

Maximal LZ78 Complexity Among Sturmian Sequences

Noble Number Extremality and Epoch Packing

Jetxel Alfredo

Independent Researcher

Scalar Resonance Research Program, March 2026

Second edition, September 2026

Abstract. We study the LZ78 phrase count $c(\alpha, L)$ of a Sturmian sequence, the coding of an irrational rotation, as a function of the rotation number α . A binary sequence with at most $n + 1$ factors of each length n parses into at most $c^*(L)$ phrases, where $c^*(L) = (9/8)^{1/3}L^{2/3} + O(L^{1/3})$ is the size of a greedily filled trie. This covers every Sturmian sequence. Any sequence at all parses into at least $\sqrt{2L} - 1$ phrases, so the growth exponent of $c(\alpha, L)$ lies between $1/2$ and $2/3$. We conjecture that it equals $2/3$ whenever the partial quotients of α are bounded, and we show that a bound of order $L^{1/3}$ on the depth of the LZ78 trie would prove it. Computation up to $L = 8 \times 10^6$ supports the conjecture. The Fibonacci word fills between 97.0 and 98.6 percent of $c^*(L)$ at every length tested. Noble slopes, whose continued fractions end in an unbroken run of ones, fill the same narrow band and overtake the Fibonacci word at about a third of all lengths. Periodic slopes $[0; M, M, \dots]$ with $M \geq 2$ sit lower and swing more widely. We read this spread through a decomposition of the parse into epochs at the convergent denominators of α .

Index Terms. Sturmian sequences, LZ78 complexity, golden ratio, continued fractions, Diophantine approximation, noble numbers, compression resistance, Fibonacci word

Second edition. The first edition remains available, and the corrections are listed at the end.

1. Introduction

The Fibonacci word, the fixed point of the substitution $A \rightarrow AB, B \rightarrow A$, is the simplest Sturmian sequence. It codes the rotation of the circle by the golden ratio $1/\varphi = (\sqrt{5} - 1)/2$. Its convergent denominators grow as slowly as any irrational allows, and its critical exponent, $2 + \varphi \approx 3.618$ (Mignosi and Pirillo [1]), is the smallest of any Sturmian word.

This paper asks how many phrases the incremental parsing of Ziv and Lempel [2], LZ78, produces on such sequences. The phrase count $c(\alpha, L)$ of a Sturmian sequence of slope α and length L is the number of distinct phrases in that parse, and a higher count means the sequence resists dictionary compression better.

Two results are proved. Every Sturmian sequence, and more generally every binary sequence with at most $n + 1$ factors of length n , parses into at most $c^*(L) = (9/8)^{1/3}L^{2/3} + O(L^{1/3})$ phrases. Every sequence parses into at least $\sqrt{2L} - 1$. Between these, computation shows the Fibonacci word within three percent of the upper bound at every length we tried, with the noble slopes

beside it. The matching lower bound of order $L^{2/3}$ is open; Section 4 reduces it to a statement about the depth of the parse tree.

The organizing tool is the epoch decomposition: the parse is cut at the convergent denominators of α , each of which marks a natural quasi-periodic scale. Sections 2 and 3 give background and the decomposition, Section 4 the bounds and the conjecture, and Sections 5 to 9 the computations.

2. Preliminaries

2.1 Sturmian Sequences

A Sturmian sequence with slope $\alpha \in (0, 1) \setminus \mathbb{Q}$ is the binary sequence $s(\alpha, n) = \lfloor (n+1)\alpha + \alpha \rfloor - \lfloor n\alpha + \alpha \rfloor$, the coding of the rotation of the unit circle by α (Morse and Hedlund [3]). Sturmian sequences have subword complexity $p(n) = n + 1$, the least possible for a sequence that is not eventually periodic. The Fibonacci word is the Sturmian sequence with slope $1/\varphi$, the fixed point of $A \rightarrow AB, B \rightarrow A$.

2.2 Continued Fractions and Convergent Denominators

Every irrational $\alpha \in (0, 1)$ has a unique continued fraction expansion $\alpha = [0; a_1, a_2, a_3, \dots]$ with partial quotients $a_k \geq 1$. The convergents p_k/q_k satisfy $q_{k+1} = a_{k+1}q_k + q_{k-1}$. For $\alpha = 1/\varphi = [0; 1, 1, 1, \dots]$ the denominators are the Fibonacci numbers, growing as $\varphi^k/\sqrt{5}$. Since every partial quotient is at least 1, no irrational has slower-growing denominators, and this is the same property that makes $1/\varphi$ the worst-approximable irrational (Hurwitz). A noble number is an irrational whose continued fraction ends in $1, 1, 1, \dots$. For the periodic expansion $[0; M, M, M, \dots]$ the denominators grow by the factor $\lambda(M) = (M + \sqrt{M^2 + 4})/2$, which increases with M and equals φ at $M = 1$.

2.3 LZ78 Parsing

The LZ78 algorithm [2] parses a sequence incrementally. Starting from an empty dictionary, it reads the longest prefix of the remaining input that is already a phrase and creates a new phrase from that match plus one more symbol. The phrases form a tree, the LZ78 trie, in which each phrase hangs from the phrase it extends. The total number of phrases is $c(\alpha, L)$. LZ77-style factorization measures something else, as Section 10 shows.

3. The Epoch Decomposition Framework

Definition 1 (Convergent epoch). The k -th convergent epoch of the LZ78 parse of $s(\alpha, \cdot)$ is the set of phrases whose starting position falls in $[q_k, q_{k+1})$. Its phrase count is Δc_k . The epoch count at length L is $K(\alpha, L) = \max\{k : q_k \leq L\}$, and the decomposition splits the total as $c(\alpha, L) = \sum_k \Delta c_k + R(\alpha, L)$, where R counts the final, incomplete epoch.

Definition 2 (Per-symbol phrase rate). The per-symbol phrase rate in epoch k is

$$\rho_k = \frac{\Delta c_k}{q_{k+1} - q_k}.$$

4. Bounds on the Phrase Count

4.1 The Upper Bound

Theorem 0. Let w be a binary sequence with at most $n + 1$ distinct factors of each length n ; every Sturmian sequence is one. Then the LZ78 parse of the first L symbols of w has at most $c^*(L)$ phrases, where $c^*(L)$ is the largest number of nodes in a trie that has at most $\ell + 1$ nodes at depth ℓ and total depth at most L . This ceiling satisfies

$$c^*(L) = (9/8)^{1/3} L^{2/3} + O(L^{1/3}), \quad (9/8)^{1/3} = \frac{3^{2/3}}{2} \approx 1.0400.$$

Proof. The phrases are distinct, and a phrase of length ℓ is a factor of w of length ℓ , so at most $\ell + 1$ phrases have length ℓ . Their lengths add up to at most L (one more symbol if the parse ends on an incomplete phrase). To fit as many phrases as possible into length L , take the shortest ones first: all $\ell + 1$ of length ℓ for $\ell = 1, \dots, D$. That gives $D(D + 3)/2$ phrases of total length $D(D + 1)(D + 2)/3$. Then $L \approx D^3/3$ and $c \approx D^2/2$, so $c^3/L^2 \rightarrow 9/8$, and the error terms are of the order of one more level of the trie, $O(L^{1/3})$. $\square$

The bound uses only the factor count, so it holds for any sequence of complexity at most $n + 1$. The first edition carried the constant $(9/4)^{1/3} \approx 1.31$ here; the elimination of D gives $(9/8)^{1/3}$. The corrected bound is close to sharp. Across 800 lengths between 10^4 and 8×10^6 , every sequence in Table 2 stays below $c^*(L)$, and the Fibonacci word fills between 97.0 and 98.6 percent of it.

4.2 The Lower Bound

Proposition 1. The LZ78 parse of any sequence of length L has at least $\sqrt{2L} - 1$ phrases.

Proof. The j -th phrase extends an earlier phrase by one symbol, and by induction every earlier phrase has length less than j , so the j -th phrase has length at most j . With c phrases, $L \leq c(c + 1)/2$, which gives $c \geq \sqrt{2L} - 1$. $\square$

Together, Theorem 0 and Proposition 1 place the growth exponent of $c(\alpha, L)$ between $1/2$ and $2/3$ for every Sturmian sequence, with no condition on α .

4.3 The Matching Lower Bound

Computation puts the exponent at $2/3$ (Table 1), and we conjecture that it is exactly that.

Conjecture 1 (depth). If the partial quotients of α are bounded, the deepest phrase in the LZ78 parse of the first L symbols of $s(\alpha, \cdot)$ has length $O(L^{1/3})$.

Conjecture 2 (exponent). If the partial quotients of α are bounded, then $c(\alpha, L) \geq c_0 L^{2/3}$ for some $c_0 > 0$.

Conjecture 1 implies Conjecture 2. Write the standard word of length q_k as $M_k = M_{k-1}^{a_k} M_{k-2}$. The tail M_{k-2} has length q_{k-2} , and while it is being parsed no phrase is longer than one plus the current trie depth, which under Conjecture 1 is $O(q_k^{1/3})$. The tail therefore contributes $\Delta_k \gtrsim q_{k-2}/q_k^{1/3}$ new phrases, and bounded partial quotients keep q_{k-2}/q_k above a fixed constant, so $\Delta_k \gtrsim q_k^{2/3}$. Summing over k gives $c(q_K) \gtrsim q_K^{2/3}$, and bounded partial quotients carry this to

every L between consecutive denominators.

The first edition presented this argument as a proof, obtaining the depth bound from the upper bound on c . The step does not go through. The upper bound limits how many phrases there are, and the trie capacity limits how many there can be at each depth, but neither limits how deep the trie reaches: a single long chain is consistent with both. The tree structure gives only $D \leq c$. The computations support Conjecture 1 all the same. For the five periodic slopes $M = 1, \dots, 5$, the maximum depth $D(L)$ stays between $1.8 L^{1/3}$ and $2.5 L^{1/3}$ from $L = 10^5$ to 8×10^6 , though for the Fibonacci word the ratio creeps upward (1.83 to 2.18), and the range is too short to rule out slow growth.

Table 1. Per-epoch fits $\Delta c_k = C q_k^\beta$ for $[0; M, M, \dots]$, first edition data at $L = 5 \times 10^6$.

M	β (fitted)	C (fitted)	R^2	Epochs
1	0.6644	0.4005	0.9978	27
2	0.6624	0.8465	0.9996	12
3	0.6611	1.2855	0.9991	7
5	0.6612	2.1229	0.9965	4

The fitted exponent, $\beta = 0.662 \pm 0.002$ for every M , is consistent with $2/3$; the small shortfall is what finite lengths produce.

5. Per-Epoch Analysis

5.1 The Prefactor

The prefactor C in Table 1 grows with M . The growth has a simple source: if the total count follows one law $c \approx K L^{2/3}$, an epoch running from q_k to $q_{k+1} \approx \lambda(M) q_k$ holds $K(\lambda(M)^{2/3} - 1)q_k^{2/3}$ phrases. Dividing each fitted C by $\lambda(M)^{2/3} - 1$ gives 1.059, 1.059, 1.056 and 1.062 for $M = 1, 2, 3, 5$. The prefactors differ because the epochs have different lengths, and once that is removed the four values agree to within one percent.

5.2 The Per-Symbol Rate

It follows that the per-symbol phrase rate at a given scale is, to within a few percent, the same for every M . Direct counts confirm this. For $M = 1, \dots, 5$ and epochs beyond $q_k = 10^3$, the ratio $\Delta c_k / (q_{k+1}^{2/3} - q_k^{2/3})$ lies between 0.90 and 1.08, centered near 1. The spread is smallest for $M = 1$ (1.01 to 1.05) and grows with M .

The first edition stated the opposite, as a lemma: that the normalized epoch rate $f(M) = \lim \Delta c_k / (M q_k^{2/3})$ is largest at $M = 1$. In that normalization Table 1 gives $f(1) \approx 0.40$ and $f(2), f(3), f(5) \approx 0.42$ to 0.43 , and direct counts to 8×10^6 place all five values in one band, 0.36 to 0.44. The lemma does not hold, and the results that relied on it are restated below as findings.

6. Noble Slopes

Noble slopes stay level with the Fibonacci word. Table 2 compares three of them, $[0; 3, 1, 1, \dots]$, $[0; 2, 2, 1, 1, \dots]$ and $[0; 4, 1, 1, \dots]$, over the same 800 lengths. Each fills 97 to 99 percent of $c^*(L)$, and each is ahead of the Fibonacci word at about a third of the lengths, by at most 1.2 percent, with the lead changing hands repeatedly. At $L = 4 \times 10^6$, for instance, $[0; 3, 1, 1, \dots]$ has 25 802 phrases to the Fibonacci word's 25 586.

So the ratio $c(\alpha, L)/c(1/\varphi, L)$ stays within about one percent of 1 for noble α , and we conjecture that it tends to 1. The first edition went further. It claimed that the difference $c(\alpha, L) - c(\beta, L)$ is bounded for two nobles with the same tail, but the measured differences grow: 104 phrases at 10^6 and 216 at 4×10^6 for $[0; 3, 1, 1, \dots]$, with a change of sign in between. Its argument assumed that the two parses agree epoch by epoch up to a constant, which was not shown, and even then a constant per epoch summed over the $O(\log L)$ epochs is not a constant.

Table 2. Share of the ceiling filled, $c(\alpha, L)/c^*(L)$, over 800 lengths from 10^4 to 8×10^6 .

Slope α	c/c^* , lowest to highest	Ahead of the Fibonacci word
$[0; 1, 1, 1, \dots]$, the Fibonacci word	97.0 % to 98.6 %	—
$[0; 3, 1, 1, \dots]$, noble	97.0 % to 98.9 %	at 37 % of lengths
$[0; 2, 2, 1, 1, \dots]$, noble	97.2 % to 98.3 %	at 33 % of lengths
$[0; 4, 1, 1, \dots]$, noble	97.2 % to 98.6 %	at 40 % of lengths
$[0; 2, 2, 2, \dots]$	94.7 % to 97.4 %	never
$[0; 3, 3, 3, \dots]$	90.7 % to 98.0 %	never (level at 6 lengths)
$[0; 4, 4, 4, \dots]$	88.8 % to 98.1 %	never
$[0; 5, 5, 5, \dots]$	89.0 % to 97.8 %	at 2 % of lengths

7. Periodic Slopes and Epoch Packing

Among the periodic slopes $[0; M, M, \dots]$, the Fibonacci word leads. It is never behind $M = 2, 3$ or 4 at any of the 800 lengths, and draws level with $M = 3$ only at six lengths near 3×10^4 . It is behind $M = 5$ at 2 percent of the lengths, never by as much as 0.1 percent. The order among $M \geq 2$ is not monotone: $M = 5$ is ahead of $M = 4$ at 69 percent of the lengths and of $M = 3$ at 46 percent, and at 8×10^6 it still leads both.

Epoch packing suggests why. The Fibonacci word's denominators grow by the smallest factor, φ , so it passes through the most epochs in any given length, and its count never falls far behind the law it follows. A slope with larger M has fewer and longer epochs, and between their boundaries its count drifts further below the ceiling, which is the widening band in Table 2. This is a heuristic that fits the data. It is not a proof.

The first edition stated two theorems here: that for $M_1 < M_2$ the count for M_1 eventually exceeds the count for M_2 , and that the Fibonacci word eventually exceeds every other Sturmian sequence. Both relied on the lemma of Section 5.2. The first is contradicted by the pairs (3, 5) and (4, 5) at the largest lengths tested, and the second by the noble slopes.

8. The Top of the Range

The computations point to a class of slopes at the top. The Fibonacci word and the noble slopes all fill 97 to 99 percent of $c^*(L)$ at every tested length, and none of them stays ahead of the others. Every other slope we tried falls further below the ceiling at some lengths. The first edition's Corollary 4, that the eventually co-optimal Sturmian sequences are exactly the nobles, describes this class. It remains a conjecture.

9. Experimental Validation

The first edition ran six experiments: thirteen rotation numbers, lengths from 5×10^3 to 5×10^6 (some to 2×10^7), and per-epoch decompositions. In the cross-sectional ranking at $L = 10^4$ and 5×10^4 , the slope $1/\varphi$ came first, the nobles clustered at the top, and bounded, unbounded and rational slopes followed in that order. A fit of the total count for $1/\varphi$ gave the exponent 0.6599.

For this edition the counts were recomputed from exact characteristic words, built from the continued fraction by the standard-word recursion, so that no floating-point rounding enters the symbols. They give the phrase counts for eight slopes at 800 lengths up to 8×10^6 (Table 2), the per-epoch counts for $M = 1, \dots, 5$ (Section 5), and the maximum trie depths (Section 4.3). The upper bound of Theorem 0 holds at every one of these lengths.

10. Discussion

10.1 The Mechanism

The first edition's reading of epoch packing, that the Fibonacci word gains by visiting the most scales, holds in a weaker form. All the periodic slopes gain phrases at nearly the same rate per symbol; the Fibonacci word differs in being steady, because its many short epochs keep its count close to the ceiling at every length.

10.2 The Exponent

The exponent of LZ78 on Sturmian sequences lies in $[1/2, 2/3]$ unconditionally, and computation places it at $2/3$. LZ77-style factorization behaves very differently. For the Fibonacci word it needs only about $\log_\varphi L$ factors: 15, 19, 24, 29 and 30 at $L = 10^3, 10^4, 10^5, 10^6$ and 2×10^6 , against $\log_\varphi L \approx 14.4, 19.1, 23.9, 28.7$ and 30.2 (compare Constantinescu and Ilie [4] on fixed points of morphisms). An LZ77 factor can copy a long earlier stretch in one step, and a word this regular offers such stretches at every scale. An LZ78 phrase adds only one symbol to an earlier phrase, so however regular the word, the count cannot fall below $\sqrt{2L}$.

10.3 Open Questions

1. Prove Conjecture 1, the $O(L^{1/3})$ depth bound, which would settle the exponent.
2. Decide whether the count difference between two noble slopes is bounded, logarithmic, or larger.
3. Find the limit behavior of $c(1/\varphi, L)/c^*(L)$. It stays between 0.970 and 0.986 up to 8×10^6 ,

and it is not clear whether it converges or keeps oscillating.

4. Prove that the Fibonacci word leads the periodic slopes $[0; M, M, \dots]$, $M \geq 2$, at all large lengths, or find a length where one of them passes it for good.

11. Conclusion

The LZ78 phrase count of any Sturmian sequence is at most $(9/8)^{1/3} L^{2/3} + O(L^{1/3})$ and at least $\sqrt{2L} - 1$. The Fibonacci word and the noble slopes run within three percent of the upper bound at every length tested, in a band narrower than any other slope we tried. The matching lower bound is open and reduces to a bound on the depth of the parse tree. In these computations the slopes that stay at the top are the nobles, the numbers equivalent to the golden ratio and the hardest to approximate by rationals.

Changes from the first edition

1. Theorem 0, upper bound. The constant is $(9/8)^{1/3} \approx 1.040$. The first edition gave $(9/4)^{1/3} \approx 1.31$, a slip in eliminating D .
2. Theorem 0, lower bound. The first edition's proof of $c \geq c_0 L^{2/3}$ inferred a depth bound from the upper bound, which does not follow. The statement is now Conjecture 2, reduced to the depth bound of Conjecture 1, and Proposition 1 gives the unconditional bound $\sqrt{2L} - 1$. The abstract's claim that all results are unconditional is withdrawn.
3. Theorem 1, noble universality. The bounded-difference claim is withdrawn: measured differences grow. The ratio statement is kept as a numerical finding.
4. Lemma 1', scale-rate monotonicity. Withdrawn. In its own normalization the data give $f(1)$ at the low end of the range, and the per-symbol rate is nearly the same for every M . The first edition's Section 5.2 said so and its Section 5.3 contradicted it; Section 5.2 was right.
5. Theorem 2, periodic ordering. Withdrawn. The pairs $(3, 5)$ and $(4, 5)$ violate it at the largest lengths tested.
6. Theorem 3 and Corollary 3, eventual maximality. Withdrawn. Noble slopes overtake the Fibonacci word at about a third of all lengths up to 8×10^6 . Corollary 4, co-optimality of the nobles, is kept as a conjecture.
7. Section 10.2. The first edition compared the $L^{2/3}$ law with a $\sqrt{L}$ law for LZ77. For the Fibonacci word, LZ77 grows only logarithmically.
8. Section 9. The first edition referred to an Appendix A that was not included. The claims that rested on it (rankings under Deflate and Gzip, and statistical learnability) are not repeated here.

Acknowledgments

This work was conducted under the Scalar Resonance Research Program using a two-agent research architecture (RESEARCH-1 for mathematical theory, RESEARCH-2 for experimental implementation). Mathematical analysis, experimental design, and the manuscript were de-

veloped with Claude (Anthropic). The corrections in this edition were found and checked by computation with Claude (Anthropic).

References

- [1] F. Mignosi and G. Pirillo, “Repetitions in the Fibonacci infinite word,” *RAIRO Theoretical Informatics and Applications*, vol. 26, no. 3, pp. 199–204, 1992.
- [2] J. Ziv and A. Lempel, “Compression of individual sequences via variable-rate coding,” *IEEE Trans. Information Theory*, vol. 24, no. 5, pp. 530–536, 1978.
- [3] M. Morse and G. A. Hedlund, “Symbolic dynamics II: Sturmian trajectories,” *American Journal of Mathematics*, vol. 62, pp. 1–42, 1940.
- [4] L. Constantinescu and L. Ilie, “The Lempel–Ziv complexity of fixed points of morphisms,” *SIAM J. Discrete Mathematics*, vol. 21, no. 2, pp. 466–481, 2007.
- [5] V. T. Sós, “On the theory of diophantine approximations. I,” *Acta Math. Acad. Sci. Hungaricae*, vol. 9, pp. 229–241, 1958.
- [6] J. Alfredo, “Channel capacity under structured self-interference: the capacity factorization theorem,” *Scalar Resonance Research Program*, 2026.
- [7] J. Alfredo, “The golden ratio as dual optimum: Diophantine irrationality and algebraic spectral closure as independent axes,” *Scalar Resonance Research Program*, 2026.
- [8] A. Carpi and A. de Luca, “Special factors, periodicity, and an application to Sturmian words,” *Acta Informatica*, vol. 36, pp. 983–1006, 2000.
- [9] S. Akiyama, “Spiral Delone sets and three distance theorem,” *Nonlinearity*, vol. 32, pp. 1973–1994, 2019.
- [10] D. Badziahin and J. Shallit, “Sturmian words and the three-gap theorem,” *J. Théorie des Nombres de Bordeaux*, 2023.
- [11] M. A. Kramer, “Golden rhythms as a theoretical framework for cross-frequency organization,” *Neuron and Behavior Data Analysis Theory*, PMC10181851, 2022.
- [12] B. Pletzer et al., “When frequencies never synchronize: the golden mean and resting EEG,” *Brain Research*, vol. 1335, pp. 91–102, 2010.
- [13] J. Berstel and P. Séébold, “Sturmian words,” in *Algebraic Combinatorics on Words*, Cambridge Univ. Press, 2002, ch. 2.
- [14] J. Justin and G. Pirillo, “On a characteristic property of Sturmian words,” *Theoretical Computer Science*, vol. 154, pp. 161–170, 1996.
- [15] J. Cassaigne, “Sequences with grouped factors,” in *Developments in Language Theory III*, 1998, pp. 211–222.