METRIC AND EXTREMAL THEORY OF ONLINE PREVIOUS-COPY COMPRESSION

LUCA BLANCHI

ABSTRACT. Let $b \geq 2$ be an integer alphabet size. We study the extremal and metric behaviour of the online previous-copy parsing complexity $\mathrm{oc}(W)$ of a finite word $W \in \{0, \ldots, b-1\}^N$. In this model a phrase is either a literal symbol or an exact copy whose source starts earlier in the word; self-overlap is allowed.

We prove the sharp universal upper bound

$$\mathrm{oc}(W) \leq (1 + o(1)) \frac{N}{\log_b N}$$

uniformly for all words W of length N, and show that this is best possible:

$$\max_{|W|=N} \mathrm{oc}(W) = (1 + o(1)) \frac{N}{\log_b N}.$$

For Lebesgue-almost every $\alpha \in [0, 1]$, if $w_N(\alpha)$ denotes the prefix of length N of the base-b expansion of α, then

$$\mathrm{oc}(w_N(\alpha)) = (1 + o(1)) \frac{N}{\log_b N}.$$

We also determine the finite-word entropy of low online previous-copy complexity. For $0 \leq \kappa \leq 1$, let

$$\mathcal{C}_N(\kappa) = \left\{ W \in \{0, \ldots, b-1\}^N : \mathrm{oc}(W) \leq \kappa \frac{N}{\log_b N} \right\}.$$

Then, for $0 < \kappa \leq 1$,

$$\lim_{N \to \infty} \frac{1}{N} \log_b |\mathcal{C}_N(\kappa)| = \kappa.$$

Consequently, the Hausdorff spectrum is exact:

$$\dim_{\mathrm{H}} \left\{ \alpha \in [0, 1] : \liminf_{N \to \infty} \frac{\mathrm{oc}(w_N(\alpha)) \log_b N}{N} \leq \kappa \right\} = \kappa \qquad (0 \leq \kappa \leq 1).$$

Finally, we record parallel extremal and almost-sure estimates for the normalized substring complexity

$$\delta(W) = \max_{1 \leq k \leq |W|} \frac{p_W(k)}{k}$$

and for the minimum string attractor size $\gamma(W)$:

$$\delta(w_N(\alpha)) = \gamma(w_N(\alpha)) = (1 + o(1)) \frac{N}{\log_b N}$$

for almost every α, and the same asymptotic scale holds in the worst case.

Date: June 2026.

1. Introduction

The online previous-copy parsing model is a structural abstraction of left-to-right dictionary compression. A finite word is parsed from left to right into phrases. A phrase is either a literal symbol or a copy of an earlier substring. The source is required to start earlier than the target, but it may overlap the target. This convention includes the self-referential behaviour familiar from variants of LZ77.

A companion paper studied this model on digit prefixes of algebraic irrational numbers and proved the lower bound

$$\mathrm{oc}(w_N(\alpha)) = \omega(\log N)$$

for every algebraic irrational α. The present note is independent of the Diophantine part. Its purpose is to determine the natural extremal and metric scale of $\mathrm{oc}(W)$ for arbitrary and typical finite words.

The main scale is

$$\frac{N}{\log_b N}.$$

This scale appears in three complementary ways.

First, every word of length N has an online previous-copy parsing with at most

$$(1 + o(1))\frac{N}{\log_b N}$$

phrases. Second, by a counting argument, this is sharp in the worst case. Third, the same asymptotic holds for almost every base-b expansion.

The proof of the universal upper bound is elementary. Choose a block length

$$L = \lfloor \log_b N - 2 \log_b \log_b N \rfloor.$$

Parsing greedily, any phrase shorter than L, except possibly near the end of the word, must start at a previously unseen block of length L. There are at most b^L such blocks, while all other phrases have length at least L. This gives

$$\mathrm{oc}(W) \le \frac{N}{L} + b^L + L + O(1).$$

The lower bounds come from counting parse descriptions. The number of words of length N admitting an online previous-copy parse with z phrases is at most

$$z \left(\frac{eN}{z}\right)^z (2bN)^z.$$

For

$$z \sim \kappa \frac{N}{\log_b N},$$

this is $b^{(\kappa + o(1))N}$. This gives both the worst-case lower bound and the entropy/Hausdorff spectrum.

The last part treats two standard repetitiveness measures: the normalized substring complexity

$$\delta(W) = \max_k p_W(k)/k$$

and the minimum string attractor size $\gamma(W)$. The inequalities

$$\delta(W) \le \gamma(W) \le \mathrm{oc}(W)$$

connect these measures to the online previous-copy model. A standard collision estimate for random words gives the matching lower bound for δ, and hence for γ, almost surely.

Throughout the paper, logarithms without subscript are natural, while $\log_b$ denotes logarithm in base b.

2. Online previous-copy parsings

Let

$$W = W[1]W[2]\cdots W[N]$$

be a finite word over the alphabet

$$\{0, \dots, b-1\}.$$

Definition 2.1. *An online previous-copy parsing of W is a factorization*

$$W = F_1 F_2 \cdots F_z$$

with boundaries

$$0 = n_0 < n_1 < \cdots < n_z = N,$$

where

$$F_j = W[n_{j-1}+1, n_j].$$

Each phrase is of one of the following two types.

First, F_j may be a literal, in which case

$$|F_j| = 1.$$

Second, F_j may be a copy. Writing

$$s = n_{j-1}, \qquad t = n_j, \qquad \ell = t - s,$$

this means that there exists a source position p with

$$1 \le p \le s$$

such that

$$W[p+h] = W[s+1+h] \qquad (0 \le h < \ell).$$

The source may overlap the target.

Let

$$\mathrm{oc}(W)$$

be the minimum number of phrases in an online previous-copy parsing of W.

Remark 2.2. *The first phrase must be a literal. The definition is structural: source positions and lengths are not charged as bits. The quantity* $\mathrm{oc}(W)$ *is therefore a phrase-complexity measure, not a bit-complexity measure.*

3. A universal upper bound

We first prove that every word has an online previous-copy parsing with about $N/\log_b N$ phrases.

Theorem 3.1 (Universal upper bound)**.** *Uniformly for all words* $W \in \{0,\dots,b-1\}^N$*,*

$$\mathrm{oc}(W) \le (1+o(1))\frac{N}{\log_b N}.$$

More precisely, for every integer $L \ge 1$*,*

$$\mathrm{oc}(W) \le \frac{N}{L} + b^L + L + O(1),$$

and with

$$L = \lfloor \log_b N - 2\log_b \log_b N \rfloor$$

this gives the displayed asymptotic.

Proof. Fix L. We construct a parsing by the following greedy rule. Suppose the current position is i. If $i+L-1 \le N$ and the block

$$W[i, i+L-1]$$

has an occurrence starting at some position $p < i$, we make a copied phrase of length L, with source p. Otherwise we make a literal phrase of length 1. Once fewer than L symbols remain, we parse the rest as literals.

Call a phrase short if it is a literal produced before the final L positions. If a short phrase starts at position $i \le N-L+1$, then the block

$$W[i, i+L-1]$$

has no earlier occurrence. Hence two different short phrase starts $i < j \le N-L+1$ give two different length-L blocks. Otherwise, if

$$W[i, i+L-1] = W[j, j+L-1],$$

then at position j the parser could have copied at least L symbols from source i, contradicting that the phrase at j was short.

There are at most b^L distinct blocks of length L. Hence the number of short phrases before the final L positions is at most b^L. The final part contributes at most L literal phrases.

All copied phrases produced by the algorithm have length L, so there are at most N/L such phrases. Therefore

$$\mathrm{oc}(W) \le \frac{N}{L} + b^L + L + O(1).$$

Now take

$$L = \lfloor \log_b N - 2\log_b \log_b N \rfloor .$$

Then

$$b^L \le \frac{N}{(\log_b N)^2},$$

and

$$\frac{N}{L} = (1+o(1))\frac{N}{\log_b N}.$$

Also

$$b^L + L = o\left(\frac{N}{\log_b N}\right).$$

Thus

$$\mathrm{oc}(W) \le (1+o(1))\frac{N}{\log_b N}.$$

□

4. Counting words with short parsings

Let

$$\mathcal{W}_{N,z} = \left\{W \in \{0,\dots,b-1\}^N : \mathrm{oc}(W) \le z\right\}.$$

Lemma 4.1 (Counting parse descriptions)**.** *For $1 \le z \le N/2$,*

$$|\mathcal{W}_{N,z}| \le z\left(\frac{eN}{z}\right)^z (2bN)^z.$$

Consequently, if

$$z_N = \left\lfloor \kappa \frac{N}{\log_b N} \right\rfloor$$

with fixed $0 < \kappa \le 1$, then

$$|\mathcal{W}_{N,z_N}| \le b^{(\kappa+o(1))N}.$$

Proof. Fix $m \le z$. We count a superset of words admitting a parsing with m phrases.

The boundaries are determined by the $m-1$ internal cut positions, so there are

$$\binom{N-1}{m-1}$$

choices.

For each phrase, choose whether it is a literal or a copy, giving at most 2^m choices. Assign to every phrase a symbol in the alphabet, even if the phrase is a copy; this gives at most b^m choices. Assign also to every phrase a source position in $\{1,\dots,N\}$, even if the phrase is a literal; this gives at most N^m choices.

Thus the number of descriptions with m phrases is at most

$$\binom{N-1}{m-1}(2bN)^m.$$

Every such description determines at most one word. Indeed, literals prescribe their symbols, and copied phrases are deterministic once the already

parsed prefix is known. If a source overlaps the target, the copied symbols are still determined progressively because the source starts earlier than the target.

Therefore

$$|\mathcal{W}_{N,z}| \leq \sum_{m \leq z} \binom{N-1}{m-1} (2bN)^m.$$

For $m \leq z \leq N/2$,

$$\binom{N-1}{m-1} \leq \left(\frac{eN}{m}\right)^m.$$

The function

$$m \mapsto \left(\frac{eN}{m}\right)^m (2bN)^m$$

is increasing for $1 \leq m \leq N/2$. Hence

$$|\mathcal{W}_{N,z}| \leq z \left(\frac{eN}{z}\right)^z (2bN)^z.$$

Now take

$$z_N = \left\lfloor \kappa \frac{N}{\log_b N} \right\rfloor.$$

Taking logarithms in base b,

$$\log_b |\mathcal{W}_{N,z_N}| \leq \log_b z_N + z_N \log_b \frac{eN}{z_N} + z_N \log_b (2bN).$$

The last term is

$$z_N \log_b N + O(z_N) = (\kappa + o(1))N.$$

The middle term is

$$z_N \log_b \frac{eN}{z_N} = O\left(\frac{N}{\log N} \log\log N\right) = o(N).$$

Also $\log_b z_N = o(N)$. Therefore

$$\log_b |\mathcal{W}_{N,z_N}| \leq (\kappa + o(1))N,$$

as claimed. □

5. Worst-case complexity

Theorem 5.1 (Worst-case asymptotics)**.**

$$\max_{W \in \{0,\dots,b-1\}^N} \mathrm{oc}(W) = (1 + o(1)) \frac{N}{\log_b N}.$$

Proof. The upper bound follows from the universal upper bound.

For the lower bound, fix $0 < \kappa < 1$, and set

$$z_N = \left\lfloor \kappa \frac{N}{\log_b N} \right\rfloor.$$

By the counting lemma,

$$|\mathcal{W}_{N,z_N}| \le b^{(\kappa+o(1))N}.$$

Since the total number of words of length N is b^N, for all large N there exists a word W with

$$\mathrm{oc}(W) > z_N.$$

Thus

$$\max_{|W|=N} \mathrm{oc}(W) \ge \kappa \frac{N}{\log_b N}(1-o(1)).$$

Letting $\kappa \uparrow 1$ gives the lower bound. □

6. Metric law for online previous-copy complexity

For $\alpha \in [0,1]$, let

$$w_N(\alpha)$$

be the prefix of length N of the base-b expansion of α. We ignore the countable set of numbers with two base-b expansions.

Theorem 6.1 (Almost-sure asymptotic)**.** *For Lebesgue-almost every $\alpha \in [0,1]$,*

$$\mathrm{oc}(w_N(\alpha)) = (1+o(1))\frac{N}{\log_b N}.$$

Equivalently,

$$\lim_{N\to\infty} \frac{\mathrm{oc}(w_N(\alpha))\log_b N}{N} = 1$$

for almost every α.

Proof. The upper bound holds for every word by the universal upper bound.

For the lower bound, fix $0 < \kappa < 1$, and set

$$z_N = \left\lfloor \kappa \frac{N}{\log_b N} \right\rfloor.$$

The set of α such that

$$\mathrm{oc}(w_N(\alpha)) \le z_N$$

is a union of base-b cylinders of length N, one for each word in $\mathcal{W}_{N,z_N}$. Each cylinder has Lebesgue measure b^{-N}. Hence its measure is at most

$$b^{-N}|\mathcal{W}_{N,z_N}| \le b^{-(1-\kappa+o(1))N}.$$

This is summable in N. By the Borel–Cantelli lemma, for almost every α, the event

$$\mathrm{oc}(w_N(\alpha)) \le \kappa \frac{N}{\log_b N}$$

occurs for only finitely many N. Therefore

$$\liminf_{N\to\infty} \frac{\mathrm{oc}(w_N(\alpha))\log_b N}{N} \ge \kappa$$

for every $0 < \kappa < 1$, and hence the liminf is at least 1. The limsup is at most 1 by the universal upper bound. □

7. Finite entropy of low-complexity words

For $0 < \kappa \leq 1$, define

$$\mathcal{C}_N(\kappa) = \left\{ W \in \{0, \dots, b-1\}^N : \mathrm{oc}(W) \leq \kappa \frac{N}{\log_b N} \right\}.$$

Theorem 7.1 (Entropy spectrum)**.** *For $0 < \kappa \leq 1$,*

$$\lim_{N\to\infty} \frac{1}{N} \log_b |\mathcal{C}_N(\kappa)| = \kappa.$$

For $\kappa \geq 1$, the limit is 1.

Proof. The upper bound for $0 < \kappa \leq 1$ follows directly from the counting lemma:

$$|\mathcal{C}_N(\kappa)| \leq b^{(\kappa+o(1))N}.$$

For the lower bound, fix $0 < \eta < \kappa$, and put

$$M = \lfloor (\kappa - \eta)N \rfloor.$$

For every word

$$U \in \{0, \dots, b-1\}^M,$$

construct a word $W(U)$ of length N as follows. First write U. Then continue periodically with period U until the word has total length N.

Different U's give different $W(U)$'s, because the first M symbols of $W(U)$ are U. Hence this construction gives b^M distinct words.

By the universal upper bound applied to U,

$$\mathrm{oc}(U) \leq (1 + o(1)) \frac{M}{\log_b M}$$

uniformly in U. After U has been parsed, the periodic continuation is obtained with one copied phrase, using source position 1 and allowing self-overlap. Thus

$$\mathrm{oc}(W(U)) \leq (1 + o(1)) \frac{M}{\log_b M} + 1.$$

Since

$$M = (\kappa - \eta + o(1))N \qquad \text{and} \qquad \log_b M \sim \log_b N,$$

we get

$$\mathrm{oc}(W(U)) \leq (\kappa - \eta + o(1)) \frac{N}{\log_b N} \leq \kappa \frac{N}{\log_b N}$$

for all sufficiently large N. Hence

$$|\mathcal{C}_N(\kappa)| \geq b^M = b^{(\kappa-\eta+o(1))N}.$$

Letting $\eta \to 0$ gives

$$\liminf_{N\to\infty} \frac{1}{N} \log_b |\mathcal{C}_N(\kappa)| \geq \kappa.$$

For $\kappa \geq 1$, the universal upper bound implies that all words are eventually counted up to $o(1)$ in the threshold, and the entropy limit is 1. □

8. Hausdorff spectrum

For $0 \le \kappa \le 1$, define

$$F_\kappa = \left\{ \alpha \in [0,1] : \liminf_{N\to\infty} \frac{\mathrm{oc}(w_N(\alpha)) \log_b N}{N} \le \kappa \right\}.$$

Theorem 8.1 (Hausdorff spectrum)**.** *For $0 \le \kappa \le 1$,*

$$\dim_{\mathrm{H}} F_\kappa = \kappa.$$

For $\kappa \ge 1$, F_κ has full Lebesgue measure and hence Hausdorff dimension 1.

Proof. We first prove the upper bound. Fix $\eta > 0$. If $\alpha \in F_\kappa$, then for infinitely many N,

$$\mathrm{oc}(w_N(\alpha)) \le (\kappa+\eta)\frac{N}{\log_b N}.$$

Thus F_κ is contained in the limsup of the union of cylinders of length N corresponding to words in

$$\mathcal{C}_N(\kappa+\eta).$$

By the entropy upper bound,

$$|\mathcal{C}_N(\kappa+\eta)| \le b^{(\kappa+\eta+o(1))N}.$$

Each cylinder has diameter comparable to b^{-N}. If $s > \kappa+\eta$, then

$$\sum_N |\mathcal{C}_N(\kappa+\eta)| b^{-sN} < \infty.$$

Hence the s-dimensional Hausdorff measure of F_κ is zero. Therefore

$$\dim_{\mathrm{H}} F_\kappa \le \kappa+\eta.$$

Letting $\eta \to 0$ gives

$$\dim_{\mathrm{H}} F_\kappa \le \kappa.$$

For the lower bound, assume $0 < \kappa \le 1$. We construct a Cantor set contained in F_κ. Choose a sequence of integers $M_j \to \infty$ growing so rapidly that the total length constructed before stage j is $o(M_j/\log M_j)$.

Suppose that before stage j a prefix of length R_{j-1} has been constructed. Choose freely a word

$$U_j \in \{0,\dots,b-1\}^{M_j}.$$

After writing U_j, continue periodically with period U_j until the total length reaches

$$N_j = R_{j-1} + \left\lfloor \frac{M_j}{\kappa} \right\rfloor.$$

Since $R_{j-1} = o(M_j)$,

$$N_j \sim \frac{M_j}{\kappa}.$$

At time N_j, the already existing prefix before U_j contributes $o(M_j/\log M_j)$ phrases. The block U_j can be parsed using the universal upper bound:

$$\mathrm{oc}(U_j) \le (1+o(1))\frac{M_j}{\log_b M_j}.$$

The periodic continuation after U_j is copied with one phrase, since self-overlap is allowed. Thus

$$\mathrm{oc}(w_{N_j}) \le (1+o(1))\frac{M_j}{\log_b M_j}.$$

Because

$$N_j \sim \frac{M_j}{\kappa} \qquad \text{and} \qquad \log_b N_j \sim \log_b M_j,$$

we obtain

$$\frac{\mathrm{oc}(w_{N_j})\log_b N_j}{N_j} \le \kappa + o(1).$$

Hence every point in the constructed Cantor set lies in F_κ.

It remains to estimate its dimension. At stage j, the number of independent choices is b^{M_j}. Since the sequence M_j grows very fast, the contribution of all previous stages is negligible compared with M_j. The cylinders at stage j have length b^{-N_j}. A standard mass distribution argument gives

$$\dim_{\mathrm{H}} \ge \liminf_{j\to\infty} \frac{\sum_{i\le j} M_i}{N_j} = \liminf_{j\to\infty} \frac{M_j + o(M_j)}{M_j/\kappa} = \kappa.$$

Thus

$$\dim_{\mathrm{H}} F_\kappa \ge \kappa.$$

The case $\kappa = 0$ follows from the upper bound and the fact that F_0 is nonempty, for example it contains eventually periodic expansions. Therefore $\dim_{\mathrm{H}} F_0 = 0$. □

9. Normalized substring complexity and string attractors

For a finite word W, let $p_W(k)$ be the number of distinct factors of W of length k. Define

$$\delta(W) = \max_{1\le k\le |W|} \frac{p_W(k)}{k}.$$

Let $\gamma(W)$ denote the size of the smallest string attractor of W. Recall that a set of positions

$$\Gamma \subseteq \{1,\dots,|W|\}$$

is a string attractor if every distinct factor of W has an occurrence crossing at least one position of Γ.

Lemma 9.1. *For every finite word W,*

$$\delta(W) \le \gamma(W) \le \mathrm{oc}(W).$$

Proof. First let Γ be a string attractor for W. Fix k. Every distinct length-k factor has an occurrence crossing some position of Γ. A fixed position can be crossed by at most k length-k factors. Therefore

$$p_W(k) \le |\Gamma| k.$$

Taking the minimum over Γ and then the maximum over k gives

$$\delta(W) \le \gamma(W).$$

Now take an online previous-copy parsing of W with phrase starts

$$n_0 + 1, n_1 + 1, \dots, n_{z-1} + 1.$$

Let Γ be this set of phrase starts. We claim that Γ is a string attractor.

Let U be any factor of W, and choose its leftmost occurrence. If this occurrence does not cross a phrase start, it lies strictly inside a single phrase. If that phrase is a literal, this is impossible unless $|U| = 1$, in which case the occurrence crosses the phrase start. If the phrase is copied, the source of the copy gives an earlier occurrence of U, contradicting the choice of the leftmost occurrence. Hence every factor has an occurrence crossing a phrase start, and Γ is an attractor. Thus

$$\gamma(W) \le z.$$

Taking the minimum over parsings gives

$$\gamma(W) \le \mathrm{oc}(W).$$

□

9.1. **Universal and worst-case bounds.**

Lemma 9.2 (Universal upper bound for δ)**.** *Uniformly for all words W of length N,*

$$\delta(W) \le (1 + o(1)) \frac{N}{\log_b N}.$$

Proof. For every k,

$$p_W(k) \le \min\{b^k, N - k + 1\} \le \min\{b^k, N\}.$$

Thus

$$\frac{p_W(k)}{k} \le \frac{\min\{b^k, N\}}{k}.$$

If $k \le \log_b N$, then the maximum of b^k/k in this range occurs at $k = \lfloor \log_b N \rfloor + O(1)$, and is

$$(1 + o(1)) \frac{N}{\log_b N}.$$

If $k \ge \log_b N$, then

$$\frac{N}{k} \le \frac{N}{\log_b N}.$$

Taking the maximum over k proves the claim. □

Theorem 9.3 (Worst-case for δ and γ)**.**

$$\max_{|W|=N} \delta(W) = (1+o(1))\frac{N}{\log_b N},$$

and

$$\max_{|W|=N} \gamma(W) = (1+o(1))\frac{N}{\log_b N}.$$

Proof. The upper bound for δ is the universal bound just proved. The upper bound for γ follows from

$$\gamma(W) \leq \mathrm{oc}(W)$$

and the universal upper bound for oc.

For the lower bounds, it is enough to show that there exist words W with

$$\delta(W) \geq (1-o(1))\frac{N}{\log_b N}.$$

This follows, for example, from the almost-sure lower bound for δ proved in the next subsection. Since

$$\delta(W) \leq \gamma(W),$$

the same examples give the lower bound for γ. □

9.2. **Almost-sure behaviour.**

Theorem 9.4 (Almost-sure behaviour of δ and γ)**.** *For Lebesgue-almost every* $\alpha \in [0,1]$*,*

$$\delta(w_N(\alpha)) = (1+o(1))\frac{N}{\log_b N},$$

and

$$\gamma(w_N(\alpha)) = (1+o(1))\frac{N}{\log_b N}.$$

Proof. The upper bound for δ is universal. The upper bound for γ follows from

$$\gamma(W) \leq \mathrm{oc}(W)$$

and the almost-sure asymptotic for oc.

It remains to prove the almost-sure lower bound for δ. Fix $\varepsilon > 0$, and set

$$k_N = \lceil (1+\varepsilon)\log_b N \rceil.$$

For a random word W_N of length N, let C_N be the number of pairs $1 \leq i < j \leq N - k_N + 1$ such that

$$W_N[i, i+k_N-1] = W_N[j, j+k_N-1].$$

For any pair $i < j$, the probability of equality is at most b^{-k_N}, even when the two blocks overlap. Indeed, if $j - i = h < k_N$, the equality forces a word of length $k_N + h$ to have period h, and the probability is at most b^{-k_N}. Therefore

$$\mathbb{E}C_N \ll N^2 b^{-k_N} \ll N^{1-\varepsilon}.$$

Choose a subsequence

$$N_m = \lfloor m^q \rfloor$$

with $q\varepsilon > 2$. By Markov's inequality,

$$\mathbb{P}(C_{N_m} > \eta N_m) \le \frac{\mathbb{E}C_{N_m}}{\eta N_m} \ll N_m^{-\varepsilon} \ll m^{-q\varepsilon}.$$

The last series is summable. By Borel–Cantelli,

$$C_{N_m} = o(N_m)$$

almost surely.

The number of distinct length-k_{N_m} factors in the prefix of length N_m is at least

$$N_m - k_{N_m} + 1 - C_{N_m} = (1 - o(1))N_m,$$

because the number of repeated occurrences is bounded by the number of colliding pairs.

Now let $N_m \le N < N_{m+1}$. Since

$$\frac{N_{m+1}}{N_m} \to 1,$$

we have $N_m \sim N$ and $\log N_m \sim \log N$. The prefix $w_N(\alpha)$ contains the prefix $w_{N_m}(\alpha)$, so

$$p_{w_N(\alpha)}(k_{N_m}) \ge (1 - o(1))N_m = (1 - o(1))N.$$

Also

$$k_{N_m} = (1 + \varepsilon + o(1)) \log_b N.$$

Thus

$$\delta(w_N(\alpha)) \ge \frac{p_{w_N(\alpha)}(k_{N_m})}{k_{N_m}} \ge (1 - o(1)) \frac{N}{(1+\varepsilon)\log_b N}.$$

Since $\varepsilon > 0$ is arbitrary, taken over a countable sequence tending to 0, we obtain

$$\delta(w_N(\alpha)) \ge (1 - o(1)) \frac{N}{\log_b N}$$

almost surely. This proves the theorem. The lower bound for γ follows from

$$\delta(W) \le \gamma(W).$$

□

10. Summary of asymptotic scales

The results above show that the online previous-copy complexity and the standard repetitiveness measures δ and γ share the same natural scale in the worst case and almost surely:

$$\frac{N}{\log_b N}.$$

More precisely,

$$\max_{|W|=N} \mathrm{oc}(W) = (1+o(1))\frac{N}{\log_b N},$$

$$\max_{|W|=N} \delta(W) = (1+o(1))\frac{N}{\log_b N},$$

and

$$\max_{|W|=N} \gamma(W) = (1+o(1))\frac{N}{\log_b N}.$$

For Lebesgue-almost every $\alpha \in [0,1]$,

$$\mathrm{oc}(w_N(\alpha)) = \delta(w_N(\alpha)) = \gamma(w_N(\alpha)) = (1+o(1))\frac{N}{\log_b N}.$$

The entropy spectrum and Hausdorff spectrum for oc are also determined:

$$\lim_{N\to\infty} \frac{1}{N}\log_b \#\left\{W \in \{0,\dots,b-1\}^N : \mathrm{oc}(W) \le \kappa\frac{N}{\log_b N}\right\} = \kappa$$

for $0 < \kappa \le 1$, and

$$\dim_{\mathrm{H}}\left\{\alpha : \liminf_{N\to\infty} \frac{\mathrm{oc}(w_N(\alpha))\log_b N}{N} \le \kappa\right\} = \kappa$$

for $0 \le \kappa \le 1$.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, ChatGPT, by OpenAI, was used to assist with mathematical drafting, formalization, review, and editing. This work is shared as a preliminary AI-assisted mathematical note. The mathematical content may have been only partially reviewed and may contain errors; it should not be treated as peer-reviewed or as a fully verified manuscript.

References

[KP18] D. Kempa and N. Prezza, *At the roots of dictionary compression: string attractors*, Proceedings of the 50th Annual ACM Symposium on Theory of Computing, 2018, 827–840.

[KNP23] T. Kociumaka, G. Navarro, and N. Prezza, *Toward a definitive compressibility measure for repetitive sequences*, IEEE Trans. Inform. Theory **69** (2023), no. 4, 2074–2092.

[Pre17] N. Prezza, *String attractors*, arXiv:1709.05314, 2017.

[ZL77] J. Ziv and A. Lempel, *A universal algorithm for sequential data compression*, IEEE Trans. Inform. Theory **23** (1977), no. 3, 337–343.