho_d(2d+7) = \rho_d(2d+8) = \rho_d(2d+9) = \rho_d(2d+10) = n-d-2$, *except for* $(d, n) = (2, 13)$ *as* $\rho_2(13) = 8$.

Proof. The fact that $\rho_{d+1}(d+2) > \rho_d(n)$ and $\rho_2(4) = 2$ implies that $\rho_d(2d) \geq d$. Thus, Lemma 9 implies that $\rho_d(2d) = d$; that is, item (i) holds as $\rho_d(n) = n - d$ for $2d \geq n$ since $\rho_d(n) = \rho_{d+1}(n+1)$ for $2d \geq n \geq d$. Similarly, the fact that $\rho_{d+1}(d+2) > \rho_d(n)$ and $\rho_2(8) = \rho_2(7) + 1 = \rho_2(6) + 2 = \rho_2(5) + 3 = 5$ implies that $\rho_d(n) \geq n - d - 1$ for $2d + 4 \geq n \geq 2d + 1$. Thus, Lemma 10 implies that $\rho_d(n) = n - d - 1$ for $2d + 4 \geq n \geq 2d + 1$; that is, item (ii) holds. The proof for item (iii) is almost the same as for item (ii) except that Lemma 11 is used instead of Lemma 10 and the base values are $\rho_2(14) = \rho_3(14) + 1 = \rho_2(12) + 2 = \rho_2(11) + 3 = \rho_2(10) + 4 = \rho_2(9) + 5 = 10$. □

See Table 1 for an illustration of Proposition 5 where the main diagonal corresponding to $n = 2d$ is in bold, the diagonals corresponding to $2d < n \leq 2d + 4$ are in italic, and the diagonals corresponding to $2d + 4 < n \leq 2d + 10$ are in bold italic. Note that these values are computationally intractable for non-trivial (d, n); i.e. to compute the largest numbers of runs over all strings of length 35 with exactly 15 symbols is beyond current computational means while Proposition 5 shows that this number is 18.

Remark 6. A generalization of the proof of Lemma 11 to higher values of $n - 2d$ may substantiate the hypothesis that $\rho_d(2d + k) - d$ is a step function independent of d. Proposition 5 might be considered as a preliminary substantiation that the number of runs in a string of length n with d symbols is at most $n - d - \lceil \log_2 \lceil (n+4-2d)/4 \rceil \rceil$ for $n \geq 2d$ as hypothesized in [8]. Assuming that run-maximal strings include binary ones and thus setting $d = 2$ in the previous inequality, the number of runs in a string of length n was hypothesized in [8] to be at most $n - \lceil \log_2 n \rceil$. A d-step approach was introduced for distinct primitively rooted squares in strings as well as hypothesized upper bounds [8]. We recall that the bound of Fraenkel and Simpson [13] was strengthened in [10] to: the number of distinct squares in a string of length n is at most $\lfloor 11n/6 \rfloor$.

2.2. Observations and auxiliary lemmas

Observation 7. *Given a string x over the alphabet $\{a_1, \ldots, a_d\}$,*

(i) *no L-root starts at position 1,*
(ii) *if an L-root of a run t starts at the position of a symbol $\boldsymbol{a_i}$:*
case (1) *there is a farther occurrence of $\boldsymbol{a_i}$ and $x = \ldots(u\underline{\boldsymbol{a_i}v)(u}\widehat{a_i}v)\ldots(ua_iv)\mu\ldots$ and for any $\nu \in u, v$,*

$$\begin{cases} a_i \preceq \nu & \text{if } \mu \prec u[1] \ \text{(i.e. } \prec \text{ is used)} \\ \nu \preceq a_i & \text{if } \mu \succ u[1] \ \text{(i.e. } \prec^{-1} \text{ is used)} \end{cases}$$

where $\widehat{a_i}$ is the farther copy of $\boldsymbol{a_i}$, the L-root is underlined, and ( )( ) . . . ( ) indicates the rightmost repetition of the run t containing $\boldsymbol{a_i}$ in its root,
case (2) *there is a previous occurrence of $\boldsymbol{a_i}$ and $x = \ldots(\widehat{a_i}w)(\underline{\boldsymbol{a_i}w})\ldots(a_iw)\mu\ldots$ and for any $\nu \in w$,*

$$\begin{cases} a_i \preceq \nu & \text{if } \mu \prec \boldsymbol{a_i} \ \text{(i.e. } \prec \text{ is used)} \\ \nu \preceq a_i & \text{if } \mu \succ \boldsymbol{a_i} \ \text{(i.e. } \prec^{-1} \text{ is used)} \end{cases}$$

where $\widehat{a_i}$ is the previous copy of $\boldsymbol{a_i}$, the L-root is underlined, and ( )( ) . . . ( ) indicates the rightmost repetition of the run t starting with $\widehat{a_i}$,

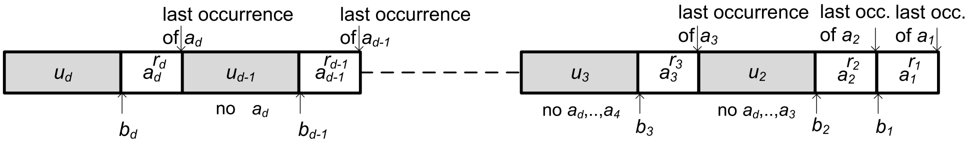

Fig. 1. Illustration of Definition 8.

(iii) *if a symbol a_i occurs only once, no L-root starts at the position of a_i,*
(iv) *if a symbol a_i occurs exactly twice, at most one L-root starts at the positions of the occurrence of a_i since they belong to at most one run,*
(v) *if a symbol a_i occurs exactly three times, at most two L-root start at the positions of the occurrences of a_i since they belong to at most two runs.*

Observation 7 leads to the following notion of redundancy: if a run has $k \geq 2$ L-roots, we consider that $k-1$ of them are redundant. For example, a_i^{k+1} has k L-roots and $k-1$ of them are redundant. The number $r(x)$ of runs in a string x is at most the number of its non-redundant L-roots.

Definition 8. Given a string x containing the symbols $\{a_1, \dots, a_d\}$ ordered by $\prec$, the string is *$\prec$-labelled* if:

(i) $a_1 \prec a_2 \prec \cdots \prec a_{d-1} \prec a_d$,
(ii) $x = u_d a_d^{r_d} u_{d-1} a_{d-1}^{r_{d-1}} \dots u_2 a_2^{r_2} a_1^{r_1}$ where u_k does not contain any symbol from $\{a_{k+1}, \dots, a_d\}$ and $r_k \geq 1$ for $k = 1, \dots, d$,
(iii) $x[b_k - 1] \prec a_k$ for $k = 1, \dots, d-1$, and $x[b_d - 1] \prec a_d$ if $d > 1$, where $b_k = 1 + (|u_k| + \cdots + |u_d|) + (r_{k+1} + \cdots + r_d)$ for $k = 1, \dots, d$.

Note that b_k is the position of the beginning of the *block of last occurrence of* a_k, i.e. $a_k^{r_k}$. Note also that $|u_k|$ may be possibly zero and that r_k is maximal as $x[b_k - 1] \prec a_k$ implies that the preceding symbol, if any, differs from a_k, see Fig. 1 for an illustration of Definition 8. Any string x can be $\prec$-labelled by a simple act of relabelling the alphabet symbols. The structure of the $\prec$-labelled strings indicates places where an L-root cannot start and where redundant L-roots might occur. The first places to look for positions with no L-roots are, of course, the beginnings of the blocks, i.e. the b_k's.

Lemma 9. *Let $\rho_d(n)$ be the largest number of runs over all strings of length n with exactly d distinct symbols, then $\rho_d(n) \leq n - d$ for $n \geq 2d$.*

Proof. Consider a $\prec$-labelled run-maximal (d, n)-string x, i.e. x is a $\prec$-labelled string with $\rho_d(n)$ runs. We show that the number of L-roots of $x = u_d a_d^{r_d} u_{d-1} a_{d-1}^{r_{d-1}} \dots u_2 a_2^{r_2} a_1^{r_1}$ is at most $n - d$, and thus $\rho_d(n) \leq n - d$, by remarking that no L-root starts at b_k for any $k \in 1..d$. Assume by contradiction that an L-root starts at b_{k_0} for some k_0. The L-root cannot correspond to case (ii)(1) of Observation 7 as there is no farther occurrence of a_{k_0} past the block $a_{k_0}^{r_{k_0}}$. Thus, the L-root must correspond to case (ii)(2). Since $x[b_{k_0} - 1] \prec x[b_{k_0}]$, the $\prec^{-1}$ order must be used, but a_{k_0} does not precede any symbol past $a_{k_0}^{r_{k_0}}$—hence a contradiction. □

Another natural place to look for no L-root is the beginning of the string, and all we need to guarantee is that b_d, the beginning of the first block $a_{a_d}^{r_{a_d}}$, does not coincide with the beginning of the string—which is guaranteed by a condition on its length.

Lemma 10. *Let $\rho_d(n)$ be the largest number of runs over all strings of length n with exactly d distinct symbols, then $\rho_d(n) \leq n - d - 1$ for $n \geq 2d + 1$.*

Proof. Consider a $\prec$-labelled run-maximal (d, n)-string x, that is x is a $\prec$-labelled string with $\rho_d(n)$ runs. We need to show that, besides the d positions corresponding to the b_k's, there is at least one position with no or a redundant L-root. The consider the following two cases. Case (i) $|u_k| = 0$ for $k = 1, \dots, d$. Then $x = a_d^{r_d} a_{d-1}^{r_{d-1}} \dots a_2^{r_2} a_1^{r_1}$ and since $n \geq 2d + 1$, there is a k_0 such that $r_{k_0} \geq 3$ and so at least one L-root is redundant. Case (ii) $|u_{k_0}| \geq 1$ for some k_0. We show that without loss of generality we can assume that $k_0 = d$. Assume otherwise that $|u_d| = 0$; that is, $x = a_d^{r_d} u_{d-1} a_{d-1}^{r_{d-1}} \dots u_2 a_2^{r_2} a_1^{r_1}$. Since $r(x) = r(x[r_d + 1..n]x[1..r_d])$, we can move $a_d^{r_d}$ to the end of the string and relabel the symbols and repeat this process until we run into the first k_0 such that $|u_{k_0}| \geq 1$. Thus, we have a run-maximal string with non-empty u_d and so $b_d > 1$ and so the number of positions with no L-roots is at least $d + 1$. □

Lemma 11. *Let $\rho_d(n)$ be the largest number of runs over all strings of length n with exactly d distinct symbols, then $\rho_d(n) \leq n - d - 2$ for $n \geq 2d + 5$.*

Proof. Consider a $\prec$-labelled run-maximal (d, n)-string x, that is x is a $\prec$-labelled string with $\rho_d(n)$ runs. Besides the d positions b_k's, we need to exhibit at least two additional positions with no or redundant L-roots.
The case when $|u_k| = 0$ for $k = 1, \dots, d$. Then is $x = a_d^{r_d} a_{d-1}^{r_{d-1}} \dots a_2^{r_2} a_1^{r_1}$ and since $n \geq 2d + 5$, there are k_0, k_1, k_2, k_3 and k_4 such that $r_{k_0} + r_{k_1} + r_{k_2} + r_{k_3} + r_{k_4} \geq 15$ and so at least 5 L-roots are redundant.

The case when $|u_{k_0}| \geq 1$ for some k_0. As in the proof of Lemma 10, we can assume that $k_0 = d$; that is $|u_d| \geq 1$, and thus $1 < b_d < b_{d-1} < \cdots < b_2 < b_1$ are positions with no L-roots. Therefore, to complete the proof, we need to exhibit one additional position with no or a redundant L-root.

We can assume that $r_k \leq 2$ for $k = 1, \ldots, d$ as otherwise one additional L-root would be redundant and the proof would be completed. Define m as the smallest k such that $|u_k| \geq 1$, i.e. $x = u_d a_d^{r_d} \ldots u_m a_m^{r_m} a_{m-1}^{r_{m-1}} \ldots a_2^{r_2} a_1^{r_1}$—note that such $m \leq d$ must exist as $|u_d| \geq 1$. Let a_ℓ the last symbol of u_m and note that $\ell \leq m-1$: if $\ell > m$, then a_ℓ would be occurring past the block of its last occurrence $a_\ell^{r_\ell}$ which is to the left of u_m, which is not possible; if $\ell = m$, then it should be a part of the block $a_m^{r_m}$. Thus, $x = \ldots \mu \boldsymbol{a}_{\boldsymbol{\ell}}^{\boldsymbol{i}} a_m^{r_m} a_{m-1}^{r_{m-1}} \ldots a_{\ell+1}^{r_{\ell+1}} a_\ell^{r_\ell} a_{\ell-1}^{r_{\ell-1}} \ldots a_2^{r_2} a_1^{r_1}$ for some $\mu \neq a_\ell$ and some $i \geq 1$. Since $i \geq 3$ would give at least one redundant L-root in one of the positions of $\boldsymbol{a}_{\boldsymbol{\ell}}^{\boldsymbol{i}}$, we can assume that i is at most 2.

We show that either there is no L-root at the beginning of $\boldsymbol{a}_{\boldsymbol{\ell}}^{\boldsymbol{i}}$ or there is a position before the beginning of $\boldsymbol{a}_{\boldsymbol{\ell}}^{\boldsymbol{i}}$ with no L-root. To do so, we assume that there is an L-root at the beginning of $\boldsymbol{a}_{\boldsymbol{\ell}}^{\boldsymbol{i}}$ and find a prior position with no L-root.

Consider that the L-root at the beginning is of $\boldsymbol{a}_{\boldsymbol{\ell}}^{\boldsymbol{i}}$ that corresponds to case (ii)(2) of Observation 7. Since the L-root starts with $\boldsymbol{a}_{\boldsymbol{\ell}}^{\boldsymbol{i}}$ followed by a_m and $\ell < m$, and so $a_\ell \prec a_m$, it follows that it is Lyndon with respect to $\prec$, and so the trailing square of the run must be followed by a symbol smaller than a_ℓ and so it must reach past the block $a_\ell^{r_\ell}$. If it reached past $a_{\ell-1}^{r_{\ell-1}}$, the suffix starting with $a_{\ell-1}$ would be lexicographically smaller than the L-root that starts with $\boldsymbol{a}_{\boldsymbol{\ell}}$—hence a contradiction. Thus, it must actually end inside the block $a_{\ell-1}^{r_{\ell-1}}$. Then the suffix starting with $a_{\ell-1}$ would be lexicographically smaller than the L-root—hence a contradiction.

Therefore, the L-root must correspond to case (ii)(1) of Observation 7, and there are only 4 possibilities for an L-root to occur at the beginning of $\boldsymbol{a}_{\boldsymbol{k}}^{\boldsymbol{i}}$ (the root is underlined):

case $(i, r_\ell) = (1, 1)$: $x = \ldots(\underbrace{a_m^{r_m} a_{m-1}^{r_{m-1}} .. \boldsymbol{a}_{\ell+1}^{\boldsymbol{r}_{\ell+1}} \underline{\boldsymbol{a}_\ell)}}_{u_m}\underline{(a_m^{r_m} a_{m-1}^{r_{m-1}} .. a_{\ell+1}^{r_{\ell+1}}} a_\ell) a_{\ell-1}^{r_{\ell-1}} .. a_2^{r_2} a_1^{r_1}$

case $(i, r_\ell) = (1, 2)$: $x = \ldots(\underbrace{a_m^{r_m} a_{m-1}^{r_{m-1}} .. \boldsymbol{a}_{\ell+1}^{\boldsymbol{r}_{\ell+1}} \underline{\boldsymbol{a}_\ell)}}_{u_m}\underline{(a_m^{r_m} a_{m-1}^{r_{m-1}} .. a_{\ell+1}^{r_{\ell+1}}} a_\ell) a_\ell a_{\ell-1}^{r_{\ell-1}} .. a_2^{r_2} a_1^{r_1}$

case $(i, r_\ell) = (2, 1)$: $x = \ldots(a_\ell \underbrace{a_m^{r_m} a_{m-1}^{r_{m-1}} .. \boldsymbol{a}_{\ell+1}^{\boldsymbol{r}_{\ell+1}} \underline{\boldsymbol{a}_\ell)(}}_{u_m}\underline{\boldsymbol{a}_\ell a_m^{r_m} a_{m-1}^{r_{m-1}} .. a_{\ell+1}^{r_{\ell+1}}} a_\ell) a_{\ell-1}^{r_{\ell-1}} .. a_2^{r_2} a_1^{r_1}$

case $(i, r_\ell) = (2, 2)$: $x = \ldots(\underbrace{a_m^{r_m} a_{m-1}^{r_{m-1}} .. \boldsymbol{a}_{\ell+1}^{\boldsymbol{r}_{\ell+1}} \underline{\boldsymbol{a}_\ell \boldsymbol{a}_\ell)}}_{u_m}\underline{(a_m^{r_m} a_{m-1}^{r_{m-1}} .. a_{\ell+1}^{r_{\ell+1}}} a_\ell a_\ell) a_{\ell-1}^{r_{\ell-1}} .. a_2^{r_2} a_1^{r_1}$

We show that in all four cases, there is no L-root at the beginning of the first $\boldsymbol{a}_{\boldsymbol{\ell+1}}^{\boldsymbol{r}_{\boldsymbol{\ell+1}}}$ (denoted in bold). First note that if there were an L-root, it would have to contain the $\boldsymbol{a}_{\boldsymbol{\ell}}$ that follows the block $\boldsymbol{a}_{\boldsymbol{\ell+1}}^{\boldsymbol{r}_{\boldsymbol{\ell+1}}}$, so the L-root would have to be determined by $\prec^{-1}$. This excludes case (ii)(1) of Observation 7 as the next available $a_{\ell+1}$ is not followed by a bigger symbol. Thus, if there were an L-root, it would correspond to case (ii)(2) of Observation 7 and then the L-root would have the same length as the L-root starting at the beginning of $\boldsymbol{a}_{\boldsymbol{\ell}}^{\boldsymbol{i}}$ as it would span from the first $\boldsymbol{a}_{\boldsymbol{\ell+1}}^{\boldsymbol{r}_{\boldsymbol{\ell+1}}}$ to the next $a_{\ell+1}^{r_{\ell+1}}$, hence they would both belong to the same run as they overlap—hence a contradiction.

We showed that there cannot be an L-root starting at the beginning of $\boldsymbol{a}_{\boldsymbol{\ell+1}}^{\boldsymbol{r}_{\boldsymbol{\ell+1}}}$. The last step is to show that the beginning of $\boldsymbol{a}_{\boldsymbol{\ell+1}}^{\boldsymbol{r}_{\boldsymbol{\ell+1}}}$ does not coincide with the beginning of the string. The only cases for which the beginning of $\boldsymbol{a}_{\boldsymbol{\ell+1}}^{\boldsymbol{r}_{\boldsymbol{\ell+1}}}$ is the beginning of the string correspond to $(i, r_\ell) = (1, 1), (1, 2)$, or $(2, 2)$ and $\ell + 1 = m = d$, but these cases cannot occur if $n \geq 2d + 5$ as detailed below:

case $(i, r_\ell) = (1, 1)$ and $\ell + 1 = m = d$: $x = (a_d^{r_d} \underline{\boldsymbol{a}_{\boldsymbol{d-1}})}\underbrace{\underline{(a_d^{r_d}} a_{d-1})}_{u_d} a_{d-2}^{r_{d-2}} \ldots a_2^{r_2} a_1^{r_1}$

implying $n = 1 + 2r_d + r_{d-1} + \ldots + r_1 \leq 3 + 2d$,

case $(i, r_\ell) = (1, 2)$ and $\ell + 1 = m = d$: $x = (a_d^{r_d} \underline{\boldsymbol{a}_{\boldsymbol{d-1}})}\underbrace{\underline{(a_d^{r_d}} a_{d-1})}_{u_d} a_{d-1} a_{d-2}^{r_{d-2}} \ldots a_2^{r_2} a_1^{r_1}$

implying $n = 1 + 2r_d + r_{d-1} + \ldots + r_1 \leq 3 + 2d$,

case $(i, r_\ell) = (2, 2)$ and $\ell + 1 = m = d$: $x = (a_d^{r_d} \underline{\boldsymbol{a}_{\boldsymbol{d-1}} \boldsymbol{a}_{\boldsymbol{d-1}})}\underbrace{\underline{(a_d^{r_d}} a_{d-1} a_{d-1})}_{u_d} a_{d-2}^{r_{d-2}} \ldots a_2^{r_2} a_1^{r_1}$

implying $n = 2 + 2r_d + r_{d-1} + \ldots + r_1 \leq 4 + 2d$. □

Acknowledgements

The authors would like to thank the anonymous referees for valuable comments and suggestions which improved the quality of the paper, and Maxime Crochemore and Robert Mercaş for informing us of the preprints [7,12]. This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada Discovery Grant program (Franek: RGPIN25112-2012, Deza: RGPIN-2015-06163), and by the Digiteo Chair C&O program (Deza).

References

[1] X. Allamigeon, P. Benchimol, S. Gaubert, M. Joswig, Long and winding central paths, 2014. arXiv:1405.4161.
[2] A. Baker, A. Deza, F. Franek, A computational framework for determining run-maximal strings, J. Discrete Algorithms 20 (2013) 43–50.
[3] H. Bannai, T. I, S. Inenaga, Y. Nakashima, M. Takeda, K. Tsuruta, The "Runs" Theorem, 2015. arXiv:1406.0263v6.
[4] M. Crochemore, An optimal algorithm for computing the repetitions in a word, Inform. Process. Lett. 12 (1981) 297–315.
[5] M. Crochemore, L. Ilie, Maximal repetitions in strings, J. Comput. System Sci. 74 (2008) 796–807.
[6] M. Crochemore, L. Ilie, L. Tinta, Towards a solution to the runs conjecture, Lecture Notes in Comput. Sci. 5029 (2008) 290–302.
[7] M. Crochemore, R. Mercaş, Fewer runs than word length, 2014. arXiv:412.4646v1.
[8] A. Deza, F. Franek, A d-step approach to the maximum number of distinct squares and runs in strings, Discrete Appl. Math. 163 (2014) 268–274.
[9] A. Deza, F. Franek, M. Jiang, A computational substantiation of the d-step approach to the number of distinct squares problem, Discrete Appl. Math. 212 (2016) 81–87.
[10] A. Deza, F. Franek, A. Thierry, How many double squares can a string contain? Discrete Appl. Math. 180 (2015) 52–69.
[11] A. Deza, T. Terlaky, Y. Zinchenko, A continuous d-step conjecture for polytopes, Discrete Comput. Geom. 41 (2009) 318–327.
[12] J. Fischer, Š. Holub, T. I, M. Lewenstein, Beyond the runs theorem, in: Proceedings of SPIRE 2015, Springer-Verlag, London, UK, 2015, pp. 277–286.
[13] A.S. Fraenkel, J. Simpson, How many squares can a string contain? J. Combin. Theory Ser. A 82 (1) (1998) 112–120.
[14] F. Franek, R.J. Simpson, W. Smyth, The maximum number of runs in a string, in: Proceedings of AWOCA 2003, Seoul, Korea, 2008, pp. 13–16.
[15] F. Franek, Q. Yang, An asymptotic lower bound for the maximal number of runs in a string, Internat. J. Found Comput. Sci. 19 (2008) 195–203.
[16] M. Giraud, Not so many runs in strings, in: Proceedings of LATA 2008, Springer, Tarragona, Spain, 2008, pp. 232–239.
[17] V. Klee, D.W. Walkup, The d-step conjecture for polyhedra of dimension $d < 6$, Acta Math. 117 (1967) 53–78.
[18] R. Kolpakov, G. Kucherov, Finding maximal repetitions in a word in linear time, in: Proceedings of FOCS 1999, IEEE Computer Society, Washington, USA, 1999, pp. 596–604.
[19] W. Matsubara, K. Kusano, H. Bannai, A. Shinohara, A series of run-rich strings, Lecture Notes in Comput. Sci. 5457 (2009) 578–587.
[20] W. Matsubara, K. Kusano, A. Ishino, H. Bannai, A. Shinohara, Lower bounds for the maximum number of runs in a string. http://www.shino.ecei.tohoku.ac.jp/runs/.
[21] W. Matsubara, K. Kusano, A. Ishino, H. Bannai, A. Shinohara, New lower bounds for the maximum number of runs in a string, in: Proceedings of PSC 2008, Prague, Czech Republic, 2008, pp. 140–145.
[22] S.J. Puglisi, R.J. Simpson, W.F. Smyth, How many runs can a string contain? Theoret. Comput. Sci. 401 (2008) 165–171.
[23] W. Rytter, The number of runs in a string: Improved analysis of the linear upper bound, Lecture Notes in Comput. Sci. 3884 (2006) 184–195.
[24] W. Rytter, The number of runs in a string, Inform. Comput. 205 (2007) 1459–1469.
[25] F. Santos, A counterexample to the Hirsch conjecture, Ann. of Math. 176 (2012) 383–412.