Autocorrelation of Strings

A comment on entries A005434 and A045690 of the Encyclopedia of Integer Sequences

Eric Rivals
LIRMM (Computer Science Department)
CNRS - Université de Montpellier
France
`rivals_AT_lirmm.fr`
http://www.lirmm.fr/~rivals

October 10, 2006

This page is also available in PDF format here and the corresponding publication on periods in strings is there. It also concerns entries A018819 and A000123 of the Encyclopedia of Integer Sequences [5].

Strings, also called word, or sequence of characters may "overlap" themselves. Example 0.1 shows how the word "abracadabra" can overlap itself.

Example 0.1 *The word* `ABRACADABRA` *admits* $\{0, 7, 11\}$ *as periods set and* $v :=$ 100000010001 *as autocorrelation.*

Offset:										0	1	2	3	4	5	6	7	8	9	0
A	B	R	A	C	A	D	A	B	R	A							⋮			⋮
							A	B	R	A	C	A	D	A	B	R	A			⋮
										A	B	R	A	C	A	D	A	B	R	A
Autocorrelation:										1	0	0	0	0	0	0	1	0	0	1

Example 0.2 *The word* `ababababa` *admits* $\{0, 2, 4, 6, 8\}$ *as periods set and has autocorrelation* $v :=$ 101010101*.*

Offset:	0	1	2	3	4	5	6	7	8
Word	a	b	a	b	a	b	a	b	a
Period 2			a	b	a	b	a	b	a
Period 4					a	b	a	b	a
Period 6							a	b	a
Period 8									a
Autocorrelation	1	0	1	0	1	0	1	0	1

An offset at which a word can overlap itself is called a *period*. Note that 0 is always a period and that the maximum period is $n-1$ if the word is of length n. Therefore, any word of length n has a non-empty set of periods, which is a subset of $[0, n-1]$. The set of period can be denoted as a set, but also as a binary vector of length n indexed from 0 until $n-1$ in which an entry equals 1 if an integer is a period of the word and 0 otherwise. This binary vector is called the *autocorrelation* of the word and has the same length as the word. Not all possible binary vector of length n are autocorrelation [3]. Examples 0.1 and 0.2 illustrate these definitions.

This page summarizes some of the results published in [4] about the nature of autocorrelations and the number κ_n of different autocorrelations of words of size n (where n is any positive integer). The sequence $(\kappa_n)_{n>0}$ corresponds to the sequence A005434 in the Encyclopedia of Integer Sequences (EOIS) [5].

This For instance, it exhibits the relation between the number of binary partitions of an integer n (sequence A018819 In the EOIS [5]) and the number of autocorrelation of length n. The first study of autocorrelation was published in the seminal article of Guibas and Odlyzko in 1981 [3].

1 Notation, Definitions and Elementary Properties

Let Σ be a finite alphabet of size σ. A sequence of n letters of Σ indexed from 0 to $n-1$ is called a *word* or a *string* of length n over Σ. We denote the *length* of a word $U := U_0U_1 \dots U_{n-1}$ by $|U|$. We denote by Σ^*, respectively by Σ^n, the set of all finite words, resp. of all words of length n, over Σ.

Definition 1 (Period) *Let $U \in \Sigma^n$ and let p be a non-negative integer with $p < n$. Then p is a* period *of U iff the suffix of length $n-p$ of U is equal to its prefix of length $n-p$. The basic period of U is its smallest non-null period, if it exists.*

We denote the *set of all periods* of U by $P(U)$. The *autocorrelation* v of U is a representation of $P(U)$. It is a binary vector of length n such that: $\forall\, 0 \leq i < n$, $v_i = 1$ iff $i \in P(U)$, and $v_i = 0$ otherwise.

Let $\Gamma_n := \{v \in \{0,1\}^n \,|\, \exists U \in \Sigma^n : v = P(U)\}$ be the set of all autocorrelations of strings in Σ^n. We denote its cardinality by κ_n. The autocorrelations in Γ_n can be partitioned according to their basic period; thus, for $0 \leq p < n$, we denote by $\Gamma_{n,p}$ the subset of autocorrelations whose basic period is p, and by $\kappa_{n,p}$ the cardinality of this set. The set inclusion defines a partial order on elements of Γ_n.

2 Structural Properties of Γ_n

First we have shown that Γ_n equipped with the set inclusion (denoted $\subseteq$) is a lattice. However, for $n > 6$, Γ_n does not satisfy the Jordan-Dedekind condition. The structure is illustrated 1.

Moreover, we have also noticed that the complete set of periods is redundant and have defined the *Irreducible Periods Set*, which is the smallest subset of the periods set whose enable to recompute the periods set. There is a one-to-one (bijective) correspondance between periods sets and Irreducible Periods Sets.

3 Enumeration of all Autocorrelations of Length n

Guibas and Odlyzko gave a predicate Ξ that determine in linear time if a binary vector is a true autocorrelation [3]. We build on the predicate Ξ to exhibit an enumaration algorithm for all autocorrelations. The algorithm is detailed in [4]. Here, we provide a C and a $C++$ implementation of this algorithm. The $C++$ implementation used bitstring to store the binary vectors and is thus able to enumerate the autocorrelations until $n = 450$ (this depends on the amount of main memory available on the computer).

C implementation	$C++$ implementation	Linux executable

Feel free to contact us per email `rivals_AT_lirmm.fr`if you need other versions (Mac OSX or Windows), or if you want the set of autocorrelations for a given n.

4 Bounds on the Number of Autocorrelations

We investigated how the number κ_n of different autocorrelations of length n grows with n. From the characterization of autocorrelation [3], we know that κ_n is independent of the alphabet size. In [3], it is shown that as $n \to \infty$,

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

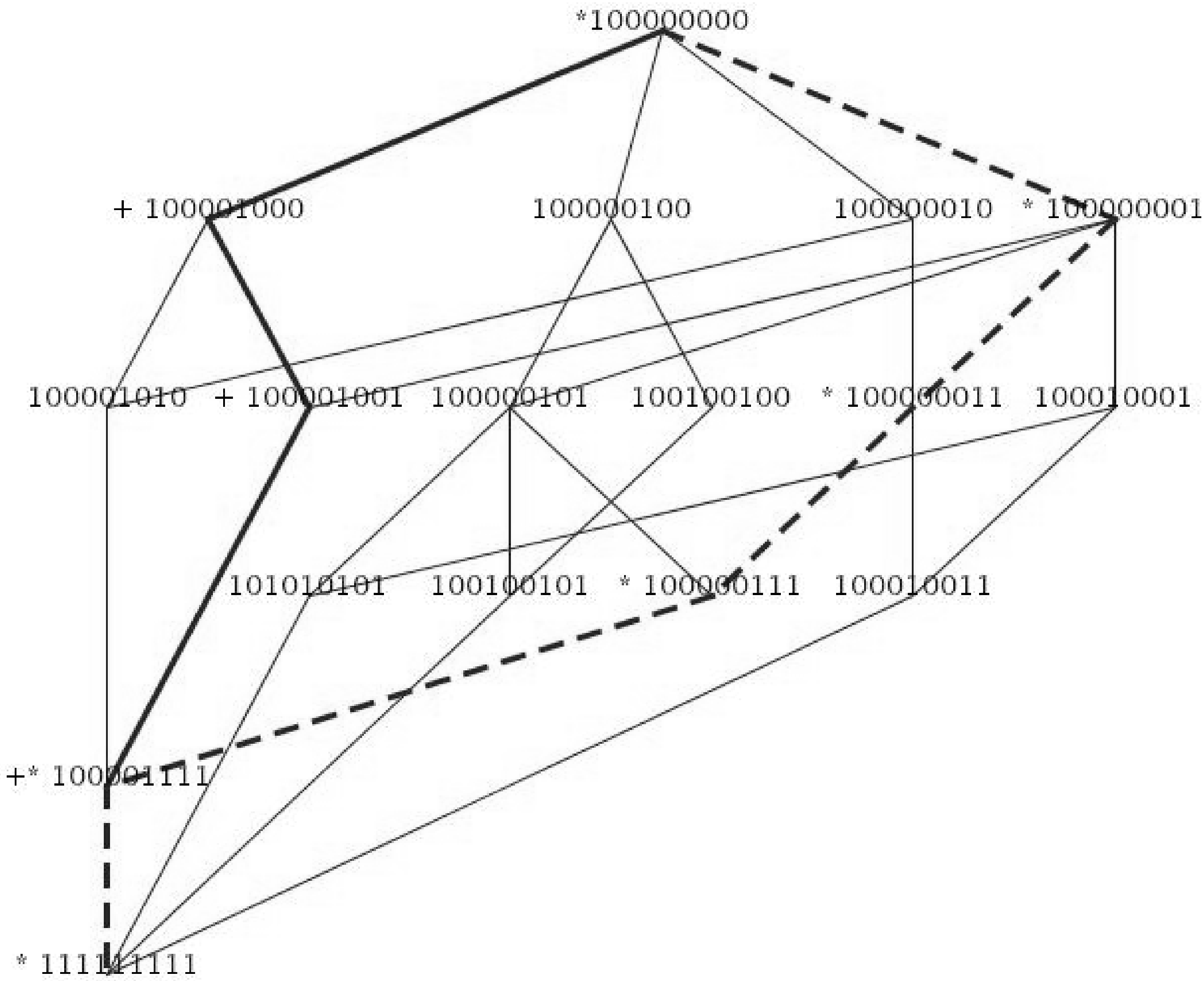

Figure 1: A representation of the lattice Γ(9). The bold-edges and dashed-edges paths shows two maximal chains of different lengths between 111111111 and 100000000. The correlations on these paths are marked with a + or a $*$, respectively.

As shown in Figure 2, these bounds are rather loose. In fact, for small n, the actual value of κ_n is below its asymptotic lower bound. While we conjecture that $\lim_{n\to\infty} \frac{\ln \kappa_n}{(\ln n)^2} = \frac{1}{2\ln 2}$, it remains an open problem to derive a tight upper bound and prove this conjecture. Our contribution is that a good lower bound for κ_n is closely related to the number of binary partitions of an integer. Both improved bounds we derive from this relationship are also shown in Figure 2.

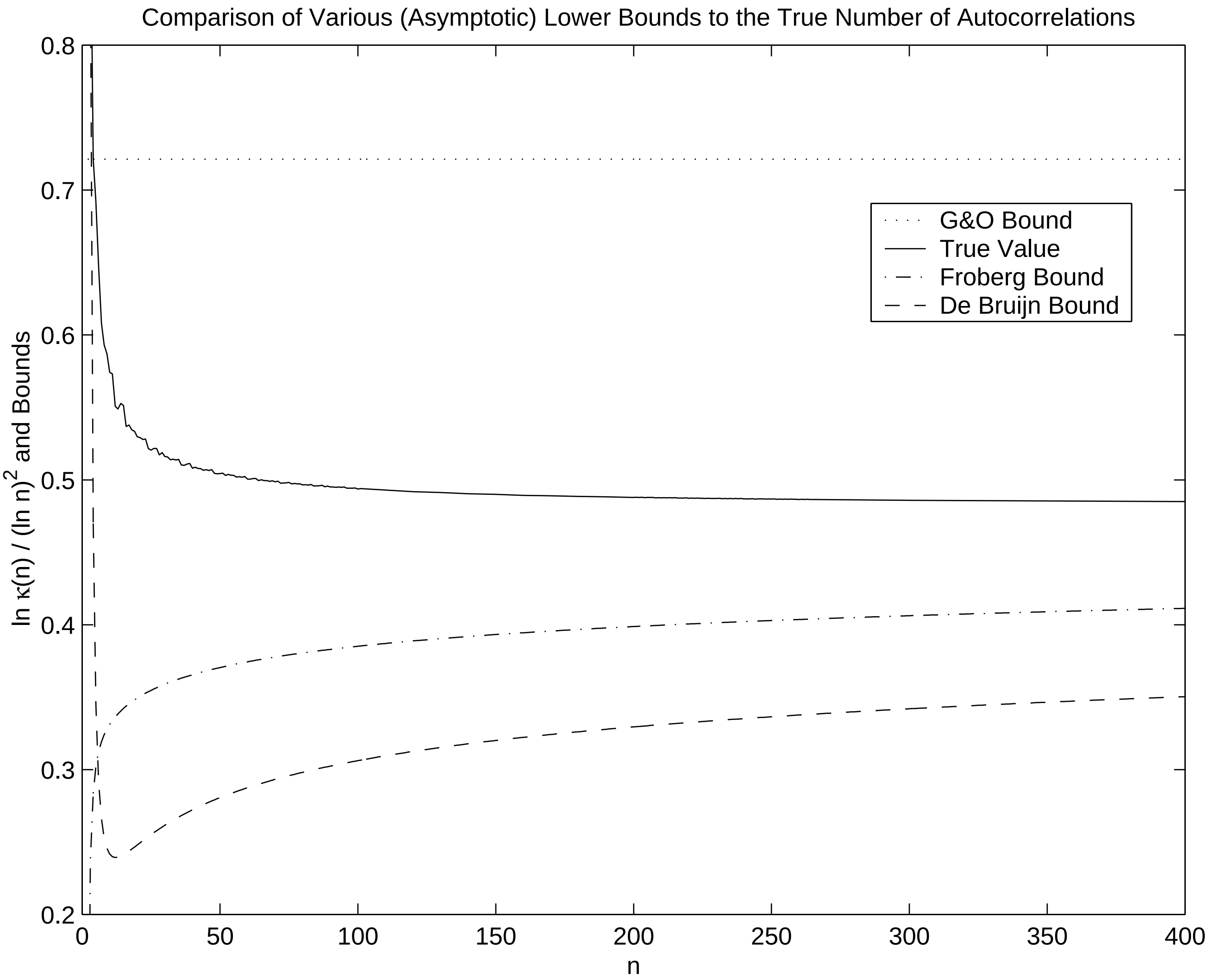

Figure 2: True values of $\ln \kappa_n/(\ln n)^2$ for $n \leq 400$, compared to Guibas & Odlyzko's (G&O) asymptotic lower bound, the improved asymptotic bound from Theorem 1 (ii) derived from DeBruijn's results, and the non-asymptotic lower bound from Theorem 1 (i) based on Fröberg's work. Both of these bounds converge to the G&O asymptotic value of $1/(2\ \ln 2)$ for $n \to \infty$. The upper bound of G&O, corresponding to the line $y = 1/(2\ \ln(3/2)) \approx 1.23$, is not visible on the figure.

We have that the number of autocorrelations of length n, κ_n, is bounded by the number of autocorrelations of length n whose basic period is larger than $n/2$. The latter equals the sum of the κ_i for i going from 0 to $\lceil n/2 \rceil - 1$. Thus, we define $L_0 := 1$, $L_1 := 1$, and, for $n \geq 2$, $L_n := \sum_{i=0}^{\lceil n/2 \rceil - 1} L_i$. By induction, $L_n \leq \kappa_n$ for all $n \geq 0$.

Now we consider a related sequence: the *number of binary partitions* B_n of an integer $n \geq 0$, i.e., the number of ways to write n as a sum of powers of 2 where the order of summands does not matter. For example, 6 can be written as such a sum in 6 different ways: 4+2, 4+1+1, 2+2+2, 2+2+1+1, 2+1+1+1+1, 1+1+1+1+1+1. Therefore $B_6 = 6$. By convention, $B_0 = 1$; furthermore $B_1 = 1$.

B_n corresponds to the sequence A018819, and B_{2n} to the sequence A000123 in the Encyclopedia of Integer Sequences [5].

We have shown that for $n \geq 1$, $L_n = 1/2 \cdot B_{n+1}$. Then building on the results of Fröberg [2] and De Bruijn [1], which both gave approximations on the number of binary partitions, we provide two lower bounds (which we refer to as Fröberg's and De Bruijn's bounds in Figure 2) for κ_n, the number of autocorrelations of length n.

Theorem 1 (Lower Bounds on κ_n) *Define $F(n)$ as follows*

$$F(n) := \sum_{k=0}^{\infty} \frac{n^k}{2^{\frac{k(k+1)}{2}} \cdot k!}. \tag{2}$$

Then:

1. *For all $n \geq 1$, $\kappa_n \geq 0.31861 \cdot F(n+1)$.*
2. *Asymptotically (with approximated constants),*

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

5 Computing the Size of Populations

The correlation of a string depends on its self-overlapping structure, but is not directly related to its characters. Hence, different strings share the same correlation. For instance over the alphabet $\{a, b\}$, take *abbabba* and *babbabb*. The *population* of a correlation v is the set of strings over Σ whose correlation is v. We wish to compute the *size of the population* of a given correlation, and by extension of all correlations. This corresponds to entry A045690 in the Encyclopedia of Integer Sequences [5].

In [3], Guibas and Odlyzko exhibit a recurrence linking the population sizes of a correlation and of its nested correlation. Here, we exhibit another recurrence which links the population size of an autocorrelation v to the population sizes of the autocorrelations it is included in. The recurrence depends on the *number of free characters* (nfc for short) of v, to be defined next.

Definition 2 (Number of Free Characters) *The nfc of a correlation v is the maximum number of positions in a string U with $P(U) = v$ that are not determined by the periods.*

To illustrate this definition, note that a correlation represents a set of equalities between the characters of a string. For example, take $v := 100001001 \in \Gamma_9$. A string $U = u_0 \dots u_8$ with $P(U) = v$ must satisfy the following set of equations: $\{u_0 = u_3 = u_5 = u_8, u_1 = u_6, u_2 = u_7\}$. Thus we can write any word U as $u_0 u_1 u_2 u_0 u_4 u_0 u_1 u_2 u_0$ for some $u_0, u_1, u_2, u_4 \in \Sigma$. So the nfc of v is 4.

The nfc is independent of Σ and can be computed from v alone. Given a correlation v and its length n, the algorithm NFC, computes the nfc of v. NFC follows the recursive structure of Predicate Ξ and requires $\Theta(n)$ time. Its pseudo-code is available here.

We now state our recurrence on the population sizes.

Theorem 2 *Let $n \in \mathbb{N}$ and let v_k be the k-th $(k = 1, \dots, \kappa_n)$ autocorrelation of Γ_n. Let ρ_k denote the number of free characters of v_k, and N_k be its population size. We have:*

$$N_k = \sigma^{\rho_k} - \sum_{j: v_k \subset v_j} N_j.$$

References

[1] N. G. DeBruijn. On Mahler's partition problem. *Proc. Akad. Wet. Amsterdam*, 51:659–669, 1948.

[2] Carl-Erik Fröberg. Accurate estimation of the number of binary partitions. *BIT*, 17:386–391, 1977.

[3] Leo J. Guibas and Andrew M. Odlyzko. Periods in strings. *J. of Combinatorial Theory series A*, 30:19–42, 1981.

[4] Eric Rivals and Sven Rahmann. Combinatorics of Periods in Strings. *J. of Combinatorial Theory series A*, 104(1):95–113, October 2003.

[5] N. J. A. Sloane. The On-Line Encyclopedia of Integer Sequences, 2004.