

Fine and Wilf's Theorem for Abelian Periods in Partial Words^{*}

F. Blanchet-Sadri¹, Amelia Tebbe², and Amy Veprauskas³

¹ Department of Computer Science, University of North Carolina,
P.O. Box 26170, Greensboro, NC 27402-6170, USA, blanchet@uncg.edu

² Department of Mathematics, St. Mary's College of Maryland,
18952 E. Fisher Rd., St. Mary's City, MD 20686-3001, USA

³ Department of Mathematics, Bryn Mawr College,
101 North Merion Avenue, Box C-775, Bryn Mawr, PA 19010-2899, USA

Abstract. Recently, Constantinescu and Ilie proved a variant of the well-known periodicity theorem of Fine and Wilf in the case of two relatively prime abelian periods, and conjectured a result for the case of two non-relatively prime abelian periods. More precisely, they proved that any full word having two coprime abelian periods p, q and length at least $2pq - 1$ has also $\gcd(p, q) = 1$ as a period. In this paper, we answer some open problems they suggested by proving that the length $2pq - 1$ is optimal and by answering affirmatively their conjecture. We also extend their study in the context of partial words, giving optimal lengths and describing an algorithm for constructing optimal words.

1 Introduction

Computing periods in words has important applications in data compression, string searching, and pattern matching algorithms. The notion of *period* is central in combinatorics on words. Although there are many fundamental results on periods of words, the one of Fine and Wilf is perhaps the best known [9]. It states that any word having two periods p, q and length at least $p + q - \gcd(p, q)$ also has the greatest common divisor of p and q , $\gcd(p, q)$, as a period. The length $p + q - \gcd(p, q)$ is *optimal* since there are examples of words with shorter length that have periods p and q but are not $\gcd(p, q)$ -periodic [5]. Extensions of Fine and Wilf's result to more than two periods are given in [4, 6, 12, 16]. In particular, Constantinescu and Ilie [6] extend Fine and Wilf's result to words having an arbitrary number of periods and prove that their lengths are optimal.

Fine and Wilf's periodicity theorem has been generalized to *partial words*, or finite sequences of symbols over a finite alphabet that may have some don't

^{*} This material is based upon work supported by the National Science Foundation under Grant No. DMS-0754154. The Department of Defense is also gratefully acknowledged. The authors would also like to thank Sean Simmons from the Department of Mathematics of the University of Texas at Austin for an insightful suggestion in proving Constantinescu and Ilie's conjecture.

care symbols or holes [1, 2, 14, 15]. Partial words are also known as spaced seeds. In [10], Ilie and Ilie give a polynomial-time heuristic algorithm to compute good (multiple) spaced seeds which are used for producing high quality sequence alignments. In [11], they give a polynomial-time heuristic algorithm to compute good neighbor seeds, an important class of spaced seeds, which save space requirements. The two main ideas is to replace sensitivity by overlap complexity, and to avoid exponential trials by swapping the matches and the don't care symbols. These ideas lead to very simple implementation, improved sensitivity, and speed several orders of magnitude faster than all previous algorithms.

Erdős raised the question whether there exist infinite abelian square-free words over a given alphabet (words in which no two adjacent subwords are permutations of each other) [8]. Infinite abelian square-free words have been constructed over alphabets of sizes as small as four [13]. In [3], Blanchet-Sadri et al. investigate the problem of avoiding abelian squares in partial words. In particular, they give lower and upper bounds for the number of letters needed to construct infinite abelian square-free partial words with finitely or infinitely many holes. In the case of one hole, they prove that the minimal alphabet size is four, while in the case of more than one hole, they prove that it is five.

The notion of *abelian period*, a generalization of the one of period (see Definition 1), is closely related with that of abelian repetition. Let $A = \{a_1, a_2, \dots, a_k\}$ be an alphabet. The number of occurrences of the letter $a_i \in A$ in the word w over A is denoted by $|w|_{a_i}$. The length of w is $|w| = \sum_{1 \leq i \leq k} |w|_{a_i}$. The *Parikh vector* of w is $\|w\| = (|w|_{a_1}, |w|_{a_2}, \dots, |w|_{a_k})$. Note that, for two words u and v , $\|u\| = \|v\|$ means that u is a permutation of v , and $\|u\| \leq \|v\|$ means that u can be obtained from v by permuting and, possibly, deleting some of its letters.

Definition 1. [7] *A word w over an alphabet A has abelian period p if $w = u_0 u_1 \dots u_m u_{m+1}$, where $m \geq 1$ and $|u_1| = |u_2| = \dots = |u_m| = p$ and $\|u_0\| \leq \|u_1\| = \|u_2\| = \dots = \|u_m\| \geq \|u_{m+1}\|$.*

For example, the word *bbaaabaaaabaaba* has abelian period 4, as it can be factorized as *b.baaa.baaa.abaa.ba*.

Constantinescu and Ilie prove a variant of Fine and Wilf's theorem in the case of two relatively prime abelian periods, while they conjecture a result for the case of non-relatively prime abelian periods [7]. More precisely, they prove that any word having two coprime periods p, q and length at least $2pq - 1$ has also $\gcd(p, q) = 1$ as a period. Among a number of problems they suggest, we investigate the following: (1) Is the length $2pq - 1$ optimal? (2) Find out what is imposed by two non-relatively prime abelian periods p and q . In particular, is it true that $\gcd(p, q) = d$ implies that the word has at most cardinality d (or the word contains at most d distinct letters)?

In this paper, we answer Problems (1) and (2) affirmatively. We also extend Constantinescu and Ilie's result in the context of partial words, giving optimal lengths and describing an algorithm for constructing optimal partial words (a partial word w with h holes and having abelian periods p, q is optimal if the length of w is one less than the optimal length for the parameters h, p and q , and

the cardinality of w is $\gcd(p, q) + 1$). In addition, we have created a World Wide Web server interface which is located at www.uncg.edu/cmp/research/finewilf6 for automated use of a program which constructs an optimal partial word with abelian periods p, q and h holes. For p and q with $\gcd(p, q) > 1$, the program produces an optimal partial word for the case where the periods “match up.”

We end this section by reviewing basic definitions on partial words. An *alphabet* A is a non-empty finite set of letters. A *partial word* over A is a finite sequence over the augmented alphabet $A_\diamond = A \cup \{\diamond\}$, where $\diamond \notin A$ plays the role of a don't care symbol or hole. More precisely, a partial word u of length n (or $|u|$) over A is a function $u : \{0, \dots, n-1\} \rightarrow A_\diamond$. For $0 \leq i < n$, if $u(i) \in A$, then i belongs to the *domain* of u , denoted by $i \in D(u)$, and if $u(i) = \diamond$, then i belongs to the *set of holes* of u , denoted by $i \in H(u)$. We refer to a partial word with an empty set of holes as a (*full*) word. The *empty* partial word is the sequence of length zero and is denoted by ε . The set of all full (respectively, partial) words over A of finite length is denoted by A^* (respectively, A_\diamond^*). For any partial word u , $u[i..j)$ is the *factor* of u that starts at position i and ends at position $j-1$. In particular, $u[0..j)$ is the *prefix* of u of length j , and $u[|u|-j..|u|)$ is the *suffix* of u of length j . A *period* of u is a positive integer p such that $u(i) = u(j)$ whenever $i, j \in D(u)$ and $i \equiv j \pmod p$ (in such a case, u is *p-periodic*). If u and v are two partial words of equal length, then u is *contained in* v , denoted by $u \subset v$, if $u(i) = v(i)$ for all $i \in D(u)$. The partial words u and v are *compatible*, denoted by $u \uparrow v$, if there exists a partial word w such that $u \subset w$ and $v \subset w$.

2 Relatively Prime Abelian Periods

Constantinescu and Ilie's result is stated as follows.

Theorem 1. [7] *If a word w has abelian periods p and q which are relatively prime and $|w| \geq 2pq - 1$, then w has period 1.*

Constantinescu and Ilie proved that the length $2pq - 1$ is an upper bound but they did not prove that it is optimal. But it is! Indeed, in Section 4 we give an algorithm for constructing non-unary words of length $2pq - 2$ that have abelian periods p and q for any coprime positive integers p, q .

Here, we repeat Constantinescu and Ilie's proof from [7] since it contains the ideas that we will use later for our own results. To prove their theorem, they first calculate how many letters in a word w with abelian periods p and q , where p and q are relatively prime and $p < q$, are needed for the two periods to first match up. For convenience, we adopt their notation. If $u_0u_1 \dots u_{m+1}$ and $v_0v_1 \dots v_{n+1}$ are factorizations of w into abelian periods p and q , respectively, they calculate how many letters are needed for the equality $u_0u_1 \dots u_i = v_0v_1 \dots v_j$ to hold for some integers i, j . They conclude that the periods match up at or before $pq - 1$ letters. After the first matching, all other matchings occur pq letters after the previous one. So, a word of length $2pq - 1$ or greater has at least two matchings.

To calculate this they first write each v_i , $1 \leq i \leq n$, in terms of u 's. They set $v_i = x_iu_{b_i+1}u_{b_i+2} \dots u_{b_{i+1}-1}y_i$, where x_i is a suffix of u_{b_i} , y_i is a prefix of $u_{b_{i+1}}$,

and u_{b_i} is the first u such that $|u_0u_1 \cdots u_{b_i}| \geq |v_0v_1 \cdots v_{i-1}|$. So, by definition, $|x_i| < p$. Both $|x_i| + |y_i| \equiv q \pmod p$ and $|x_{i+1}| + |y_i| = p$ hold. Subtracting the first from the second we get $|x_{i+1}| \equiv |x_i| - q \pmod p$ and, by induction on r , $r \geq 1$, we obtain $|x_{i+r-1}| \equiv |x_i| - (r-1)q \pmod p$. In the case where $i = 1$ we get $|x_r| \equiv |x_1| - (r-1)q \pmod p$. Letting $r = ((|x_1| (q^{-1} \pmod p)) \pmod p) + 1$ we obtain $|x_r| \equiv 0 \pmod p$. So $x_r = \varepsilon$ and $r \leq p$.

Hence $v_0v_1 \cdots v_{r-1} = u_0u_1 \cdots u_{b_r}$ where $|v_0v_1 \cdots v_{r-1}| = |v_0| + (r-1)q$ and $1 \leq |v_0| \leq q$. Since $r \leq p$ we get $|v_0| + (r-1)q \leq pq$. However, if $|v_0| = q$ then $|x_{r-1}| \equiv |x_0| - (r-1)q \pmod p$ which implies $|x_{r-1}| \equiv 0 \pmod p$ and so $x_{r-1} = \varepsilon$. This means that $v_0v_1 \cdots v_{r-2} = u_0u_1 \cdots u_{b_r}$ and $|v_0v_1 \cdots v_{r-2}| = |v_0| + (r-2)q \leq q(p-1)$. So, if $|v_0| = q$ we obtain the equality q letters sooner and so the value is largest when $|v_0| \neq q$. This implies, however, that $|v_0| + (r-1)q \leq pq - 1$. So the first matching occurs at or before $pq - 1$ letters. Note that the first matching occurs at exactly $pq - 1$ letters when $|u_0| = p - 1$ and $|v_0| = q - 1$.

For any integers i and j , $1 \leq i \leq n$, $1 \leq j \leq m$, $\alpha = \|v_i\|$ and $\beta = \|u_j\|$ have the same non-zero components. Further, since there are q letters in v_i , the sum of the non-zero components in α is equal to q . Denote α_l (respectively, β_l) to be the number of times the letter a_l occurs within one abelian q (respectively, p) period. Now the number of times a_l occurs in the subword $v_rv_{r+1} \cdots v_{r+p-1} = u_{b_r+1}u_{b_r+2} \cdots u_{b_r+q}$, which is the subword between the first two matchings of p and q , is $\alpha_l p = \beta_l q$ times. Combining these facts, if α has more than one non-zero component, then some component, say α_i , is less than q . So $\frac{p}{q} = \frac{\beta_i}{\alpha_i}$ which implies that $\frac{p}{q}$ is reducible, a contradiction. Hence α can only contain one non-zero component, so w has period 1.

Using similar logic, we examine the case when w is a partial word over $A = \{a_1, \dots, a_k\}$, where the number of occurrences of a_i in w is denoted by $|w|_{a_i}$, while the *Parikh vector* of w by $\|w\| = (|w|_{a_1}, \dots, |w|_{a_k})$.

Definition 2. A partial word w over an alphabet A has abelian period p if $w = u_0u_1 \cdots u_mu_{m+1}$, where $m \geq 1$ and $|u_1| = |u_2| = \cdots = |u_m| = p$ and $|u_0|, |u_{m+1}| \leq p$ and there exists a full word v over A , $|v| = p$, such that for all $0 \leq i \leq m+1$, $\|u_i\| \leq \|v\|$.

We now extend Theorem 1 to apply to partial words.

Theorem 2. Let w be a partial word with an arbitrary number of holes h . If w has abelian periods p and q which are relatively prime and $|w| \geq (h+2)pq - 1$, then w has period 1.

Proof. As mentioned earlier, since $\gcd(p, q) = 1$ the abelian periods p and q first match up at or before $pq - 1$ letters and the subsequent matches occur every pq letters later. A partial word w with $|w| \geq (h+2)pq - 1$ contains at least $h+2$ matchings of p and q , and $h+1$ subwords between these matchings. For $0 \leq i \leq h$, let

$$w_{i+2} = v_{r+ip}v_{r+ip+1} \cdots v_{r+(i+1)p-1} = u_{b_r+iq+1}u_{b_r+iq+2} \cdots u_{b_r+(i+1)q}$$

be the subword between the $(i + 1)$ st and $(i + 2)$ nd matching points of p and q . Since we have only h holes, one of the subwords w_2, w_3, \dots, w_{h+2} does not contain any hole. Examining this subword which is full, we get by the argument given in the proof of Theorem 1 that the Parikh vector of any u_l or v_l within this subword cannot contain more than one non-zero component. So for any u_i and v_i in w we have $\|u_i\| \leq \|u_l\|$ and $\|v_i\| \leq \|v_l\|$. Therefore all u_i and v_i in w contain at most one non-zero component, so w has period 1. \square

Further, we claim that the length $(h + 2)pq - 1$ is optimal for h holes as our algorithm in Section 4 constructs non-unary partial words with h holes of length $(h + 2)pq - 2$ that have abelian periods p and q for any coprime positive integers p, q .

3 Non-relatively Prime Abelian Periods

As observed in [7], Fine and Wilf's theorem cannot in general be extended to non-relatively prime abelian periods. That is, if $\gcd(p, q) = d, d > 1$, then the two abelian periods p and q cannot impose the abelian period d no matter how long the word is. For example, the infinite word $(aabbcc.abc|abc.aabbcc)^\omega$ has abelian periods $p = 6$ and $q = 9$ but does not have abelian period $\gcd(p, q) = 3$ (note that if v is a non-empty finite word, then we denote by v^ω the unique infinite word w such that w has period $|v|$ and $w(0) \cdots w(|v| - 1) = v$). Constantinescu and Ilie raise the question of whether or not a word w with abelian periods p and q such that $\gcd(p, q) = d, d > 1$, has at most cardinality d . In this section, we answer their question affirmatively (see Theorem 3 and Theorem 5). In the case where $\|u_0\| - \|v_0\| = \mu d$ for some integer $\mu \geq 0$ (here $u_0 u_1 \cdots u_{m+1}, v_0 v_1 \cdots v_{n+1}$ are factorizations of w into p and q , respectively) the length $2\text{lcm}(p, q) - 1$ turns out to be optimal. For the remainder of this section, we assume that all words are full and that p, q are integers satisfying $p < q, \gcd(p, q) = d > 1$, and $p = dp', q = dq'$ (note that $p' > 1$).

Lemma 1. *For a word w with abelian periods p and q such that $\gcd(p, q) = d, d > 1, p$ and q match up if and only if $\|u_0\| - \|v_0\| = \mu d$ for some integer $\mu \geq 0$.*

Proof. Let us first suppose that $\|u_0\| - \|v_0\| = \mu d$ for some integer $\mu \geq 0$. Since our argument does not depend on which length is greater, we may assume that $\|v_0\| \geq \|u_0\|$. Consider the subword $w' = u_1 \cdots u_{m+1} = u'_0 u'_1 \cdots u'_{m+1}$ where $u'_0 = \varepsilon, u'_1 = u_1, \dots, u'_{m+1} = u_{m+1}$, and $w' = v'_0 v'_1 \cdots v'_{n+1}$ where $v'_0 = v_0[\|u_0\|..|v_0|), v'_1 = v_1, \dots, v'_{n+1} = v_{n+1}$. Note that $|v'_0| = \mu d$. We have that the p periods end at length $sp = d(sp')$, for some integer $s \geq 1$, and the q periods end at length $\mu d + tq = d(\mu + tq')$, for some integer $t \geq 0$. Since d divides both of these, they match up when $\mu + tq' = sp'$ which is possible for any μ since p' and q' are coprime. Thus, periods p and q match up. The other direction is proved similarly. \square

Lemma 2. *If a word w has abelian periods p and q with $\gcd(p, q) = d, d > 1, |w| \geq 2\text{lcm}(p, q) - 1$, and $\|u_0\| - \|v_0\| = \mu d$ for some integer $\mu \geq 0$, then the abelian periods p and q match up in at least two places.*

Proof. First suppose that $|v_0| \geq |u_0|$. Consider, as in Lemma 1, the subword $w' = u_1 \cdots u_{m+1} = v'_0 v_1 \cdots v_{n+1}$, where $v'_0 = v_0[|u_0|..|v_0|)$. Note that $|v'_0| = \mu d$. Now treat each d letters as if they were one letter. For $p = dp'$ and $q = dq'$, our new “periods” are p' and q' with $|v'_0| = \mu \leq q'$. If $\mu = 0$ or $\mu = q'$, then p' and q' first match up at the zeroth position and their next matching occurs $p'q'$ letters later. Further, since p' and q' are coprime, p' and q' match up at or before $p'q' - 1$ by Theorem 1. Now, assume $0 < \mu < q'$. Then p' and q' match up when $\mu + tq' = sp'$ for some integers $1 \leq t < p'$ and $s \geq 1$. Since $\mu < q'$, $\mu + tq' \leq q'p' - 1$ which implies $p'q' > sp'$, and so $s < q'$ which means $s \leq q' - 1$. So, p' and q' match up at or before length $p'(q' - 1)$. Multiplying by d we have that periods p and q first match up at or before length $dp'(q' - 1) = dp'q' - dp'$. Our next matching occurs $q'p = p'q$ letters later. So our first two matchings occur at or before length $2dp'q' - dp' \leq |w|$. Now, since we have a bound for w' , we need to determine what is the maximum length for w . The longest u_0 can be is $p - 1$. Note that if instead we suppose that $|u_0| > |v_0|$ then we would have considered the subword $w' = v_1 \cdots v_{n+1}$. We would conclude that the minimum length of w' would be the same and that the length of v_0 would still be less than or equal to $p - 1$. So, for two matchings to be contained in w we need $|w| \geq 2dp'q' - dp' + p - 1 = 2dp'q' - 1 = 2\text{lcm}(p, q) - 1$. \square

Theorem 3. *If a word w has abelian periods p and q with $\gcd(p, q) = d$, $d > 1$, and $\||u_0| - |v_0|\| = \mu d$ for some integer $\mu \geq 0$, then w has at most cardinality d for $|w| \geq 2\text{lcm}(p, q) - 1$.*

Proof. By Lemma 2, w contains at least two matchings of p and q . Now, we use these matchings to prove that the cardinality of w is at most d . Suppose w has cardinality $d + 1$. Let $p = dp'$, $q = dq'$, for p', q' coprime. After p and q first match up our second matching occurs $q'p = p'q = \text{lcm}(p, q)$ letters later. As in Section 2, let $v_0 v_1 \cdots v_{r-1} = u_0 u_1 \cdots u_{b_r}$ be the first matching, where r and b_r are positive integers. Then the next matching is $v_0 \cdots v_{r-1} v_r v_{r+1} \cdots v_{r+p'-1} = u_0 \cdots u_{b_r} u_{b_r+1} u_{b_r+2} \cdots u_{b_r+q'}$. Consider $v_r \cdots v_{r+p'-1} = u_{b_r+1} \cdots u_{b_r+q'}$. For a letter a_l in w let α_l represent the number of times that letter occurs in one q period and β_l represent the number of times that letter occurs in one p period. Then we have $\alpha_l p' = \beta_l q'$. This implies that $\frac{\alpha_l}{\beta_l} = \frac{q'}{p'}$. But $\gcd(p', q') = 1$ so we must have $q' \mid \alpha_l$ and $p' \mid \beta_l$. Therefore, for $\alpha_l \neq 0$ and $\beta_l \neq 0$ we must have $\alpha_l \geq q'$ and $\beta_l \geq p'$. Let our letters be indexed such that a_1, \dots, a_{d+1} are the letters with non-zero components. So we have $q = \sum_{l=1}^{d+1} \alpha_l \geq (d+1)q'$ and $p = \sum_{l=1}^{d+1} \beta_l \geq (d+1)p'$. This gives $p \geq (d+1)p'$ and $q \geq (d+1)q'$, a contradiction. Hence the cardinality of w is at most d . \square

This bound is also optimal. Indeed, letting $p = 6$ and $q = 9$, the word

$$cbbaa.abd|acb.aabbd.|aabbcd.abc|abd.aabbc$$

of length 34 has abelian periods p and q and cardinality 4.

We now discuss the case where $\||u_0| - |v_0|\| \neq \mu d$ for any integer $\mu \geq 0$. Here the situation becomes more complicated. We believe that the optimal length for

words where the abelian periods do not match up is shorter than the optimal length for when they do match up.

Conjecture 1. If a word w has abelian periods p and q with $\gcd(p, q) = d$, $d > 1$, and $\|u_0\| - \|v_0\| \neq \mu d$ for any integer $\mu \geq 0$, then w has at most cardinality d for $|w| \geq 2 \operatorname{lcm}(p, q) - 2$.

If Conjecture 1 is true, then a word w having abelian periods $p = dp'$ and $q = dq'$ with $\gcd(p, q) = d$, $d \geq 1$, has at most cardinality d for $|w| \geq 2 \operatorname{lcm}(p, q) - 1$. We now prove Conjecture 1 for the case where $q = \gamma p + d$ for some γ .

Lemma 3. *For a word w with abelian periods p and q , where $\gcd(p, q) = d > 1$ and $\|u_0\| - \|v_0\| \neq \mu d$ for any integer $\mu \geq 0$, if w contains at least p' full q periods and $q = \gamma p + d$ for some integer $\gamma \geq 1$, then there exists at least one q period in w that contains γ full p periods.*

Proof. Suppose v_1 contains only $\gamma - 1$ full p periods. Then v_1 can be written as $x_1 u_{b_1+1} \cdots u_{b_1+\gamma-1} y_1$, where x_1 is a suffix of u_{b_1} and y_1 is a prefix of $u_{b_1+\gamma}$. Since $|x_1|, |y_1| < p$ and $|x_1 u_{b_1+1} \cdots u_{b_1+\gamma-1} y_1| = \gamma p + d$ we have that $d < |x_1|, |y_1| < p$. So, let $|x_1| = d + \delta$ and $|y_1| = p + d - |x_1| = p - \delta$ for some integer $0 < \delta < p - d$. Now consider $v_2 = x_2 \cdots y_2$ factorized similarly, where $|x_2| = p - |y_1| = \delta$. If $|x_2| \leq d$ then v_2 contains γ full p periods. Otherwise, $|y_2| = \gamma p + d - (\gamma - 1)p - |x_2| = p + d - \delta$. By induction, if none of $v_1, \dots, v_{p'-1}$ contains γ full p periods, then $|x_i| = \delta - (i - 2)d$ for $1 \leq i \leq p'$. We get $|x_{p'}| = \delta - (p' - 2)d = \delta - p + 2d < p - d - p + 2d = d$. So, we have at least one of $v_1, \dots, v_{p'}$ containing γ full p periods. \square

Theorem 4. *For a word w with abelian periods p and q , where $\gcd(p, q) = d > 1$ and $\|u_0\| - \|v_0\| \neq \mu d$ for any integer $\mu \geq 0$, if $|w| \geq 2 \operatorname{lcm}(p, q) - 2$ and $q = \gamma p + d$ for some integer $\gamma \geq 1$, then w has cardinality at most d .*

Proof. Suppose w has cardinality $d + 1$, and $q = \gamma p + d = \gamma dp' + d$. Since $|w| \geq 2 \operatorname{lcm}(p, q) - 2$, w contains at least $2p' - 1$ full q periods, i.e. $n \geq 2p' - 1$. Then there exists a letter a such that $\alpha_a = \gamma \beta_a$. Suppose w has h letters a_l with $\alpha_l = \gamma \beta_l$, and $d + 1 - h$ letters with $\alpha_l > \gamma \beta_l$, where $h \geq 1$ (we assume without loss of generality that $a = a_1$).

By Lemma 3, some q period must contain γ full p periods. Let v_z be this first v_i containing γ full p periods. We can write v_z as $x_z u_{b_z+1} \cdots u_{b_z+\gamma} y_z$ and v_{z-1} as $x_{z-1} u_{b_{z-1}+1} \cdots u_{b_{z-1}+\gamma-1} y_{z-1}$. Since we have γ full p periods in v_z and $\alpha_a = \gamma \beta_a$, $|x_z|_a = |y_z|_a = 0$. Further, since $y_{z-1} x_z$ forms a full p period, $|y_{z-1}|_a = \beta_a$, which implies $|x_{z-1}|_a = 0$. As we work backwards through w , this pattern continues with $|y_{z-z'}|_a = \beta_a$ and $|x_{z-z'}|_a = 0$, until the word ends or we reach another q period containing γ full p periods, which occurs in the p' th previous q period. If we reach this period, $v_{z-p'}$, we have $|y_{z-p'}|_a = \beta_a$. However, then we would have that $|v_{z-p'}|_a \geq (\gamma + 1)\beta_a$ which is a contradiction. Further we note that since our periods do not match up, if v_0 were a full q period, we would have $|x_0| \geq 1$. So, our word w must end here with $|v_0| \leq q - 1 - \gamma \beta_a \leq q - 2$. We can similarly argue that our word must end with the $v_{z+p'}$ subword. Since w contains $2p' - 1$

full q periods, either $v_{p'-1}$ or $v_{p'}$ must be a subword containing γ full p periods. Without loss of generality, we let $v_{p'}$ be this subword.

Let $v_{p'} = x_{p'} u_{b_{p'+1}} \cdots u_{b_{p'+\gamma}} y_{p'}$, where $x_{p'}$ is a suffix of $u_{b_{p'}}$ and $y_{p'}$ is a prefix of $u_{b_{p'+\gamma+1}}$. Let $|x_{p'}|_{a_l} + |y_{p'}|_{a_l} = g_l$ and $|x_{p'}|_{a_l} = h_l$. So, $|y_{p'}|_{a_l} = g_l - h_l$, $|v_{p'}| = \gamma\beta_l + g_l$, and $\sum_{l=1}^{d+1} g_l = d$. Then $|y_{p'-1}|_{a_l} = \beta_l - h_l$ and $|x_{p'-1}|_{a_l} = g_l + h_l$. By induction, we get $|x_{p'-z'}|_{a_l} = z'g_l + h_l$. We also have $|x_{p'+1}|_{a_l} = \beta_l - g_l + h_l$ and $|y_{p'+1}|_{a_l} = 2g_l - h_l$. By induction, we get $|y_{p'+z'}|_{a_l} = (z' + 1)g_l - h_l$. So, $\beta_l \geq |x_1|_{a_l} = (p' - 1)g_l + h_l = g_l p' - g_l + h_l$ and $\beta_l \geq |y_{2p'-1}|_{a_l} = g_l p' - h_l$. Since $0 \leq h_l \leq g_l$, $\max\{g_l p' - g_l + h_l, g_l p' - h_l\} \geq g_l p' - \lfloor \frac{g_l}{2} \rfloor$. So, $\beta_l \geq g_l p' - \lfloor \frac{g_l}{2} \rfloor$. Each letter with $\gamma\beta_l = \alpha_l$ must have $\beta_l \geq 1$.

Suppose $\alpha_l = \gamma\beta_l$ for $l = 1, \dots, h$ and $\alpha_l > \gamma\beta_l$ for $l = h + 1, \dots, d + 1$. So, $p = \sum_{l=1}^{d+1} \beta_l \geq h + \sum_{l=h+1}^{d+1} \beta_l \geq h + \sum_{l=h+1}^{d+1} (g_l p' - \lfloor \frac{g_l}{2} \rfloor) = h + d p' - \sum_{l=h+1}^{d+1} \lfloor \frac{g_l}{2} \rfloor$. So, $\sum_{l=h+1}^{d+1} \lfloor \frac{g_l}{2} \rfloor \geq h$. Suppose $h = 1$. Then we have d letters with $\alpha_l > \gamma\beta_l$. So, for $h + 1 \leq l \leq d + 1$, $g_l = 1$. So, $\sum_{l=h+1}^{d+1} \lfloor \frac{g_l}{2} \rfloor = 0 < 1$, a contradiction. Thus $h > 1$. Therefore, a word with $d + 1$ letters has at most $2p' - 2$ full q periods. We must still have either $v_{p'-1}$ or $v_{p'}$ containing γ full p periods. Therefore, w has length at most $(2p' - 2)q + q - 1 + q - 2 = 2 \operatorname{lcm}(p, q) - 3$. \square

We can prove the following result for the general case where $q = \gamma p + sd$ for some γ, s . Although the length in Theorem 5 may not be optimal, Theorem 3 and Theorem 5 answer affirmatively Constantinescu and Ilie's conjecture.

Theorem 5. *For a word w with abelian periods p and q , where $\gcd(p, q) = d > 1$ and $\|u_0\| - \|v_0\| \neq \mu d$ for any integer $\mu \geq 0$, if $|w| > 2(p' - 1) \operatorname{lcm}(p, q) - 2$, then w has cardinality at most d .*

Proof. We can write $q = \gamma p + sd$ where $\gcd(p', s) = 1$, $0 < s < p'$, $\gamma > 0$. There exists an integer r such that $0 < r < p'$ and $rs = tp' + 1$ for some integer t . Then note that w has abelian period rq , where $rq = (t + r\gamma)p + d$. It then follows by Theorem 4 that since $\gcd(rq, p) = d$ and since $|w| > 2(p' - 1) \operatorname{lcm}(p, q) - 2 \geq 2r \operatorname{lcm}(p, q) - 2 = 2 \operatorname{lcm}(p, rq) - 2$ that w has cardinality at most d . \square

4 Constructing Optimal Partial Words

We start our discussion with optimal full words. First, suppose that $p < q$ and $\gcd(p, q) = 1$. We would like to construct a word w over the alphabet $A = \{a, b\}$ such that w has abelian periods p and q , and w has length $2pq - 2$. Let α_a and α_b be the number of times the letters a and b , respectively, occur within one abelian q period and let β_a and β_b be the number of times the letters a and b , respectively, occur within one abelian p period.

For our word to have optimal length, the periods p and q must match up after exactly $pq - 1$ letters. So $|u_0| = |u_{m+1}| = p - 1$ and $|v_0| = |v_{n+1}| = q - 1$, where $u_0 u_1 \cdots u_{m+1}$ and $v_0 v_1 \cdots v_{n+1}$ are factorizations of w into abelian periods p and q , respectively. For simplicity we assume $\beta_a \geq \beta_b$ and, whenever possible, we place a 's before b 's. For example, letting $p = 3$ and $q = 7$, the word

$$w = aa.aab.b|aa.aab.ab|a.aab.aab.|aab.aab.a|ab.aab.aa|b.aab.aa$$

of length $2pq - 2 = 40$ is optimal. We can write $w = w_1w_2$, where $w_1 = w[0..pq - 1] = w_{1,9}w_{1,8}w_{1,7}w_{1,6}w_{1,5}w_{1,4}w_{1,3}w_{1,2}w_{1,1}$ and $w_2 = w[pq - 1..2pq - 2] = w_{2,1}w_{2,2}w_{2,3}w_{2,4}w_{2,5}w_{2,6}w_{2,7}w_{2,8}w_{2,9}$, where $w_{1,1} = w_{2,1} = aab$, $w_{1,2} = w_{2,2} = aab$, $w_{1,3} = w_{2,3} = a$, $w_{1,4} = w_{2,4} = ab$, etc. More generally, subwords in w are created by both the p and q periods as follows: if we look at the two subwords on each side of the first matching point, denote the first subword to the left of the matching $w_{1,1}$ and the first subword to the right $w_{2,1}$ and continue this labeling outward. We will write $w_1 = \text{rev}_{p,q}(w_2)$. Note that in w_2 , we have $\beta_bq - \alpha_bp = \pm 1$. So, in order to construct an optimal word the key is to determine for which values of β_b and α_b we have $\beta_bq - \alpha_bp = \pm 1$.

Further, we can extend this idea to construct an optimal word w with abelian periods $p = dp'$ and $q = dq'$ such that $\gcd(p, q) = d$ in the case where p and q have matching points. In this case, the key is to determine for which values of α and β we have $\beta q' - \alpha p' = \pm 1$. Algorithm 1, given later, gives a construction for optimal words when p and q have matching points. Further, if Conjecture 1 is true, then Algorithm 1, when $h = 0$, gives a construction for all optimal words.

Based on Algorithm 1, we give some closed forms.

Theorem 6. *Let p, q, p', q', d, γ be positive integers such that $p = dp'$, $q = dq' = \gamma p + d$, and $\gcd(p, q) = d$. If $\|u_0\| - \|v_0\| = \mu d$ for some integer $\mu \geq 0$, then $w = u_0u_1 \cdots = v_0v_1 \cdots v_{2p'-1}$ is an optimal word of cardinality $d + 1$, length $2\text{lcm}(p, q) - 2$, having abelian periods p and q , that can be constructed by Algorithm 1, where*

$$\begin{aligned} v_0 &= a_1^{p'} \cdots a_{d-1}^{p'} a_d^{p'-1} (a_1^{p'} \cdots a_{d-1}^{p'} a_d^{p'-1} a_{d+1})^{\gamma-1} a_1 \cdots a_{d-1} a_{d+1} | \\ v_i &= |a_1^{p'-i} \cdots a_{d-i}^{p'-i} a_d^{p'-i} (a_1^{p'} \cdots a_{d-1}^{p'} a_d^{p'-1} a_{d+1})^{\gamma-1} a_1^{i+1} \cdots a_{d-1}^{i+1} a_d^i a_{d+1} | \\ v_{p'} &= |(a_1^{p'} \cdots a_{d-1}^{p'} a_d^{p'-1} a_{d+1})^\gamma a_1 \cdots a_{d-1} a_d | \\ v_{p'+j} &= |a_1^{p'-j} \cdots a_{d-1}^{p'-j} a_d^{p'-1-j} a_{d+1} (a_1^{p'} \cdots a_{d-1}^{p'} a_d^{p'-1} a_{d+1})^{\gamma-1} a_1^{j+1} \cdots a_{d-1}^{j+1} a_d^{j+1} | \\ v_{2p'-1} &= |a_1 \cdots a_{d-1} a_{d+1} (a_1^{p'} \cdots a_{d-1}^{p'} a_d^{p'-1} a_{d+1})^{\gamma-1} a_1^{p'} \cdots a_{d-1}^{p'} a_d^{p'-1} \end{aligned}$$

for $0 < i < p'$ and $0 < j < p' - 1$.

Proof. Since $q' = \gamma p' + 1$, we have from Algorithm 1 that $\beta = 1$ and $\alpha = \gamma$, and therefore $\|u_1\| = (p', p', \dots, p', p' - 1, 1)$ and $\|v_1\| = (q', q', \dots, q', q' - \gamma, \gamma)$. Also, the v_j subword after the matching, denoted v_z , contains γ full p periods followed by one of each of the letters a_1, \dots, a_{d-1}, a_d . So,

$$v_z = (a_1^{p'} \cdots a_{d-1}^{p'} a_d^{p'-1} a_{d+1})^\gamma a_1 \cdots a_{d-1} a_d$$

Since $q = \gamma p + d$ and from the proof of Theorem 3 we know that w contains only one matching of p and q , w can contain only two full q periods, v_{z-1} and v_z , containing γ full p periods. Based on $\|u_1\|$ and $\|v_1\|$, we find that

$$v_{z+1} = a_1^{p'-1} a_2^{p'-1} \cdots a_{d-1}^{p'-1} a_d^{p'-2} a_{d+1} (a_1^{p'} a_2^{p'} \cdots a_{d-1}^{p'} a_d^{p'-1} a_{d+1})^{\gamma-1} a_1^2 \cdots a_{d-1}^2 a_d^2$$

By induction, we can show that for $0 \leq i \leq p' - 2$,

$$v_{z+i} = a_1^{p'-i} \cdots a_{d-1}^{p'-i} a_d^{p'-1-i} a_{d+1} (a_1^{p'} \cdots a_{d-1}^{p'} a_d^{p'-1} a_{d+1})^{\gamma-1} a_1^{i+1} \cdots a_{d-1}^{i+1} a_d^{i+1}$$

Thus, $v_{z+p'-2} = a_1^2 \cdots a_{d-1}^2 a_d a_{d+1} (a_1^{p'} \cdots a_{d-1}^{p'} a_d^{p'-1} a_{d+1})^{\gamma-1} a_1^{p'-1} \cdots a_{d-1}^{p'-1} a_d^{p'-1}$. So based on $\|u_1\|$, we know $v_{z+p'-1}$ must begin with $a_1 \cdots a_{d-1} a_{d+1}$. In order to complete the q period, we must add $q' - 1 = \gamma p'$ each of the letters a_1, \dots, a_{d-1} , $\gamma(p' - 1) + 1$ a_d 's, and $\gamma - 1$ a_{d+1} 's. Since we can only add $\gamma - 1$ a_{d+1} 's, we can only add $\gamma - 1$ full p periods. Now, in order to complete our q period $v_{z+p'-1}$, we still need $\gamma p' - (\gamma - 1)p' = p'$ each of the letters a_1, \dots, a_{d-1} , $\gamma(p' - 1) + 1 - (\gamma - 1)(p' - 1) = p'$ a_d 's, and no a_{d+1} 's. So, we add all the letters from the next p period, except a_{d+1} . We now have q' each of a_1, \dots, a_{d-1} , $\gamma p' - \gamma$ a_d 's, and γ a_{d+1} 's. So from $\|u_1\|$ we can only add an a_{d+1} , but from $\|v_1\|$ we can only add an a_d . These conflict so our word must end here with

$$v_{z+p'-1} = a_1 \cdots a_{d-1} a_{d+1} (a_1^{p'} \cdots a_{d-1}^{p'} a_d^{p'-1} a_{d+1})^{\gamma-1} a_1^{p'} \cdots a_{d-1}^{p'} a_d^{p'-1}$$

and $|v_z \cdots v_{z+p'-1}| = (p' - 1)q + q - 1$. We can similarly argue that for $0 \leq i \leq z - 1$, $v_i = a_1^{p'-i} \cdots a_{d-i}^{p'-i} a_d^{p'-i} (a_1^{p'} \cdots a_{d-1}^{p'} a_d^{p'-1} a_{d+1})^{\gamma-1} a_1^{i+1} \cdots a_{d-1}^{i+1} a_d^i a_{d+1}$, and also that $v_0 = a_1^{p'} \cdots a_{d-1}^{p'} a_d^{p'-1} (a_1^{p'} \cdots a_{d-1}^{p'} a_d^{p'-1} a_{d+1})^{\gamma-1} a_1 \cdots a_{d-1} a_{d+1}$, and $|v_0 \cdots v_{z-1}| = (p' - 1)q + q - 1$. This gives us $z = p'$ and $|v_0 \cdots v_{2p'-1}| = 2p'q - 2 = 2\text{lcm}(p, q) - 2$. \square

It should be noted that, if Conjecture 1 is true, the Parikh vectors that create optimal length words for when the abelian periods do not match up are the same as the Parikh vectors for when they do match up. In other words, for any p and q , if we calculate their Parikh vectors based on Algorithm 1, we can construct a word of length $2\text{lcm}(p, q) - 3$ in which p and q do not match up. The word $w = ab.aaa|b.aaab.a|aab.aba|a.aaab.b|aaa.ab$ has abelian periods $p = 4$ and $q = 6$ and length $2\text{lcm}(p, q) - 3$.

Theorem 7. *Let p, q, p', q', d, γ be positive integers such that $p = dp'$, $q = dq' = \gamma p + d$, and $\text{gcd}(p, q) = d > 1$. If $\|u_0\| - \|v_0\| \neq \mu d$ for any integer $\mu \geq 0$, then $w = u_0 u_1 \cdots = v_0 v_1 \cdots v_{2p'-1}$ is an optimal word of cardinality $d + 1$, length $2\text{lcm}(p, q) - 3$, having abelian periods p and q , that can be constructed by Algorithm 1, where*

$$\begin{aligned} v_0 &= a_1^{p'} a_2^{p'} \cdots a_{d-1}^{p'} a_d^{p'-1} (a_1^{p'} \cdots a_{d-1}^{p'} a_d^{p'-1} a_{d+1})^{\gamma-1} a_2 \cdots a_{d-1} a_{d+1} | \\ v_i &= |a_1^{p'+1-i} a_2^{p'-i} \cdots a_d^{p'-i} (a_1^{p'} \cdots a_{d-1}^{p'} a_d^{p'-1} a_{d+1})^{\gamma-1} a_1^i a_2^{i+1} \cdots a_{d-1}^{i+1} a_d^i a_{d+1} | \\ v_{p'} &= |a_1 (a_1^{p'} \cdots a_{d-1}^{p'} a_d^{p'-1} a_{d+1})^\gamma a_2 a_3 \cdots a_{d-1} a_d | \\ v_{p'+j} &= |a_1^{p'+1-j} a_2^{p'-j} \cdots a_{d-1}^{p'-j} a_d^{p'-1-j} a_{d+1} (a_1^{p'} \cdots a_{d-1}^{p'} a_d^{p'-1} a_{d+1})^{\gamma-1} \\ &\quad a_1^j a_2^{j+1} \cdots a_d^{j+1} | \\ v_{2p'-1} &= |a_1^2 a_2 \cdots a_{d-1} a_{d+1} (a_1^{p'} \cdots a_{d-1}^{p'} a_d^{p'-1} a_{d+1})^{\gamma-1} a_1^{p'-1} a_2^{p'} \cdots a_{d-1}^{p'} a_d^{p'-1} \end{aligned}$$

for $0 < i < p'$ and $0 < j < p' - 1$.

Algorithm 1 Constructing an optimal partial word for two abelian periods

Input: Non-negative integer h and abelian periods $p < q$

Output: An optimal partial word w with h holes and length $(h + 2) \text{lcm}(p, q) - 2$

1. $d \leftarrow \text{gcd}(p, q)$, $p' \leftarrow \frac{p}{d}$ and $q' \leftarrow \frac{q}{d}$
 2. Find smallest positive integer β and corresponding positive integer α such that $\beta q' - \alpha p' = \pm 1$
 3. Define Parikh vectors for periods p and q with distinct letters a_1, \dots, a_{d+1}
 $U \leftarrow (p', p', \dots, p', p' - \beta, \beta)$ and $V \leftarrow (q', q', \dots, q', q' - \alpha, \alpha)$
 4. Generate subword w_2 of w from position $\text{lcm}(p, q) - 1$ up to position $2 \text{lcm}(p, q) - 3$
 - (a) $U' \leftarrow U$ // U' represents the number of each letter left to be filled into the current p period
 - (b) $w_2 \leftarrow \varepsilon$ and $L \leftarrow 0$
 - (c) **while** $L < \text{lcm}(p, q) - q$
 $V' \leftarrow V$ // V' represents the number of each letter left to be filled into the current q period
 $w_2 \leftarrow w_2 u'$ // u' is some word with $\|u'\| = U'$
 $l \leftarrow |u'|$ // l represents the number of letters added to the current q period
 $L \leftarrow L + l$
 $V' \leftarrow V' - U'$ and $U' \leftarrow U'$
while $l + p < q$
 $w_2 \leftarrow w_2 u$ // u is some word with $\|u\| = U$
 $l \leftarrow l + p$ and $L \leftarrow L + p$
 $V' \leftarrow V' - U'$
 $w_2 \leftarrow w_2 v'$ and $U' \leftarrow U' - V'$ // v' is some word with $\|v'\| = V'$
 $L \leftarrow L + |v'|$
 - (d) $V' \leftarrow V$
 - (e) $w_2 \leftarrow w_2 u'$ and $l \leftarrow |u'|$ // u' is some word with $\|u'\| = U'$
 - (f) $V' \leftarrow V' - U'$ and $U' \leftarrow U'$
 - (g) **while** $l + p < q$
 $w_2 \leftarrow w_2 u$ // u is some word with $\|u\| = U$
 $l \leftarrow l + p$
 $V' \leftarrow V' - U'$
 - (h) **for** $i = 1$ **to** $d + 1$
for $j = 1$ **to** $\min\{\#_{a_i}(U'), \#_{a_i}(V')\}$
 $w_2 \leftarrow w_2 a_i$
 5. $w_1 \leftarrow \text{rev}_{p', q'}(w_2)$
 6. $w \leftarrow w_1 w_2 (\diamond w_2)^h$
-

Algorithm 1 outputs optimal partial words with h holes when p and q match up by constructing the subword w_2 after the first matching and then concatenating $w_1 = \text{rev}_{p', q'}(w_2)$ with w_2 , and then with $(\diamond w_2)^h$. For instance, on input $p = 6$, $q = 10$ and $h = 4$, Algorithm 1 outputs the optimal word $w_1 w_2 (\diamond w_2)^4$ of length 178, where $w_1 = cbaaa.bbba|ca.cbbaaa.ba|cbaa.cbbaaa.|$ and $w_2 = aaabbc.aabc|ab.aaabbc.ac|aabb.aaabc.$

Remark 1. The position in which a hole is placed within a subword contained between two matching points to construct an optimal partial word does not

matter so long as the hole represents letter a in terms of p and letter b in terms of q , say, where a, b are distinct. To see this, if we add one more letter to an optimal full word, creating a second matching point between p and q , we have $\beta_a q' - \alpha_a p' = \pm 1$ and $\beta_b q' - \alpha_b p' = \mp 1$. So, when we add a hole this way, $\beta_a q' - \alpha_a p' + \beta_b q' - \alpha_b p' = 0$, we can complete the first matching and continue the word from the new matching. Otherwise, we still cannot construct an optimal partial word longer than a full one. The same applies for each hole that we add.

References

1. J. Berstel and L. Boasson. Partial words and a theorem of Fine and Wilf. *Theoretical Computer Science*, 218:135–141, 1999.
2. F. Blanchet-Sadri. *Algorithmic Combinatorics on Partial Words*. Chapman & Hall/CRC Press, Boca Raton, FL, 2008.
3. F. Blanchet-Sadri, J. I. Kim, R. Mercas, W. Severa, and S. Simmons. Abelian square-free partial words. In A.-H. Dediu, H. Fernau, and C. Martín-Vide, editors, *LATA 2010, 4th International Conference on Language and Automata Theory and Applications, Trier, Germany*, volume 6031 of *Lecture Notes in Computer Science*, pages 94–105, Berlin, Heidelberg, 2010. Springer-Verlag.
4. M. G. Castelli, F. Mignosi, and A. Restivo. Fine and Wilf’s theorem for three periods and a generalization of Sturmian words. *Theoretical Computer Science*, 218:83–94, 1999.
5. C. Choffrut and J. Karhumäki. Combinatorics of Words. In G. Rozenberg and A. Salomaa, editors, *Handbook of Formal Languages*, volume 1, pages 329–438. Springer-Verlag, Berlin, 1997.
6. S. Constantinescu and L. Ilie. Generalised Fine and Wilf’s theorem for arbitrary number of periods. *Theoretical Computer Science*, 339:49–60, 2005.
7. S. Constantinescu and L. Ilie. Fine and Wilf’s theorem for abelian periods. *Bulletin of the European Association for Theoretical Computer Science*, 89:167–170, 2006.
8. P. Erdős. Some unsolved problems. *Magyar Tudományok Akadémia Matematikai Kutató Intézete Közl.*, 6:221–254, 1961.
9. N. J. Fine and H. S. Wilf. Uniqueness theorems for periodic functions. *Proceedings of the American Mathematical Society*, 16:109–114, 1965.
10. L. Ilie and S. Ilie. Multiple spaced seeds for homology search. *Bioinformatics*, 23:2969–2977, 2007.
11. L. Ilie and S. Ilie. Fast computation of neighbor seeds. *Bioinformatics*, 25:822–823, 2009.
12. J. Justin. On a paper by Castelli, Mignosi, Restivo. *Theoretical Informatics and Applications*, 34:373–377, 2000.
13. V. Keränen. Abelian squares are avoidable on 4 letters. In W. Kuich, editor, *ICALP 1992, 19th International Colloquium on Automata, Languages and Programming*, volume 623 of *Lecture Notes in Computer Science*, pages 41–52, Berlin, 1992. Springer-Verlag.
14. A. M. Shur and Y. V. Gamzova. Partial words and the interaction property of periods. *Izvestiya RAN*, 68(2):191–214, 2004.
15. W. F. Smyth and S. Wang. A new approach to the periodicity lemma on strings with holes. *Theoretical Computer Science*, 410:4295–4302, 2009.
16. R. Tijdeman and L. Zamboni. Fine and Wilf words for any periods. *Indagationes Mathematicae*, 14:135–147, 2003.