Optimal Bounds for Computing α-gapped Repeats

Maxime Crochemore, Roman Kolpakov, Gregory Kucherov

Optimal bounds for computing α-gapped repeats

Maxime Crochemore[1], Roman Kolpakov[2,⋆], and Gregory Kucherov[3]

[1] King's College London, London WC2R 2LS, UK and Université Paris-Est, France,
Maxime.Crochemore@kcl.ac.uk

[2] Lomonosov Moscow State University, Leninskie Gory, Moscow, 119992 Russia,
foroman@mail.ru

[3] LIGM/CNRS, Université Paris-Est, 77454 Marne-la-Vallée, France,
Gregory.Kucherov@univ-mlv.fr

Abstract. Following (Kolpakov et al., 2013; Gawrychowski and Manea, 2015), we continue the study of α-*gapped repeats* in strings, defined as factors uvu with $|uv| \leq \alpha|u|$. Our main result is the $O(\alpha n)$ bound on the number of *maximal* α-gapped repeats in a string of length n, previously proved to be $O(\alpha^2 n)$ in (Kolpakov et al., 2013). For a closely related notion of maximal δ-subrepetition (maximal factors of exponent between $1+\delta$ and 2), our result implies the $O(n/\delta)$ bound on their number, which improves the bound of (Kolpakov et al., 2010) by a $\log n$ factor.

We also prove an algorithmic time bound $O(\alpha n + S)$ (S size of the output) for computing all maximal α-gapped repeats. Our solution, inspired by (Gawrychowski and Manea, 2015), is different from the recently published proof by (Tanimura et al., 2015) of the same bound. Together with our bound on S, this implies an $O(\alpha n)$-time algorithm for computing all maximal α-gapped repeats.

1 Introduction

Notation and basic definitions. Let $w = w[1]w[2] \dots w[n] = w[1\,..\,n]$ be an arbitrary word. The length n of w is denoted by $|w|$. For any $1 \leq i \leq j \leq n$, word $w[i] \dots w[j]$ is called a *factor* of w and is denoted by $w[i\,..\,j]$. Note that notation $w[i\,..\,j]$ denotes two entities: a word and its occurrence starting at position i in w. To underline the second meaning, we will sometimes use the term *segment*. Speaking about the equality between factors can also be ambiguous, as it may mean that the factors are identical words or identical segments. If two factors u, v are identical words, we call them *equal* and denote this by $u = v$. To express that u and v are the same segment, we use the notation $u \equiv v$. For any $i = 1 \dots n$, factor $w[1\,..\,i]$ (resp. $w[i\,..\,n]$) is a *prefix* (resp. *suffix*) of w. By *positions* on w we mean indices $1, 2, \dots, n$ of letters in w. For any factor $v \equiv w[i\,..\,j]$ of w, positions i and j are called respectively *start position* and *end position* of v and denoted by $beg(v)$ and $end(v)$ respectively. Let u, v be two factors of w.

[⋆] The author was partially supported by Russian Foundation for Fundamental Research (Grant 15-07-03102).

Factor u *is contained* in v iff $beg(v) \leq beg(u)$ and $end(u) \leq end(v)$. Letter $w[i]$ *is contained* in v iff $beg(v) \leq i \leq end(v)$.

A positive integer p is called a *period* of w if $w[i] = w[i+p]$ for each $i = 1, \dots, n-p$. We denote by $per(w)$ the *smallest period* of w and define the *exponent* of w as $exp(w) = |w|/per(w)$. A word is called *periodic* if its exponent is at least 2. Occurrences of periodic words are called *repetitions.*

Repetitions, squares, runs. Patterns in strings formed by repeated factors are of primary importance in word combinatorics [1] as well as in various applications such as string matching algorithms [2,3], molecular biology [4], or text compression [5]. The simplest and best known example of such patterns is a factor of the form uu, where u is a nonempty word. Such repetitions are called *squares.* Squares have been extensively studied. While the number of all square occurrences can be quadratic (consider word $\mathtt{a}^n$), it is known that the number of *primitively-rooted* squares is $O(n \log n)$ [3], where a square uu is primitively-rooted if the exponent of u is not an integer greater than 1. An optimal $O(n \log n)$-time algorithm for finding all primitively-rooted squares was proposed in [6].

Repetitions can be seen as a natural generalization of squares. A repetition in a given word is called *maximal* if it cannot be extended by at least one letter to the left nor to the right without changing (increasing) its minimal period. More precisely, a repetition $r \equiv w[i \,..\, j]$ in w is called *maximal* if it satisfies the following conditions:

1. $w[i-1] \neq w[i-1+per(r)]$ if $i > 1$,
2. $w[j+1-per(r)] \neq w[j+1]$ if $j < n$.

For example, word `cababaaa` has two maximal repetitions: `ababa` and `aaa`. Maximal repetitions are usually called *runs* in the literature. Since any repetition is contained in some run, the set of all runs can be considered as a compact encoding of all repetitions in the word, and can then be used to efficiently infer various useful properties related to repetitions [7]. For any word w, we denote by $\mathcal{R}(w)$ the number of maximal repetitions in w and by $\mathcal{E}(w)$ the sum of exponents of all maximal repetitions in w. Let $\mathcal{R}(n) = \max_{|w|=n} \mathcal{R}(w)$ and $\mathcal{E}(n) = \max_{|w|=n} \mathcal{E}(w)$. The following statements are proved in [8].

Theorem 1. *$\mathcal{E}(n) = O(n)$.*

Corollary 1. *$\mathcal{R}(n) = O(n)$.*

A series of papers (e.g., [9,10]) focused on more precise upper bounds on $\mathcal{E}(n)$ and $\mathcal{R}(n)$ trying to obtain the best possible constant factor behind the O-notation. A breakthrough in this direction was recently made in [11] where the so-called "runs conjecture" $\mathcal{R}(n) < n$ was proved. To the best of our knowledge, the currently best upper bound $\mathcal{R}(n) \leq \frac{22}{23}n$ on $\mathcal{R}(n)$ is shown in [12].

On the algorithmic side, an $O(n)$-time algorithm for finding all runs in a word of length n was proposed in [8] for the case of constant-size alphabet. Another $O(n)$-time algorithm, based on a different approach, has been proposed in [11].

The $O(n)$ time bound holds for the (polynomially-bounded) integer alphabet as well, see, e.g., [11]. However, for the case of unbounded-size alphabet where characters can only be tested for equality, the lower bound $\Omega(n \log n)$ on computing all runs has been known for a long time [13]. It is an interesting open question (raised over 20 years ago in [14]) whether the $O(n)$ bound holds for an unbounded linearly-ordered alphabet. Some results related to this question have recently been obtained in [15].

Gapped repeats and subrepetitions. Another natural generalization of squares are factors of the form uvu where u and v are nonempty words. We call such factors *gapped repeats*. For a gapped repeat uvu, the left (resp. right) occurrence of u is called the *left* (resp. *right*) *copy*, and v is called the *gap*. The *period* of this gapped repeat is $|u| + |v|$. For a gapped repeat π, we denote the length of copies of π by $c(\pi)$ and the period of π by $p(\pi)$. Note that a gapped repeat $\pi = uvu$ may have different periods, and $per(\pi) \le p(\pi)$. For example, in string `cabacaabaa`, segment `abacaaba` corresponds to two gapped repeats having copies `a` and `aba` and periods 7 and 5 respectively. Gapped repeats forming the same segment but having different periods are considered distinct. This means that to specify a gapped repeat it is generally not sufficient to specify its segment. If u', u'' are equal non-overlapping factors and u' occurs to the left of u'', then by (u', u'') we denote the gapped repeat with left copy u' and right copy u''. For a given gapped repeat (u', u''), equal factors $u'[i\mathinner{.\,.}j]$ and $u''[i\mathinner{.\,.}j]$, for $1 \le i \le j \le |u'|$, of the copies u', u'' are called *corresponding factors* of repeat (u', u'').

For any real $\alpha > 1$, a gapped repeat π is called *α-gapped* if $p(\pi) \le \alpha c(\pi)$. Maximality of gapped repeats is defined similarly to repetitions. A gapped repeat $(w[i'\mathinner{.\,.}j'], w[i''\mathinner{.\,.}j''])$ in w is called *maximal* if it satisfies the following conditions:

1. $w[i'-1] \ne w[i''-1]$ if $i' > 1$,
2. $w[j'+1] \ne w[j''+1]$ if $j'' < n$.

In other words, a gapped repeat π is maximal if its copies cannot be extended to the left nor to the right by at least one letter without breaking its period $p(\pi)$. As observed in [16], any α-gapped repeat is contained either in a (unique) maximal α-gapped repeat with the same period, or in a (unique) maximal repetition with a period which is a divisor of the repeat's period. For example, in the above string `cabacaabaa`, gapped repeat (`ab`)`aca`(`ab`) is contained in maximal repeat (`aba`)`ca`(`aba`) with the same period 5. In string `cabaaabaaa`, gapped repeat (`ab`)`aa`(`ab`) with period 4 is contained in maximal repetition `abaaabaaa` with period 4. Since all maximal repetitions can be computed efficiently in $O(n)$ time (see above), the problem of computing all α-gapped repeats in a word can be reduced to the problem of finding all maximal α-gapped repeats.

Several variants of the problem of computing gapped repeats have been studied earlier. In [17], it was shown that all maximal gapped repeats with a gap length belonging to a specified interval can be found in time $O(n \log n + S)$, where n is the word length and S is output size. In [18], an algorithm was proposed for finding all gapped repeats with a fixed gap length d running in time $O(n \log d + S)$. In [16], it was proved that the number of maximal α-gapped repeats in a word of length n is bounded by $O(\alpha^2 n)$ and all maximal α-gapped repeats can be found in $O(\alpha^2 n)$ time for the case of integer alphabet. A new approach to computing gapped repeats was recently proposed in [19, 20]. In particular, in [19] it is shown that the longest α-gapped repeat in a word of length n over an integer alphabet can be found in $O(\alpha n)$ time. Finally, in a recent paper [21], an algorithm is proposed for finding all maximal α-gapped repeats in $O(\alpha n + S)$ time where S is the output size, for a constant-size alphabet. The algorithm uses an approach previously introduced in [22].

Recall that repetitions are segments with exponent at least 2. Another way to approach gapped repeats is to consider segments with exponent smaller than 2, but strictly greater than 1. Clearly, such a segment corresponds to a gapped repeat $\pi = uvu$ with $per(\pi) = p(\pi) = |u|+|v|$. We will call such factors (segments) *subrepetitions*. More precisely, for any δ, $0 < \delta < 1$, by a δ-subrepetition we mean a factor v that satisfies $1 + \delta \leq exp(v) < 2$. Again, the notion of maximality straightforwardly applies to subrepetitions as well: maximal subrepetitions are defined exactly in the same way as maximal repetitions. The relationship between maximal subrepetitions and maximal gapped repeats was clarified in [16]. Directly from the definitions, a maximal subrepetition π in a string w corresponds to a maximal gapped repeat with $p(\pi) = per(\pi)$. Futhermore, a maximal δ-subrepetition corresponds to a maximal $\frac{1}{\delta}$-gapped repeat. However, there may be more maximal $\frac{1}{\delta}$-gapped repeats than maximal δ-subrepetitions, as not every maximal $\frac{1}{\delta}$-gapped repeat corresponds to a maximal δ-subrepetition.

Some combinatorial results on the number of maximal subrepetitions in a string were obtained in [23]. In particular, it was proved that the number of maximal δ-subrepetitions in a word of length n is bounded by $O(\frac{n}{\delta} \log n)$. In [16], an $O(n/\delta^2)$ bound on the number of maximal δ-subrepetitions in a word of length n was obtained. Moreover, in [16], two algorithms were proposed for finding all maximal δ-subrepetitions in the word running respectively in $O(\frac{n \log \log n}{\delta^2})$ time and in $O(n \log n + \frac{n}{\delta^2} \log \frac{1}{\delta})$ expected time, over the integer alphabet. In [22], it is shown that all subrepetitions with the largest exponent (over all subrepetitions) can be found in an overlap-free string in time $O(n)$, for a constant-size alphabet.

Our results. In the present work we improve the results of [16] on maximal gapped repeats: we prove an asymptotically tight bound of $O(\alpha n)$ on the number of maximal α-gapped repeats in a word of length n (Section 2). From our bound, we also derive an $O(n/\delta)$ bound on the number of maximal δ-subrepetitions occurring in a word, which improves the bound of [23] by a $\log n$ factor. Then, based on the algorithm of [19], we obtain an asymptotically optimal $O(\alpha n)$ time bound for computing all maximal α-gapped repeats in a word (Section 3). Note that this bound follows from the recently published paper [21] that presents an $O(\alpha n + S)$ algorithm for computing all maximal α-gapped repeats. In this work, we present an alternative algorithm with the same bound that we obtained independently.

2 Number of maximal repeats and subrepetitions

In this section, we obtain an improved upper bound on the number of maximal gapped repeats and subrepetitions in a string w. Following the general approach of [16], we split all maximal gapped repeats into three categories according to periodicity properties of repeat's copy: periodic, semiperiodic and ordinary repeats. Bounds for periodic and semiperiodic repeats are directly borrowed from [16], while for ordinary repeats, we obtain a better bound.

Periodic repeats. We say that a maximal gapped repeat is *periodic* if its copies are periodic strings (i.e. of exponent at least 2). The set of all periodic maximal α-gapped repeats in w is denoted by $\mathcal{PP}_\alpha$. The following bound on the size of $\mathcal{PP}_\alpha$ was been obtained in [16, Corollary 6].

Lemma 1. *$|\mathcal{PP}_k| = O(kn)$ for any natural $k > 1$.*

Semiperiodic repeats. A maximal gapped repeat is called *prefix* (*suffix*) *semiperiodic* if the copies of this repeat are not periodic, but have a prefix (suffix) which is periodic and its length is at least half of the copy length. A maximal gapped repeat is *semiperiodic* if it is either prefix or suffix semiperiodic. The set of all semiperiodic α-gapped maximal repeats is denoted by $\mathcal{SP}_\alpha$. In [16, Corollary 8], the following bound was obtained on the number of semiperiodic maximal α-gapped repeats.

Lemma 2 ([16]). *$|\mathcal{SP}_k| = O(kn)$ for any natural $k > 1$.*

Ordinary repeats. Maximal gapped repeats which are neither periodic nor semiperiodic are called *ordinary.* The set of all ordinary maximal α-gapped repeats in the word w is denoted by $\mathcal{OP}_\alpha$. In the rest of this section, we prove that the cardinality of $\mathcal{OP}_\alpha$ is $O(\alpha n)$. For simplicity, assume that α is an integer number k.

To estimate the number of ordinary maximal k-gapped repeats, we use the following idea from [24]. We represent a maximal repeat $\pi \equiv (u', u'')$ from $\mathcal{OP}_k$ by a triple (i, j, c) where $i = beg(u')$, $j = beg(u')$ and $c = c(\pi) = |u'| = |u''|$. Such triples will be called *points.* Obviously, π is uniquely defined by values i, j and c, therefore two different repeats from $\mathcal{OP}_k$ can not be represented by the same point.

For any two points (i', j', c'), (i'', j'', c'') we say that point (i', j', c') *covers* point (i'', j'', c'') if $i' \le i'' \le i' + c'/6$, $j' \le j'' \le j' + c'/6$, $c' \ge c'' \ge \frac{2c'}{3}$. A point is *covered* by a repeat π if this it is covered by the point representing π. By $V[\pi]$ we denote the set of all points covered by a repeat π. We show that any point can not be covered by two different repeats from $\mathcal{OP}_k$.

Lemma 3. *Two different repeats from $\mathcal{OP}_k$ cannot cover the same point.*

Proof. Let $\pi_1 \equiv (u_1', u_1'')$, $\pi_2 \equiv (u_2', u_2'')$ be two different repeats from $\mathcal{OP}_k$ covering the same point (i, j, c). Denote $c_1 = c(\pi_1)$, $c_2 = c(\pi_2)$, $p_1 = \mathit{per}(\pi_1)$, $p_2 = \mathit{per}(\pi_2)$. Without loss of generality we assume $c_1 \geq c_2$. From $c_1 \geq c \geq \frac{2c_1}{3}$, $c_2 \geq c \geq \frac{2c_2}{3}$ we have $c_1 \geq c_2 \geq \frac{2c_1}{3}$, i.e. $c_2 \leq c_1 \leq \frac{3c_2}{2}$. Note that $w[i]$ is contained in both left copies u_1', u_2', i.e. these copies overlap. If $p_1 = p_2$, then repeats π_1 and π_2 must coincide due to the maximality of these repeats. Thus, $p_1 \neq p_2$. Denote $\Delta = |p_1 - p_2| > 0$. From $\mathit{beg}(u_1') \leq i \leq \mathit{beg}(u_1') + c_1/6$ and $\mathit{beg}(u_1'') \leq j \leq \mathit{beg}(u_1'') + c_1/6$ we have

$$(j - i) - c_1/6 \leq p_1 \leq (j - i) + c_1/6.$$

Analogously, we have

$$(j - i) - c_2/6 \leq p_2 \leq (j - i) + c_2/6.$$

Thus $\Delta \leq (c_1 + c_2)/6$ which, together with inequality $c_1 \leq \frac{3c_2}{2}$, implies $\Delta \leq \frac{5c_2}{12}$.

First consider the case when one of the copies u_1', u_2' is contained in the other, i.e. u_2' is contained in u_1'. In this case, u_1'' contains some factor $\widehat{u}_2''$ corresponding to the factor u_2' in u_1'. Since $\mathit{beg}(u_2'') - \mathit{beg}(u_2') = p_2$, $\mathit{beg}(\widehat{u}_2'') - \mathit{beg}(u_2') = p_1$ and $u_2'' = \widehat{u}_2'' = u_2'$, we have

$$|\mathit{beg}(u_2'') - \mathit{beg}(\widehat{u}_2'')| = \Delta,$$

so Δ is a period of u_2'' such that $\Delta \leq \frac{5}{12}c_2 = \frac{5}{12}|u_2''|$. Thus, u_2'' is periodic which contradicts that π_2 is not periodic.

Now consider the case when u_1', u_2' are not contained in one another. Denote by z' the overlap of u_1' and u_2'. Let z' be a suffix of u_k' and a prefix of u_l' where $k, l = 1, 2$, $k \neq l$. Then u_k'' contains a suffix z'' corresponding to the suffix z' in u_k', and u_l'' contains a prefix $\widehat{z}''$ corresponding to the prefix z' in u_l'. Since $\mathit{beg}(z'') - \mathit{beg}(z') = p_k$ and $\mathit{beg}(\widehat{z}'') - \mathit{beg}(z') = p_l$ and $z'' = \widehat{z}'' = z'$, we have

$$|\mathit{beg}(z'') - \mathit{beg}(\widehat{z}'')| = |p_k - p_l| = \Delta,$$

therefore Δ is a period of z'. Note that in this case

$$\mathit{beg}(u_k') < \mathit{beg}(u_l') \leq i \leq \mathit{beg}(u_k') + c_k/6,$$

therefore $0 < \mathit{beg}(u_l') - \mathit{beg}(u_k') \leq c_k/6$. Thus

$$|z'| = c_k - (\mathit{beg}(u_l') - \mathit{beg}(u_k')) \geq \frac{5}{6}c_k \geq \frac{5}{6}c_2.$$

From $\Delta \leq \frac{5}{12}c_2$ and $c_2 \leq \frac{6}{5}|z'|$ we obtain $\Delta \leq |z'|/2$. Thus, z' is a periodic suffix of u_k' such that $|z'| \geq \frac{5}{6}|u_k'|$, i.e. π_k is either suffix semiperiodic or periodic which contradicts $\pi_k \in \mathcal{OP}_k$.

Denote by $\mathcal{Q}_k$ the set of all points (i, j, c) such that $1 \leq i, j, c \leq n$ and $i < j \leq i + (\frac{3}{2}k + \frac{1}{4})c$.

Lemma 4. *Any point covered by a repeat from $\mathcal{OP}_k$ belongs to $\mathcal{Q}_k$.*

Proof. Let a point (i, j, c) be covered by some repeat $\pi \equiv (u', u'')$ from $\mathcal{OP}_k$. Denote $c' = c(\pi)$. Note that $w[i]$ and $w[j]$ are contained respectively in u' and u'' and $n > c' \geq c \geq \frac{2c'}{3} > 0$, so inequalities $1 \leq i, j, c \leq n$ and $i < j$ are obvious. Note also that

$$j \leq beg(u'') + c'/6 = beg(u') + per(\pi) + c'/6 \leq i + kc' + c'/6,$$

therefore, taking into account $c' \leq \frac{3c}{2}$, we have $j \leq i + (\frac{3}{2}k + \frac{1}{4})c$.

From Lemmas 3 and 4, we obtain

Lemma 5. *$|\mathcal{OP}_k| = O(nk)$.*

Proof. Assign to each point (i, j, c) the weight $\rho(i, j, c) = 1/c^3$. For any finite set A of points, we define

$$\rho(A) = \sum_{(i,j,c)\in A} \rho(i, j, c) = \sum_{(i,j,c)\in A} \frac{1}{c^3}.$$

Let π be an arbitrary repeat from $\mathcal{OP}_k$ represented by a point (i', j', c'). Then

$$\rho(V[\pi]) = \sum_{i'\leq i\leq i'+c'/6} \; \sum_{j'\leq j\leq j'+c'/6} \; \sum_{2c'/3\leq c\leq c'} \frac{1}{c^3}$$
$$> \frac{c'^2}{36} \sum_{2c'/3\leq c\leq c'} \frac{1}{c^3}.$$

Using a standard estimation of sums by integrals, one can deduce that $\sum_{2c'/3\leq c\leq c'} \frac{1}{c^3} \geq \frac{5}{32}\frac{1}{c'^2}$ for any c'. Thus, for any π from $\mathcal{OP}_k$

$$\rho(V[\pi]) > \frac{1}{36}\frac{5}{36} = \Omega(1).$$

Therefore,

$$\sum_{\pi\in\mathcal{OP}_k} \rho(V[\pi]) = \Omega(|\mathcal{OP}_k|). \tag{1}$$

Note also that

$$\rho(\mathcal{Q}_k) \leq \sum_{i=1}^{n} \; \sum_{i<j\leq i+(\frac{3}{2}k+\frac{1}{4})c} \; \sum_{c=1}^{n} \frac{1}{c^3}$$
$$< n(\frac{3}{2}k + \frac{1}{4})c \sum_{c=1}^{n} \frac{1}{c^3} < 2nk \sum_{c=1}^{n} \frac{1}{c^2} < 2nk \sum_{c=1}^{\infty} \frac{1}{c^2} = \frac{nk\pi^2}{3}.$$

Thus,

$$\rho(\mathcal{Q}_k) = O(nk). \tag{2}$$

By Lemma 4, any point covered by repeats from $\mathcal{OP}_k$ belongs to $\mathcal{Q}_k$. On the other hand, by Lemma 3, each point of $\mathcal{Q}_k$ can not be covered by two repeats from $\mathcal{OP}_k$. Therefore,

$$\sum_{\pi \in \mathcal{OP}_k} \rho(V[\pi]) \leq \rho(\mathcal{Q}_k).$$

Thus, using 1 and 2, we conclude that $|\mathcal{OP}_k| = O(nk)$.

Putting together Lemma 1, Lemma 2, and Lemma 5, we obtain that for any integer $k \geq 2$, the number of maximal k-gapped repeats in w is $O(nk)$. The bound straightforwardly generalizes to the case of real $\alpha > 1$. Thus, we conclude with

Theorem 2. *For any $\alpha > 1$, the number of maximal α-gapped repeats in w is $O(\alpha n)$.*

Note that the bound of Theorem 2 is asymptotically tight. To see this, it is enough to consider word $w_k = (0110)^k$. It is easy to check that for a big enough α and $k = \Omega(\alpha)$, w_k contains $\Theta(\alpha|w_k|)$ maximal α-gapped repeats whose copies are single-letter words.

We now use Theorem 2 to obtain an upper bound on the number of maximal δ-subrepetitions. The following proposition, shown in [16, Proposition 3], follows from the fact that each maximal δ-subrepetition defines at least one maximal $1/\delta$-gapped repeat (cf. Introduction).

Proposition 1 ([16]). *For $0 < \delta < 1$, the number of maximal δ-subrepetitions in a string is no more then the number of maximal $1/\delta$-gapped repeats.*

Theorem 2 combined with Proposition 1 immediately imply the following upper bound for maximal δ-subrepetitions that improves the bound of [23] by a $\log n$ factor.

Theorem 3. *For $0 < \delta < 1$, the number of maximal δ-subrepetitions in w is $O(n/\delta)$.*

The $O(n/\delta)$ bound on the number of maximal δ-subrepetitions is asymptotically tight, at least on an unbounded alphabet : word $\mathtt{ab}_1\mathtt{ab}_2 \ldots \mathtt{ab}_k$ contains $\Omega(n/\delta)$ maximal δ-subrepetitions for $\delta \leq 1/2$.

3 Computing all maximal α-gapped repeats

We now turn to the algorithmic question how to efficiently compute all maximal α-gapped repeats in a given word. Recall (cf Introduction) that an algorithm with running time $O(\alpha^2 n + S)$ has been proposed in [16] for this problem, which becomes $O(\alpha^2 n)$-time taken into account the bound on S. On the other hand, it was shown in [19] that computing the *longest* α-gapped repeat can be done in time $O(\alpha n)$. It is therefore a natural question whether all maximal α-gapped repeats can be computed in time $O(\alpha n + S)$. Here we answer this question positively. Together with the the $S = O(\alpha n)$ bound of Theorem 2, this implies the following result.

Theorem 4. *For a fixed $\alpha > 1$, all maximal α-gapped repeats in a word of length n over a constant alphabet can be computed in $O(\alpha n)$ time.*

The proof of Theorem 4 can be found in the full version of this work [25]. It is based on a case analysis and uses ideas of [19].

We note that independently of our work, another $O(\alpha n + S)$-time algorithm for computing all maximal α-gapped repeats has been recently announced in [21].

Note that, as mentioned earlier, a word can contain $\Theta(\alpha n)$ maximal α-gapped repeats, and therefore the $O(\alpha n)$ time bound stated in Theorem 4 is asymptotically optimal.

4 Concluding remarks

In this work, we proved the tight $O(\alpha n)$ bound on the number of maximal α-gapped repeats in a word. We note that while submitting this paper, manuscript [26] appeared that proves that the number of maximal α-gapped repeats is bounded by $18\alpha n$. From our bound, we obtain an $O(n/\delta)$ bound on the number of maximal δ-subrepetitions in a word, which improves the bound of [23] by a $\log n$ factor. We also presented an $O(\alpha n)$-time algorithm (obtained independently from [21]) for computing all maximal α-gapped repeat in a word.

Besides gapped repeats we can also consider gapped palindromes which are factors of the form uvu^R, where u and v are nonempty words and u^R is the reversal of u [27]. A gapped palindrome uvu^R in a word w is called *maximal* if $w[end(u) + 1] \neq w[beg(u^R) - 1]$ and $w[beg(u) - 1] \neq w[end(u^R) + 1]$ for $beg(u) > 1$ and $end(u^R) < |w|$. A maximal gapped palindrome uvu^R is α-gapped if $|u| + |v| \leq \alpha|u|$ [19]. It can be shown analogously to the results of this paper that for $\alpha > 1$ the number of maximal α-gapped palindromes in a word of length n is bounded by $O(\alpha n)$ and for the case of constant alphabet, all these palindromes can be found in $O(\alpha n)$ time[4].

In this paper, we consider maximal α-gapped repeats with $\alpha > 1$. However, this notion can be formally generalized to the case of $\alpha \leq 1$. In particular, maximal 1-gapped repeats are maximal repeats whose copies are adjacent or overlapping. It is easy to see that such repeats form runs whose minimal periods are divisors of the periods of these repeats. Moreover, each run in a word is formed by at least one maximal 1-gapped repeat, therefore the number of runs in a word is not greater than the number of maximal 1-gapped repeats. More precisely, each run r is formed by $\lfloor exp(r)/2 \rfloor$ distinct maximal 1-gapped repeats. Thus, if a word contains runs with exponent greater than or equal to 4 then the number of maximal 1-gapped repeats is strictly greater than the number

[4] Note that in [19], the number of maximal α-gapped palindromes was conjectured to be $O(\alpha^2 n)$.

of runs. However, using an easy modification of the proof of “runs conjecture” from [11], it can be also proved the number of maximal 1-gapped repeats in a word is strictly less than the length of the word. Moreover, denoting by $\mathcal{R}_1(n)$ the maximal possible number of maximal 1-gapped repeats in words of length n, we conjecture that $\mathcal{R}(n) = \mathcal{R}_1(n)$ since known words with a large number of runs have no runs with big exponents. We can also consider the case of $\alpha < 1$ for repeats with overlapping copies and, in particular, the case of maximal $1/k$-gapped repeats where k is integer greater than 1. It is easy to see that such repeats form runs with exponents greater than or equal to $k + 1$. It is known from [11, Theorem 11] that the number of such runs in a word of length n is less than n/k, and it seems to be possible to modify the proof of this fact to prove that the number of maximal $1/k$-gapped repeats in the word is also less than $n/k = \alpha n$. These observations together with results of computer experiments for the case of $\alpha > 1$ leads to a conjecture that for any $\alpha > 0$, the number maximal α-gapped repeats in a word of length n is actually less than αn. This generalization of the “runs conjecture” constitutes an interesting open problem. Another interesting open question is whether the obtained $O(n/\delta)$ bound on the number of maximal δ-subrepetitions is asymptotically tight for the case of constant alphabet.

References

1. Lothaire, M.: Combinatorics on Words. Addison Wesley (1983)
2. Galil, Z., Seiferas, J.I.: Time-space-optimal string matching. J. Comput. Syst. Sci. **26**(3) (1983) 280–294
3. Crochemore, M., Rytter, W.: Sqares, cubes, and time-space efficient string searching. Algorithmica **13**(5) (1995) 405–425
4. Gusfield, D.: Algorithms on Strings, Trees, and Sequences - Computer Science and Computational Biology. Cambridge University Press (1997)
5. Storer, J.A.: Data Compression: Methods and Theory. Computer Science Press (1988)
6. Crochemore, M.: An optimal algorithm for computing the repetitions in a word. Inf. Process. Lett. **12**(5) (1981) 244–250
7. Crochemore, M., Iliopoulos, C.S., Kubica, M., Radoszewski, J., Rytter, W., Walen, T.: Extracting powers and periods in a string from its runs structure. In Chávez, E., Lonardi, S., eds.: String Processing and Information Retrieval - 17th International Symposium, SPIRE 2010, Los Cabos, Mexico, October 11-13, 2010. Proceedings. Volume 6393 of Lecture Notes in Computer Science., Springer (2010) 258–269
8. Kolpakov, R., Kucherov, G.: On maximal repetitions in words. J. Discrete Algorithms **1**(1) (2000) 159–186
9. Crochemore, M., Ilie, L., Tinta, L.: Towards a solution to the ”runs” conjecture. In Ferragina, P., Landau, G.M., eds.: Combinatorial Pattern Matching, 19th Annual Symposium, CPM 2008, Pisa, Italy, June 18-20, 2008, Proceedings. Volume 5029 of Lecture Notes in Computer Science., Springer (2008) 290–302
10. Crochemore, M., Kubica, M., Radoszewski, J., Rytter, W., Walen, T.: On the maximal sum of exponents of runs in a string. J. Discrete Algorithms **14** (2012) 29–36

11. Bannai, H., I, T., Inenaga, S., Nakashima, Y., Takeda, M., Tsuruta, K.: A new characterization of maximal repetitions by Lyndon trees. CoRR **abs/1406.0263** (2014) intermediate version presented to SODA'2015.
12. Fischer, J., Holub, S., I, T., Lewenstein, M.: Beyond the runs theorem. CoRR **abs/1502.04644** (2015)
13. Main, M., Lorentz, R.: An $O(n \log n)$ algorithm for finding all repetitions in a string. J. of Algorithms **5**(3) (1984) 422–432
14. Breslauer, D.: Efficient string algorithmics. PhD thesis, Columbia University (1992)
15. Kosolobov, D.: Lempel-Ziv factorization may be harder than computing all runs. In Mayr, E.W., Ollinger, N., eds.: 32nd International Symposium on Theoretical Aspects of Computer Science, STACS 2015, March 4-7, 2015, Garching, Germany. Volume 30 of LIPIcs., Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik (2015) 582–593
16. Kolpakov, R., Podolskiy, M., Posypkin, M., Khrapov, N.: Searching of gapped repeats and subrepetitions in a word. CoRR **abs/1309.4055** (2013) presented to CPM'2014.
17. Brodal, G.S., Lyngs, R.B., Pedersen, C.N.S., Stoye, J.: Finding maximal pairs with bounded gap. J. Discrete Algorithms **1**(1) (2000) 77–104
18. Kolpakov, R.M., Kucherov, G.: Finding repeats with fixed gap. In: SPIRE. (2000) 162–168
19. Gawrychowski, P., Manea, F.: Longest α-gapped repeat and palindrome. In Kosowski, A., Walukiewicz, I., eds.: Fundamentals of Computation Theory - 20th International Symposium, FCT 2015, Gdańsk, Poland, August 17-19, 2015, Proceedings. Volume 9210 of Lecture Notes in Computer Science., Springer (2015) 27–40
20. Dumitran, M., Manea, F.: Longest gapped repeats and palindromes. In Italiano, G.F., Pighizzini, G., Sannella, D., eds.: Mathematical Foundations of Computer Science 2015 - 40th International Symposium, MFCS 2015, Milan, Italy, August 24-28, 2015, Proceedings, Part I. Volume 9234 of Lecture Notes in Computer Science., Springer (2015) 205–217
21. Tanimura, Y., Fujishige, Y., I, T., Inenaga, S., Bannai, H., Takeda, M.: A faster algorithm for computing maximal α-gapped repeats in a string. In Iliopoulos, C.S., Puglisi, S.J., Yilmaz, E., eds.: String Processing and Information Retrieval - 22nd International Symposium, SPIRE 2015, London, UK, September 1-4, 2015, Proceedings. Volume 9309 of Lecture Notes in Computer Science., Springer (2015) 124–136
22. Badkobeh, G., Crochemore, M., Toopsuwan, C.: Computing the maximal-exponent repeats of an overlap-free string in linear time. In Calderón-Benavides, L., González-Caro, C.N., Chávez, E., Ziviani, N., eds.: String Processing and Information Retrieval - 19th International Symposium, SPIRE 2012, Cartagena de Indias, Colombia, October 21-25, 2012. Proceedings. Volume 7608 of Lecture Notes in Computer Science., Springer (2012) 61–72
23. Kolpakov, R., Kucherov, G., Ochem, P.: On maximal repetitions of arbitrary exponent. Inf. Process. Lett. **110**(7) (2010) 252–256
24. Kolpakov, R.: On primary and secondary repetitions in words. Theor. Comput. Sci. **418** (2012) 71–81
25. Crochemore, M., Kolpakov, R., Kucherov, G.: Optimal searching of gapped repeats in a word. CoRR **abs/1509.01221** (2015)
26. Gawrychowski, P., I, T., Inenaga, S., Köppl, D., Manea, F.: Efficiently finding all maximal $\$\alpha\$$-gapped repeats. CoRR **abs/1509.09237** (2015)

27. Kolpakov, R., Kucherov, G.: Searching for gapped palindromes. Theor. Comput. Sci. **410**(51) (2009) 5365–5373