McMaster University

Advanced Optimization Laboratory

Title:

A d-step analogue for runs on strings

Authors:

Antoine Deza and Frantisek Franek

Advanced Optimization Laboratory,
Department of Computing and Software,
McMaster University, Hamilton, Ontario, Canada.
Email: deza, franek@mcmaster.ca

AdvOL-Report No. 2010/2

March 2010, Hamilton, Ontario, Canada

A d-step analogue for runs on strings

Antoine Deza and Frantisek Franek

March 31, 2010

Abstract

Kolpakov and Kucherov conjectured in 1999 that the number of runs in a string is at most its length n. Let us denote $\rho_d(n)$ the maximum number of runs over all strings of length n containing exactly d distinct symbols. We show that in order to prove $\rho_d(n) \leq n - d$ for all n and d, it is enough to prove $\rho_d(2d) \leq d$ for all d. We also show that if $\rho_d(2d) = \rho_d(2d+1)$ for all d, then we have $\rho_d(n) \leq n - d - 1$ for all $n > 2d$, and the string admitting $\rho_d(2d)$ runs is, up to relabelling, unique. In other words, we show that that in order to prove the number of runs in a string of length $n \geq 5$ is at most $n - 3$, it is enough to prove that $\rho_d(2d) = \rho_d(2d+1)$ for all d.

1 Introduction and Preliminary

A run, a maximal fractional repetition in a string of length n, was conceptually introduced by Main in 1989 [7]. The term was coined by Iliopoulos, Moore, and Smyth in 1997 [4]. Though there may be at most $O(n \log n)$ repetitions in a string [1], it was hoped that the more concise notation of runs would eliminate the need to list all repetitions. In 1999, Kolpakov and Kucherov [6] showed that the number of runs $\boldsymbol{r}(x)$ in a string x is $O(n)$ and conjectured that the maximum number of runs in a string is at most n. Let us denote $\rho(n)$ the maximum number of runs over all strings of length n. Several authors have estimated upper and lower bounds for $\rho(n)$, see Crochemore and Ilie [2], respectively Matsubara et al. [8], for upper, respectively lower, bounds, and references therein.

Preliminary 1. *If a string contains one symbol occurring exactly twice, then this symbol can occur in at most one run.*

Proof. Let us assume that the two $\mathtt{a}$'s occur in two distinct runs r_1 and r_2. Hence they occur in a square $s_1 = [u_1\mathtt{a}v_1][u_1\mathtt{a}v_1]$ of r_1, and in a square $s_2 = [u_2\mathtt{a}v_2][u_2\mathtt{a}v_2]$ of r_2. Since the substring between the two $\mathtt{a}$'s is unique,we can assume, without loss of generality, that v_1 is a prefix of v_2, then $v_2 = v_1v_1'$ and $s_2 = [u_2\mathtt{a}v_1v_1'][u_2\mathtt{a}v_1v_1']$. Again, by uniqueness of the substring between the two $\mathtt{a}$'s, $v_1u_1 = v_1v_1'u_2$, that is: $u_1 = v_1'u_2$. Thus $s_1 = [v_1'u_2\mathtt{a}v_1][v_1'u_2\mathtt{a}v_1]$ and $s_2 = [u_2\mathtt{a}v_1v_1'][u_2\mathtt{a}v_1v_1']$ but s_1 and s_2 have the same period and therefore are in the same run, which is a contradiction. The case when the two $\mathtt{a}$'s are adjacent is trivial. □

2 Main Results

Using an approach and presentation inspired by similar results used in combinatorial and continuous optimization, we highlight the role of strings of length being exactly twice the number of distinct symbols. Let us denote $\rho_d(n)$ the maximum number of runs over all strings of length n containing exactly d distinct symbols. We show in Proposition 2 that in order to prove $\rho_d(n) \leq n - d$ for all n and d it is enough to prove the bound for the special case $n = 2d$ for all d. In other words, $\rho_d(2d) \leq d$ for all d implies that the maximum number $\rho(n)$ of runs over all strings of length $n \geq 3$ satisfies $\rho(n) \leq n - 2$. This result is analogue to the equivalency between the Hirsch conjecture and the d-step conjecture dealing with the maximum diameter of polytope defined by n facets in dimension d, see [5]. Additional propositions are apparently specific to runs on strings and focus on the conjectured equality between $\rho_d(2d)$ and $\rho_d(2d + 1)$. Proposition 5 characterizes uniquely strings of length $2d$ and containing exactly d distinct symbols admitting $\rho_d(2d)$ runs given that $\rho_d(2d) = \rho_d(2d + 1) = d$. Proposition 4 implies that the maximum number $\rho(n)$ of runs over all strings of length $n \geq 5$ satisfies $\rho(n) \leq n - 3$ if $\rho_d(2d) = \rho_d(2d + 1) = d$. Finally, Proposition 6 implies that the maximum number $\rho(n)$ of runs over all strings of length $n \geq 5$ satisfies $\rho(n) \leq n - 3$ if $\rho_d(2d) = \rho_d(2d + 1)$. All the properties presented in the paper were computationally checked for small instances of n and d. We first present a few lemmas dealing with elementary properties of the function $\rho_d(n)$.

2.1 Elementary properties of the function $\rho_d(n)$

Lemma 1. *For $n \geq d$, $\rho_d(n) \leq \rho_{d+1}(n + 1)$.*

Proof. Consider a string x of length n containing exactly d distinct symbols and satisfying $\rho_d(n) = \boldsymbol{r}(x)$. The string y obtained from x by appending a new symbol satisfies $\boldsymbol{r}(x) = \boldsymbol{r}(y) \leq \rho_{d+1}(n + 1)$, thus $\rho_d(n) \leq \rho_{d+1}(n + 1)$. □

Lemma 2. *For $2 \leq d \leq n \leq 2d$, $\rho_d(n) \geq n - d$.*

Proof. Consider the string $x^*_{d,n} = (\texttt{aabbcc...})$ containing exactly d distinct symbols and of length $n \leq 2d$ where the last $2d - n$ symbols occur exactly once and the first $n - d$ symbols occur exactly twice and are adjacent. For $2 \leq d \leq n \leq 2d$, $\boldsymbol{r}(x^*_{d,n}) = n - d$, thus $\rho_d(n) \geq n - d$. □

Lemma 3. *For $2 \leq d$, $\rho_d(2d + 1) \geq d$.*

Proof. Consider the string $x^*_{d,2d+1} = (\texttt{aabbcc...a})$ containing exactly d distinct symbols and of length $2d + 1$ which is obtained by appending to $x^*_{d,2d}$ its first symbol. We have $\boldsymbol{r}(x^*_{d,2d+1}) = d$, thus $\rho_d(2d + 1) \geq d$. □

Lemma 4. *For $n \geq d$, $\rho_{d+1}(n + 2) - \rho_d(n) \geq 1$.*

Proof. Consider a string x of length n containing exactly d distinct symbols and satisfying $\rho_d(n) = \boldsymbol{r}(x)$. The string y obtained from x by appending a pair of adjacent new symbol satisfies $\boldsymbol{r}(x) + 1 = \boldsymbol{r}(y) \leq \rho_{d+1}(n + 2)$, thus $1 \leq \rho_{d+1}(n + 2) - \rho_d(n)$. □

Assume the entries for $\rho_d(n)$ are listed in a table T where d is the index for the rows and $n-d$ the index for the columns. Lemma 1 means that the entries are increasing with d within any column, Lemma 2 means that any entry on or below the main diagonal is at least $n-d$, and Lemma 3 means that any entry immediately to the right of the main diagonal is at least $n-d-1$. Lemma 4 means that the entries increase by at least one along any diagonal. For an illustration see Table 1 where the entries on the main diagonal are in bold and the entries immediately to the right of the main diagonal are in italic.

		$n-d$							
		1	2	3	4	5	6	7	8
	1	**1**	*1*	1	1	1	1	1	1
	2	1	**2**	*2*	3	4	5	5	6
d	3	1	2	**3**	*3*	4	5	6	6
	4	1	2	3	**4**	*4*	5	6	7
	5	1	2	3	4	**5**	*5*	6	7
	6	1	2	3	4	5	**6**	*6*	7
	7	1	2	3	4	5	6	**7**	*7*

Table 1: Entries computed for $\rho_d(n)$ with $1 \le d \le 7$ and $1 \le n-d \le 8$

Remark 1. *Some hypothesized properties dealing with the maximal number of runs in a string can be restated in term of the Table T . For example, the property that the number of runs can only decrease if the number of distinct symbol increases would mean that the entries of Table T increase along any counter-diagonal, that is $\rho_{d+1}(n) \le \rho_d(n)$ for $n \ge d \ge 2$. In other words, the maximum along any counter-diagonal is achieved for $d = 2$, i.e. for binary strings. Similarly, the property that the number of runs can increase by at most 2 if the length increases by 1 would mean that the entries of Table T increase by at most 2 along any rows, that is $\rho_d(n+1) \le \rho_d(n) + 2$ for $n \ge d \ge 2$.*

2.2 A d-step analogue and additional properties

Proposition 1. *We have ($\rho_d(2d) = d$ for $d \ge 2$) implies ($\rho_{d+k}(2d+k) = \rho_d(2d) = d$ for $k \ge 1$ and $d \ge 2$).*

Proof. Assume $\rho_d(2d) = d$ for $d \ge 2$. Consider the induction hypothesis (H_d): $\rho_{d'+k}(2d'+k) = d'$ for $k \ge 1$ and $2 \le d' \le d$. We first prove that (H_d) implies (H_{d+1}). For $k \ge 1$ consider a string x containing exactly $d+1+k$ symbols and of length $n = 2d+2+k$ and satisfying $r(x) = \rho_{d+1+k}(2d+2+k)$. Note that $r(x) \ge d+1$ by Lemma 2. Such string x must contain at least one symbol occurring only once; let `a` be this symbol.

(i) If `a` occurs at the first or last position, then $\boldsymbol{r}(x) = \boldsymbol{r}(y)$ where y is obtained by removing `a` from x. Since y is of length $n = 2d+1+k$ and contains exactly $d+k$ distinct symbols, we have $\rho_{d+1+k}(2d+2+k) = \boldsymbol{r}(x) = \boldsymbol{r}(y) \le \rho_{d+k}(2d+1+k)$. Thus, by Lemma 1,

$\rho_{d+1+k}(2d+2+k) = \rho_{d+k}(2d+1+k)$.

(ii) If `a` does not occur at the first or last position, then $\boldsymbol{r}(x) = \boldsymbol{r}(x_1) + \boldsymbol{r}(x_2)$ where x_1 and x_2 are the parts of x respectively on the left and right of `a`. Assume that x_1 has length n_1 and d_1 symbols, and x_2 has length n_2 and d_2 symbols. We have $\boldsymbol{r}(x) = \boldsymbol{r}(x_1) + \boldsymbol{r}(x_2) \leq \rho_{d_1}(n_1) + \rho_{d_2}(n_2)$. We now prove that $\rho_{d_i}(n_i) \leq n_i - d_i$ for $i = 1$ and 2. We can assume that $d_i \geq 2$ as the case $d_i = 1$ is trivial. Assume first that $n_i \leq 2d_i$, by (H_d), we have $\rho_{d_i}(n_i) = n_i - d_i$. Assume then that $n_i > 2d_i$, by Lemma 1, we have $\rho_{d_i}(n_i) \leq \rho_{n_i-d_i}(2n_i-2d_i) = n_i-d_i$ since $\rho_d(2d) = d$ for $d \geq 2$. Thus, since $\rho_{d_i}(n_i) \leq n_i-d_i$, we have $\boldsymbol{r}(x) \leq \rho_{d_1}(n_1) + \rho_{d_2}(n_2) \leq n_1 + n_2 - d_1 - d_2 = 2d + 1 + k - (d_1 + d_2)$. Since all symbols except `a` occur in either x_1 or x_2, we have $d_1 + d_2 \geq d + k$ and thus $\boldsymbol{r}(x) \leq 2d+1+k-(d+k) = d+1 = \rho_{d+1}(2d+2)$. Note that since $\boldsymbol{r}(x) = \rho_{d+1+k}(2d+2+k) \geq d+1$, we must have $\rho_{d+1+k}(2d+2+k) = \rho_{d+1}(2d+2) = d+1$, $d_1+d_2 = d+k$, and $\rho_{d_i}(n_i) = n_i-d_i$.

Items (i) and (ii) show that $\rho_{d+k}(2d + k)$ is not increasing with k, that is by Lemma 1, $\rho_{d+k}(2d+k) = \rho_d(2d) = d$ for $k \geq 1$. To complete the proof, one can easily check $\rho_d(d+1) = 1$ and that $\rho_d(d + 2) = 2$, that is, the base cases (H_1) and (H_2) hold. □

Proposition 2. *The following 2 statements are equivalent*

(a) $\rho_d(n) \leq n - d$ *for* $n \geq d \geq 2$

(b) $\rho_d(2d) \leq d$ *for* $d \geq 2$

Proof. Since (a) trivially implies (b), we need to prove that (b) implies (a). We note that, by Lemma 2, (b) is equivalent to: $\rho_d(2d) = d$ for $d \geq 2$ and that, Proposition 1 can be restated as $(\rho_d(2d) = d \text{ for } d \geq 2) \Rightarrow (\rho_d(n) = n - d \text{ for } 2 \leq d \leq n \leq 2d)$. Therefore, by Lemma 1, $(\rho_d(2d) = d \text{ for } d \geq 2) \Rightarrow (\rho_d(n) \leq n - d \text{ for all } n \geq d \geq 2)$. □

Corollary 1. *We have* $\rho_d(2d) \leq d$ *for* $d \geq 2$ *implies that the maximum number* $\rho(n)$ *of runs over all strings of length* $n \geq 3$ *satisfies* $\rho(n) \leq n - 2$.

Corollary 2 illustrates that the hypothesized property that the maximal number of runs increases by at most 1 if the number of distinct symbols increases by 1, implies that the number of runs in a string is at most $n - 2$.

Corollary 2. *We have* $\rho_d(n) \leq \rho_{d+1}(n) + 1$ *for* $2d \geq n \geq d \geq 1$ *implies that* $\rho_d(n) \leq n - d$ *for* $n \geq d \geq 2$*; that is that the maximum number* $\rho(n)$ *of runs over all strings of length* $n \geq 3$ *satisfies* $\rho(n) \leq n - 2$.

Proof. Note that $\rho_d(d + 2) = 2$, and that this entry is $d - 2$ steps away from the entry for $\rho_d(2d)$ on a counter-diagonal on the Table where d is the index for the rows and $n - d$ the index for the columns, yielding that $\rho_d(2d) \leq d$. □

Proposition 3. *We have* $(\rho_d(2d) \leq d$ and $\rho_d(2d + 1) \leq d$ *for* $d \geq 2)$ *implies that a string of length* $2d$ *and containing exactly* d *distinct symbols admitting* $\rho_d(2d)$ *runs contains each symbol exactly twice.*

Proof. Note that $\rho_d(2d) \leq d$ and $\rho_d(2d+1) \leq d$ implies $\rho_d(2d) = \rho_d(2d+1) = d$ by Lemma 2 and 3. We prove by contradiction that x cannot contain a symbol occurring exactly once, and therefore all symbols should occur exactly twice. Assume $\mathtt{a}$ occurs exactly once.

(i) If $\mathtt{a}$ occurs at the first or last position, then $\boldsymbol{r}(x) = \boldsymbol{r}(y)$ where y is obtained by removing $\mathtt{a}$ from x. Since y is of length $2d-1$ and contains exactly $d-1$ distinct symbols, we have $d = \rho_d(2d) = \boldsymbol{r}(x) = \boldsymbol{r}(y) \leq \rho_{d-1}(2d-1) = d-1$ which is a contradiction.

(ii) If $\mathtt{a}$ does not occur at the first or last position, then $\boldsymbol{r}(x) = \boldsymbol{r}(x_1)+\boldsymbol{r}(x_2)$ where x_1 and x_2 are the parts of x respectively on the left and right of $\mathtt{a}$. Assume that x_1 has length n_1 and d_1 symbols, and x_2 has length n_2 and d_2 symbols. As noted in the proof of Proposition 1, $\boldsymbol{r}(x) = d$ implies that $d_1 + d_2 = d - 1$ and $\rho_{d_i}(n_i) = n_i - d_i$ which implies that $n_i \leq 2d_i$ since $\rho_{d_i}(n_i) \leq \rho_{n_i-d_i-1}(2n_i - 2d_i - 1) = n_i - d_i - 1$ for $n_i > 2d_i$, and yields the following contradiction: $2d - 1 = n_1 + n_2 \leq 2(d_1 + d_2) = 2d - 2$. □

Proposition 4. *We have* ($\rho_d(2d) \leq d$ and $\rho_d(2d+1) \leq d$ *for* $d \geq 2$) *implies*

(*a*) $\rho_d(n) = n - d$ *for* $2d \geq n \geq d \geq 2$

(*b*) $\rho_d(n) \leq n - d - 1$ *for* $n > 2d \geq 4$

Proof. Similarly to Proposition 2, (a) and (b) are direct consequence that the main diagonal of table T would maximize $\rho_d(n)$ within a column with constant $n-d$, and that the entries right above the main diagonal are larger than any entries with a lower d and same $n-d$. □

Corollary 3. *We have* ($\rho_d(2d) \leq d$ and $\rho_d(2d+1) \leq d$ *for* $d \geq 2$) *implies that the maximum number* $\rho(n)$ *of runs over all strings of length* $n \geq 5$ *satisfies* $\rho(n) \leq n - 3$.

Proposition 5. *We have* ($\rho_d(2d) \leq d$ and $\rho_d(2d+1) \leq d$ *for* $d \geq 2$) *implies that a string of length* $2d$ *and containing exactly* d *distinct symbols admitting* $\rho_d(2d)$ *runs is, up to relabelling, unique and equal to* $x^*_{d,2d} = (\mathtt{aabbcc}\ldots)$.

Proof. Let x be a string of length $2d$ and containing exactly d distinct symbols such that $\boldsymbol{r}(x) = \rho_d(2d) = d$. By Proposition 3, each symbol occur exactly twice. We prove by contradiction that each pair of symbol are adjacent. Let $\mathtt{a}$ be non-adjacent, then x can de decomposed in at most 3 parts with $\boldsymbol{r}(x) \leq \boldsymbol{r}(x_1)+\boldsymbol{r}(x_2)+\boldsymbol{r}(x_3)+1$ since, by Preliminary 1, $\mathtt{a}$ can occur in at most one run. As noted in the proof of Proposition 1, $\boldsymbol{r}(x) = d$ implies that $d_1 + d_2 + d_3 = d - 1$ and therefore $\mathtt{a}$ occur in no run which lead to the following contradiction: $d = \boldsymbol{r}(x) = n_1 + n_2 + n_3 - (d_1 + d_2 + d_3) = 2d - 2 - (d - 1) = d - 1$. □

A similar discussion shows that $\rho_d(2d) \leq d$ and $\rho_d(2d+1) \leq d$ for all $d \geq 2$ implies that, up to relabelling and reordering of the positions of the symbols occurring exactly ones, $x^*_{d,n}$ is the unique string x of length n and containing exactly d distinct symbols such that $\boldsymbol{r}(x) = \rho_d(n) = n - d$ for $2d > n \geq 4$. The entries of $\rho_d(n)$ for small n and d including the one reported in Table 1 were compiled by Andrew Baker who maintains a webpage with known entries and computationally verified that strings admitting $\rho_d(n)$ runs for small d and $n \leq 2d$ are essentially unique.

Proposition 6. *We have $\rho_d(2d) = \rho_d(2d+1)$ for $d \geq 2$ implies $\rho_d(2d) = \rho_d(2d+1) = d$ for $d \geq 2$.*

Proof. Assume $\rho_d(2d) = \rho_d(2d+1)$ for $d \geq 2$. Consider the induction hypothesis (H_d): $\rho_{d'+k}(2d'+k) = d'$ for $k \geq 0$ and $2 \leq d' \leq d$. Note that (H_d) implies that $\rho_d(2d') = \rho_d(2d'+1) = d'$ for $2 \leq d' \leq d$. Note also that the base cases (H_1) and (H_2) hold. We prove that (H_d) implies (H_{d+1}).

(i) First consider a string x containing exactly $d+1$ symbols and of length $n = 2d+2$ and satisfying $r(x) = \rho_{d+1}(2d+2)$.

(i_1) Assume that x contains a symbol occurring only once; let `a` be this symbol.

If `a` occurs at the first or last position, then $\boldsymbol{r}(x) = \boldsymbol{r}(y)$ where y is obtained by removing `a` from x. Since y is of length $2d+1$ and contains exactly d distinct symbols, we have $d+1 \leq \rho_{d+1}(2d+2) = \boldsymbol{r}(x) = \boldsymbol{r}(y) \leq \rho_d(2d+1) = d$ which is a contradiction.

If `a` does not occur at the first or last position, then $\boldsymbol{r}(x) = \boldsymbol{r}(x_1) + \boldsymbol{r}(x_2)$ where x_1 and x_2 are the parts of x respectively on the left and right of `a`. Assume that x_1 has length n_1 and d_1 symbols, and x_2 has length n_2 and d_2 symbols. We can assume that $d_i \geq 2$ as the case $d_i = 1$ is trivial. We have $\rho_{d_i}(n_i) = n_i - d_i$ if $n_i \leq 2d_i$ by (H_d). Since $n_1 + n_2 = 2d+1$ and $d_1 + d_2 \geq d$, we have $n_i - d_i \leq 2d+1-d = d+1$. Thus, for $n_i > 2d_i$, by Lemma 1, we have $\rho_{d_i}(n_i) \leq \rho_{n_i-d_i-1}(2n_i - 2d_i - 1) = n_i - d_i - 1$ by (H_d) and $\rho_d(2d) = \rho_d(2d+1)$ for $d \geq 2$. In other words, we have $\rho_{d+1}(2d+2) = r(x) \leq \rho_{d_1}(n_1) + \rho_{d_2}(n_2) \leq n_1 + n_2 - (d_1 + d_2) \leq d+1$, that is $\rho_{d+1}(2d+2) = d+1$ by Lemma 2, and thus $\rho_{d+1}(2d+2) = \rho_{d+1}(2d+3) = d+1$.

(i_2) Assume that all symbols of x occur exactly twice. The case when the pair of `a` are adjacent is trial, so we can assume non-adjacency. Then, similarly to the proof of Proposition 4, x can de decomposed in at most 3 parts with $\boldsymbol{r}(x) \leq \boldsymbol{r}(x_1) + \boldsymbol{r}(x_2) + \boldsymbol{r}(x_3) + 1$ since, by Preliminary 1, `a` can occur in at most one run. As in the previous case, we have $\boldsymbol{r}(x_i) = n_i - d_i$ if $n_i \leq 2d_i$ by (H_d). Since $n_1 + n_2 + n_3 = 2d$ and $d_1 + d_2 + d_3 \geq d$, we have $n_i - d_i \leq 2d - d = d$. Thus, if $n_i > 2d_1$, $\boldsymbol{r}(x_i) \leq \rho_{d_i}(n_i) \leq \rho_{n_i-d_i-1}(2n_i - 2d_i - 1) = n_i - d_i - 1$. In other words, we have $\rho_{d+1}(2d+2) = r(x) \leq \rho_{d_1}(n_1) + \rho_{d_2}(n_2) + \rho_{d_3}(n_3) \leq d+1$, that is $\rho_{d+1}(2d+2) = d+1$ by Lemma 2, and thus $\rho_{d+1}(2d+2) = \rho_{d+1}(2d+3) = d+1$.

(ii) to complete the proof we consider, for $k \geq 1$, a string x containing exactly $d+1+k$ symbols and of length $n = 2d+2+k$ and satisfying $r(x) = \rho_{d+1+k}(2d+2+k)$. The approach

is similar to the one used for the proof of Proposition 1. Note that $r(x) \geq d+1$ by Lemma 2. Such string x much contain at least one symbol occurring only once; let $\mathtt{a}$ be this symbol.

(ii_1) If $\mathtt{a}$ occurs at the first or last position, then $\boldsymbol{r}(x) = \boldsymbol{r}(y)$ where y is obtained by removing $\mathtt{a}$ from x. Since y is of length $n = 2d+1+k$ and contains exactly $d+k$ distinct symbols, we have $\rho_{d+1+k}(2d+2+k) = \boldsymbol{r}(x) = \boldsymbol{r}(y) \leq \rho_{d+k}(2d+1+k)$. Thus, by Lemma 1, $\rho_{d+1+k}(2d+2+k) = \rho_{d+k}(2d+1+k)$.

(ii_2) If $\mathtt{a}$ does not occur at the first or last position, then $\boldsymbol{r}(x) = \boldsymbol{r}(x_1) + \boldsymbol{r}(x_2)$ where x_1 and x_2 are the parts of x respectively on the left and right of $\mathtt{a}$. Assume that x_1 has length n_1 and d_1 symbols, and x_2 has length n_2 and d_2 symbols. We have $\boldsymbol{r}(x) = \boldsymbol{r}(x_1) + \boldsymbol{r}(x_2) \leq \rho_{d_1}(n_1) + \rho_{d_2}(n_2)$. We can assume that $d_i \geq 2$ as the case $d_i = 1$ is trivial. Assume first that $n_i \leq 2d_i$, by (H_d), we have $\rho_{d_i}(n_i) = n_i - d_i$. Since $n_1 + n_2 = 2d+1+k$ and $d_1 + d_2 \geq d+k$, we have $n_i - d_i \leq 2d+1+k-(d+k) = d+1$. Thus, for $n_i > 2d_i$, by Lemma 1, we have $\rho_{d_i}(n_i) \leq \rho_{n_i-d_i-1}(2n_i - 2d_i - 1) = n_i - d_i - 1$ by (H_d) and $\rho_d(2d) = \rho_d(2d+1)$ for $d \geq 2$. In other words, $\rho_{d+1+k}(2d+2+k) = \boldsymbol{r}(x) \leq \rho_{d_1}(n_1) + \rho_{d_2}(n_2) \leq n_1 + n_2 - (d_1 + d + 2) \leq 2d+1+k-(d+k) = d+1 = \rho_{d+1}(2d+2)$.

Items (ii_1) and (ii_2) show that $\rho_{d+1+k}(2d+2+k)$ is not increasing with k, that is by Lemma 2, $\rho_{d+1+k}(2d+2+k) = \rho_{d+1}(2d+2) = d+1$ for $k \geq 1$. □

Corollary 4. *We have $\rho_d(2d) = \rho_d(2d+1)$ for $d \geq 2$ implies that the maximum number $\rho(n)$ of runs over all strings of length $n \geq 5$ satisfies $\rho(n) \leq n-3$.*

Remark 2. *As mentioned as the beginning of Section 2, the approach taken in this paper is inspired by the Hirsch conjecture which deals with the behaviour of the function $\Delta(d,n)$, the maximum possible diameter over all d-dimensional polytopes with n facets. The Hirsch conjecture, first posed in 1957, states that $\Delta(d,n) \leq n-d$. It is known that $\Delta(d,n) \leq \Delta(d+1,n+1)$ and that $\Delta(d+k,2d+k) = \Delta(d,2d)$ for $k \geq 1$. In other words, if the entries for $\Delta(d,n)$ are listed in a table T where d is the index for the rows and $n-d$ the index for the columns, then the maximum of $\Delta(d,n)$ within a column is achieved on the main diagonal and all entries below an entry on the main diagonal are equal to that entry. The importance of the main diagonal entries is underlined by the so-called d-step conjecture stating that $\Delta(d,2d) \leq d$ for all $d \geq 2$, see [5]. Note that the d-cube has diameter d and therefore $\Delta(d,2d) \geq d$, in other words, the string $x^*_{d,2d}$ can be seen as an analogue of the d-cube.*

Remark 3. *The value of $\Delta(d,n)$ provides a lower bound for the worst case behaviour for simplex methods. The simplex and central-path following primal-dual interior point methods are currently the most computationally successful algorithms for linear optimization. The curvature of a polytope, defined as the largest possible total curvature of the associated central path, can be regarded as the continuous analogue of its diameter. Considering the largest curvature $\Lambda(d,n)$, Deza et al. [3] proved the following continuous analogue of the equivalency between the Hirsch conjecture and the d-step conjecture: if $\Lambda(d,2d) = \mathcal{O}(d)$ for all d, then $\Lambda(d,n) = \mathcal{O}(n)$.*

3 Acknowledgments

The authors would like to thanks Maxime Crochemore, Jakub Radoszewski, and Jamie Simpson for helpful suggestions and pointing out an error in a preliminary version of the paper. Thanks also to Andrew Baker who computationally checked most of the properties for small instances. This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada, MITACS, and by the Canada Research Chairs program.

References

[1] M. Crochemore: *An optimal algorithm for computing the repetitions in a word*. Information Processing Letters 12 (1981) 297–315.

[2] M. Crochemore, L. Ilie, and L. Tinta: *The "runs" conjecture*.
HP `http://www.csd.uwo.ca/~ilie/runs.html`

[3] A. Deza, T. Terlaky, and Y. Zinchenko: *A continuous d-step conjecture for polytopes*. Discrete and Computational Geometry 41 (2009) 318 327.

[4] C. S. Iliopoulos, D. Moore, and W. F. Smyth; *A characterization of the squares in a fibonacci string*. Theoretical Computer Science 172 (1997) 281–291.

[5] V. Klee and D. W. Walkup; *The d-step conjecture for polyhedra of dimension* $d < 6$. Acta Mathematica 117 (1967) 53–78.

[6] R. M. Kolpakov and G. Kucherov: *Finding maximal repetitions in a word in linear time*, In Proceedings of the 40^{th} Symposium on Foundations of Computer Science (1999) 596–604.

[7] M. G. Main: *Detecting leftmost maximal periodicities*. Discrete Applied Mathematics 25 (1989) 145–153.

[8] W. Matsubara, K. Kusano, A. Ishino, H. Bannai, and A. Shinohara: *Lower bounds for the maximum number of runs in a string*.
HP `http://www.shino.ecei.tohoku.ac.jp/runs/`

Antoine Deza and Frantisek Franek
Advanced Optimization Laboratory,
Department of Computing and Software,
McMaster University, Hamilton, Ontario, Canada.
Email: deza, franek@mcmaster.ca