entropy

MDPI

Article

Is Natural Language a Perigraphic Process? The Theorem about Facts and Words Revisited

Łukasz Dębowski

Institute of Computer Science, Polish Academy of Sciences, ul. Jana Kazimierza 5, 01-248 Warszawa, Poland; ldebowsk@ipipan.waw.pl; Tel.: +48-22-3800-553

Received: 4 January 2018; Accepted: 24 January 2018; Published: 26 January 2018

Abstract: As we discuss, a stationary stochastic process is nonergodic when a random persistent topic can be detected in the infinite random text sampled from the process, whereas we call the process strongly nonergodic when an infinite sequence of independent random bits, called probabilistic facts, is needed to describe this topic completely. Replacing probabilistic facts with an algorithmically random sequence of bits, called algorithmic facts, we adapt this property back to ergodic processes. Subsequently, we call a process perigraphic if the number of algorithmic facts which can be inferred from a finite text sampled from the process grows like a power of the text length. We present a simple example of such a process. Moreover, we demonstrate an assertion which we call the theorem about facts and words. This proposition states that the number of probabilistic or algorithmic facts which can be inferred from a text drawn from a process must be roughly smaller than the number of distinct word-like strings detected in this text by means of the Prediction by Partial Matching (PPM) compression algorithm. We also observe that the number of the word-like strings for a sample of plays by Shakespeare follows an empirical stepwise power law, in a stark contrast to Markov processes. Hence, we suppose that natural language considered as a process is not only non-Markov but also perigraphic.

Keywords: stationary processes; PPM code; mutual information; power laws; algorithmic information theory; natural language

MSC: 94A17, 60G10, 94A29, 68Q30, 68T50

1. Introduction

One of the motivating assumptions of information theory [1–3] is that communication in natural language can be reasonably modeled as a discrete stationary stochastic process, namely, an infinite sequence of discrete random variables with a well defined time-invariant probability distribution. The same assumption is made in several practical applications of computational linguistics, such as speech recognition [4] or part-of-speech tagging [5]. Whereas state-of-the-art stochastic models of natural language are far from being satisfactory, we may ask a more theoretically oriented question, namely:

> What can be some general mathematical properties of natural language treated as a stochastic process, in view of empirical data?

In this paper, we will investigate a question of whether it is reasonable to assume that natural language communication is a *perigraphic* process.

To recall, a stationary process is called ergodic if the relative frequencies of all finite substrings in the infinite text generated by the process converge in the long run with probability one to some constants—the probabilities of the respective strings. Now, some basic linguistic intuition suggests that natural language does not satisfy this property, cf. ([3], Section 6.4). Namely, we can probably agree that there is a variation of topics of texts in natural language, and these topics can be empirically distinguished by counting relative frequencies of certain substrings called keywords. Hence, we expect that the relative frequencies of keywords in a randomly selected text in natural language are random variables depending on the random text topic. In the limit, for an infinitely long text, we may further suppose that the limits of relative frequencies of keywords persist to be random, and if this is true then natural language is not ergodic, i.e., it is nonergodic.

In this paper, we will entertain first a stronger hypothesis, namely, that natural language communication is strongly nonergodic. Informally speaking, a stationary process will be called strongly nonergodic if its random persistent topic has to be described using an infinite sequence of probabilistically independent binary random variables, called probabilistic facts. Like nonergodicity, strong nonergodicity is not empirically verifiable if we only have a single infinite sequence of data. However, replacing probabilistic facts with an algorithmically random sequence of bits, called algorithmic facts, we can adapt the property of strong nonergodicity back to ergodic processes. Subsequently, we will call a process *perigraphic* if the number of algorithmic facts which can be inferred from a finite text sampled from the process grows like a power of the text length. It is a general observation that perigraphic processes have uncomputable distributions.

It is interesting to note that *perigraphic* processes can be singled out by some statistical properties of the texts they generate. We will exhibit a proposition, which we call the theorem about facts and words. Suppose that we have a finite text drawn from a stationary process. The theorem about facts and words says that the number of independent probabilistic or algorithmic facts that can be reasonably inferred from the text must be roughly smaller than the number of distinct word-like strings detected in the text by some standard data compression algorithm called the Prediction by Partial Matching (PPM) code [6,7]. It is important to stress that in this theorem we do not relate the numbers all facts and all word-like strings, which would sound trivial, but we compare only the numbers of independent facts and distinct word-like strings.

Having the theorem about facts and words, we can also discuss some empirical data. Since the number of distinct word-like strings for texts in natural language follows an empirical stepwise power law, in a stark contrast to Markov processes, consequently, we suppose that the number of inferrable random facts for natural language also follows a power law. That is, we suppose that natural language is not only non-Markov but also *perigraphic*.

Whereas in this paper we fill several important missing gaps and provide an overarching narration, the basic ideas presented in this paper are not so new. The starting point was a corollary of Zipf's law and a hypothesis by Hilberg. Zipf's law is an empirical observation that in texts in natural language, the frequencies of words obey a power law decay when we sort the words according to their decreasing frequencies [8,9]. A corollary of this law, called Heaps' law [10–13], states that the number of distinct words in a text in natural language grows like a power of the text length. In contrast to these simple empirical observations, Hilberg's hypothesis is a less known conjecture about natural language that the entropy of a text chunk of an increasing length [14] or the mutual information between two adjacent text chunks [15–18] obey also a power law growth. In Ref. [19], it was heuristically shown that, if Hilberg's hypothesis for mutual information is satisfied for an arbitrary stationary stochastic process, then texts drawn from this process satisfy also a kind of Heaps' law if we detect the words using the grammar-based codes [20–23]. This result is a historical antecedent of the theorem about facts and words.

Another important step was a discovery of some simple strongly nonergodic processes, satisfying the power law growth of mutual information, called Santa Fe processes, discovered by Dębowski in August 2002, but first reported only in [24]. Subsequently, in Ref. [25], a completely formal proof of the theorem about facts and words for strictly minimal grammar-based codes [23,26] was provided. The respective related theory of natural language was later reviewed in [27,28] and supplemented by a discussion of Santa Fe processes in [29]. A drawback of this theory at that time was that strictly minimal grammar-based codes used in the statement of the theorem about facts and words are not computable in a polynomial time [26]. This precluded an empirical verification of the theory.

To state the relative novelty, in this paper, we are glad to announce a new stronger version of the theorem about facts and words for a somewhat more elegant definition of inferrable facts and the PPM code, which is computable almost in a linear time. For the first time, we also present two cases of the theorem: one for strongly nonergodic processes, applying Shannon information theory, and one for general stationary processes, applying algorithmic information theory. Having these results, we can supplement them finally with a rudimentary discussion of some empirical data.

The organization of this paper is as follows. In Section 2, we discuss some properties of ergodic and nonergodic processes. In Section 3, we define strongly nonergodic processes and we present some examples of them. Analogically, in Section 4, we discuss perigraphic processes. In Section 5, we discuss two versions of the theorem about facts and words. In Section 6, we discuss some empirical data and we suppose that natural language may be a perigraphic process. In Section 7, we offer concluding remarks. Moreover, three appendices follow the body of the paper. In Appendix A, we prove the first part of the theorem about facts and words. In Appendix B, we prove the second part of this theorem. In Appendix C, we show that that the number of inferrable facts for the Santa Fe processes follows a power law.

2. Ergodic and Nonergodic Processes

We assume that the reader is familiar with some probability measure theory [30]. For a real-valued random variable Y on a probability space $(\Omega, \mathcal{J}, P)$, we denote its expectation

$$\mathbf{E}\, Y := \int Y dP. \tag{1}$$

Consider now a discrete stochastic process $(X_i)_{i=1}^{\infty} = (X_1, X_2, \dots)$, where random variables X_i take values from a set $\mathbb{X}$ of countably many distinct symbols, such as letters with which we write down texts in natural language. We denote blocks of consecutive random variables $X_j^k := (X_j, \dots, X_k)$ and symbols $x_j^k := (x_j, \dots, x_k)$. Let us define a binary random variable telling whether some string x_1^n has occurred in sequence $(X_i)_{i=1}^{\infty}$ on positions from i to $i + n - 1$,

$$\Phi_i(x_1^n) := \mathbf{1}\left\{X_i^{i+n-1} = x_1^n\right\}, \tag{2}$$

where

$$\mathbf{1}\{\phi\} = \begin{cases} 1, & \text{if } \phi \text{ is true,} \\ 0, & \text{if } \phi \text{ is false.} \end{cases} \tag{3}$$

The expectation of this random variable,

$$\mathbf{E}\, \Phi_i(x_1^n) = P(X_i^{i+n-1} = x_1^n), \tag{4}$$

is the probability of the chosen string according to the considered probability measure P, whereas the arithmetic average of consecutive random variables $\frac{1}{m}\sum_{i=1}^{m} \Phi_i(x_1^n)$ is the relative frequency of the same string in a finite sequence of random symbols X_1^{m+n-1}.

Process $(X_i)_{i=1}^{\infty}$ is called *stationary* (with respect to a probability measure P) if expectations $\mathbf{E}\, \Phi_i(x_1^n)$ do not depend on position i for any string x_1^n. In this case, we have the following well known theorem, which establishes that the limiting relative frequencies of strings x_1^n in infinite sequence $(X_i)_{i=1}^{\infty}$ exist almost surely, i.e., with probability 1:

Theorem 1 (ergodic theorem, cf. e.g., [31])**.** *For any discrete stationary process* $(X_i)_{i=1}^{\infty}$, *there exist limits*

$$\Phi(x_1^n) := \lim_{m\to\infty} \frac{1}{m} \sum_{i=1}^{m} \Phi_i(x_1^n) \text{ almost surely,} \tag{5}$$

with expectations $\mathbf{E}\,\Phi(x_1^n) = \mathbf{E}\,\Phi_i(x_1^n)$.

In general, limits $\Phi(x_1^n)$ are random variables depending on a particular value of infinite sequence $(X_i)_{i=1}^{\infty}$. It is quite natural, however, to require that the relative frequencies of strings $\Phi(x_1^n)$ are almost surely constants, equal to the expectations $\mathbf{E}\,\Phi_i(x_1^n)$. Subsequently, process $(X_i)_{i=1}^{\infty}$ will be called *ergodic* (with respect to a probability measure P) if limits $\Phi(x_1^n)$ are almost surely constant for any string x_1^n. The standard definition of an ergodic process is more abstract but is equivalent to this statement ([31], Lemma 7.15).

The following examples of ergodic processes are well known:

1. Process $(X_i)_{i=1}^{\infty}$ is called *IID* (independent identically distributed) if

$$P(X_1^n = x_1^n) = \pi(x_1)\dots\pi(x_n). \tag{6}$$

All IID processes are ergodic.

2. Process $(X_i)_{i=1}^{\infty}$ is called *Markov* (of order 1) if

$$P(X_1^n = x_1^n) = \pi(x_1)p(x_2|x_1)\dots p(x_n|x_{n-1}). \tag{7}$$

A Markov process is ergodic in particular if

$$p(x_i|x_{i-1}) > c > 0. \tag{8}$$

For a sufficient and necessary condition, see ([32], Theorem 7.16).

3. Process $(X_i)_{i=1}^{\infty}$ is called *hidden Markov* if $X_i = g(S_i)$ for a certain Markov process $(S_i)_{i=1}^{\infty}$ and a function g. A hidden Markov process is ergodic in particular if the underlying Markov process is ergodic.

Whereas IID and Markov processes are some basic models in probability theory, hidden Markov processes are of practical importance in computational linguistics [4,5]. Hidden Markov processes as considered there usually satisfy condition (8) and therefore they are ergodic.

Let us call a probability measure P stationary or ergodic, respectively, if the process $(X_i)_{i=1}^{\infty}$ is stationary or ergodic with respect to the measure P. Suppose that we have a stationary measure P that generates some data $(X_i)_{i=1}^{\infty}$. We can define a new random measure F equal to the relative frequencies of blocks in the data $(X_i)_{i=1}^{\infty}$. It turns out that the measure F is almost surely ergodic. Formally, we have this proposition.

Theorem 2 (cf. ([33], Theorem 9.10))**.** *Any process* $(X_i)_{i=1}^{\infty}$ *with a stationary measure* P *is almost surely ergodic with respect to the random measure* F *given by*

$$F(X_1^n = x_1^n) := \Phi(x_1^n). \tag{9}$$

Moreover, from the random measure F, we can obtain the stationary measure P by integration, $P(X_1^n = x_1^n) = \mathbf{E}\,F(X_1^n = x_1^n)$. The following result asserts that this integral representation of measure P is unique.

Theorem 3 (ergodic decomposition, cf. ([33], Theorem 9.12))**.** *Any stationary probability measure P can be represented as*

$$P(X_1^n = x_1^n) = \int F(X_1^n = x_1^n) d\nu(F), \tag{10}$$

where ν is a unique measure on stationary ergodic measures.

In other words, stationary ergodic measures are some building blocks from which we can construct any stationary measure. For a stationary probability measure P, the particular values of the random ergodic measure F are called the ergodic components of measure P.

Consider for instance, a Bernoulli(θ) process with measure

$$F_\theta(X_1^n = x_1^n) = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}, \tag{11}$$

where $x_i \in \{0, 1\}$ and $\theta \in [0, 1]$. This measure will be contrasted with the measure of a mixture Bernoulli process with parameter θ uniformly distributed on interval $[0, 1]$,

$$\begin{aligned} P(X_1^n = x_1^n) &= \int_0^1 F_\theta(X_1^n = x_1^n) d\theta \\ &= \frac{1}{n+1}\left[\binom{n}{\sum_{i=1}^n x_i}\right]^{-1}. \end{aligned} \tag{12}$$

Measure (11) is a measure of an IID process and is therefore ergodic, whereas measure (12) is a mixture of ergodic measures and hence it is nonergodic.

3. Strongly Nonergodic Processes

According to our definition, a process is ergodic when the relative frequencies of any strings in a random sample in the long run converge to some constants. Consider now the following thought experiment. Suppose that we select a random book from a library. In [34], it was observed that there is hardly any book that contains both the word *lemma* and the word *love*, namely, there are some keywords that are specific to particular topics of texts. We can pursue this idea one little step farther. Counting the relative frequencies of keywords, such as *lemma* for a text on mathematics and *love* for a romance, we can effectively recognize the topic of the book. Simply put, the relative frequencies of some keywords will be higher for books concerning some topics, whereas they will be lower for books concerning other topics. Hence, in our thought experiment, we expect that the relative frequencies of keywords are some random variables with values depending on the particular topic of the randomly selected book. Since keywords are just some particular strings, we may conclude that the stochastic process that models natural language should be nonergodic.

The above thought experiment provides another perspective onto nonergodic processes. According to the following theorem, a process is nonergodic when we can effectively distinguish in the limit at least two random topics in it. In the statement, function $f : \mathbb{X}^* \to \{0, 1, 2\}$ assumes values 0 or 1 when we can identify the topic, whereas it takes value 2 when we are not certain which topic a given text is about.

Theorem 4 (cf. [24])**.** *A stationary discrete process $(X_i)_{i=1}^\infty$ is nonergodic if and only if there exists a function $f : \mathbb{X}^* \to \{0, 1, 2\}$ and a binary random variable Z such that $0 < P(Z = 0) < 1$ and*

$$\lim_{n\to\infty} P(f(X_i^{i+n-1}) = Z) = 1 \tag{13}$$

for any position $i \in \mathbb{N}$.

A binary variable Z satisfying condition (13) will be called a *probabilistic fact*. A probabilistic fact tells which of two topics the infinite text generated by the stationary process is about. It is a kind of a random switch which is preset before we start scanning the infinite text; compare a similar wording in [35]. To keep the proofs simple, here we only give a new elementary proof of the " $\Longrightarrow$ " statement of Theorem 4. The proof of the " $\Longleftarrow$ " part applies some measure theory and follows the idea of Theorem 9 from [24] for strongly nonergodic processes, which we will discuss in the next paragraph.

Proof. (only $\Longrightarrow$) Suppose that process $(X_i)_{i=1}^{\infty}$ is nonergodic. Then, there exists a string x_1^k such that $\Phi \neq \mathbf{E}\,\Phi$ for $\Phi := \Phi(x_1^k)$ with some positive probability. Hence, there exists a real number y such that $P(\Phi = y) = 0$ and

$$P(\Phi > y) = 1 - P(\Phi < y) \in (0,1). \tag{14}$$

Define $Z := \mathbf{1}\{\Phi > y\}$ and $f(X_i^{i+n-1}) := Z_{in} := \mathbf{1}\{\Phi_{in} > y\}$, where

$$\Phi_{in} := \frac{1}{n-k+1} \sum_{j=i}^{i+n-k} \Phi_j(x_1^k). \tag{15}$$

Since $\lim_{n\to\infty} \Phi_{in} = \Phi$ almost surely and Φ satisfies (14), convergence $\lim_{n\to\infty} Z_{in} = Z$ also holds almost surely. Applying the Lebesgue dominated convergence theorem, we obtain

$$\begin{aligned}\lim_{n\to\infty} P(f(X_i^{i+n-1}) = Z) &= \lim_{n\to\infty} \mathbf{E}\,[Z_{in}Z + (1-Z_{in})(1-Z)] \\ &= \mathbf{E}\left[Z^2 + (1-Z)^2\right] = 1. \end{aligned}\tag{16}$$

□

As for books in the natural language, we may have an intuition that the pool of available book topics is extremely large and contains many more topics than just two. For this reason, we may need not a single probabilistic fact Z but rather a sequence of probabilistic facts Z_1, Z_2, … to specify the topic of a random book completely. Formally, stationary processes requiring an infinite sequence of independent uniformly distributed probabilistic facts to describe the topic of an infinitely long text will be called strongly nonergodic.

Definition 1 (cf. [24,25]). *A stationary discrete process $(X_i)_{i=1}^{\infty}$ is called* strongly nonergodic *if there exist a function $g : \mathbb{N} \times \mathbb{X}^* \to \{0,1,2\}$ and a binary IID process $(Z_k)_{k=1}^{\infty}$ such that $P(Z_k = 0) = P(Z_k = 1) = 1/2$ and*

$$\lim_{n\to\infty} P(g(k; X_i^{i+n-1}) = Z_k) = 1 \tag{17}$$

for any position $i \in \mathbb{N}$ and any index $k \in \mathbb{N}$.

As we have stated above, for a strongly nonergodic process, there is an infinite number of independent probabilistic facts $(Z_k)_{k=1}^{\infty}$ with a uniform distribution on the set $\{0,1\}$. Formally, these probabilistic facts can be assembled into a single real random variable $T = \sum_{k=1}^{\infty} 2^{-k} Z_k$, which is uniformly distributed on the unit interval $[0,1]$. The value of variable T identifies the topic of a random infinite text generated by the stationary process. Thus, for a strongly nonergodic process, we have a continuum of available topics which can be incrementally identified from any sufficiently long text. Put formally, according to Theorem 9 from [24], a stationary process is strongly nonergodic if and only if its shift-invariant σ-field contains a nonatomic sub-σ-field. We note in passing that in [24] strongly nonergodic processes were called *uncountable description processes*.

In view of Theorem 9 from [24], the mixture Bernoulli process (12) is some example of a strongly nonergodic process. In this case, the parameter θ plays the role of the random variable $T = \sum_{k=1}^{\infty} 2^{-k} Z_k$.

Showing that condition (17) is satisfied for this process in an elementary fashion is a tedious exercise. Hence, let us present now a simpler guiding example of a strongly nonergodic process, which we introduced in [24,25] and called the Santa Fe process. Let $(Z_k)_{k=1}^{\infty}$ be a binary IID process with $P(Z_k = 0) = P(Z_k = 1) = 1/2$. Let $(K_i)_{i=1}^{\infty}$ be an IID process with K_i assuming values in natural numbers with a power-law distribution

$$P(K_i = k) \propto \frac{1}{k^{\alpha}}, \quad \alpha > 1. \tag{18}$$

The *Santa Fe process* with exponent α is a sequence $(X_i)_{i=1}^{\infty}$, where

$$X_i = (K_i, Z_{K_i}) \tag{19}$$

are pairs of a random number K_i and the corresponding probabilistic fact Z_{K_i}. The Santa Fe process is strongly nonergodic since condition (17) holds for example for

$$g(k; x_1^n) = \begin{cases} 0, & \text{if for all } 1 \le i \le n,\, x_i = (k, z) \implies x_i = (k, 0), \\ 1, & \text{if for all } 1 \le i \le n,\, x_i = (k, z) \implies x_i = (k, 1), \\ 2, & \text{else.} \end{cases} \tag{20}$$

Simply speaking, function $g(k; \cdot)$ returns 0 or 1 when an unambiguous value of the second constituent can be read off from pairs $x_i = (k, \cdot)$ and returns 2 when there is some ambiguity. Condition (17) is satisfied since

$$\begin{aligned} P(g(k; X_i^{i+n-1}) = Z_k) \quad &= P(K_i = k \text{ for some } 1 \le i \le n) \\ &= 1 - (1 - P(K_i = k))^n \xrightarrow[n \to \infty]{} 1. \end{aligned} \tag{21}$$

Some salient property of the Santa Fe process is the power law growth of the expected number of probabilistic facts, which can be inferred from a finite text drawn from the process. Consider a strongly nonergodic process $(X_i)_{i=1}^{\infty}$. The set of initial independent probabilistic facts inferrable from a finite text X_1^n will be defined as

$$U(X_1^n) := \{ l \in \mathbb{N} : g(k; X_1^n) = Z_k \text{ for all } k \le l \}. \tag{22}$$

In other words, we have $U(X_1^n) = \{1, 2, \ldots, l\}$, where l is the largest number such that $g(k; X_1^n) = Z_k$ for all $k \le l$. To capture the power-law growth of an arbitrary function $s : \mathbb{N} \to \mathbb{R}$, we will denote the Hilberg exponent defined

$$\operatorname*{hilb}_{n \to \infty} s(n) := \limsup_{n \to \infty} \frac{\log^{+} s(2^n)}{\log 2^n}, \tag{23}$$

where $\log^{+} x := \log(x+1)$ for $x \ge 0$ and $\log^{+} x := 0$ for $x < 0$, cf. [36]. In contrast to Ref. [36], for technical reasons, we define the Hilberg exponent only for an exponentially sparse subsequence of terms $s(2^n)$ rather than all terms $s(n)$. Moreover, in [36], the Hilberg exponent was considered only for mutual information $s(n) = \mathbb{I}(X_1^n; X_{n+1}^{2n})$, defined later in Equation (51). We observe that for the exact power law growth $s(n) = n^{\beta}$ with $\beta \ge 0$ we have $\operatorname{hilb}_{n \to \infty} s(n) = \beta$. More generally, the Hilberg exponent captures an asymptotic power-law growth of the sequence. As shown in Appendix C, for the Santa Fe process with exponent α, we have the asymptotic power-law growth

$$\operatorname*{hilb}_{n \to \infty} \mathbf{E} \operatorname{card} U(X_1^n) = 1/\alpha \in (0, 1). \tag{24}$$

This property distinguishes the Santa Fe process from the mixture Bernoulli process (12), for which the respective Hilberg exponent is zero, as we discuss in Section 6.

4. Perigraphic Processes

Is it possible to demonstrate by a statistical investigation of texts that natural language is really strongly nonergodic and satisfies a condition similar to (24)? In the thought experiment described in the beginning of the previous section, we have ignored the issue of constructing an infinitely long text. In reality, every book with a well defined topic is finite. If we want to obtain an unbounded collection of texts, we need to assemble a corpus of different books and it depends on our assembling criteria whether the books in the corpus will concern some persistent random topic. Moreover, if we already have a *single* infinite sequence of books generated by some stationary source and we estimate probabilities as relative frequencies of blocks of symbols in this sequence, then, by Theorem 2, we will obtain an ergodic probability measure almost surely.

In this situation, we may ask whether the idea of the power-law growth of the number of inferrable probabilistic facts can be translated somehow to the case of ergodic measures. Some straightforward method to apply is to replace the sequence of independent uniformly distributed probabilistic facts $(Z_k)_{k=1}^{\infty}$, being random variables, with an algorithmically random sequence of particular binary digits $(z_k)_{k=1}^{\infty}$. Such digits z_k will be called *algorithmic facts* in contrast to variables Z_k being called *probabilistic facts*.

Let us recall some basic concepts. For a discrete random variable X, let $P(X)$ denote the random variable that takes value $P(X = x)$ when X takes value x. We will introduce the pointwise entropy

$$\mathbb{H}(X) := -\log P(X), \tag{25}$$

where log stands for the natural logarithm. The prefix-free Kolmogorov complexity $K(u)$ of a string u is the length of the shortest self-delimiting program written in binary digits that prints out string u ([37], Chapter 3). $K(u)$ is the founding concept of the algorithmic information theory and is an analogue of the pointwise entropy. To keep our notation analogical to (25), we will write the algorithmic entropy

$$\mathbb{H}_a(u) := K(u) \log 2. \tag{26}$$

If the probability measure is computable, then the algorithmic entropy is close to the pointwise entropy. On the one hand, by the Shannon–Fano coding for a computable probability measure, the algorithmic entropy is less than the pointwise entropy plus a constant which depends on the probability measure and the dimensionality of the distribution ([37], Corollary 4.3.1). Formally,

$$\mathbb{H}_a(X_1^n) \le \mathbb{H}(X_1^n) + 2\log n + C_P, \tag{27}$$

where $C_P \ge 0$ is a certain constant depending on the probability measure P. On the other hand, since the prefix-free Kolmogorov complexity is also the length of a prefix-free code, we have

$$\mathbf{E}\,\mathbb{H}_a(X_1^n) \ge \mathbf{E}\,\mathbb{H}(X_1^n). \tag{28}$$

It is also true that $\mathbb{H}_a(X_1^n) \ge \mathbb{H}(X_1^n)$ for sufficiently large n almost surely ([38], Theorem 3.1). Thus, we have shown that the algorithmic entropy is in some sense close to the pointwise entropy, for a computable probability measure.

Next, we will discuss the difference between probabilistic and algorithmic randomness. Whereas for an IID sequence of random variables $(Z_k)_{k=1}^{\infty}$ with $P(Z_k = 0) = P(Z_k = 1) = 1/2$ we have

$$\mathbb{H}(Z_1^k) = k \log 2, \tag{29}$$

similarly an infinite sequence of binary digits $(z_k)_{k=1}^{\infty}$ is called algorithmically random (in the Martin-Löf sense) when there exists a constant $C \geq 0$ such that

$$\mathbb{H}_a(z_1^k) \geq k \log 2 - C \tag{30}$$

for all $k \in \mathbb{N}$ ([37], Theorem 3.6.1). The probability that the aforementioned sequence of random variables $(Z_k)_{k=1}^{\infty}$ is algorithmically random equals 1—for example by ([38], Theorem 3.1), so algorithmically random sequences are typical realizations of sequence $(Z_k)_{k=1}^{\infty}$.

Let $(X_i)_{i=1}^{\infty}$ be a stationary process. We observe that generalizing condition (17) in an algorithmic fashion does not make much sense. Namely, condition

$$\lim_{n\to\infty} P(g(k; X_i^{i+n-1}) = z_k) = 1 \tag{31}$$

is trivially satisfied for any stationary process for a certain computable function $g : \mathbb{N} \times \mathbb{X}^* \to \{0, 1, 2\}$ and an algorithmically random sequence $(z_k)_{k=1}^{\infty}$. It turns out so since there exists a computable function $\omega : \mathbb{N} \times \mathbb{N} \to \{0, 1\}$ such that $\lim_{n\to\infty} \omega(k; n) = \Omega_k$, where $(\Omega_k)_{k=1}^{\infty}$ is the binary expansion of the halting probability $\Omega = \sum_{k=1}^{\infty} 2^{-k}\Omega_k$, which is a lower semi-computable algorithmically random sequence ([37], Section 3.6.2).

In spite of this negative result, the power-law growth of the number of inferrable algorithmic facts corresponds to some nontrivial property. For a computable function $g : \mathbb{N} \times \mathbb{X}^* \to \{0, 1, 2\}$ and an algorithmically random sequence of binary digits $(z_k)_{k=1}^{\infty}$, which we will call *algorithmic facts*, the set of initial algorithmic facts inferrable from a finite text X_1^n will be defined as

$$U_a(X_1^n) := \{l \in \mathbb{N} : g(k; X_1^n) = z_k \text{ for all } k \leq l\} . \tag{32}$$

Subsequently, we will call a process perigraphic if the expected number of algorithmic facts which can be inferred from a finite text sampled from the process grows asymptotically like a power of the text length.

Definition 2. *A stationary discrete process $(X_i)_{i=1}^{\infty}$ is called* perigraphic *if*

$$\operatorname*{hilb}_{n\to\infty} \mathbf{E} \operatorname{card} U_a(X_1^n) > 0 \tag{33}$$

for some computable function $g : \mathbb{N} \times \mathbb{X}^* \to \{0, 1, 2\}$ and an algorithmically random sequence of binary digits $(z_k)_{k=1}^{\infty}$.

Perigraphic processes can be ergodic. The proof of Theorem A10 from Appendix C can be easily adapted to show that some example of a perigraphic process is the Santa Fe process with sequence $(Z_k)_{k=1}^{\infty}$ replaced by an algorithmically random sequence of binary digits $(z_k)_{k=1}^{\infty}$. To be very concrete, the example of a perigraphic process can be process $(X_i)_{i=1}^{\infty}$ with

$$X_i = (K_i, \Omega_{K_i}) \tag{34}$$

where $(\Omega_k)_{k=1}^{\infty}$ is the binary expansion of the halting probability and $(K_i)_{i=1}^{\infty}$ is an IID process with K_i assuming values in natural numbers with the power-law distribution (18). This process is not only perigraphic but also IID and hence ergodic.

We can also easily show the following proposition.

Theorem 5. *Any perigraphic process $(X_i)_{i=1}^{\infty}$ has an uncomputable measure P.*

Proof. Assume that a perigraphic process $(X_i)_{i=1}^{\infty}$ has a computable measure P. By inequalities (A25) and (A26) from Appendix A, we have

$$\operatorname*{hilb}_{n\to\infty} \mathbf{E} \operatorname{card} U_a(X_1^n) \le \operatorname*{hilb}_{n\to\infty} \mathbf{E} \left[\mathbb{H}_a(X_1^n) - \mathbb{H}(X_1^n)\right]. \tag{35}$$

Since, for a computable measure P, we also have inequality (27), then

$$\operatorname*{hilb}_{n\to\infty} \mathbf{E} \operatorname{card} U_a(X_1^n) = 0. \tag{36}$$

Since we have obtained a contradiction with the assumption that the process is perigraphic, measure P cannot be computable. □

5. Theorem about Facts and Words

In this section, we will present a result about stationary processes, which we call the theorem about facts and words. This proposition states that the expected number of independent probabilistic or algorithmic facts inferrable from the text drawn from a stationary process must be roughly less than the expected number of distinct word-like strings detectable in the text by a simple procedure involving the PPM compression algorithm. This result states, in particular, that an asymptotic power law growth of the number of inferrable probabilistic or algorithmic facts as a function of the text length produces a statistically measurable effect, namely, an asymptotic power law growth of the number of word-like strings.

To state the theorem about facts and words formally, we need first to discuss the PPM code. The general idea of the PPM code comes from Refs. [6,7], developed independently. This compression scheme was called the PPM code in [7], which stands for "Prediction by Partial Matching" and prevails in the literature, whereas it was called measure R in [6,39]. Whereas Ref. [7] focused on practical applications to data compression and earned most of the fame, in Refs. [6,39], one can find a few results that matter for theoretical considerations. Let us denote strings of symbols $x_j^k := (x_j, \ldots, x_k)$, adopting an important convention that x_j^k is the empty string for $k < j$. In the following, we consider strings over a finite alphabet, say, $x_i \in \mathbb{X} = \{1, \ldots, D\}$. We define the frequency of a substring w_1^k in a string x_1^n as

$$N(w_1^k|x_1^n) := \sum_{i=1}^{n-k+1} \mathbf{1}\left\{x_i^{i+k-1} = w_1^k\right\}. \tag{37}$$

Now, we will define the PPM probabilities in a way that is closer to the conventions of paper [6,39] than to the conventions of Ref. [7]. In particular, in Equation (38), we consider frequencies of strings x_{i-k}^i and x_{i-k}^{i-1} in different strings, x_1^{i-1} and x_1^{i-2}, respectively, in the numerator and in the denominator to guarantee the proper normalization according to our definition of $N(w_1^k|x_1^n)$.

Definition 3 (cf. [6,7])**.** *For $x_1^n \in \mathbb{X}^n$ and $k \in \{-1, 0, 1, \ldots\}$, we put*

$$\mathrm{PPM}_k(x_i|x_1^{i-1}) := \begin{cases} \dfrac{1}{D}, & i \le k, \\ \dfrac{N(x_{i-k}^i|x_1^{i-1}) + 1}{N(x_{i-k}^{i-1}|x_1^{i-2}) + D}, & i > k. \end{cases} \tag{38}$$

Quantity $\mathrm{PPM}_k(x_i|x_1^{i-1})$ *is called the* conditional PPM probability *of order k of symbol x_i given string x_1^{i-1}. Next, we put*

$$\mathrm{PPM}_k(x_1^n) := \prod_{i=1}^{n} \mathrm{PPM}_k(x_i|x_1^{i-1}). \tag{39}$$

Quantity $\mathrm{PPM}_k(x_1^n)$ *is called the* PPM probability *of order k of string* x_1^n*. Finally, we put*

$$\mathrm{PPM}(x_1^n) := \frac{6}{\pi^2} \sum_{k=-1}^{\infty} \frac{\mathrm{PPM}_k(x_1^n)}{(k+2)^2}. \tag{40}$$

Quantity $\mathrm{PPM}(x_1^n)$ *is called the (total)* PPM probability *of the string* x_1^n*.*

Quantity $\mathrm{PPM}_k(x_1^n)$ is an incremental approximation of the unknown true probability of the string x_1^n, assuming that the string has been generated by a Markov process of order k. In contrast, quantity $\mathrm{PPM}(x_1^n)$ is a mixture of such Markov approximations for all finite orders. In general, the PPM probabilities are probability distributions over strings of a fixed length. That is:

- $\mathrm{PPM}_k(x_i|x_1^{i-1}) > 0$ and $\sum_{x_i \in \mathbb{X}} \mathrm{PPM}_k(x_i|x_1^{i-1}) = 1$,
- $\mathrm{PPM}_k(x_1^n) > 0$ and $\sum_{x_1^n \in \mathbb{X}^n} \mathrm{PPM}_k(x_1^n) = 1$,
- $\mathrm{PPM}(x_1^n) > 0$ and $\sum_{x_1^n \in \mathbb{X}^n} \mathrm{PPM}(x_1^n) = 1$.

In the following, we define an analogue of the pointwise entropy

$$\mathbb{H}_{\mathrm{PPM}}(x_1^n) := -\log \mathrm{PPM}(x_1^n). \tag{41}$$

Quantity $\mathbb{H}_{\mathrm{PPM}}(x_1^n)$ will be called the length of the PPM code for the string x_1^n. By nonnegativity of the Kullback–Leibler divergence, we have for any random block X_1^n that

$$\mathbf{E}\, \mathbb{H}_{\mathrm{PPM}}(X_1^n) \geq \mathbf{E}\, \mathbb{H}(X_1^n). \tag{42}$$

The length of the PPM code or the PPM probability, respectively, have two notable properties. First, the PPM probability is a universal probability, i.e., in the limit, the length of the PPM code consistently estimates the entropy rate of a stationary source. Second, the PPM probability can be effectively computed, i.e., the summation in definition (40) can be rewritten as a finite sum. Let us state these two results formally.

Theorem 6 (cf. [39])**.** *The PPM probability is universal in expectation, i.e., we have*

$$\lim_{n\to\infty} \frac{1}{n} \mathbf{E}\, \mathbb{H}_{\mathrm{PPM}}(X_1^n) = \lim_{n\to\infty} \frac{1}{n} \mathbf{E}\, \mathbb{H}(X_1^n) \tag{43}$$

for any stationary process $(X_i)_{i=1}^{\infty}$*.*

For stationary ergodic processes, the above claim follows by an iterated application of the ergodic theorem as shown, e.g., in Theorem 1.1 from [39] for the measure R, which is a slight modification of the PPM probability. To generalize the claim for nonergodic processes, one can use the ergodic decomposition theorem, but the exact proof requires too large of a theoretical overload to be presented within the framework of this paper.

Theorem 7. *The PPM probability can be effectively computed, i.e., we have*

$$\mathrm{PPM}(x_1^n) = \frac{6}{\pi^2} \sum_{k=0}^{L(x_1^n)} \frac{\mathrm{PPM}_k(x_1^n)}{(k+2)^2} + \left(1 - \frac{6}{\pi^2} \sum_{k=0}^{L(x_1^n)} \frac{1}{(k+2)^2}\right) D^{-n}, \tag{44}$$

where

$$L(x_1^n) = \max \left\{ k : N(w_1^k|x_1^n) > 1 \textit{ for some } w_1^k \right\} \tag{45}$$

is the maximal repetition of string x_1^n*.*

Proof. We have $N(x_{i-k}^{i-1}|x_1^{i-2}) = 0$ for $k > L(x_1^i)$. Hence, $\mathrm{PPM}_k(x_1^n) = D^{-n}$ for $k > L(x_1^n)$ and, in view of this, we obtain the claim. □

Maximal repetition as a function of a string was studied, e.g., in [40,41]. Since the PPM probability is a computable probability distribution, then, by (27) for a certain constant C_{PPM}, we have

$$\mathbb{H}_a(X_1^n) \le \mathbb{H}_{\mathrm{PPM}}(X_1^n) + 2\log n + C_{\mathrm{PPM}}. \tag{46}$$

Let us denote the length of the PPM code of order k,

$$\mathbb{H}_{\mathrm{PPM}_k}(x_1^n) := -\log \mathrm{PPM}_k(x_1^n). \tag{47}$$

As we can easily see, the code length $\mathbb{H}_{\mathrm{PPM}}(x_1^n)$ is approximately equal to the minimal code length $\mathbb{H}_{\mathrm{PPM}_k}(x_1^n)$ where the minimization goes over $k \in \{-1, 0, 1, \ldots\}$. Thus, it is meaningful to consider this definition of the PPM order of an arbitrary string.

Definition 4. *The* PPM order $G_{\mathrm{PPM}}(x_1^n)$ *is the smallest G such that*

$$\mathbb{H}_{\mathrm{PPM}_G}(x_1^n) \le \mathbb{H}_{\mathrm{PPM}_k}(x_1^n) \text{ for all } k \ge -1. \tag{48}$$

Theorem 8. *We have* $G_{\mathrm{PPM}}(x_1^n) \le L(x_1^n)$.

Proof. It follows by $\mathrm{PPM}_k(x_1^n) = D^{-n} = \mathrm{PPM}_{-1}(x_1^n)$ for $k > L(x_1^n)$. □

Let us divert for a short while from the PPM code definition. The set of distinct substrings of length m in string x_1^n is

$$V(m|x_1^n) := \left\{ y_1^m : x_{t+1}^{t+m} = y_1^m \text{ for some } 0 \le t \le n-m \right\}. \tag{49}$$

The cardinality of set $V(m|x_1^n)$ as a function of substring length m is called the subword complexity of string x_1^n [40]. Now let us apply the concept of the PPM order to define some special set of substrings of an arbitrary string x_1^n. The set of distinct PPM words detected in x_1^n will be defined as the set $V(m|x_1^n)$ for $m = G_{\mathrm{PPM}}(x_1^n)$, i.e.,

$$V_{\mathrm{PPM}}(x_1^n) := V(G_{\mathrm{PPM}}(X_1^n)|x_1^n). \tag{50}$$

Let us define the pointwise mutual information

$$\mathbb{I}(X;Y) := \mathbb{H}(X) + \mathbb{H}(Y) - \mathbb{H}(X,Y) \tag{51}$$

and the algorithmic mutual information

$$\mathbb{I}_a(u;v) := \mathbb{H}_a(u) + \mathbb{H}_a(v) - \mathbb{H}_a(u,v). \tag{52}$$

Now, we may write down the theorem about facts and words. The theorem states that the Hilberg exponent for the expected number of initial independent inferrable facts is less than the Hilberg exponent for the expected mutual information and this is less than the Hilberg exponent for the expected number of distinct detected PPM words plus the PPM order. (The PPM order is usually much less than the number of distinct PPM words.)

Theorem 9 (facts and words I, cf. [25])**.** *Let $(X_i)_{i=1}^{\infty}$ be a stationary strongly nonergodic process over a finite alphabet. We have inequalities*

$$\begin{aligned} \underset{n\to\infty}{\text{hilb}}\ \mathbf{E}\ \text{card}\, U(X_1^n) &\le \underset{n\to\infty}{\text{hilb}}\ \mathbf{E}\, \mathbb{I}(X_1^n; X_{n+1}^{2n}) \\ &\le \underset{n\to\infty}{\text{hilb}}\ \mathbf{E}\ \left[G_{\text{PPM}}(X_1^n) + \text{card}\, V_{\text{PPM}}(X_1^n)\right]. \end{aligned} \tag{53}$$

Proof. The claim follows by conjunction of Theorem A2 from Appendix A and Theorem A8 from Appendix B. □

Theorem 9 also has an algorithmic version, for ergodic processes in particular.

Theorem 10 (facts and words II)**.** *Let $(X_i)_{i=1}^{\infty}$ be a stationary process over a finite alphabet. We have inequalities*

$$\begin{aligned} \underset{n\to\infty}{\text{hilb}}\ \mathbf{E}\ \text{card}\, U_a(X_1^n) &\le \underset{n\to\infty}{\text{hilb}}\ \mathbf{E}\, \mathbb{I}_a(X_1^n; X_{n+1}^{2n}) \\ &\le \underset{n\to\infty}{\text{hilb}}\ \mathbf{E}\ \left[G_{\text{PPM}}(X_1^n) + \text{card}\, V_{\text{PPM}}(X_1^n)\right]. \end{aligned} \tag{54}$$

Proof. The claim follows by conjunction of Theorem A3 from Appendix A and Theorem A8 from Appendix B. □

The theorem about facts and words previously proven in [25] differs from Theorem 9 in three aspects. First of all, the theorem in [25] did not apply the concept of the Hilberg exponent and compared $\liminf_{n\to\infty}$ with $\limsup_{n\to\infty}$ rather than $\limsup_{n\to\infty}$ with $\limsup_{n\to\infty}$. Second, the number of inferrable facts was defined as a functional of the process distribution rather than a random variable depending on a particular text. Third, the number of words was defined using a minimal grammar-based code rather than the concept of the PPM order. Minimal grammar-based codes are not computable in a polynomial time in contrast to the PPM order. Thus, we may claim that Theorem 9 is stronger than the theorem about facts and words previously proven in [25]. Moreover, applying Kolmogorov complexity and algorithmic randomness to formulate and prove Theorem 10 is a new idea.

It is an interesting question whether we have an almost sure version of Theorems 9 and 10, namely, whether

$$\begin{aligned} \text{hilb}_{n\to\infty}\, \text{card}\, U(X_1^n) &\le \text{hilb}_{n\to\infty}\, \mathbb{I}(X_1^n; X_{n+1}^{2n}) \\ &\le \text{hilb}_{n\to\infty} \left[G_{\text{PPM}}(X_1^n) + \text{card}\, V_{\text{PPM}}(X_1^n)\right] \text{ almost surely} \end{aligned} \tag{55}$$

for strongly nonergodic processes, or

$$\begin{aligned} \text{hilb}_{n\to\infty}\, \text{card}\, U_a(X_1^n) &\le \text{hilb}_{n\to\infty}\, \mathbb{I}_a(X_1^n; X_{n+1}^{2n}) \\ &\le \text{hilb}_{n\to\infty} \left[G_{\text{PPM}}(X_1^n) + \text{card}\, V_{\text{PPM}}(X_1^n)\right] \text{ almost surely} \end{aligned} \tag{56}$$

for general stationary processes. We leave this question as an open problem.

6. Hilberg Exponents and Empirical Data

It is advisable to show that the Hilberg exponents considered in Theorem 9 can assume any value in range $[0, 1]$ and the difference between them can be arbitrarily large. We adopt a convention that the set of inferrable probabilistic facts is empty for ergodic processes, $U(X_1^n) = \emptyset$. With this remark in mind, let us inspect some examples of processes.

First of all, for Markov processes and their strongly nonergodic mixtures, of any order k, but, over a finite alphabet, we have

$$\underset{n\to\infty}{\text{hilb}}\ \mathbf{E}\ \text{card}\, U(X_1^n) = \underset{n\to\infty}{\text{hilb}}\ \mathbf{E}\, \mathbb{I}(X_1^n; X_{n+1}^{2n}) = 0. \tag{57}$$

This happens to be so since the sufficient statistic of text X_1^n for predicting text X_{n+1}^{2n} is the maximum likelihood estimate of the transition matrix, the elements of which can assume at most $(n+1)$ distinct values. Hence, $\mathbf{E}\,\mathbb{I}(X_1^n;X_{n+1}^{2n}) \le D^{k+1}\log(n+1)$, where D is the cardinality of the alphabet and k is the Markov order of the process. Similarly, it can be shown for these processes that the PPM order satisfies $\lim_{n\to\infty} G_{\text{PPM}}(X_1^n) \le k$. Hence, the number of PPM words, which satisfies inequality $\operatorname{card} V_{\text{PPM}}(X_1^n) \le D^{G_{\text{PPM}}(X_1^n)}$, is also bounded above. In consequence, for Markov processes and their strongly nonergodic mixtures, of any order but over a finite alphabet, we obtain

$$\operatorname*{hilb}_{n\to\infty}\left[G_{\text{PPM}}(X_1^n)+\operatorname{card} V_{\text{PPM}}(X_1^n)\right]=0 \text{ almost surely.} \tag{58}$$

In contrast, Santa Fe processes are strongly nonergodic mixtures of some IID processes over an infinite alphabet. Being mixtures of IID processes over an infinite alphabet, they need not satisfy condition (58). In fact, as shown in [25,29] and Appendix C, for the Santa Fe process with exponent α, we have the asymptotic power-law growth

$$\operatorname*{hilb}_{n\to\infty}\mathbf{E}\operatorname{card} U(X_1^n)=\operatorname*{hilb}_{n\to\infty}\mathbf{E}\,\mathbb{I}(X_1^n;X_{n+1}^{2n})=1/\alpha\in(0,1). \tag{59}$$

The same equality for the number of inferrable probabilistic facts and the mutual information is also satisfied by a stationary coding of the Santa Fe process into a finite alphabet (see [29]).

Let us also note that, whereas the theorem about facts and words provides an inequality of Hilberg exponents, this inequality can be strict. To provide some substance, in [29], we have constructed a modification of the Santa Fe process that is ergodic and over a finite alphabet. For this modification, we have only the power-law growth of mutual information

$$\operatorname*{hilb}_{n\to\infty}\mathbf{E}\,\mathbb{I}(X_1^n;X_{n+1}^{2n})=1/\alpha\in(0,1). \tag{60}$$

Since, in this case, $\operatorname{hilb}_{n\to\infty}\mathbf{E}\operatorname{card} U(X_1^n)=0$, then the difference between the Hilberg exponents for the number of inferrable probabilistic facts and the number of PPM words can be an arbitrary number in range $(0,1)$.

Now, we are in a position to discuss some empirical data. In this case, we cannot directly measure the number of facts and the mutual information, but we can compute the PPM order and count the number of PPM words. In Figure 1, we have presented data for a collection of 35 plays by William Shakespeare (downloaded from the Project Gutenberg, https://www.gutenberg.org/) and a random permutation of characters appearing in this collection of texts. The random permutation of characters is an IID process over a finite alphabet, so, in this case, we obtain

$$\operatorname*{hilb}_{n\to\infty}\operatorname{card} V_{\text{PPM}}(x_1^n)=0. \tag{61}$$

In contrast, for the plays of Shakespeare, we seem to have a stepwise power law growth of the number of distinct PPM words. Thus, we may suppose that, for natural language, we have more generally

$$\operatorname*{hilb}_{n\to\infty}\operatorname{card} V_{\text{PPM}}(x_1^n)>0. \tag{62}$$

If relationship (62) holds true, then natural language cannot be a Markov process of any order. Moreover, in view of the striking difference between observations (61) and (62), we may suppose that the number of inferrable probabilistic or algorithmic facts for texts in natural language also obeys a power-law growth. Formally speaking, this condition would translate to natural language being strongly nonergodic or perigraphic. We note that this hypothesis arises only as a form of a weak inductive inference since formally we cannot deduce condition (33) from mere condition (62), regardless of the amount of data supporting condition (62).

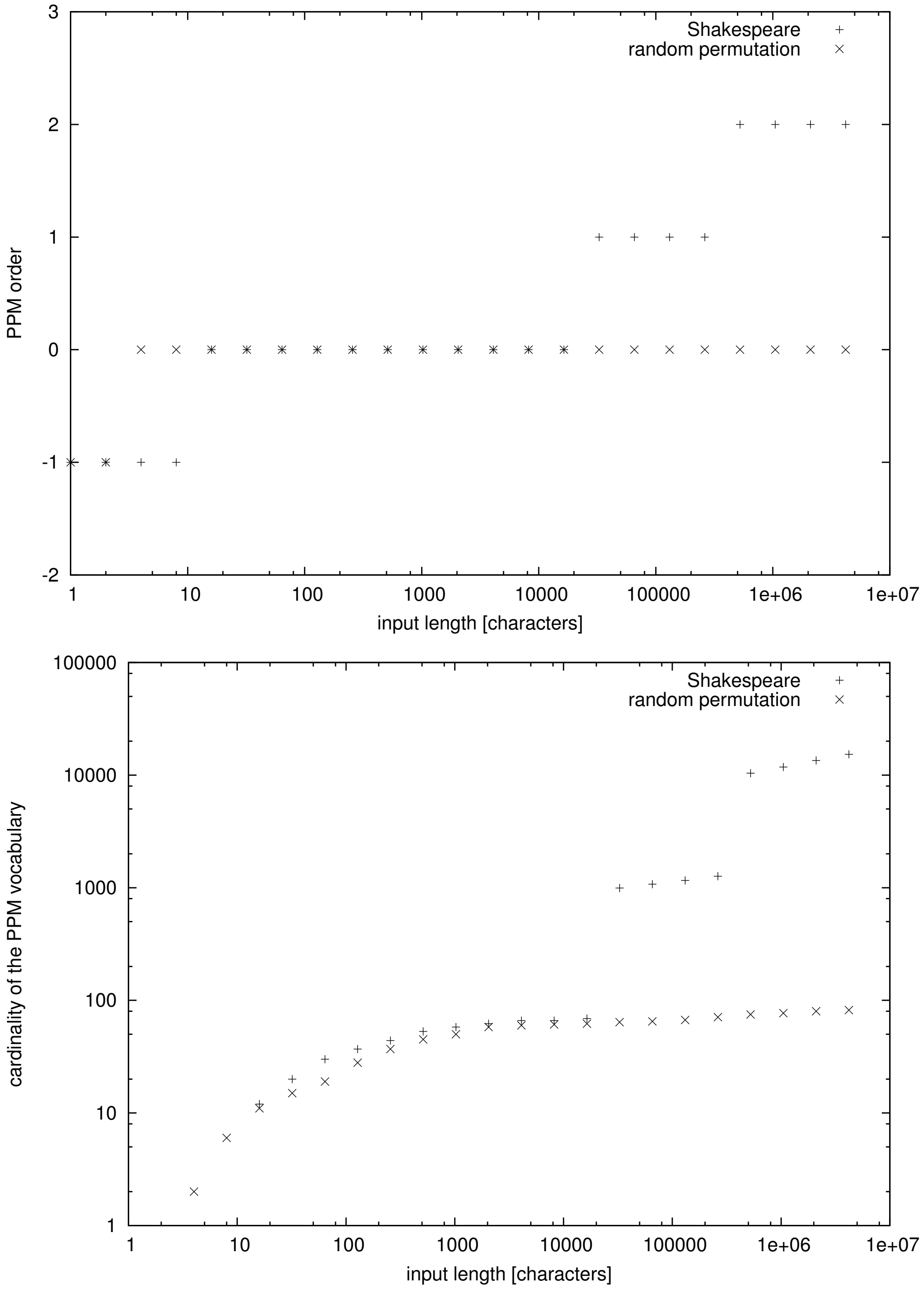

Figure 1. The PPM order $G_{\mathrm{PPM}}(x_1^n)$ and the cardinality of the PPM vocabulary card $V_{\mathrm{PPM}}(x_1^n)$ versus the input length n for William Shakespeare's First Folio/35 Plays and a random permutation of the text's characters.

7. Conclusions

In this article, a stationary process has been called strongly nonergodic if some persistent random topic can be detected in the process and an infinite number of independent binary random variables, called probabilistic facts, is needed to describe this topic completely. Replacing probabilistic facts with an algorithmically random sequence of bits, called algorithmic facts, we have adapted this property back to ergodic processes. Subsequently, we have called a process perigraphic if the number of algorithmic facts which can be inferred from a finite text sampled from the process grows like a power of the text length.

We have demonstrated an assertion, which we call the theorem about facts and words. This proposition states that the number of independent probabilistic or algorithmic facts which can be inferred from a text drawn from a process must be roughly smaller than the number of distinct word-like strings detected in this text by means of the PPM compression algorithm. We have exhibited two versions of this theorem: one for strongly nonergodic processes, applying the Shannon information theory, and one for ergodic processes, applying the algorithmic information theory.

Subsequently, we have exhibited an empirical observation that the number of distinct word-like strings grows like a stepwise power law for a collection of plays by William Shakespeare, in stark contrast to Markov processes. This observation does not rule out that the number of probabilistic or algorithmic facts inferrable from texts in natural language also grows like a power law. Hence, we have supposed that natural language is a perigraphic process.

We suppose that the path of the future related research should lead through a further analysis of the theorem about facts and words, and demonstrating an almost sure version of this statement. It is also an important, still unresolved question whether theoretical analysis of effective universal coding algorithms and their rates of convergence to the entropy rate can contribute to some definite statements about natural language treated as a stochastic process. We realize that the results of this paper as far as the linguistic theory is concerned may be still too inconclusive. As we see it, the main merit of this paper lies in linking some concepts in the Shannon information theory and the algorithmic information theory and providing some linguistic interpretations of them.

Acknowledgments: We wish to thank Jacek Koronacki, Jan Mielniczuk, Vladimir Vovk, and Boris Ryabko for helpful comments.

Conflicts of Interest: The author declares no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

IID	independent identically distributed
PPM	prediction by partial matching

Appendix A. Facts and Mutual Information

In the appendices, we will make use of several kinds of information measures.

1. First, there are four pointwise Shannon information measures:

   - entropy
     $\mathbb{H}(X) = -\log P(X)$,
   - conditional entropy
     $\mathbb{H}(X|Z) := -\log P(X|Z)$,
   - mutual information
     $\mathbb{I}(X;Y) := \mathbb{H}(X) + \mathbb{H}(Y) - \mathbb{H}(X,Y)$,
   - conditional mutual information
     $\mathbb{I}(X;Y|Z) := \mathbb{H}(X|Z) + \mathbb{H}(Y|Z) - \mathbb{H}(X,Y|Z)$,

   where $P(X)$ is the probability of a random variable X and $P(X|Z)$ is the conditional probability of a random variable X given a random variable Z. The above definitions make sense for discrete-valued random variables X and Y and an arbitrary random variable Z. If Z is a discrete-valued random variable, then also $\mathbb{H}(X,Z) - \mathbb{H}(Z) = \mathbb{H}(X|Z)$ and $\mathbb{I}(X;Z) = \mathbb{H}(X) - \mathbb{H}(X|Z)$.
2. Moreover, we will use four algorithmic information measures:

- entropy
  $\mathbb{H}_a(x) = K(x)\log 2,$
- conditional entropy
  $\mathbb{H}_a(x|z) := K(x|z)\log 2,$
- mutual information
  $\mathbb{I}_a(x;y) := \mathbb{H}_a(x) + \mathbb{H}_a(y) - \mathbb{H}_a(x,y),$
- conditional mutual information
  $\mathbb{I}_a(x;y|z) := \mathbb{H}_a(x|z) + \mathbb{H}_a(y|z) - \mathbb{H}_a(x,y|z),$

where $K(x)$ is the prefix-free Kolmogorov complexity of an object x and $K(x|z)$ is the prefix-free Kolmogorov complexity of an object x given an object z. In the above definitions, x and y must be finite objects (finite texts), whereas z can be also an infinite object (an infinite sequence). If z is a finite object, then $\mathbb{H}_a(x,z) - \mathbb{H}_a(z) \overset{\pm}{=} \mathbb{H}_a(x|z,K(z))$ rather than being equal to $\mathbb{H}_a(x|z)$, where $\overset{\pm}{=}$, $\overset{+}{<}$, and $\overset{+}{>}$ are the equality and the inequalities up to an additive constant ([37], Theorem 3.9.1). Hence,

$$\begin{aligned}\mathbb{H}_a(x) - \mathbb{H}_a(x|z) + \mathbb{H}_a(K(z)) \overset{+}{>} \mathbb{I}_a(x;z) &\overset{\pm}{=} \mathbb{H}_a(x) - \mathbb{H}_a(x|z,K(z)) \\ &\overset{+}{>} \mathbb{H}_a(x) - \mathbb{H}_a(x|z). \end{aligned} \tag{A1}$$

In the following, we will prove a result for Hilberg exponents.

Theorem A1. *For a function $\mathfrak{G} : \mathbb{N} \to \mathbb{R}$, define $\mathfrak{J}(n) := 2\mathfrak{G}(n) - \mathfrak{G}(2n)$. If the limit $\lim_{n\to\infty} \mathfrak{G}(n)/n = \mathfrak{g}$ exists and is finite, then*

$$\operatorname*{hilb}_{n\to\infty} [\mathfrak{G}(n) - n\mathfrak{g}] \le \operatorname*{hilb}_{n\to\infty} \mathfrak{J}(n), \tag{A2}$$

with an equality if $\mathfrak{J}(2^n) \overset{+}{>} 0$ for all but finitely many n.

Proof. The proof makes use of the telescope sum

$$\sum_{k=0}^{\infty} \frac{\mathfrak{J}(2^{k+n})}{2^{k+1}} = \mathfrak{G}(2^n) - 2^n\mathfrak{g}. \tag{A3}$$

Denote $\delta := \operatorname{hilb}_{n\to\infty} \mathfrak{J}(n)$. Since $\operatorname{hilb}_{n\to\infty} (\mathfrak{G}(n) - n\mathfrak{g}) \le 1$, it is sufficient to prove inequality (A2) for $\delta < 1$. In this case, $\mathfrak{J}(2^n) \le 2^{(\delta+\epsilon)n}$ for all but finitely many n for any $\epsilon > 0$. Then, for $\epsilon < 1 - \delta$, by the telescope sum (A3), we obtain for sufficiently large n that

$$\mathfrak{G}(2^n) - 2^n\mathfrak{g} \le \sum_{k=0}^{\infty} \frac{2^{(\delta+\epsilon)(k+n)}}{2^{k+1}} \le 2^{(\delta+\epsilon)n} \sum_{k=0}^{\infty} 2^{(\delta+\epsilon-1)k-1} = \frac{2^{(\delta+\epsilon)n}}{2(1-2^{\delta+\epsilon-1})}. \tag{A4}$$

Since ϵ can be taken arbitrarily small, we obtain (A2).

Now assume that $\mathfrak{J}(2^n) \overset{+}{>} 0$ for all but finitely many n. By the telescope sum (A3), we have $\mathfrak{J}(2^n)/2 \overset{+}{<} \mathfrak{G}(2^n) - 2^n\mathfrak{g}$ for sufficiently large n. Hence,

$$\delta \le \operatorname*{hilb}_{n\to\infty} (\mathfrak{G}(n) - n\mathfrak{g}) \tag{A5}$$

Combining this with (A2), we obtain $\operatorname{hilb}_{n\to\infty} (\mathfrak{G}(n) - n\mathfrak{g}) = \delta$. □

For any stationary process $(X_i)_{i=1}^{\infty}$ over a finite alphabet, there exists a limit

$$h := \lim_{n\to\infty} \frac{\mathbf{E}\,\mathbb{H}(X_1^n)}{n} = \mathbf{E}\,\mathbb{H}(X_1|X_2^{\infty}), \tag{A6}$$

called the entropy rate of process $(X_i)_{i=1}^{\infty}$ [3]. By (28), (43), and (46), we also have

$$h = \lim_{n\to\infty} \frac{\mathbf{E}\,\mathbb{H}_a(X_1^n)}{n}. \tag{A7}$$

Moreover, for a stationary process, the mutual information satisfies

$$\mathbf{E}\,\mathbb{I}(X_1^n; X_{n+1}^{2n}) = 2\mathbf{E}\,\mathbb{H}(X_1^n) - \mathbf{E}\,\mathbb{H}(X_1^{2n}) \geq 0, \tag{A8}$$

$$\mathbf{E}\,\mathbb{I}_a(X_1^n; X_{n+1}^{2n}) = 2\mathbf{E}\,\mathbb{H}_a(X_1^n) - \mathbf{E}\,\mathbb{H}_a(X_1^{2n}) \overset{+}{>} 0. \tag{A9}$$

Hence, by Theorem A1, we obtain

$$\operatorname*{hilb}_{n\to\infty} \left[\mathbf{E}\,\mathbb{H}(X_1^n) - hn\right] = \operatorname*{hilb}_{n\to\infty} \mathbf{E}\,\mathbb{I}(X_1^n; X_{n+1}^{2n}), \tag{A10}$$

$$\operatorname*{hilb}_{n\to\infty} \left[\mathbf{E}\,\mathbb{H}_a(X_1^n) - hn\right] = \operatorname*{hilb}_{n\to\infty} \mathbf{E}\,\mathbb{I}_a(X_1^n; X_{n+1}^{2n}). \tag{A11}$$

Subsequently, we will prove the initial parts of Theorems 9 and 10, i.e., the two versions of the theorem about facts and words. The probabilistic statement for strongly nonergodic processes goes first.

Theorem A2 (facts and mutual information I)**.** *Let $(X_i)_{i=1}^{\infty}$ be a stationary strongly nonergodic process over a finite alphabet. We have inequality*

$$\operatorname*{hilb}_{n\to\infty} \mathbf{E}\operatorname{card} U(X_1^n) \leq \operatorname*{hilb}_{n\to\infty} \mathbf{E}\,\mathbb{I}(X_1^n; X_{n+1}^{2n}). \tag{A12}$$

Proof. Let us write $S_n := \operatorname{card} U(X_1^n)$. Observe that

$$\begin{aligned} \mathbf{E}\,\mathbb{H}(Z_1^{S_n}|S_n) &= -\textstyle\sum_{s,w} P(S_n = s, Z_1^s = w) \log P(Z_1^s = w|S_n = s) \\ &\geq -\textstyle\sum_{s,w} P(S_n = s, Z_1^s = w) \log \frac{P(Z_1^s=w)}{P(S_n=s)} \\ &= -\textstyle\sum_{s,w} P(S_n = s, Z_1^s = w) \log \frac{2^{-s}}{P(S_n=s)} \\ &= (\log 2)\mathbf{E}\, S_n - \mathbf{E}\,\mathbb{H}(S_n), \end{aligned} \tag{A13}$$

$$\begin{aligned} \mathbf{E}\,\mathbb{H}(S_n) &\leq (\mathbf{E}\, S_n + 1)\log(\mathbf{E}\, S_n + 1) - \mathbf{E}\, S_n \log \mathbf{E}\, S_n \\ &= \log(\mathbf{E}\, S_n + 1) + \mathbf{E}\, S_n \log \frac{\mathbf{E}\, S_n + 1}{\mathbf{E}\, S_n} \\ &\leq \log(\mathbf{E}\, S_n + 1) + 1, \end{aligned} \tag{A14}$$

where the second row of inequalities follows by the maximum entropy bound from ([3], Lemma 13.5.4). Hence, by the inequality

$$\mathbf{E}\,\mathbb{H}(X|Y) \leq \mathbf{E}\,\mathbb{H}(X|f(Y)) \tag{A15}$$

for a measurable function f, we obtain that

$$\begin{aligned} \mathbf{E}\,\mathbb{H}(X_1^n) - \mathbf{E}\,\mathbb{H}(X_1^n|Z_1^\infty) &\geq \mathbf{E}\,\mathbb{H}(X_1^n|S_n) - \mathbf{E}\,\mathbb{H}(X_1^n|Z_1^\infty, S_n) - \mathbf{E}\,\mathbb{H}(S_n) \\ &\geq \mathbf{E}\,\mathbb{H}(X_1^n|S_n) - \mathbf{E}\,\mathbb{H}(X_1^n|Z_1^{S_n}, S_n) - \mathbf{E}\,\mathbb{H}(S_n) \\ &= \mathbf{E}\,\mathbb{I}(X_1^n; Z_1^{S_n}|S_n) - \mathbf{E}\,\mathbb{H}(S_n) \\ &\geq \mathbf{E}\,\mathbb{H}(Z_1^{S_n}|S_n) - \mathbf{E}\,\mathbb{H}(Z_1^{S_n}|X_1^n, S_n) - \mathbf{E}\,\mathbb{H}(S_n) \\ &= \mathbf{E}\,\mathbb{H}(Z_1^{S_n}|S_n) - \mathbf{E}\,\mathbb{H}(S_n) \\ &\geq (\log 2)\mathbf{E}\,S_n - 2\mathbf{E}\,\mathbb{H}(S_n) \\ &\geq (\log 2)\mathbf{E}\,S_n - 2\left[\log(\mathbf{E}\,S_n + 1) + 1\right]. \end{aligned} \tag{A16}$$

Now, we observe that

$$\mathbf{E}\,\mathbb{H}(X_1^n|Z_1^\infty) \geq \mathbf{E}\,\mathbb{H}(X_1^n|X_{n+1}^\infty) = hn \tag{A17}$$

since the sequence of random variables Z_1^∞ is a measurable function of the sequence of random variables X_{n+1}^∞, as shown in [24,25]. Hence, we have

$$\mathbf{E}\,\mathbb{H}(X_1^n) - \mathbf{E}\,\mathbb{H}(X_1^n|Z_1^\infty) \leq \mathbf{E}\,\mathbb{H}(X_1^n) - hn. \tag{A18}$$

By inequalities (A17) and (A18) and equality (A10), we obtain inequality (A12). □

The algorithmic version of the theorem about facts and words follows roughly the same idea, with some necessary adjustments.

Theorem A3 (facts and mutual information II). *Let $(X_i)_{i=1}^\infty$ be a stationary process over a finite alphabet. We have inequality*

$$\underset{n\to\infty}{\operatorname{hilb}}\ \mathbf{E}\ \operatorname{card} U_a(X_1^n) \leq \underset{n\to\infty}{\operatorname{hilb}}\ \mathbf{E}\,\mathbb{I}_a(X_1^n; X_{n+1}^{2n}). \tag{A19}$$

Proof. Let us write $S_n := \operatorname{card} U_a(X_1^n)$. Observe that

$$\begin{aligned} \mathbb{H}_a(z_1^{S_n}|S_n) &\overset{+}{>} \mathbb{H}_a(z_1^{S_n}) - \mathbb{H}_a(S_n) \\ &\overset{+}{=} (\log 2)S_n - C - \mathbb{H}_a(S_n), \end{aligned} \tag{A20}$$

$$\mathbb{H}_a(S_n) \overset{+}{<} 2\log(S_n + 1), \tag{A21}$$

$$\begin{aligned} \mathbb{H}_a(K(z_1^{S_n})) &\overset{+}{<} 2\log(K(z_1^{S_n}) + 1) \\ &\overset{+}{<} 2\log(S_n + 1), \end{aligned} \tag{A22}$$

where the first row of inequalities follows by the algorithmic randomness of z_1^∞, whereas the second and the third row of inequalities follow by the bounds $K(n) \overset{+}{<} 2\log_2(n+1)$ for $n \geq 0$ and $K(z_1^k) \overset{+}{<} 2k$. Moreover, for any a computable function f, there exists a constant $C_f \geq 0$ such that

$$\mathbb{H}_a(x|y) \overset{+}{<} \mathbb{H}_a(x|f(y)) + C_f. \tag{A23}$$

Hence, we obtain that

$$\begin{aligned} \mathbb{H}_a(X_1^n)-\mathbb{H}_a(X_1^n|z_1^\infty) &\stackrel{+}{>} \mathbb{H}_a(X_1^n|S_n)-\mathbb{H}_a(X_1^n|z_1^\infty,S_n)-\mathbb{H}_a(S_n)\\ &\stackrel{+}{>} \mathbb{H}_a(X_1^n|S_n)-\mathbb{H}_a(X_1^n|z_1^{S_n},S_n)-\mathbb{H}_a(S_n)\\ &\stackrel{+}{>} \mathbb{I}_a(X_1^n;z_1^{S_n}|S_n)-\mathbb{H}_a(K(z_1^{S_n}))-\mathbb{H}_a(S_n)\\ &\stackrel{+}{>} \mathbb{H}_a(z_1^{S_n}|S_n)-\mathbb{H}_a(z_1^{S_n}|X_1^n,K(X_1^n),S_n)\\ &\quad -\mathbb{H}_a(K(z_1^{S_n}))-\mathbb{H}_a(S_n)\\ &\stackrel{+}{>} \mathbb{H}_a(z_1^{S_n}|S_n)-C_g-\mathbb{H}_a(K(z_1^{S_n}))-\mathbb{H}_a(S_n)\\ &\stackrel{+}{>} (\log 2)S_n-C-C_g-\mathbb{H}_a(K(z_1^{S_n}))-2\mathbb{H}_a(S_n)\\ &\stackrel{+}{>} (\log 2)S_n-6\log(S_n+1)-C-C_g. \end{aligned} \tag{A24}$$

Since $-\mathbf{E}\,\log(S_n+1)\geq-\log(\mathbf{E}\,S_n+1)$ by the Jensen inequality, then

$$\mathbf{E}\,\mathbb{H}_a(X_1^n)-\mathbf{E}\,\mathbb{H}_a(X_1^n|z_1^\infty) \stackrel{+}{>} (\log 2)\mathbf{E}\,S_n-6\log(\mathbf{E}\,S_n+1)-C-C_g. \tag{A25}$$

Now, we observe that

$$\mathbf{E}\,\mathbb{H}_a(X_1^n|z_1^\infty)\geq\mathbf{E}\,\mathbb{H}(X_1^n)\geq hn \tag{A26}$$

since the conditional prefix-free Kolmogorov complexity with the second argument fixed is the length of a prefix-free code. Hence, we have

$$\mathbf{E}\,\mathbb{H}_a(X_1^n)-\mathbf{E}\,\mathbb{H}_a(X_1^n|z_1^\infty)\leq\mathbf{E}\,\mathbb{H}_a(X_1^n)-hn. \tag{A27}$$

By inequalities (A25) and (A27) and equality (A11), we obtain inequality (A19). □

Appendix B. Mutual Information and PPM Words

In this appendix, we will investigate some algebraic properties of the length of the PPM code to be used for proving the second part of the theorem about facts and words. First of all, it can be seen that

$$\mathbb{H}_{\mathrm{PPM}_k}(x_1^n)=\begin{cases} n\log D, & k=-1,\\ k\log D+\displaystyle\sum_{u\in\mathbb{X}^k}\log\frac{(N(u|x_1^{n-1})+D-1)!}{(D-1)!\prod_{a=1}^D N(ua|x_1^n)!}, & k\geq 0. \end{cases} \tag{A28}$$

Expression (A28) can be further rewritten using notation

$$\log^* n:=\begin{cases} 0, & n=0,\\ \log n!-n\log n+n, & n\geq 1, \end{cases} \tag{A29}$$

$$\mathfrak{H}(n_1,\dots,n_l):=\begin{cases} \sum_{i=1:n_i>0}^l n_i\log\left(\frac{\sum_{j=1}^l n_j}{n_i}\right), & \text{if } n_j>0 \text{ exists},\\ 0, & \text{else}, \end{cases} \tag{A30}$$

$$\mathfrak{K}(n_1,\dots,n_l):=\sum_{i=1}^l\log^* n_i-\log^*\left(\sum_{i=1}^l n_i\right). \tag{A31}$$

Then, for $k\geq 0$, we define

$$\mathbb{H}_{\mathrm{PPM}_k^0}(x_1^n):=\sum_{u\in\mathbb{X}^k}\mathfrak{H}\left(N(u1|x_1^n),\dots,N(uD|x_1^n)\right), \tag{A32}$$

$$\begin{aligned}\mathbb{H}_{\mathrm{PPM}^1_k}(x_1^n) &:= \sum_{u\in\mathbb{X}^k}\mathfrak{H}\left(N(u|x_1^{n-1}),D-1\right)\\ &\quad-\sum_{u\in\mathbb{X}^k}\mathfrak{K}\left(N(u1|x_1^n),\dots,N(uD|x_1^n),D-1\right).\end{aligned} \tag{A33}$$

As a result for $k\geq 0$, we obtain

$$\mathbb{H}_{\mathrm{PPM}_k}(x_1^n)=k\log D+\mathbb{H}_{\mathrm{PPM}^0_k}(x_1^n)+\mathbb{H}_{\mathrm{PPM}^1_k}(x_1^n). \tag{A34}$$

In the following, we will analyze the terms on the right-hand side of (A34).

Theorem A4. *For $k\geq 0$ and $n\geq 1$, we have*

$$\tilde{D}\operatorname{card}V(k|x_1^{n-1})\leq\mathbb{H}_{\mathrm{PPM}^1_k}(x_1^n)<D\operatorname{card}V(k|x_1^{n-1})\,(2+\log n)\,, \tag{A35}$$

where $\tilde{D}:=-D\log\left(D^{-1}\right)!>0$.

Proof. Observe that $\mathfrak{H}(0,D-1)=\mathfrak{K}(0,\dots,0,D-1)=0$. Hence, the summation in $\mathbb{H}_{\mathrm{PPM}^1_k}(x_1^n)$ can be restricted to $u\in\mathbb{X}^k$ such that $N(u|x_1^{n-1})\geq 1$. Consider such a u and write $N=N(u|x_1^{n-1})$ and $N_a=N(ua|x_1^n)$.

Since $\mathfrak{H}(n_1,\dots,n_l)\geq 0$ and $\mathfrak{K}(n_1,\dots,n_l)\geq 0$ (the second inequality follows by subadditivity of $\log^* n$), we obtain first

$$\begin{aligned}\mathfrak{H}\left(N,D-1\right)-\mathfrak{K}\left(N_1,\dots,N_D,D-1\right) &\leq\mathfrak{H}\left(N,D-1\right)\\ &=N\log\left(1+\tfrac{D-1}{N}\right)+(D-1)\log\left(1+\tfrac{N}{D-1}\right)\\ &\leq N\cdot\tfrac{D-1}{N}+(D-1)\log\left(1+\tfrac{N}{D-1}\right)\\ &=(D-1)\left[1+\log\left(1+\tfrac{N}{D-1}\right)\right]<D\left(2+\log n\right),\end{aligned} \tag{A36}$$

where we use $\log(1+x)\leq x$ and $N<n$. On the other hand, function $\log^* n$ is concave so by $\sum_{a=1}^D N_a=N$ and the Jensen inequality for $\log^* n$, we obtain

$$\begin{aligned}&\mathfrak{H}\left(N,D-1\right)-\mathfrak{K}\left(N_1,\dots,N_D,D-1\right)\geq\mathfrak{F}\left(N,D\right):\\ &=N\log\left(1+\tfrac{D-1}{N}\right)+(D-1)\log\left(1+\tfrac{N}{D-1}\right)\\ &\quad+\log^*(N+D-1)-\log^*(D-1)-D\log^*\left(N/D\right)\\ &=\log(N+D-1)!-\log(D-1)!-D\log\left(N/D\right)!-N\log D\\ &=\log\tfrac{(N+D-1)!}{(D-1)!(N/D)!^D D^N}\geq 0,\end{aligned} \tag{A37}$$

since

$$\begin{aligned}(N/D)!^D D^N &=N^D(N-D)^D(N-2D)^D\dots D^D\\ &\leq(N+D-1)(N+D-2)\dots D=\tfrac{(N+D-1)!}{(D-1)!}.\end{aligned} \tag{A38}$$

Moreover, function $\mathfrak{F}\left(N,D\right)$ is growing in argument N. Hence,

$$\mathfrak{F}\left(N,D\right)\geq\mathfrak{F}\left(1,D\right)=-D\log\left(D^{-1}\right)!. \tag{A39}$$

Summing inequalities (A36) and (A39) over $u\in\mathbb{X}^k$ such that $N(u|x_1^n)\geq 1$, we obtain the claim. □

The mutual information is defined as a difference of entropies. Replacing the entropy with an arbitrary function $\mathbb{H}_Q(u)$, we obtain this quantity:

Definition A1. *The* Q pointwise mutual information *is defined as*

$$\mathbb{I}_Q(u;v) := \mathbb{H}_Q(u) + \mathbb{H}_Q(v) - \mathbb{H}_Q(uv). \tag{A40}$$

We will show that the PPM^0_k pointwise mutual information cannot be positive.

Theorem A5. *For* $n_i = \sum_{j=1}^{l} n_{ij}$*, where* $n_{ij} \geq 0$*, we have*

$$\mathfrak{H}(n_1, \dots, n_k) \geq \sum_{j=1}^{l} \mathfrak{H}(n_{1j}, \dots, n_{kj}). \tag{A41}$$

Proof. Write $N := \sum_{i=1}^{k} \sum_{j=1}^{l} n_{ij}$, $p_{ij} := n_{ij}/N$, $q_i := \sum_{j=1}^{l} p_{ij}$, and $r_j := \sum_{i=1}^{k} p_{ij}$. We observe that

$$\mathfrak{H}(n_1, \dots, n_k) - \sum_{j=1}^{l} \mathfrak{H}(n_{1j}, \dots, n_{kj}) = N \sum_{i=1}^{k} \sum_{j=1}^{l} p_{ij} \log \frac{p_{ij}}{q_i r_j}, \tag{A42}$$

which is N times the Kullback–Leibler divergence between distributions $\{p_{ij}\}$ and $\{q_i r_j\}$ and thus is nonnegative. □

Theorem A6. *For* $k \geq 0$*, we have*

$$\mathbb{I}_{\text{PPM}^0_k}(x_1^n; x_{n+1}^{n+m}) \leq 0. \tag{A43}$$

Proof. Consider $k \geq 0$. For $u \in \mathbb{X}^k$ and $a \in \mathbb{X}$, we have

$$N(ua|x_1^{n+m}) = N(ua|x_1^n) + N(ua|x_{n-k}^{n+k}) + N(ua|x_{n+1}^{n+m}). \tag{A44}$$

Thus, using Theorem A5, we obtain

$$\begin{aligned} \mathfrak{H}\left(N(u1|x_1^{n+m}), \dots, N(uD|x_1^{n+m})\right) \quad & \geq \mathfrak{H}\left(N(u1|x_1^n), \dots, N(uD|x_1^n)\right) \\ & + \mathfrak{H}\left(N(u1|x_{n-k}^{n+k}), \dots, N(uD|x_{n-k}^{n+k})\right) \\ & + \mathfrak{H}\left(N(u1|x_{n+1}^{n+m}), \dots, N(uD|x_{n+1}^{n+m})\right). \end{aligned} \tag{A45}$$

Since the second term on the right-hand side is greater than or equal zero, we may omit it and summing the remaining terms over all $u \in \mathbb{X}^k$ we obtain the claim. □

Now, we will show that the PPM pointwise mutual information between two parts of a string is roughly bounded above by the cardinality of the PPM vocabulary of the string multiplied by the logarithm of the string length.

Theorem A7. *We have*

$$\begin{aligned} \mathbb{I}_{\text{PPM}}(x_1^n; x_{n+1}^{n+m}) \quad & \leq 1 + 4 \log\left[G_{\text{PPM}}(x_1^{n+m}) + 2\right] + \left[G_{\text{PPM}}(x_1^{n+m}) + 1\right] \log D \\ & + 2D \operatorname{card} V_{\text{PPM}}(x_1^{n+m}) \left[2 + \log(n+m)\right]. \end{aligned} \tag{A46}$$

Proof. Consider $k \geq 0$. By Theorems A4 and A6, we obtain

$$\begin{aligned} \mathbb{I}_{\text{PPM}_k}(x_1^n; x_{n+1}^{n+m}) \quad & = k \log D + \mathbb{I}_{\text{PPM}^0_k}(x_1^n; x_{n+1}^{n+m}) + \mathbb{I}_{\text{PPM}^1_k}(x_1^n; x_{n+1}^{n+m}) \\ & \leq k \log D + D \operatorname{card} V(k|x_1^n) \left[2 + \log n\right] \\ & \quad + D \operatorname{card} V(k|x_{n+1}^{n+m}) \left[2 + \log m\right] \\ & \leq k \log D + 2D \operatorname{card} V(k|x_1^{n+m}) \left[2 + \log(n+m)\right]. \end{aligned} \tag{A47}$$

In contrast, $\mathbb{I}_{\text{PPM}_{-1}}(x_1^n; x_{n+1}^{n+m}) = 0$. Now, let $G = G_{\text{PPM}}(x_1^{n+m})$. Since

$$\mathbb{H}_{\text{PPM}}(x_1^{n+m}) \geq \mathbb{H}_{\text{PPM}_G}(x_1^{n+m}) \tag{A48}$$

and

$$\mathbb{H}_{\text{PPM}}(u) \leq \mathbb{H}_{\text{PPM}_k}(u) + 1/2 + 2\log(k+2) \tag{A49}$$

for any $u \in \mathbb{X}^*$ and $k \geq -1$, we obtain

$$\begin{aligned} \mathbb{I}_{\text{PPM}}(x_1^n; x_{n+1}^{n+m}) &\leq \mathbb{I}_{\text{PPM}_G}(x_1^n; x_{n+1}^{n+m}) + 1 + 4\log(G+2) \\ &\leq 1 + 4\log(G+2) + (G+1)\log D \\ &\quad +2D \operatorname{card} V(G|x_1^{n+m}) \left[2 + \log(n+m)\right]. \end{aligned} \tag{A50}$$

Hence, the claim follows. □

Consequently, we may prove the second part of Theorems 9 and 10, i.e., the theorems about facts and words.

Theorem A8 (mutual information and words)**.** *Let $(X_i)_{i=1}^{\infty}$ be a stationary process over a finite alphabet. We have inequalities*

$$\begin{aligned} \text{hilb}_{n\to\infty} \mathbf{E}\, \mathbb{I}(X_1^n; X_{n+1}^{2n}) &\leq \text{hilb}_{n\to\infty} \mathbf{E}\, \mathbb{I}_a(X_1^n; X_{n+1}^{2n}) \\ &\leq \text{hilb}_{n\to\infty} \mathbf{E} \left[G_{\text{PPM}}(X_1^n) + \operatorname{card} V_{\text{PPM}}(X_1^n)\right]. \end{aligned} \tag{A51}$$

Proof. By Theorem A7, we obtain

$$\underset{n\to\infty}{\text{hilb}}\, \mathbf{E}\, \mathbb{I}_{\text{PPM}}(X_1^n; X_{n+1}^{2n}) \leq \underset{n\to\infty}{\text{hilb}}\, \mathbf{E} \left[G_{\text{PPM}}(X_1^n) + \operatorname{card} V_{\text{PPM}}(X_1^n)\right]. \tag{A52}$$

In contrast, Theorems 6 and A1 and inequalities (28) and (46) yield

$$\begin{aligned} \text{hilb}_{n\to\infty} \left[\mathbf{E}\, \mathbb{H}(X_1^n) - hn\right] &\leq \text{hilb}_{n\to\infty} \left[\mathbf{E}\, \mathbb{H}_a(X_1^n) - hn\right] \\ &\leq \text{hilb}_{n\to\infty} \left[\mathbf{E}\, \mathbb{H}_{\text{PPM}}(X_1^n) - hn\right] \\ &\leq \text{hilb}_{n\to\infty} \mathbf{E}\, \mathbb{I}_{\text{PPM}}(X_1^n; X_{n+1}^{2n}). \end{aligned} \tag{A53}$$

Hence, by equalities (A10) and (A11), we obtain inequality (A51). □

Appendix C. Hilberg Exponents for Santa Fe Processes

We begin with a general observation for Hilberg exponents. In [36], this result was discussed only for the Hilberg exponent of mutual information.

Theorem A9 (cf. [36])**.** *For a sequence of random variables $Y_n \geq 0$, we have*

$$\underset{n\to\infty}{\text{hilb}}\, Y_n \leq \underset{n\to\infty}{\text{hilb}}\, \mathbf{E}\, Y_n \text{ almost surely.} \tag{A54}$$

Proof. Denote $\delta := \text{hilb}_{n\to\infty} \mathbf{E}\, Y_n$. From the Markov inequality, we have

$$\begin{aligned} \sum_{k=1}^{\infty} P\left(\frac{Y_{2^k}}{2^{k(\delta+\epsilon)}} \geq 1\right) &\leq \sum_{k=1}^{\infty} \frac{\mathbf{E}\, Y_{2^k}}{2^{k(\delta+\epsilon)}} \\ &\leq A + \sum_{k=1}^{\infty} \frac{2^{k(\delta+\epsilon/2)}}{2^{k(\delta+\epsilon)}} < \infty, \end{aligned} \tag{A55}$$

where $A < \infty$. Hence, by the Borel–Cantelli lemma, we have $Y_{2^k} < 2^{k(\delta+\epsilon)}$ for all but finitely many n almost surely. Since we can choose ϵ arbitrarily small, in particular, we obtain inequality (A54). □

In [29,36], it was shown that the Santa Fe process with exponent α satisfies equalities

$$\underset{n\to\infty}{\text{hilb}}\, \mathbb{I}(X_{-n+1}^{0};X_1^n) = 1/\alpha \text{ almost surely,} \tag{A56}$$

$$\underset{n\to\infty}{\text{hilb}}\, \mathbf{E}\, \mathbb{I}(X_{-n+1}^{0};X_1^n) = 1/\alpha. \tag{A57}$$

We will now show a similar result for the number of probabilistic facts inferrable from the Santa Fe process almost surely and in expectation. Since Santa Fe processes are processes over an infinite alphabet, we cannot apply the theorem about facts and words.

Theorem A10. *For the Santa Fe process with exponent α, we have*

$$\underset{n\to\infty}{\text{hilb}}\, \text{card}\, U(X_1^n) = 1/\alpha \textit{ almost surely,} \tag{A58}$$

$$\underset{n\to\infty}{\text{hilb}}\, \mathbf{E}\, \text{card}\, U(X_1^n) = 1/\alpha. \tag{A59}$$

Proof. First, we obtain

$$\begin{aligned} P(\text{card}\, U(X_1^n) \le m_n) &\le \textstyle\sum_{k=1}^{m_n} P(g(k;X_1^n) \neq Z_k) = \sum_{k=1}^{m_n} \left[1 - P(K_i = k)\right]^n \\ &\le m_n \left[1 - \frac{m_n^{-\alpha}}{\zeta(\alpha)}\right]^n \le m_n \exp\left(-n m_n^{-\alpha}/\zeta(\alpha)\right), \end{aligned} \tag{A60}$$

where $\zeta(\alpha) := \sum_{k=1}^{\infty} k^{-\alpha}$ is the zeta function. Put now $m_n = n^{1/\alpha-\epsilon}$ for an $\epsilon > 0$. It is easy to observe that $\sum_{n=1}^{\infty} P(\text{card}\, U(X_1^n) \le m_n) < \infty$. Hence, by the Borel–Cantelli lemma, we have inequality $\text{card}\, U(X_1^n) > m_n$ for all but finitely many n almost surely.

Second, we obtain

$$\begin{aligned} P(\text{card}\, U(X_1^n) \ge M_n) &\le \tfrac{n!}{(n-M_n)!} \textstyle\prod_{k=1}^{M_n} P(K_i = k) \\ &= \tfrac{n!}{(n-M_n)!(M_n!)^{\alpha}[\zeta(\alpha)]^{M_n}}. \end{aligned} \tag{A61}$$

Recalling from Appendix B that $\log n! = n(\log n - 1) + \log^* n$, where $\log^* n \le \log(n+2)$ is subadditive, we obtain

$$\begin{aligned} \log P(\text{card}\, U(X_1^n) \ge M_n) \le{}& n(\log n - 1) - (n - M_n)\left[\log(n - M_n) - 1\right] \\ &- \alpha M_n(\log M_n - 1) + \log^* M_n - M_n \log \zeta(\alpha) \\ \le{}& M_n\left[\log n - \alpha(\log M_n - 1) - \log \zeta(\alpha)\right] + \log^* M_n \end{aligned} \tag{A62}$$

by $\log n \le \log(n - M_n) + \frac{M_n}{n}$. Put now $M_n = Cn^{1/\alpha}$ for a $C > e[\zeta(\alpha)]^{-1/\alpha}$. We obtain

$$P(\text{card}\, U(X_1^n) \ge M_n) \le (Cn^{1/\alpha} + 2)\exp(-\delta n^{1/\alpha}), \tag{A63}$$

where $\delta > 0$ so $\sum_{n=1}^{\infty} P(\text{card}\, U(X_1^n) \ge M_n) < \infty$. Hence, by the Borel–Cantelli lemma, we have inequality $\text{card}\, U(X_1^n) < M_n$ for all but finitely many n almost surely. Combining this result with the previous result yields equality (A58).

To obtain equality (A59), we invoke Theorem A9 for the lower bound, whereas, for the upper bound, we observe that

$$\mathbf{E}\, \text{card}\, U(X_1^n) \le M_n + nP(\text{card}\, U(X_1^n) \ge M_n), \tag{A64}$$

where the last term decays according to the stretched exponential bound (A63) for $M_n = Cn^{1/\alpha}$. □

References

1. Shannon, C. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *30*, 379–423, 623–656.
2. Shannon, C. Prediction and entropy of printed English. *Bell Syst. Tech. J.* **1951**, *30*, 50–64.
3. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; Wiley: Hoboken, NJ, USA, 2006.
4. Jelinek, F. *Statistical Methods for Speech Recognition*; The MIT Press: Cambridge, MA, USA, 1997.
5. Manning, C.D.; Schütze, H. *Foundations of Statistical Natural Language Processing*; The MIT Press: Cambridge, MA, USA, 1999.
6. Ryabko, B. Twice-universal coding. *Probl. Inf. Transm.* **1984**, *20*, 173–177.
7. Cleary, J.G.; Witten, I.H. Data compression using adaptive coding and partial string matching. *IEEE Trans. Commun.* **1984**, *32*, 396–402.
8. Zipf, G.K. *The Psycho-Biology of Language: An Introduction to Dynamic Philology*, 2nd ed.; The MIT Press: Cambridge, MA, USA, 1965.
9. Mandelbrot, B. Structure formelle des textes et communication. *Word* **1954**, *10*, 1–27.
10. Kuraszkiewicz, W.; Łukaszewicz, J. The number of different words as a function of text length. *Pamięt. Lit.* **1951**, *42*, 168–182. (In Polish)
11. Guiraud, P. *Les Caractères Statistiques du Vocabulaire*; Presses Universitaires de France: Paris, France, 1954.
12. Herdan, G. *Quantitative Linguistics*; Butterworths: London, UK, 1964.
13. Heaps, H.S. *Information Retrieval—Computational and Theoretical Aspects*; Academic Press: Cambridge, MA, USA, 1978.
14. Hilberg, W. Der bekannte Grenzwert der redundanzfreien Information in Texten—Eine Fehlinterpretation der Shannonschen Experimente? *Frequenz* **1990**, *44*, 243–248.
15. Ebeling, W.; Nicolis, G. Entropy of Symbolic Sequences: The Role of Correlations. *Europhys. Lett.* **1991**, *14*, 191–196.
16. Ebeling, W.; Pöschel, T. Entropy and long-range correlations in literary English. *Europhys. Lett.* **1994**, *26*, 241–246.
17. Bialek, W.; Nemenman, I.; Tishby, N. Complexity through nonextensivity. *Physica A* **2001**, *302*, 89–99.
18. Crutchfield, J.P.; Feldman, D.P. Regularities unseen, randomness observed: The entropy convergence hierarchy. *Chaos* **2003**, *15*, 25–54.
19. Dębowski, Ł. On Hilberg's law and its links with Guiraud's law. *J. Quant. Linguist.* **2006**, *13*, 81–109.
20. Wolff, J.G. Language acquisition and the discovery of phrase structure. *Lang. Speech* **1980**, *23*, 255–269.
21. De Marcken, C.G. Unsupervised Language Acquisition. Ph.D. Thesis, Massachussetts Institute of Technology, Cambridge, MA, USA, 1996.
22. Kit, C.; Wilks, Y. Unsupervised Learning of Word Boundary with Description Length Gain. In *Proceedings of the Computational Natural Language Learning ACL Workshop, Bergen*; Osborne, M., Sang, E.T.K., Eds.; The Association for Computational Linguistics: Stroudsburg, PA, USA, 1999; pp. 1–6.
23. Kieffer, J.C.; Yang, E. Grammar-based codes: A new class of universal lossless source codes. *IEEE Trans. Inf. Theory* **2000**, *46*, 737–754.
24. Dębowski, Ł. A general definition of conditional information and its application to ergodic decomposition. *Statist. Probab. Lett.* **2009**, *79*, 1260–1268.
25. Dębowski, Ł. On the Vocabulary of Grammar-Based Codes and the Logical Consistency of Texts. *IEEE Trans. Inf. Theory* **2011**, *57*, 4589–4599.
26. Charikar, M.; Lehman, E.; Lehman, A.; Liu, D.; Panigrahy, R.; Prabhakaran, M.; Sahai, A.; Shelat, A. The Smallest Grammar Problem. *IEEE Trans. Inf. Theory* **2005**, *51*, 2554–2576.
27. Dębowski, Ł. Excess entropy in natural language: present state and perspectives. *Chaos* **2011**, *21*, 037105.
28. Dębowski, Ł. The Relaxed Hilberg Conjecture: A Review and New Experimental Support. *J. Quantit. Linguist.* **2015**, *22*, 311–337.
29. Dębowski, Ł. Mixing, Ergodic, and Nonergodic Processes with Rapidly Growing Information between Blocks. *IEEE Trans. Inf. Theory* **2012**, *58*, 3392–3401.
30. Billingsley, P. *Probability and Measure*; Wiley: Hoboken, NJ, USA, 1979.
31. Gray, R.M. *Probability, Random Processes, and Ergodic Properties*; Springer: Berlin/Heidelberg, Germany, 2009.
32. Breiman, L. *Probability*; SIAM: Philadephia, PA, USA, 1992.
33. Kallenberg, O. *Foundations of Modern Probability*; Springer: Berlin/Heidelberg, Germany, 1997.

34. Yaglom, A.M.; Yaglom, I.M. Probability and Information. In *Theory and Decision Library*; Springer: Berlin/Heidelberg, Germany, 1983.
35. Gray, R.M.; Davisson, L.D. The ergodic decomposition of stationary discrete random processses. *IEEE Trans. Inf. Theory* **1974**, *20*, 625–636.
36. Dębowski, Ł. Hilberg Exponents: New Measures of Long Memory in the Process. *IEEE Trans. Inf. Theory* **2015**, *61*, 5716–5726.
37. Li, M.; Vitányi, P.M.B. *An Introduction to Kolmogorov Complexity and Its Applications*, 3rd ed.; Springer: Berlin/Heidelberg, Germany, 2008.
38. Barron, A.R. Logically Smooth Density Estimation. Ph.D. Thesis, Stanford University, Stanford, CA, USA, 1985.
39. Ryabko, B. Applications of Universal Source Coding to Statistical Analysis of Time Series. In *Selected Topics in Information and Coding Theory*; Series on Coding and Cryptology; Woungang, I., Misra, S., Misra, S.C., Eds.; World Scientific Publishing: Singapore, 2010.
40. De Luca, A. On the combinatorics of finite words. *Theor. Comput. Sci.* **1999**, *218*, 13–39.
41. Dębowski, Ł. Maximal Repetitions in Written Texts: Finite Energy Hypothesis vs. Strong Hilberg Conjecture. *Entropy* **2015**, *17*, 5903–5919.