Convergence of the number of period sets in strings

Eric Rivals[1], Michelle Sweering[2], and Pengfei Wang[1]

[1]LIRMM, Université Montpellier, CNRS, Montpellier, France, rivals@lirmm.fr, pengfei.wang@lirmm.fr
[2]CWI, Amsterdam, The Netherlands, michelle.sweering@cwi.nl

September 30, 2022

Abstract

Consider words of length n. The set of all periods of a word of length n is a subset of $\{0, 1, 2, \ldots, n-1\}$. However, any subset of $\{0, 1, 2, \ldots, n-1\}$ is not necessarily a valid set of periods. In a seminal paper in 1981, Guibas and Odlyzko have proposed to encode the set of periods of a word into an n long binary string, called an autocorrelation, where a one at position i denotes a period of i. They considered the question of recognizing a valid period set, and also studied the number of valid period sets for length n, denoted κ_n. They conjectured that $\ln(\kappa_n)$ asymptotically converges to a constant times $\ln^2(n)$. If improved lower bounds for $\ln(\kappa_n)/\ln^2(n)$ were proposed in 2001, the question of a tight upper bound has remained opened since Guibas and Odlyzko's paper. Here, we exhibit an upper bound for this fraction, which implies its convergence and closes this long standing conjecture. Moreover, we extend our result to find similar bounds for the number of correlations: a generalization of autocorrelations which encodes the overlaps between two strings.

Key words: autocorrelation; period; border; combinatorics; correlation; periodicity; upper bound; asymptotic convergence

1 Introduction

A linear word can overlap itself if one of its prefixes is equal to one of its suffixes. The corresponding prefix (or suffix) is called a border and the shift needed to match the prefix to the suffix is called a period. The dual notions of period and border are critical concepts in word combinatorics: important definitions such as periodic and primitive words, or the normal form of a word rely on them. These concepts play a role in key results of the field like the Critical Factorization Theorem [10]. In computer science, in the field of string algorithms (a.k.a., stringology), pattern matching algorithms heavily exploit borders/periods to optimize the search of occurrences of a word in a text [17]. These notions also play a role in statistics. The set of periods of a word controls how two occurrences of the same word can overlap in a text. Hence, the set of periods (or autocorrelation) is a key variable to study the statistics of word occurrences in random texts (waiting time, distance between successive occurrences, etc.) [15]. The notion of autocorrelation has been extended to describe how two distinct words can have overlapping occurrences in the same text. This has been used for instance to study the number of missing words in random texts [12], or to design procedure for testing pseudo-random number generators [11].

Autocorrelations are the binary vector representations of the set of periods of a string. The concept of autocorrelation was introduced by Guibas and Odlyzko in [6]. They give the characterization of autocorrelations and prove the following bounds on κ_n - the cardinality of the set Γ_n of autocorrelations of strings of length n.

$$\frac{1}{2\ln(2)} + o(1) \leq \frac{\ln(\kappa_n)}{\ln^2(n)} \leq \frac{1}{2\ln(3/2)} + o(1)$$

They conjecture that $\ln(\kappa_n)$ is asymptotic to a constant times $\ln^2(n)$. Rivals and Rahmann [14], later on give the combinatorial structure of autocorrelations set Γ_n and improve the lower bound on κ_n as follows.

$$\frac{\ln(\kappa_n)}{\ln^2(n)} \geq \frac{1}{2\ln(2)}\left(1 - \frac{\ln(\ln(n))}{\ln(n)}\right)^2 + \frac{0.4139}{\ln(n)} - \frac{1.47123\ln(\ln(n))}{\ln^2(n)} + O\left(\frac{1}{\ln^2(n)}\right).$$

However, to date, no one has focused on improving the upper bound on κ_n. In this work, we apply the notion of irreducible period set introduced by Rivals and Rahmann [13, 14] to prove that

$$\frac{\ln(\kappa_n)}{\ln^2(n)} \leq \frac{1}{2\ln(2)} + \frac{3}{2\ln(2)\ln(n)} \quad \forall n \in \mathbb{N}_{\geq 2}.$$

Together with known asymptotic lower bounds [14], we find that

$$\frac{\ln \kappa_n}{\ln^2(n)} \to \frac{1}{2\ln(2)} \quad \text{as} \quad n \to \infty,$$

thus resolving the conjecture of Guibas and Odlyzko.

In their paper about autocorrelations [6] Guibas and Odlyzko also introduced the notion of correlation between strings. For two strings u and v the *correlation* of u over v is a binary vector indicating all overlaps between suffixes of u and prefixes of v. In particular, an autocorrelation is the correlation of a string with itself. We show that the number of correlations between two strings of length n, denoted by δ_n, has the same asymptotic convergence behavior as the number of autocorrelations of strings of length n, that is

$$\frac{\ln \delta_n}{\ln^2(n)} \to \frac{1}{2\ln(2)} \quad \text{as} \quad n \to \infty.$$

1.1 Related works

Apart from previously cited articles that deal with combinatorics of period sets, some related works exist in the literature.

For instance, the question of the average period of a random word has been investigated in [9]. Clearly, the number of periods of a word of length n is comprised between one and n. A recent work exhibits an upper bound on the number of periods of a word in function of its *initial critical exponent* – a characteristic of the word related to its degree of periodicity [5], but this has not been used to bound the number of period sets. Last, the combinatorics of period sets has also been investigated in depth for a generalization of the notion of words, called *partial words* [3]. In partial words, some positions may contain a *don't care* symbol, which remove some constraints of equality between positions. To study the combinatorics of period sets, Blanchet-Sadri *et al.* defined weak and strong periods, and prove several important theorems [1], including lower and upper bounds on the number of binary and ternary autocorrelations [3, 2]. However, the cardinality of period sets differ between normal words and partial words, and the upper bound for normal words cannot be deduced those for partial words.

2 Preliminaries

A string $u = u[0\,..\,n-1] \in \Sigma^n$ is a sequence of n letters over a finite alphabet Σ. For any $0 \leq i \leq j \leq n-1$, we denote the substring starting at position i and ending at position j with $u[i\,..\,j]$. In particular, $u[0\,..\,j]$ denotes a prefix and $u[i\,..\,n-1]$ a suffix of u. Throughout this paper, all our strings and vectors will be zero-indexed.

2.1 Periodicity

In this subsection, we define the concepts of period, period set, basic period and autocorrelation and review some useful results.

Definition 2.1 (Period)**.** String $u = u[0\,..\,n-1]$ has a period $p \in \{1, \ldots, n-1\}$ if and only if for any $0 \leq i, j \leq n-1$ such that $i \equiv j \mod p$, we have $u[i] = u[j]$. Moreover, we consider $p = 0$ a period for all strings with length n.

An equivalent definition is the following.

Definition 2.2 (Period)**.** String $u = u[0\,..\,n-1]$ has period $p \in \{0, 1, \ldots, n-1\}$ if and only if $u[0\,..\,n-p-1] = u[p\,..\,n-1]$, i.e. for all $0 \leq i \leq n-p-1$, we have $u[i] = u[i+p]$.

The smallest non-zero period of u is called its *basic period*. The *period set* of a string u is the set of all its periods and is denoted by $P(u)$. We will now prove some useful properties about periods, which we will need later on.

Lemma 2.1. *Let p be a period of $u \in \Sigma^n$ and $k \in \mathbb{Z}_{\geq 0}$ such that $kp < n$. Then kp is also a period of u.*

Proof. If $p = 0$ or $k = 0$, the statement trivially holds. Suppose $p \in \{1, \ldots, n-1\}$ and $k > 0$. If $i, j \in \{0, \ldots, n-1\}$ such that $i \equiv j \mod kp$, then we also have $i \equiv j \mod p$, and hence $u[i] = u[j]$ by Definition 2.1. This shows kp is a period of u by Definition 2.1. □

Lemma 2.2. *Let p be a period of $u \in \Sigma^n$ and q a period of the suffix $w = u[p \mathinner{.\,.} n-1]$. Then $p+q$ is a period of u. Moreover, $p+kq$ is also a period of u for all $k \in \mathbb{Z}_{\geq 0}$ with $p+kq < n$.*

Proof. By Definition 2.2 of period, the fact that p is a period of u implies $u[0 \mathinner{.\,.} n-p-1] = u[p \mathinner{.\,.} n-1]$, while q is a period of w implies $w[0 \mathinner{.\,.} n-p-q-1] = w[q \mathinner{.\,.} n-p-1]$. As w is the suffix of u starting at position p, we can combine the above results to find that

$$\begin{aligned} u[0 \mathinner{.\,.} n-p-q-1] &= u[p \mathinner{.\,.} n-q-1] = w[0 \mathinner{.\,.} n-p-q-1] \\ &= w[q \mathinner{.\,.} n-p-1] = u[p+q \mathinner{.\,.} n-1], \end{aligned}$$

which indicates that $p+q$ is a period of u. Moreover, if $p+iq$ is a period of u for some $i \in \mathbb{N}$, then we can similarly show that $p+(i+1)q$ is also a period of u if $p+(i+1)q < n$. It follows by induction that $p+kq$ is a period of u for all $k \in \mathbb{N}$ with $p+kq < n$. The case $k=0$ is trivial. □

Lemma 2.3. *Let p, q be periods of $u \in \Sigma^n$ with $0 \leq q \leq p$. Then the prefix (suffix) of length $n-q$ has the period $p-q$.*

Proof. Since p, q be periods of $u \in \Sigma^n$ with $0 \leq q \leq p$, we have

$$\begin{aligned} u[0 \mathinner{.\,.} n-p-1] &= u[p \mathinner{.\,.} n-1] && \text{(by periodicity } p\text{)} \\ &= u[p-q \mathinner{.\,.} n-q-1] && \text{(by periodicity } q\text{).} \end{aligned}$$

It follows that $u[0 \mathinner{.\,.} n-q-1]$ has period $p-q$. Similarly the length $n-q$ suffix of u also has period $p-q$. □

Lemma 2.4. *Suppose p is a period of $u \in \Sigma^n$ and there exists a substring v of u of length at least p and with period r, where $r|p$. Then r is also a period of u.*

Proof. If $p=0$, then $r=0$ and the lemma trivially holds.

Otherwise p is non-zero. Let $i, j \in [0, n-1]$ with $i \equiv j \mod r$. We can write $v = u[h \mathinner{.\,.} k]$ with $0 \leq h < k \leq n-1$. Since v has length at least p, there exist $i', j' \in [h, k]$ such that $i \equiv i' \mod p$ and $j \equiv j' \mod p$. By Definition 2.1 of period, we have $u[i] = u[i']$ and $u[j] = u[j']$. Note that $i' \equiv i \equiv j \equiv j' \mod r$, because $r \mid p$. Applying Definition 2.1 again, we obtain $u[i'] = u[j']$. It follows that $u[i] = u[i'] = u[j'] = u[j]$. Therefore r is a period of u. □

We will also use the famous Fine and Wilf theorem [4], a.k.a. the periodicity lemma, for which a short proof was provided by Halava and colleagues [8].

Theorem 2.5 (Fine and Wilf)**.** *Let p, q be periods of $u \in \Sigma^n$. If $n \geq p+q-\gcd(p,q)$, then $\gcd(p,q)$ is a period of u.*

2.2 Autocorrelation

We now give a formal definition of an autocorrelation.

Definition 2.3 (Autocorrelation)**.** For every string $u \in \Sigma^n$, its autocorrelation is the string $s \in \{0,1\}^n$ such that

$$s[i] = \begin{cases} 1 & \text{if } i \text{ is a period of } u \\ 0 & \text{otherwise} \end{cases} \qquad \forall i \in \{0, \ldots, n-1\}.$$

pos.	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	
u	a	b	b	a	a	b	b	a	-	-	-	-	-	-	-	s
u	a	b	b	a	a	b	b	a	-	-	-	-	-	-	-	1
	-	a	b	b	a	a	b	b	a	-	-	-	-	-	-	0
	-	-	a	b	b	a	a	b	b	a	-	-	-	-	-	0
	-	-	-	a	b	b	a	a	b	b	a	-	-	-	-	0
	-	-	-	-	a	b	b	a	a	b	b	a	-	-	-	1
	-	-	-	-	-	a	b	b	a	a	b	b	a	-	-	0
	-	-	-	-	-	-	a	b	b	a	a	b	b	a	-	0
	-	-	-	-	-	-	-	a	b	b	a	a	b	b	a	1

Table 1: The top string is u. The blue numbers are its periods. The corresponding shifted the first row marked with the blue are periods. The last column contains the autocorrelation of u.

To illustrate this concept, consider the following example (detailed in Table 1).

Example 1. We consider the string $u = \texttt{abbaabba}$. Its period set is $P(u) = \{0, 4, 7\}$, its basic period is 4 and its autocorrelation is $s = 10001001$. See Table 1.

Guibas and Odlyzko [6] show that *any alphabet of size at least two will give rise to the same set of correlations* (Corollary 5.1). Autocorrelations have many other useful properties [6, 14]. We show the most significant one for our work.

Lemma 2.6. *If $s \in \{0,1\}^n$ is an autocorrelation and $s[i] = 1$, then $s[i \mathinner{..} n-1]$ is the autocorrelation of $u[i \mathinner{..} n-1]$*

Proof. Note that $s[i] = 1$ means: i is a period of u. Suppose $s[i+j] = 1$. Then $i+j$ is a period of u. Thus $u[i \mathinner{..} n-1]$ has period $(i+j) - i = j$ by Lemma 2.3. Conversely, suppose $u[i \mathinner{..} n-1]$ has period $(i+j) - i = j$. Then $i+j$ is a period of u by Lemma 2.2. Thus $s[i+j] = 1$. Combining these results, we find that $s[i+j] = 1$ if and only of j is a period of $u[i \mathinner{..} n-1]$, and equivalently $s[i \mathinner{..} n-1]$ is the autocorrelation of $u[i \mathinner{..} n-1]$. $\square$

2.3 Irreducible Period Set

To prove the upper bound on the number of autocorrelations, we use the notion of irreducible period set as introduced by Rivals and Rahmann [14]. An irreducible period set is the minimum subset of a period set that determines the period set using the Forward Propagation Rule. Before formally introducing the irreducible period set, we will first explain what forward propagation is.

Lemma 2.7 (Forward Propagation Rule)**.** *Let $p \leq q$ be periods of a string u of length n and let $k \in \mathbb{Z}_{\geq 0}$ such that $p + k(q-p) < n$. Then $p + k(q-p)$ is a period of $u[0 \mathinner{..} n-1]$.*

Proof. It follows from Lemma 2.3 that $u[p \mathinner{..} n-1]$ has period $q - p$. Applying Lemma 2.2 we find that $u[0 \mathinner{..} n-1]$ has period $p + k(q-p)$ for all $k \in \mathbb{Z}_{\geq 0}$. $\square$

The forward closure $FC_n(S)$ of a set $S \subseteq \{0, \ldots, n-1\}$ (not necessarily a period set, typically a subset of one) is the closure of S under the forward propagation rule.

Definition 2.4 (Forward Closure)**.** Let $S \subseteq \{0, \ldots, n-1\}$. Its forward closure $FC_n(S)$ is the minimum superset of S such that for any $p, q \in FC_n(S)$ and $k \geq 0$ with $p \leq q$ and $p + k(q-p) < n$, we have

$$p + k(q-p) \in FC_n(S).$$

We can now define the irreducible period set.

Definition 2.5 (Irreducible Period Set)**.** Let P be the period set of a string $u \in \Sigma^n$. An irreducible period set of P is a minimal subset $R(P) \subseteq P$ with forward closure P.

Observe that there exists an irreducible period set for any period set P, because $FC_n(P) = P$ by the forward propagation rule. We will now give a useful characterization of an irreducible period set as the set of periods which are not in the forward closure of the set of all smaller periods. Consequently, every period set has exactly one irreducible period set. For a given string length n, we will denote the set of all irreducible period sets by Λ_n. The bijective relation between period sets and irreducible period sets, implies that $|\Gamma_n| = |\Lambda_n|$.

Lemma 2.8. *Let P be the period set of a string $u \in \Sigma^n$ and $R(P)$ an irreducible period set of P. Then*

$$R(P) = \{q \in P \mid q \notin FC_n(P \cap [0, q-1])\}.$$

Proof. Let $p \in P$. We will prove the two alternative cases separately:

(a) $p \notin \{q \in P \mid q \notin FC_n(P \cap [0, q-1])\} \implies p \notin R(P)$ and

(b) $p \in \{q \in P \mid q \notin FC_n(P \cap [0, q-1])\} \implies p \in R(P)$.

(a) Suppose $p \notin \{q \in P \mid q \notin FC_n(P \cap [0, q-1])\}$, or equivalently $p \in FC_n(P \cap [0, p-1])$. Then

$$\begin{aligned} p \in FC_n(P \cap [0, p-1]) &= FC_n(FC_n(R(P)) \cap [0, p-1]) \\ &\subseteq FC_n(FC_n(R(P) \cap [0, p-1])) \\ &= FC_n(R(P) \cap [0, p-1]) \\ &\subseteq FC_n(R(P) \setminus \{p\}). \end{aligned}$$

It follows that $FC_n(R(P) \setminus \{p\}) = FC_n(R(P))$. By minimality of irreducible period sets, we have $p \notin R(P)$.

(b) Suppose on the other hand that $p \notin FC_n(P \cap [0, p-1])$. Then

$$p \notin FC_n(P \setminus \{p\}) \supseteq FC_n(R(P) \setminus \{p\})$$

either. However, as $p \in P$ and $P = FC_n(R(P))$, it follows that $p \in R(P)$.

□

3 Asymptotic convergence of κ_n

In this section, we present a new upper bound on κ_n, the number of distinct autocorrelations of strings of length n. Moreover, we shall prove that $\ln(\kappa_n)$ asymptotically converges to $c \cdot \ln^2(n)$, where $c = \frac{1}{2\ln(2)}$.

Theorem 3.1 (Upper bound on κ_n)**.** *For all $n \in \mathbb{N}_{\geq 2}$ we have*

$$\frac{\ln(\kappa_n)}{\ln^2(n)} \leq \frac{1}{2\ln(2)} + \frac{3}{2\ln(2)\ln(n)}.$$

Proof. To prove this theorem, we need several lemmas.

Lemma 3.2. *Let $u \in \Sigma^n$ with autocorrelation s, period set P and irreducible period set $R(P) = \{0 = a_0 < \ldots < a_i < \ldots < a_k < n\}$. Then for all $0 \leq i \leq k$, there exists $q_i \in \{1, \ldots, n - a_i\}$ such that*

1. *$q_i \leq n/2^i$, and*
2. *$a_i + q_i = n$ or $a_i + q_i$ is in the forward closure of $\{a_0, \ldots, a_i\}$.*

Proof. We will prove this by induction.

Basis By picking $q_0 = n \in \{1, \ldots, n - a_0\}$, we satisfy both $q_0 \leq n/2^0$ and $a_0 + q_0 = n$.

Hypothesis For some $1 \leq i < k$, there exists a $q_i \in \{1, \ldots, n - a_i\}$ such that

1. $q_i \leq n/2^i$, and
2. $a_i + q_i = n$ or $a_i + q_i$ is in the forward closure of $\{a_0, \ldots, a_i\}$.

Step We first note that if $n - a_{i+1} \leq n/2^{i+1}$, then we can pick $q_{i+1} = n - a_{i+1}$. Suppose on the other hand that $n - a_{i+1} > n/2^{i+1}$. We distinguish two cases.

- If $a_i + q_i = n$, then

$$\begin{aligned} a_{i+1} - a_i &= (n - a_i) - (n - a_{i+1}) \\ &< n/2^i - n/2^{i+1} \\ &= n/2^{i+1} \\ &< n - a_{i+1}. \end{aligned}$$

  Thus, we can pick $q_{i+1} = a_{i+1} - a_i \in \{1, \ldots, n - a_{i+1}\}$, since

  1. it satisfies $q_{i+1} \leq n/2^{i+1}$ and
  2. $a_{i+1} + q_{i+1} = a_i + 2(a_{i+1} - a_i)$ is in the forward closure of $\{a_0, \ldots, a_{i+1}\}$.

- If $a_i + q_i$ is in the forward closure of $\{a_0, \ldots, a_i\}$, then

$$a_i + \lambda q_i = a_i + \lambda(a_i + q_i - a_i)$$

is in the forward closure of $\{a_0, \ldots, a_i\}$ for all integers $0 \leq \lambda \leq (n-1-a_i)/q_i$. Since a_{i+1} is an irreducible period, there must exist an integer $\lambda_0 \in [0, (n-1-a_i)/q_i]$ such that

$$a_i + \lambda_0 q_i < a_{i+1} < a_i + (\lambda_0 + 1)q_i.$$

In other words, a_{i+1} is comprised between two successive, non-irreducible periods generated from a_i and q_i using the FPR. We pick

$$q_{i+1} = \min(a_{i+1} - (a_i + \lambda_0 q_i), (a_i + (\lambda_0 + 1)q_i) - a_{i+1})$$

and note that

$$\begin{aligned} q_{i+1} &\leq \frac{a_{i+1} - (a_i + \lambda_0 q_i) + (a_i + (\lambda_0 + 1)q_i) - a_{i+1}}{2} \\ &= q_i/2 \\ &\leq n/2^{i+1}. \end{aligned}$$

It follows that $a_{i+1} + q_{i+1} < n$. Furthermore, either $a_{i+1} + q_{i+1} = (a_i + \lambda_0 q_i) + 2(a_{i+1} - (a_i + \lambda_0 q_i))$ or $a_{i+1} + q_{i+1} = a_i + (\lambda_0 + 1)(a_i + q_i - a_i)$. Hence, $a_{i+1} + q_{i+1}$ is in the forward closure of $\{a_0, \ldots, a_{i+1}\}$. Therefore q_{i+1} has all desired properties.

Conclusion For all $0 \leq i \leq k$, there exists $q_i \in \{1, \ldots, n - a_i\}$ such that

1. $q_i \leq n/2^i$, and
2. $a_i + q_i = n$ or $a_i + q_i$ is in the forward closure of $\{a_0, \ldots, a_i\}$.

□

Lemma 3.3. *Let $R(P) = \{0 = a_0 < a_1 < \ldots < a_k\}$ be the irreducible period set of a string of length n. Then $k \leq \log_2(n)$.*

Proof. It follows from the Lemma 3.2 that there exists an integer $q_k \in \{1, \ldots, n - a_k\}$ such that $n/2^k \geq q_k$. Hence $k \leq \log_2(n)$. □

To count the number of irreducible period sets, we count the number of possibilities for each a_i with $1 \leq i \leq k$. We know that $a_0 = 0$ is fixed. The other a_i take values in the set $\{1, \ldots, n-1\}$.

Lemma 3.4. *Let $0 \leq i \leq k-1$. Then a_{i+1} can take at most $2^{1-i}n - 1$ possible values given $a_0, \ldots, a_i$.*

Proof. Let q_i be defined as in Lemma 3.2. We distinguish 3 cases:

1. If $a_{i+1} \leq a_i + q_i$, there are at most $q_i - 1 \leq n/2^i - 1$ possible values for a_{i+1} (note that $a_{i+1} \neq a_i + q_i$, because a_{i+1} cannot be in the forward closure of $\{a_0, \ldots, a_i\}$, nor can it be equal to n).
2. If $a_{i+1} \geq n - q_i$, there are at most $q_i \leq n/2^i$ possible values for a_{i+1}.
3. In the remaining case, $a_{i+1} \in [a_i + q_i + 1, n - q_i - 1]$.

Let us first show, that case 3 is impossible. For the sake of contradiction, assume we are in case 3. Since $a_i + q_i < n$, we know that $a_i + q_i$ is in the forward closure of $\{a_0, \ldots, a_i\}$ (by property 2 from Lemma 3.2). Hence, q_i is a period of $u[a_i \mathinner{..} n-1]$. Moreover $a_{i+1} - a_i$ is also a period of $u[a_i \mathinner{..} n-1]$. By the Fine and Wilf theorem, it follows that

(a) either $n - a_i < q_i + (a_{i+1} - a_i) - \gcd(q_i, a_{i+1} - a_i)$

(b) or $\gcd(q_i, a_{i+1} - a_i)$ is a period of $u[a_i \mathinner{..} n-1]$.

We are not in subcase (a) since by hypothesis $a_{i+1} \leq n - q_i - 1$. Suppose we are in subcase (b). Note that $a_i + \gcd(q_i, a_{i+1} - a_i) \leq a_i + q_i < a_{i+1}$ and that a_{i+1} is in the forward propagation of $\{a_0, \ldots, a_i, a_i + \gcd(q_i, a_{i+1} - a_i)\}$. It follows that a_{i+1} is not an irreducible period, which is a contradiction. Therefore both subcases (a) and (b) are impossible.

Summing other all three cases, we conclude that, given $a_0, \ldots, a_i$, there are at most

$$(n/2^i - 1) + n/2^i + 0 = 2^{1-i}n - 1$$

possibilities for a_{i+1}. □

Note that the bound of Lemma 3.4 is not tight: indeed, there are $n - 1$ possible values for a_1, while the lemma gives an upper bound of $2n-1$. However, this bound suffices to prove our asymptotic result. Since an autocorrelation is uniquely defined by its irreducible period set, it suffices to count the possible such sets $\{a_0, \ldots, a_k\}$ for all possible values of k. Recall that a_0 is fixed at 0 and that $k \leq \log_2(n)$ by Lemma 3.4. We thus derive a bound on the total number of autocorrelations by taking the product of all possibilities for a_{i+1} with i going from 0 to $k-1$ and sum this over all integers k from 1 to $\lfloor \log_2(n) \rfloor$, as follows:

$$\begin{aligned}
\kappa_n = |\Gamma_n| = |\Lambda_n| &\leq \sum_{k=1}^{\lfloor \log_2(n) \rfloor} \prod_{i=0}^{k-1} \left(2^{1-i}n - 1\right) \\
&\leq \sum_{k=1}^{\lfloor \log_2(n) \rfloor} \left(\left(2^{2-k}n - 1\right) \prod_{i=0}^{k-2} 2^{1-i}n \right).
\end{aligned}$$

Writing $2^{2-k} \prod_{i=0}^{k-2} 2^{1-i}n$ and $\prod_{i=0}^{k-2} 2^{1-i}n$ in exponential form, we get

$$\begin{aligned}
\kappa_n \leq \sum_{k=1}^{\lfloor \log_2(n) \rfloor} \Bigg(&\exp\left(\frac{-k(k-3)\ln(2)}{2} + k\ln(n) \right) \\
&- \exp\left(\frac{-(k-1)(k-4)\ln(2)}{2} + (k-1)\ln(n) \right) \Bigg).
\end{aligned}$$

Observe that this is a telescoping sum, so all but two terms cancel out.

$$\kappa_n \leq \exp\left(\frac{-\lfloor \log_2(n) \rfloor (\lfloor \log_2(n) \rfloor - 3)\ln(2)}{2} + \lfloor \log_2(n) \rfloor \ln(n) \right) - 1$$

Since $\frac{d}{dk}\left(\frac{-k(k-3)\ln(2)}{2} + k\ln(n)\right) = \frac{(-2k+3)\ln(2)}{2} + \ln(n)$ is positive for all $k \leq \log_2(n)$, we have

$$\begin{aligned}\kappa_n &< \exp\left(\frac{\ln(n)(3-\ln(n))}{2\ln(2)} + \frac{\ln^2(n)}{\ln(2)}\right) \\ &= \exp\left(\frac{3\ln(n)}{2\ln(2)} + \frac{\ln^2(n)}{2\ln(2)}\right).\end{aligned}$$

Taking the natural logarithm of both sides and dividing by $\ln^2(n)$, we get that

$$\frac{\ln(\kappa_n)}{\ln^2(n)} \leq \frac{1}{2\ln(2)} + \frac{3}{2\ln(2)\ln(n)},$$

thereby proving Theorem 3.1.

□

Corollary 3.4.1 (Asymptotic Convergence of κ_n)**.** *Let κ_n be the number of autocorrelations of length n. Then*

$$\frac{\ln \kappa_n}{\ln^2(n)} \to \frac{1}{2\ln(2)} \quad \text{as} \quad n \to \infty.$$

Proof. It follows from Theorem 3.1 that for $n \in \mathbb{N}_{\geq 2}$

$$\frac{\ln(\kappa_n)}{\ln^2(n)} \leq \frac{1}{2\ln(2)} + \frac{3}{2\ln(2)\ln(n)} = \frac{1}{2\ln(2)} + o(1).$$

The lower bound for κ_n from Theorem 5.1 in [14] indicates that asymptotically

$$\begin{aligned}\frac{\ln(\kappa_n)}{\ln^2(n)} &\geq \frac{1}{2\ln(2)}\left(1 - \frac{\ln(\ln(n))}{\ln(n)}\right)^2 + \frac{0.4139}{\ln(n)} - \frac{1.47123\ln(\ln(n))}{\ln^2(n)} + O\left(\frac{1}{\ln^2(n)}\right) \\ &= \frac{1}{2\ln(2)} - O\left(\frac{\ln(\ln(n))}{\ln(n)}\right).\end{aligned}$$

Combining this lower bound with our upper bound, we obtain

$$\frac{1}{2\ln(2)} - O\left(\frac{\ln\ln n}{\ln n}\right) \quad \leq \quad \frac{\ln \kappa_n}{\ln^2(n)} \quad \leq \quad \frac{1}{2\ln(2)} + o(1).$$

Using the classic *sandwich theorem*, we conclude that

$$\frac{\ln \kappa_n}{\ln^2(n)} \to \frac{1}{2\ln(2)} \quad \text{as} \quad n \to \infty$$

thereby proving the conjecture by Guibas and Odlyzko. □

The known values of κ_n are recorded in entry A005434 (see `https://oeis.org/A005434`) of the On-Line Encyclopedia of Integer Sequences [16]. Because, the enumeration of Γ_n takes exponential time, the list of κ_n values is limited to a few hundreds. In Figure 1, we compare the values of κ_n with the so-called Fröberg lower bound from [14], the upper bound of Guibas and Odlyzko [6], our new upper bound. The figure illustrates the improvement brought by the new upper bound compared to that given by Guibas and Odlyzko [6]. At $n = 500$, the lower bound, our new upper bound and the values of κ_n clearly differ, meaning the sequences are far from convergence at $n = 500$.

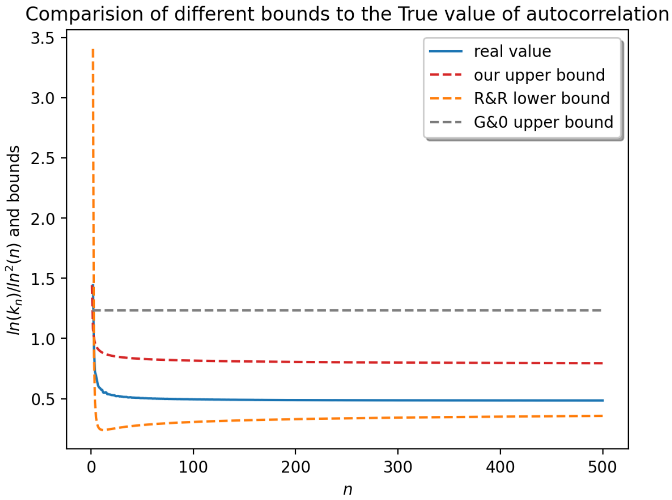

Figure 1: The values of $\ln k_n / \ln^2(n)$ for $n \leq 500$ are compared to: the upper bound of Guibas & Odlyzko [6], the Fröberg lower bound [14], and our upper bound. Our upper bound seems not so tight: the reason might be that n is small, as $\ln 500 \approx 6.2$.

4 Correlation

In this section, we show that the number of correlations between two strings of length n has the same asymptotic convergence behavior as the the number of autocorrelations of strings of length n.
In [7], Guibas and Odlyzko introduced the notion of *correlation* of two strings: it encodes the offset of possible overlaps between these two strings. In [6], the same authors investigate the self-overlaps of a string, which is then encoded in an *autocorrelation*. Before we start, let us define precisely the notion of correlation (which is illustrated in Table 2).

Definition 4.1 (Correlation)**.** For every pair of strings $(u, v) \in \Sigma^n \times \Sigma^m$, the correlation of u over v is the vector $t \in \{0,1\}^n$ such that

$$t[k] = \begin{cases} 1 & \text{if } u[i] = v[j] \text{ for all } i \in \{0, \dots, n-1\}, j \in \{0, \dots, m-1\} \\ & \text{with } i = j + k, \\ 0 & \text{otherwise} \end{cases}$$

for all $k \in \{0, \dots, n-1\}$.

Intuitively, we can find correlations as follows. For each index $i \in \{0, \dots, n-1\}$ we write v below u starting under the ith character of u. Then the ith element of the correlation is 1, if all pairs of characters that are directly above each other match, and 0 otherwise. See Table 2 for an example.

Observe, that if $v \in \Sigma^m$ is longer than $u \in \Sigma^n$, then the correlation of u over v equals the correlation of u over $v[0 \,.\,.\, n-1]$. Conversely, any binary vector $t \in \{0,1\}^n$ is the correlation of $u = t \in \{0,1\}^n$ over $v = 1 \in \{0,1\}^1$. Therefore

pos.	0	1	2	3	4	5	6	7	8	9	10	
u	`a`	`a`	`b`	`b`	`a`	`a`	-	-	-	-	-	t
v	`b`	`a`	`a`	`b`	`a`	`a`	-	-	-	-		0
	-	`b`	`a`	`a`	`b`	`a`	`a`	-	-	-	-	0
	-	-	`b`	`a`	`a`	`b`	`a`	`a`	-	-	-	0
	-	-	-	`b`	`a`	`a`	`b`	`a`	`a`	-	-	1
	-	-	-	-	`b`	`a`	`a`	`b`	`a`	`a`	-	0
	-	-	-	-	-	`b`	`a`	`a`	`b`	`a`	`a`	0

Table 2: The correlation of $u = \texttt{aabbaa}$ over $v = \texttt{baabaa}$ is $t = 000100$.

we will restrict ourselves to the interesting case where both strings have the same length.

Let Δ_n be the set of all correlations between two strings of the same length n and let δ_n be the cardinality of Δ_n. We can characterize Δ_n as follows.

Lemma 4.1. *The set of correlations of length n is of the form*

$$\Delta_n = \left\{ 0^{(n-j)} s_j \;\mid\; s_j \in \Gamma_j,\; j \in [0, n] \right\},$$

where Γ_j is the set of autocorrelations of length j.

Proof. Let $t = 0^{(n-j)} s_j$ with s_j the autocorrelation of some string w of length $0 \leq j \leq n$. Without loss of generality w does not start with the letter $\texttt{a}$. Let $u = \texttt{a}^{(n-j)} w$ and $v = w\texttt{b}^{(n-j)}$. Observe that the correlation of u over v is precisely $0^{(n-j)} s_j = t$. Therefore

$$\left\{ 0^{(n-j)} s_j \;\mid\; s_j \in \Gamma_j,\; j \in [0, n] \right\} \subseteq \Delta_n.$$

Conversely, let $u, v \in \Sigma^n$ and let t' be the correlation of u over v. We can write t' in the form $0^{(n-j)} s_j$, where s_j is a binary string starting with 1 (or is empty). If s_j is the empty string, then it is the only autocorrelation of length 0. Otherwise, there is a 1 at position $n - j$, which indicates that $u[n-j\,..\,n-1] = v[0\,..\,j-1]$. Moreover, s_j is the correlation of $u[n-j\,..\,n-1]$ over v. It follows that s_j is exactly the autocorrelation of $u[n-j\,..\,n-1] = v[0\,..\,j-1]$. Therefore

$$\Delta_n \subseteq \left\{ 0^{(n-j)} s_j \;\mid\; s_j \in \Gamma_j,\; j \in [0, n] \right\}.$$

□

In the above characterization, we consider strings over a finite alphabet and found that a correlation depends on some autocorrelation. As it is known that Γ_n is independent of the alphabet size (provided $|\Sigma| > 1$), the reader may wonder whether the number of correlations depends on it. In the Appendix, we show that the set of correlations for equally long strings is independent of the alphabet size, provided that Σ is not unary.

Now we have characterized Δ_n, we can easily deduce its cardinality.

Lemma 4.2. *Let κ_n be the number of autocorrelations of length n and δ_n the number of correlations between two strings of length n. Then*

$$\delta_n = \sum_{j=0}^{n} \kappa_j.$$

Proof. Since autocorrelations do not start with a zero, no two strings of the form $0^{(n-j)}s_j$ with $s_j \in \Gamma_j$ and $j \in [0,n]$ are the same. Therefore

$$\delta_n = |\Delta_n| = \left\{0^{(n-j)}s_j \;\mid\; s_j \in \Gamma_j,\; j \in [0,n]\right\} = \sum_{j=0}^{n} |\Gamma_j| = \sum_{j=0}^{n} \kappa_j.$$

□

Theorem 4.3 (Asymptotic Convergence of δ_n)**.** *Let δ_n be the number of correlations between two strings of length n. Then*

$$\frac{\ln \delta_n}{\ln^2(n)} \to \frac{1}{2\ln(2)} \quad \textit{as} \quad n \to \infty.$$

Proof. From Lemma 3.4 we know that for all $n \in \mathbb{N}_{\geq 2}$

$$\ln(\kappa_n) \leq \frac{\ln^2(n)}{2\ln(2)} + \frac{3\ln(n)}{2\ln(2)}.$$

It follows that for all $n \in \mathbb{N}_{\geq 2}$ we have

$$\begin{aligned}
\frac{\ln(\delta_n)}{\ln^2(n)} &= \ln\left(\sum_{i=0}^{n} \kappa_n\right) / \ln^2(n) \\
&\leq \ln\left(2 + (n-1)\exp\left(\frac{\ln^2(n)}{2\ln(2)} + \frac{3\ln(n)}{2\ln(2)}\right)\right) / \ln^2(n) \\
&\leq \left(\frac{\ln^2(n)}{2\ln(2)} + \frac{3\ln(n)}{2\ln(2)} + \ln(n)\right) / \ln^2(n) \\
&= \frac{1}{2\ln(2)} + o(1) \quad \text{as} \quad n \to \infty.
\end{aligned}$$

Conversely, using the fact that $\delta_n \geq \kappa_n$, we find

$$\frac{\ln \delta_n}{\ln^2(n)} \geq \frac{\ln \kappa_n}{\ln^2(n)} = \frac{1}{2\ln(2)} + o(1) \quad \text{as} \quad n \to \infty.$$

Again, by the sandwich theorem we conclude

$$\frac{\ln \delta_n}{\ln^2(n)} \to \frac{1}{2\ln(2)} \quad \text{as} \quad n \to \infty.$$

□

Acknowledgement

This work is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 956229 and the Netherlands Organisation for Scientific Research (NWO) through Gravitation-grant NETWORKS-024.002.003.

References

[1] Francine Blanchet-Sadri and S. Duncan. Partial words and the critical factorization theorem. *Journal of Combinatorial Theory, Series. A*, 109(2):221–245, 2005.

[2] Francine Blanchet-Sadri, Justin Fowler, Joshua D. Gafni, and Kevin H. Wilson. Combinatorics on partial word correlations. *J. Comb. Theory, Ser. A*, 117(6):607–624, 2010.

[3] Francine Blanchet-Sadri, Joshua D. Gafni, and Kevin H. Wilson. Correlations of partial words. In Wolfgang Thomas and Pascal Weil, editors, *STACS 2007, 24th Annual Symposium on Theoretical Aspects of Computer Science, Aachen, Germany, February 22-24, 2007, Proceedings*, volume 4393 of *Lecture Notes in Computer Science*, pages 97–108. Springer, 2007.

[4] Nathan J Fine and Herbert S Wilf. Uniqueness theorems for periodic functions. *Proceedings of the American Mathematical Society*, 16(1):109–114, 1965.

[5] Daniel Gabric, Narad Rampersad, and Jeffrey Shallit. An inequality for the number of periods in a word. *International Journal of Foundations of Computer Science*, 32(05):597–614, Jun 2021.

[6] Leonidas J. Guibas and Andrew M. Odlyzko. Periods in strings. *Journal of Combinatorial Theory, Series. A*, 30:19–42, 1981.

[7] Leonidas J. Guibas and Andrew M. Odlyzko. String overlaps, pattern matching, and nontransitive games. *Journal of Combinatorial Theory, Series A*, 30(2):183–208, 1981.

[8] Vesa Halava, Tero Harju, and Lucian Ilie. Periods and binary words. *Journal of Combinatorial Theory, Series A*, 89(2):298–303, 2000.

[9] Stepan Holub and Jeffrey O. Shallit. Periods and borders of random words. In Nicolas Ollinger and Heribert Vollmer, editors, *33rd Symposium on Theoretical Aspects of Computer Science, STACS 2016, February 17-20, 2016, Orléans, France*, volume 47 of *LIPIcs*, pages 44:1–44:10. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2016.

[10] M. Lothaire, editor. *Combinatorics on Words*. Cambridge University Press, second edition, 1997.

[11] Ora E. Percus and Paula A. Whitlock. Theory and Application of Marsaglia's Monkey Test for Pseudorandom Number Generators. *ACM Transactions on Modeling and Computer Simulation*, 5(2):87–100, April 1995.

[12] Sven Rahmann and Eric Rivals. On the distribution of the number of missing words in random texts. *Combinatorics, Probability and Computing*, 12(01), Jan 2003.

[13] Eric Rivals and Sven Rahmann. Combinatorics of Periods in Strings. In F. Orejas, P. Spirakis, and J. van Leuween, editors, *Proc. of the 28th ICALP*, volume 2076 of *LNCS*, pages 615–626. Springer Verlag, 2001.

[14] Eric Rivals and Sven Rahmann. Combinatorics of periods in strings. *Journal of Combinatorial Theory, Series A*, 104(1):95–113, 2003.

[15] Stéphane Robin, François Rodolphe, and Sophie Schbath. *DNA, Words and Models*. Cambrigde University Press, 2005.

[16] Neil J. A. Sloane. The on-line encyclopedia of integer sequences. Published electronically at https://oeis.org, 2022.

[17] William F. Smyth. *Computating Pattern in Strings*. Pearson - Addison Wesley, 2003.

Appendix

Guibas and Odlyzko showed that for every autocorrelation, there exists a string over a binary alphabet with that autocorrelation [6]. A nice alternative constructive proof appears in [8]. We will now show that the same holds for arbitrary correlations of equally long strings.

Corollary 4.3.1. *For any $t \in \Delta_n$, there exist $u, v \in \{\mathtt{a}, \mathtt{b}\}^n$ such that the correlation of u over v is t.*

Proof. Let t be the correlation of u' over v' with $u', v' \in \Sigma^n$. By Lemma 4.1, we can write $t = 0^{(n-j)}s_j$, where $s_j \in \{0,1\}^j$ is the autocorrelation of $u'[n-j\,..\,n-1] = v'[0\,..\,j-1]$. By the result of Guibas and Odlyzko, we know that there also exists some binary string $w \in \{\mathtt{a}, \mathtt{b}\}^j$ with the same autocorrelation. Without loss of generality this vector starts with $\mathtt{b}$. It follows that the constructed strings $u = \mathtt{a}^{(n-j)}w$ and $v = w\mathtt{b}^{(n-j)}$, which have a correlation of t by the proof of Lemma 4.1, use the same binary alphabet. □

We conclude that the number of correlations between strings of equal length is alphabet-independent (i.e. every alphabet of size at least 2 gives rise to the same set of correlations).

Remark. Such a binary string w can be constructed from $u'[n-j\,..\,n-1]$ in linear time using the algorithm of Halava, Harju and Ilie [8]. Therefore u and v can also be constructed in linear time given u' and v'.