Algorithmica (2025) 87:690–711
https://doi.org/10.1007/s00453-025-01295-y

Convergence of the Number of Period sets in Strings

Eric Rivals[1] · Michelle Sweering[2] · Pengfei Wang[1]

Received: 24 June 2024 / Accepted: 17 January 2025 / Published online: 21 February 2025

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

Keywords Autocorrelation · Period · Border · Combinatorics · Correlation · Periodicity · Upper bound · Asymptotic convergence

1 Introduction

A word overlaps 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

A preliminary version of this work appeared in the proceeding of the 50th International Colloquium on Automata, Languages, and Programming (ICALP 2023) [1].

✉ Eric Rivals
rivals@lirmm.fr

Michelle Sweering
michelle.sweering@cwi.nl

Pengfei Wang
pengfei.wang@lirmm.fr

[1] LIRMM, Université Montpellier, CNRS, Montpellier, France

[2] CWI, Amsterdam, The Netherlands

to the suffix is called a period. The dual notions of period and border are critical concepts in combinatorics on words: important definitions such as periodic and primitive words, and 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 [2, 3]. In computer science, in the field of string algorithms (a.k.a. stringology), pattern-matching algorithms heavily exploit borders/periods to optimize the search for occurrences of a word in a text [4]. For clarity, note that the terms *word* and *string* both mean a sequence of letters taken from an alphabet. 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, the distance between successive occurrences, etc.) [5]. The notion of autocorrelation was extended to describe how two distinct words can have overlapping occurrences in the same text. This is used, for instance, to study the number of missing words in random texts [6] or to design procedures for testing pseudo-random number generators [7].

An autocorrelation is a binary vector representing the set of periods of a word. The concept of autocorrelation was introduced by Guibas and Odlyzko in [8]. They gave a characterization of autocorrelations and proved the following bounds on κ_n—the cardinality of the set Γ_n of autocorrelations of words of length n;

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

They conjectured that $\ln \kappa_n$ is asymptotic to a constant times $(\ln n)^2$. Later, Rivals and Rahmann [9] studied the combinatorial structure of the set of autocorrelations Γ_n, and improved the lower bound on κ_n as follows:

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

However, up to now, there has been no improvement on the upper bound on κ_n. In this work (Sect. 3), we apply the notion of irreducible period sets introduced by Rivals and Rahmann [9, 10] to prove that

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

Together with known asymptotic lower bounds [9], we find that

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

thus resolving the conjecture of Guibas and Odlyzko.

In their paper about autocorrelations [8], Guibas and Odlyzko also introduced the notion of the correlation of two words. For two words 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 word with itself. In Sect. 4, we first characterize the set, denoted by Δ_n, of distinct correlations between words of length n, and then show that the ratio $\frac{\ln \delta_n}{(\ln n)^2}$, where δ_n denotes the cardinality of Δ_n, has the same asymptotic convergence behavior as the ratio $\frac{\ln \kappa_n}{(\ln n)^2}$ (i.e., the ratio involving the number of autocorrelations of words of length n), that is

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

The seminal characterization of autocorrelations in Γ_n from [8] includes a recursive predicate that certifies whether a binary vector is an autocorrelation of length n. The *basic period* of an autocorrelation is defined as its smallest nonzero period if one exists, or n otherwise. This predicate partitions autocorrelations of Γ_n in two cases: when the basic period of an autocorrelation is smaller than or equal to $\lfloor \frac{n}{2} \rfloor$ it is in case a; if not, it is in case b. Enumeration of Γ_n for $n \leq 60$ shows that autocorrelations are unequally distributed among cases a and b. Thus, in Sect. 5 we study the cardinalities, denoted by κ_n^a and κ_n^b, of the two subsets of Γ_n corresponding to case a and to case b, respectively. For both κ_n^a and κ_n^b, we demonstrate a convergence result similar to the one obtained for κ_n in Theorem 3.4.1. Section 5 is new compared to the previous version of this work presented at ICALP 2023 [1]. Finally, we conclude in Sect. 6 with related open questions.

1.1 Related Works

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

For instance, the question of the average period of a random word was investigated in [11]. Clearly, the number of periods of a word of length n lies between 1 and n. A recent work exhibits an upper bound on the number of periods of a word as a function of its *initial critical exponent*—a characteristic of the word related to its degree of periodicity [12], but this was not used to bound the number of period sets. Finally, the combinatorics of period sets was also investigated in depth for a generalization of the notion of words, called *partial words* [13]. In partial words, some positions may contain a *don't care* symbol, which removes some constraints of equality between positions. To study the combinatorics of period sets, Blanchet-Sadri et al. defined weak and strong periods, and proved several important theorems [14], including lower and upper bounds on the number of binary and ternary autocorrelations [13, 15]. However, the cardinality of the family of period sets differs between normal words and partial words, and a tight upper bound for normal words cannot be deduced from that for partial words. Several works investigate sets of words with constraints (either absence or presence) on their mutual overlaps: mutually bordered (overlapping) pairs of words are studied in [16], while algorithms for constructing a set of mutually unbordered words (also called cross-bifix-free words or non-overlapping words) are provided in [17–22]. These algorithms aim to compute *non-overlapping codes*, which are used for frame synchronization in network communication or for DNA-based data storage [17, 23].

2 Preliminaries

Let Σ be a finite *alphabet*: a set of *letters* of cardinality σ. A sequence of elements of Σ is called a *string* or a *word*. Let us denote by Σ^n the set of all strings of length n over Σ, with $n \in \mathbb{N}$. For a string x, $|x|$ denotes the *length* of x. For two strings x, y, we let xy denote their concatenation and the k-fold concatenation of x with itself by x^k for any integer $k > 0$. A string $u = u[0\ddot{n} - 1] \in \Sigma^n$ is a sequence of n letters over a finite alphabet Σ. For any $0 \leq i \leq j \leq n-1$, we let $u[i\ddot{j}]$ denote the substring starting at position i and ending at position j. In particular, $u[0\ddot{j}]$ denotes a prefix, $u[i\ddot{n} - 1]$ denotes a suffix, and $u[j, j] = u[j]$ denotes the j-th letter of u. Throughout this paper, all our strings and vectors are indexed starting at position 0.

2.1 Periodicity

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

Definition 2.1 (Period) The string u has a period $p \in \{1, \ldots, n-1\}$ if and only if for any $0 \leq i, j \leq n-1$ such that $i \equiv j \pmod{p}$, we have $u[i] = u[j]$. Moreover, we consider that $p = 0$ is a period of any string.

An equivalent definition is the following.

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

Recall that the *basic period* of u is the smallest non-zero period of a string. We let $\pi(u)$ denote the basic period of u. The *period set* of a string u is the set of all its periods and is denoted by $P(u)$. We now list 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 $i \equiv j \pmod{kp}$, then we also have $i \equiv j \pmod{p}$, 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 $w = u[p\ddot{n} - 1]$. 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 $u[0\ddot{n} - p - 1] = u[p\ddot{n} - 1]$, while q is a period of w implies $w[0\ddot{n} - p - q - 1] = w[q\ddot{n} - p - 1]$.

As w is the suffix of u starting at position p, we can combine the above results to find that

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

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 $0 \le q \le p$. Then the prefix and the suffix of length $(n - q)$ have the period $(p - q)$.*

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

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

It follows that $u[0\ddot{n} - q - 1]$ has period $p - q$. Similarly the suffix of u of length $(n - q)$ 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 $r \mid p$. 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 $i \equiv j \pmod{r}$. We can write $v = u[h\ddot{k}]$ with $0 \le h < k \le n - 1$. Since v has length at least p, there exist $i', j' \in [h, k]$ such that $i \equiv i' \pmod{p}$ and $j \equiv j' \pmod{p}$. By Definition 2.1 of period, we have $u[i] = u[i']$ and $u[j] = u[j']$. Note that $i' \equiv i \equiv j \equiv j' \pmod{r}$, 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 [24], a.k.a. the periodicity lemma, for which a short proof was provided by Halava and colleagues [25].

Theorem 2.5 *(Fine and Wilf) Let p, q be periods of $u \in \Sigma^n$. If $n \ge p + q - \gcd(p, q)$, 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 for all $i \in \{0, \ldots, n - 1\}$,

$$s[i] = \begin{cases} 1, & \text{if } i \text{ is a period of } u; \\ 0, & \text{otherwise.} \end{cases}$$

Table 1 This table illustrates the set of period and the autocorrelation of the word u = abbaabba of length 8. The first copy of the word u is shown on the second line. Another copy of u is displayed on (each) line $(3+i)$ shifted by i positions to the right, with i going from 0 to 7. If the suffix of the copy of u matches the prefix of the first copy u on line 2, then i is a period, and both the line and the corresponding position/shift (on the first line) are colored in blue. The last column contains the autocorrelation of u, with 1 bits corresponding to periods colored in blue

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

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

Example 1 We consider the word u = abbaabba of length 8. 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 [8] show that *any alphabet of size at least two gives rise to the same set of correlations* (Corollary 5.1). Autocorrelations have many other useful properties [8, 9]. The most significant one for our work is the following.

Lemma 2.6 *(Lemma 3.1 [8] and Theorem 1.3 [9]) If $s \in \{0,1\}^n$ is an autocorrelation and $s[i] = 1$, then $s[i..n-1]$ is the autocorrelation of $u[i..n-1]$.*

Proof Note that $s[i] = 1$ means that i is a period of u. Suppose $s[i+j] = 1$. Then $i + j$ is a period of u. Thus $u[i..n-1]$ has period $(i+j) - i = j$ by Lemma 2.3. Conversely, suppose $u[i..n-1]$ 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 if j is a period of $u[i..n-1]$, and equivalently $s[i..n-1]$ is the autocorrelation of $u[i..n-1]$. □

2.3 Irreducible Period Set

To prove the upper bound on the number of autocorrelations, we use the notion of irreducible period sets as introduced by Rivals and Rahmann [9]. 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 first explain what forward propagation is.

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

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

The forward closure $\mathrm{FC}_n(S)$ of a set $S \subseteq \{0, \dots, n-1\}$ (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 $S \subseteq \{0, \dots, n-1\}$. Its forward closure $\mathrm{FC}_n(S)$ is the minimum superset of S such that for every $p, q \in \mathrm{FC}_n(S)$ and $k \geq 0$ with $p \leq q$ and $p+k(q-p) < n$, we have

$$p + k(q-p) \in \mathrm{FC}_n(S).$$

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 $\mathrm{FC}_n(P) = P$ by the forward propagation rule. We now give a useful characterization of an irreducible period set as the set of periods that are not in the forward closure of the set of all smaller periods. Consequently, every period set has exactly one irreducible period set, whose elements we call *irreducible periods*.

Recall that for a given length n, we let Γ_n denote the set of all period sets. Formally stated Γ_n is defined as

$$\Gamma_n = \{S \subseteq \{0, 1, \dots, n-1\} : \exists u \in \Sigma^n \text{ such that } P(u) = S\}.$$

As in [9], for a given length n, we let Λ_n denote the set of all irreducible period sets:

$$\Lambda_n = \{T \subseteq \{0, 1, \dots, n-1\} : \exists u \in \Sigma^n \text{ such that } R(P(u)) = T\}.$$

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*

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

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

1. $p \notin \{q \in P \mid q \notin \mathrm{FC}_n(P \cap [0, q-1])\} \implies p \notin R(P)$ and
2. $p \in \{q \in P \mid q \notin \mathrm{FC}_n(P \cap [0, q-1])\} \implies p \in R(P)$.

1. Suppose $p \notin \{q \in P \mid q \notin \mathrm{FC}_n(P \cap [0, q-1])\}$, or equivalently $p \in \mathrm{FC}_n(P \cap [0, p-1])$. Since $FC_n(R(P) \cap [0, p-1] \subseteq FC_n(R(P)) \cap [0, p-1])$,

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

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

2. Suppose on the other hand that $p \notin \mathrm{FC}_n(P \cap [0, p-1])$. Then $p \notin \mathrm{FC}_n(P \setminus \{p\})$. As

$$\mathrm{FC}_n(P \setminus \{p\}) \supseteq \mathrm{FC}_n(R(P) \setminus \{p\}),$$

then $p \notin \mathrm{FC}_n(R(P) \setminus \{p\})$. However, as $p \in P$ and $P = \mathrm{FC}_n(R(P))$, 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 prove that $\ln \kappa_n$ converges to $c \cdot (\ln n)^2$, 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*

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

To prove this theorem, we need several lemmas.

Lemma 3.2 *Let $u \in \Sigma^n$ with period set P and irreducible period set $R(P) = \{0 = a_0 < \cdots < a_i < \cdots < a_k < n\}$. 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 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 $0 \leq i < k$, 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 $\{a_0, \ldots, a_i\}$.

Induction 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 $\{a_0, \dots, a_{i+1}\}$.

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

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

is in the forward closure of $\{a_0, \dots, a_i\}$ for all integers $0 \leq \lambda \leq (n-1-a_i)/q_i$. Since a_{i+1} is an irreducible period, there exists 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 bounded below and above by two successive, non-irreducible periods generated from a_i and q_i using the FPR (or $n \leq a_i + (\lambda_0+1)q_i$). We pick

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

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 $a_{i+1} + q_{i+1} < n$. Consequently, 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 $\{a_0, \dots, a_{i+1}\}$. Therefore q_{i+1} has all desired properties.

The lemma now follows, by induction on i. For all $0 \leq i \leq k$, there exists $q_i \in \{1, \dots, n - a_i\}$ 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 $\{a_0, \dots, a_i\}$.

□

Lemma 3.3 *Let $R(P) = \{0 = a_0 < a_1 < \cdots < a_k\}$ 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 $q_k \in \{1, \dots, n - a_k\}$ 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 $\{1, \dots, n-1\}$.

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 $a_0, \dots, a_i$.*

Proof Let q_i be defined as in Lemma 3.2. We distinguish three 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 $u[a_i \ddot{n} - 1]$. Moreover $a_{i+1} - a_i$ is also a period of $u[a_i \ddot{n} - 1]$. 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 $u[a_i \ddot{n} - 1]$.

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 because $a_i + \gcd(q_i, a_{i+1} - a_i)$ is a period by Lemma 2.2. Therefore both subcases (a) and (b) are impossible by Lemma 3.2. Summing over cases 1 and 2 (since case 3 is impossible), 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.

Proof (of Theorem 3.1) 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 3.3. 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 $2^{2-k}n\prod_{i=0}^{k-2}2^{1-i}n$ and $\prod_{i=0}^{k-2}2^{1-i}n$ in exponential form, we get

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

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

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

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 $k\le \log_2 n$, we have

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

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

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

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*

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

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

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

The lower bound for κ_n from Theorem 5.1 in [9] indicates that asymptotically

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

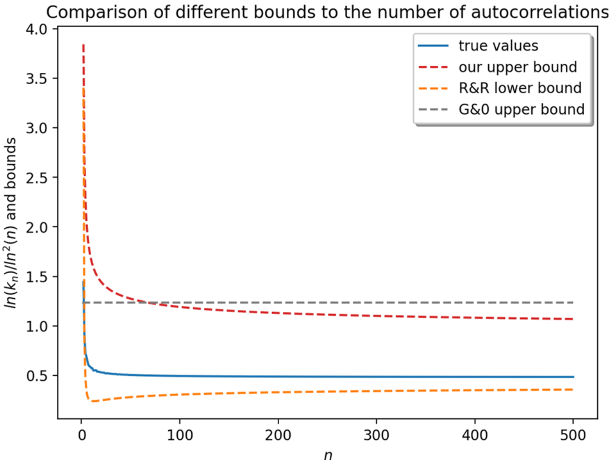

Fig. 1 The true values of $\ln k_n/(\ln n)^2$ for $n \leq 500$ are compared to the upper bound of Guibas & Odlyzko (G&O upper bound) [8], the Fröberg lower bound (R&R lower bound) [9], and our upper bound. Our upper bound seems not so tight: the reason is that n is small, as $\ln 500 \approx 6.2$

Combining this lower bound with our upper bound, we obtain

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

Using the classic *sandwich theorem*, we conclude that

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

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 [26]. Because the enumeration of Γ_n takes exponential time, the list of κ_n values is limited to a few hundred. In Fig. 1, we compare the values of κ_n with the so-called Fröberg lower bound from [9], the upper bound of Guibas and Odlyzko [8], and our new upper bound. The figure illustrates the improvement brought by the new upper bound compared to that given by Guibas and Odlyzko [8]. 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 $n = 500$.

4 Characterization of Correlations and Growth of Δ_n

In this section, we show that the number of correlations between two strings of length n has the same asymptotic convergence behavior as the number of autocorrelations of strings of length n.

Table 2 The correlation of word u = aabbaa over word v = baabaa (both of length 6) is t = 000100. This table is organized as Table 1—see the corresponding caption for details

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

In [27], Guibas and Odlyzko introduced the notion of *correlation* of two strings: it encodes the offset of possible overlaps between these two strings. In [8], the same authors investigate the self-overlaps of a string, which are then encoded in an *autocorrelation*. Before we start, let us precisely define 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 for all $k \in \{0, \ldots, n-1\}$,

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

Intuitively, we can find correlations as follows. For each index $i \in \{0, \ldots, n-1\}$ 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 $v[0\ddot{n} - 1]$. 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 we 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. 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 j with $0 \le j \le n$. Without loss of generality, w does not start with the letter a. Let

$u = \mathtt{a}^{(n-j)}w$ and $v = w\mathtt{b}^{(n-j)}$. 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 $u[n - j\ddot{n} - 1] = v[0\ddot{j} - 1]$. Moreover, s_j is the correlation of $u[n - j\ddot{n} - 1]$ over v. It follows that s_j is exactly the autocorrelation of $u[n - j\ddot{n} - 1] = v[0\ddot{j} - 1]$. Therefore

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

□

4.1 Alphabet Independence of Δ_n

In the above characterization, we consider strings over a finite alphabet and find 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. Below, we show that the set of correlations for equally long strings is independent of the alphabet size, provided that Σ is not unary.

Guibas and Odlyzko showed that for every autocorrelation, there exists a string over a binary alphabet with that autocorrelation [8]. A nice alternative constructive proof appears in [25]. Let us now show that the same holds for arbitrary correlations of equally long strings.

Corollary 4.1.1 *For every $t \in \Delta_n$, 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 $u'[n - j\ddot{n} - 1] = v'[0\ddot{j} - 1]$. 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, we can assume that w 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 $u'[n - j\ddot{n} - 1]$ in linear time using the algorithm of Halava, Harju, and Ilie [25]. Therefore u and v can also be constructed in linear time given u' and v'.

4.2 Growth of the Cardinality of Δ_n as a Function of the Word Length

Now that we have characterized Δ_n, let us give the formula to count correlations. Let δ_n be the cardinality of Δ_n. The characterization of Lemma 4.1 implies the following formula for δ_n.

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

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

□

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

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

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

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

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

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

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

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

Again, by the sandwich theorem, we conclude

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

□

5 Convergence Behaviors of κ_n^a and κ_n^b

In [8], Guibas and Odlyzko exhibit a recursive predicate, denoted by Ξ, which characterizes if a binary vector of length n belongs to Γ_n or not. The predicate distinguishes two cases, denoted by a and b, depending on whether the potential basic period (that is, the position of the second leftmost position with a one) is smaller than or equal to $\lfloor \frac{n}{2} \rfloor$ (case a) or not (case b). Thus, the predicate Ξ partitions Γ_n into two subsets of autocorrelations that do satisfy this condition or do not. Let us denote these two subsets by Γ_n^a and Γ_n^b respectively, and their respective cardinalities by κ_n^a and by κ_n^b. Hence, we have $\Gamma_n = \Gamma_n^a \sqcup \Gamma_n^b$ and $\kappa_n = \kappa_n^a + \kappa_n^b$.

The predicate Ξ emphasizes that the most important variable of an autocorrelation is its basic period, since in case a it imposes constraints on larger periods. This raises the question of comparing the growth with n of Γ_n^a and Γ_n^b; in particular, do the ratios $\frac{\ln(\kappa_n^a)}{(\ln n)^2}$ and $\frac{\ln \kappa_n^b}{(\ln n)^2}$ converge as does $\frac{\ln(\kappa_n)}{(\ln n)^2}$?

Figure 2, which plots κ_n^a, κ_n^b, and κ_n as a function of the word length n for $n \leq 60$, suggests that the three cardinalities have similar exponential growth. However, if κ_n^b can be shown to be non-decreasing, this is not the case for κ_n^a, which sometimes decreases. Figure 3 displays an example for $n = 57$ of the distribution of the number of period sets (on y-axis) with a given basic period (on x-axis). Clearly, the counts for basic periods smaller than $\lfloor \frac{n}{2} \rfloor = 27$, which correspond to case a, follow a different distribution than those whose basic periods lie strictly above 27. This example suggests that perhaps the functions κ_n^a and κ_n^b could behave quite differently.

Clearly, both κ_n^a and κ_n^b are both bounded from above by κ_n. In the sequel, we show that the abovementioned ratios converge towards $\frac{1}{2\ln(2)}$ when n tends to ∞, and thus that they admit the same limit as $\frac{\ln(\kappa_n)}{(\ln n)^2}$ (see Corollary 3.4.1). In both cases, we exploit lower bounds that also converge towards this limit and use the sandwich theorem.

5.1 Convergence Behavior of κ_n^b

In this subsection, we show that $\ln \kappa_n^b / (\ln n)^2$ converges asymptotically to $\frac{1}{2\ln(2)}$.

Theorem 5.1

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

Proof To get a lower bound for κ_n^b, we use Theorem 5.1 from [9], which originally states a lower bound for κ_n. In their proof, the authors only consider autocorrelations of case b: they show for any integer $n > 0$ that $\kappa_n^b \geq \frac{1}{2} \cdot B_{n+1}$, where B_n is the number

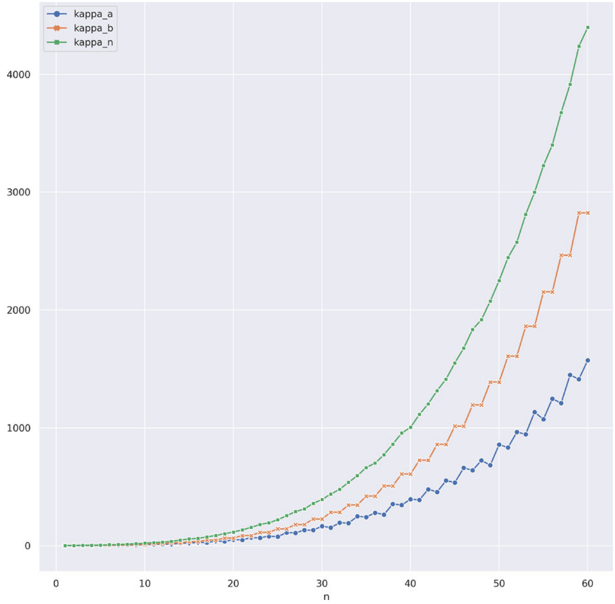

Fig. 2 Exponential growth of κ_n^a, κ_n^b, and their sum κ_n

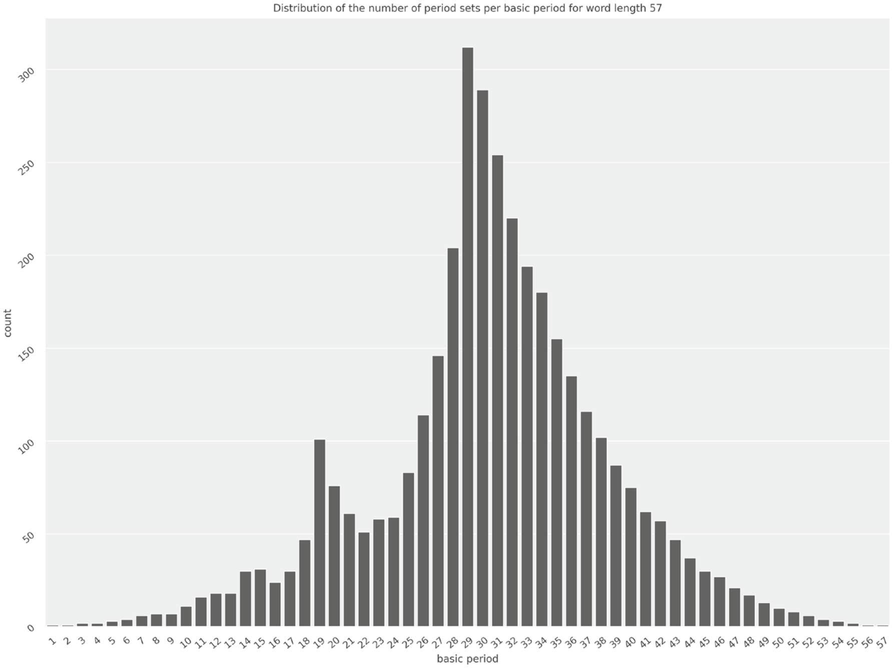

Fig. 3 The distribution of the number of autocorrelations of length $n = 57$ as a function of their basic period

of binary partitions of n (i.e., the number of ways to write n as an unordered sum of powers of 2). The number B_n of binary partitions of n corresponds to the entry A018819 of the Encyclopedia of Integer Sequences [26]. Reformulating Theorem 5.1 from [9], we get that κ_n^b asymptotically satisfies

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

Combining the above lower bound and the upper bound from Theorem 3.1, we obtain

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

Using the *sandwich theorem*, we conclude that $\frac{\ln \kappa_n^b}{(\ln n)^2} \rightarrow \frac{1}{2\ln(2)}$ when n tends to ∞. □

5.2 Convergence Behavior of κ_n^a

To demonstrate the convergence of the ratio of $\ln(\kappa_n^a)/(\ln n)^2$, we first derive a lower bound for κ_n^a stated in Corollary 5.2.1. For any word u, let us denote the autocorrelation of u by $s(u)$. The idea behind Lemma 5.2 is to identify a subset of Γ_{3n}^a in which each autocorrelation includes as a substring a distinct autocorrelation of Γ_n. Thus, it allows bounding κ_{3n}^a from below by κ_n^a. The same idea is also used to get a lower bound for κ_{3n-1}^a and for κ_{3n-2}^a.

Lemma 5.2 *Let $n \geq 3$ and let w be any word of Σ^n. We let w_1 and w_2 denote, respectively, the prefixes of length $n-1$ and $n-2$ of w. Let us define three subsets of autocorrelations of length $3n$, $3n-1$, and $3n-2$, and their cardinalities as follows:*

- $\Gamma_{3n}^c = \{s(w^3) : w \in \Sigma^n\}$ *and its cardinality* κ_{3n}^c.
- $\Gamma_{3n-1}^c = \{s(w^2w_1) : w \in \Sigma^n\}$ *and its cardinality* κ_{3n-1}^c.
- $\Gamma_{3n-2}^c = \{s(w^2w_2) : w \in \Sigma^n\}$ *and its cardinality* κ_{3n-2}^c.

Then the following three set inclusions and three corresponding inequalities between cardinalities hold:

1. $\Gamma_{3n}^c \subset \Gamma_{3n}^a$,
2. $\Gamma_{3n-1}^c \subset \Gamma_{3n-1}^a$,
3. $\Gamma_{3n-2}^c \subset \Gamma_{3n-2}^a$,

4. $\kappa_{3n}^a \geq \kappa_{3n}^c \geq \kappa_n$,
5. $\kappa_{3n-1}^a \geq \kappa_{3n-1}^c \geq \kappa_{n-1}$,
6. $\kappa_{3n-2}^a \geq \kappa_{3n-2}^c \geq \kappa_{n-2}$.

Proof • First, we prove the three set inclusions. Let $s(w^3)$ be an element of Γ_{3n}^c. Since w^3 is a power of the word $w \in \Sigma^n$, n and $2n$ are periods of w^3. Hence, $\pi(w^3)$, the basic period of w^3, satisfies $\pi(w^3) \leq n \leq \lfloor\frac{3n}{2}\rfloor$, and thus $s(w^3)$ belongs to Γ_{3n}^a. This yields $\Gamma_{3n}^c \subset \Gamma_{3n}^a$. Applying the same reasoning to an element of Γ_{3n-1}^c

or Γ^c_{3n-2}, we get $\pi(w^2w_1) \le n \le \lfloor\frac{3n-1}{2}\rfloor$ and $\pi(w^2w_2) \le n \le \lfloor\frac{3n-2}{2}\rfloor$, which imply $\Gamma^c_{3n-1} \subset \Gamma^a_{3n-1}$ and $\Gamma^c_{3n-2} \subset \Gamma^a_{3n-2}$, respectively.

- Proof of the three inequalities. Note that the left-hand sides of the inequalities are consequences of the set inclusions. Thus, we focus on their right-hand sides. For any w in Σ^n, we claim that $s(w)$ is a suffix of $s(w^3)$, $s(w_1)$ is a suffix of $s(w^2w_1)$, and $s(w_2)$ is a suffix of $s(w^2w_2)$. Let us show the last claim (i.e., the case of $s(w_2)$); the proof of two others are similar. First note that w_2 is a border of w^2w_2, since w_2 is a prefix of w. Thus, by the property of borders (see Lemma 1.3.1 of [4, Chap. 1]), all other borders shorter than $|w_2|$ of w^2w_2 are only and necessarily borders of w_2. Thus, the autocorrelation of w_2 is a suffix of that of w^2w_2. It follows that $\kappa_{n-2} \le \kappa^c_{3n-2}$, what we wanted. Proving the right-hand sides of the two other inequalities goes along the same line.

□

The incremental algorithm to enumerate Γ_n from Γ_{n-1} given in [28] implies that for every length $n > 0$, we have $\kappa_n \le 2\kappa_{n-1}$. Combining this with the inequalities of Lemma 5.2, we obtain a bound for κ^a_n stated in Corollary 5.2.1. When $n \in [1, 2]$ one can check that $\kappa^a_n \ge \kappa_{\lceil\frac{n}{3}\rceil}$.

Corollary 5.2.1 *For all $n \in \mathbb{N}^+$, we have $\kappa^a_n \ge \kappa_{\lceil\frac{n}{3}\rceil}/4$.*

Note that when $n \to \infty$, the asymptotic lower bound of $\frac{\ln(\kappa_{\lceil\frac{n}{3}\rceil})}{(\ln\lceil\frac{n}{3}\rceil)^2}$ is the same as that of $\frac{\ln(\kappa_n)}{(\ln n)^2}$. Thus, we obtain the desired convergence result.

Theorem 5.3

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

Proof Combining the upper bound $\kappa^a_n \le \kappa_n$ and the lower bound from Corollary 5.2.1, we get that there exists some $N \in \mathbb{N}$ such that for $n > N$, the following inequalities are true

$$\frac{\ln(\frac{\kappa_{\lceil\frac{n}{3}\rceil}}{4})}{(\ln\lceil\frac{n}{3}\rceil)^2} = \frac{\ln(\frac{\kappa_{\lceil\frac{n}{3}\rceil}}{4})}{(\ln n - O(1))^2} \le \frac{\ln(\kappa^a_n)}{(\ln n)^2} \le \frac{\ln(\kappa_n)}{(\ln n)^2} \le \frac{1}{2\ln(2)} + \frac{3}{2\ln n}.$$

From the lower bound from Theorem 5.1 in [9] and from Theorem 3.1, asymptotically, we have

$$\frac{1}{2\ln(2)} - O\left(\frac{\ln\ln n}{\ln n}\right) \le \frac{\ln\kappa^a_n}{(\ln n)^2} \le \frac{1}{2\ln(2)} + o(1)$$

and using the *sandwich theorem* we conclude that

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

□

6 Conclusion

Although this work closes a long-standing conjecture about the combinatorics of period sets (a.k.a. autocorrelations), we wish to mention that its topic still offers several interesting open questions regarding the sets Γ_n and Δ_n. Each string has a unique period set, but many strings can share the same. In other words, the mapping from Σ^n to Γ_n is surjective. Two recursive algorithms were proposed to compute the so-called population size of a period set in general, i.e., the number of strings of Σ^n sharing a given period set [8, 9], but a closed-form expression for computing population sizes remains open. Note that a specific recurrence is known for counting bifix-free (a.k.a. unbordered) words, which correspond to autocorrelation 10 . . . 0 [29]. Globally, an issue is to understand how the σ^n words of Σ^n are distributed among the κ_n autocorrelations of Γ_n, and how this distribution evolves with the alphabet cardinality, σ, and with the string length, n. A similar question arises regarding the relation of pairs of strings to correlations [30], i.e, the mapping between $\Sigma^n \times \Sigma^n$ to Δ_n, which is of interest for pattern matching and vocabulary statistics [6, 27, 31].

The definition of correlation for an ordered pair of string (u, v) is asymmetrical (see Definition 4.1 on page 13). An alternative could be to define a vector of length $2n - 1$ that combines the correlations for both pairs (u, v) and (v, u) (for this, it suffices to allow k to lie in the range $[-(n - 1), n - 1]$ in Definition 4.1). Let us term this vector the *bidirectional correlation* of (u, v). Now, for some length n, consider t and t' two correlations of Δ_n. It turns out that the combination of t and t' does not necessarily form a bidirectional correlation. Thus, the characterization of the set of bidirectional correlations of pairs of strings of length n arises as a challenging open question.

Despite some algorithmic progress, efficient enumeration algorithms for both Γ_n and Δ_n remain open [28]. In this work, we provide an upper bound on the logarithm of the number of period sets for strings of length n, which leads to a proof for the Guibas and Odlyzko conjecture on the convergence of the ratio $\frac{\ln \kappa_n}{(\ln n)^2}$. However, as illustrated in Fig. 1, there is room to improve the bounds on κ_n, especially the upper bound. Recall that period sets of Γ_n are partitioned according to their basic period into cases a and b. The lower bound for κ_n is in fact a lower bound for κ_n^b—see the Proof of Theorem 5.1. Furthermore, we conjecture that $\kappa_n^a < \kappa_n^b$ for $n > 6$ and that κ_n is strictly increasing from $n > 0$, as suggested by initial values of κ_n^a, κ_n^b, and κ_n—see Fig. 2.

Acknowledgements We thank the anonymous reviewers for their detailed comments and suggestions. 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.

Author Contributions All authors wrote and reviewed the manuscript.

Data Availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no Conflict of interest.

References

1. Rivals, E., Sweering, M., Wang, P.: Convergence of the Number of Period Sets in Strings. In: Etessami, K., Feige, U., Puppis, G. (eds.) 50th International Colloquium on Automata, Languages, and Programming (ICALP 2023). Leibniz International Proceedings in Informatics (LIPIcs), 261. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany (2023). https://doi.org/10.4230/LIPIcs.ICALP.2023.100 . 100:1–100:14
2. Lothaire, M. (ed.): Combinatorics on Words, 2nd end. Cambridge University Press, New York (1997)
3. Lothaire, M.: Algebraic Combinatorics on Words. Cambridge University Press, Cambridge (2005)
4. Smyth, W.F.: Computing Pattern in Strings. Pearson-Addison Wesley, Boston (2003)
5. Robin, S., Rodolphe, F., Schbath, S.: DNA, Words and Models. Cambridge University Press, New York (2005)
6. Rahmann, S., Rivals, E.: On the distribution of the number of missing words in random texts. Comb. Probab. Comput. (2003). https://doi.org/10.1017/s0963548302005473
7. Percus, O.E., Whitlock, P.A.: Theory and application of Marsaglia's monkey test for pseudorandom number generators. ACM Trans. Model. Comput. Simul. **5**(2), 87–100 (1995). https://doi.org/10.1145/210330.210331
8. Guibas, L.J., Odlyzko, A.M.: Periods in strings. J. Comb. Theory Ser. A **30**, 19–42 (1981). https://doi.org/10.1016/0097-3165(81)90038-8
9. Rivals, E., Rahmann, S.: Combinatorics of periods in strings. J. Combin. Theory Ser. A **104**(1), 95–113 (2003). https://doi.org/10.1016/s0097-3165(03)00123-7
10. Rivals, E., Rahmann, S.: Combinatorics of Periods in Strings. In: Orejas, F., Spirakis, P., van Leuween, J. (eds.) ICALP 2001, Proc. of the 28th International Colloquium on Automata, Languages and Programming, (ICALP), July 8-12, 2001. Lecture Notes in Computer Science, 2076, 615–626. Springer, Hersonissos, Creta, Greece (2001). https://doi.org/10.1007/3-540-48224-5_51
11. Holub, S., Shallit, J.O.: Periods and borders of random words. In: Ollinger, N., Vollmer, H. (eds.) STACS 2016, Proc. of the 33rd Symposium on Theoretical Aspects of Computer Science, (STACS), February 17-20, 2016. LIPIcs, 44, 1–10. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, Orléans, France (2016). https://doi.org/10.4230/LIPIcs.STACS.2016.44
12. Gabric, D., Rampersad, N., Shallit, J.: An inequality for the number of periods in a word. Int. J. Found. Comput. Sci. **32**(05), 597–614 (2021). https://doi.org/10.1142/s0129054121410094
13. Blanchet-Sadri, F., Gafni, J.D., Wilson, K.H.: Correlations of partial words. In: Thomas, W., Weil, P. (eds.) STACS 2007, Proc. of the 24th Annual Symposium on Theoretical Aspects of Computer Science, Aachen, Germany, February 22-24, 2007. Lecture Notes in Computer Science, 4393, 97–108. Springer, Aachen, Germany (2007). https://doi.org/10.1007/978-3-540-70918-3_9
14. Blanchet-Sadri, F., Duncan, S.: Partial words and the critical factorization theorem. J. Combin. Theory Ser. A **109**(2), 221–245 (2005). https://doi.org/10.1016/j.jcta.2004.09.002
15. Blanchet-Sadri, F., Fowler, J., Gafni, J.D., Wilson, K.H.: Combinatorics on partial word correlations. J. Combin. Theory Ser. A **117**(6), 607–624 (2010). https://doi.org/10.1016/j.jcta.2010.03.001
16. Gabric, D.: Mutual borders and overlaps. IEEE Trans. Inf. Theory **68**(10), 6888–6893 (2022). https://doi.org/10.1109/TIT.2022.3167935
17. Bilotta, S., Pergola, E., Pinzani, R.: A new approach to cross-bifix-free sets. IEEE Trans. Inf. Theory **58**(6), 4058–4063 (2012). https://doi.org/10.1109/TIT.2012.2189479
18. Bajic, D., Loncar-Turukalo, T.: A simple suboptimal construction of cross-bifix-free codes. Cryptogr. Commun. **6**(6), 27–37 (2014). https://doi.org/10.1007/s12095-013-0088-8
19. Bilotta, S.: Variable-length non-overlapping codes. IEEE Trans. Inf. Theory **63**(10), 6530–6537 (2017). https://doi.org/10.1109/TIT.2017.2742506
20. Blackburn, S.R.: Non-overlapping codes. IEEE Trans. Inf. Theory **61**, 4890–4894 (2015). https://doi.org/10.1109/TIT.2015.2456634
21. Levenshtein, V.I.: Maximum number of words in codes without overlaps. Probl. Inf. Transm. **6**(4), 355–357 (1970)
22. Stanovnik, L., Moškon, M., Mraz, M.: In search of maximum non-overlapping codes. Des. Codes Crypt. **92**(5), 1299–1326 (2024). https://doi.org/10.1007/s10623-023-01344-z
23. Yazdi, S.M.H.T., Kiah, H.M., Gabrys, R., Milenkovic, O.: Mutually uncorrelated primers for dna-based data storage. IEEE Trans. Inf. Theory **64**(9), 6283–6296 (2018). https://doi.org/10.1109/TIT.2018.2792488

24. Fine, N.J., Wilf, H.S.: Uniqueness theorems for periodic functions. Proc. Am. Math. Soc. **16**(1), 109–114 (1965). https://doi.org/10.1090/S0002-9939-1965-0174934-9
25. Halava, V., Harju, T., Ilie, L.: Periods and binary words. J. Combin. Theory Ser. A **89**(2), 298–303 (2000). https://doi.org/10.1006/jcta.1999.3014
26. Sloane, N.J.A.: The On-Line Encyclopedia of Integer Sequences (2022). https://oeis.org
27. Guibas, L.J., Odlyzko, A.M.: String overlaps, pattern matching, and nontransitive games. J. Combin. Theory Ser. A **30**(2), 183–208 (1981). https://doi.org/10.1016/0097-3165(81)90005-4
28. Rivals, E.: Incremental computation of the set of period sets. In: Královič, R., Kůrková, V. (eds.) SOFSEM 2025: Theory and Practice of Computer Science. SOFSEM 2025. Lecture Notes in Computer Science, vol 15539. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-82697-9_19
29. Nielsen, P.: A note on bifix-free sequences (Corresp.). IEEE Trans. Inf. Theory **19**(5), 704–706 (1973). https://doi.org/10.1109/TIT.1973.1055065
30. Rivals, E., Wang, P.: Counting overlapping pairs of strings. ArXiv arXiv:2405.09393 (2024)
31. Cakir, I., Chryssaphinou, O., Månsson, M.: On A conjecture by Eriksson concerning overlap in strings. Comb. Probab. Comput. **8**(5), 429–440 (1999). https://doi.org/10.1017/S0963548399003806