On the Vocabulary of Grammar-Based Codes and the Logical Consistency of Texts

Łukasz Dębowski

Abstract—**The article presents a new interpretation for Zipf's law in natural language which relies on two areas of information theory. We reformulate the problem of grammar-based compression and investigate properties of strongly nonergodic stationary processes. The motivation for the joint discussion is to prove a proposition with a simple informal statement: If an n-letter long text describes n^β independent facts in a random but consistent way then the text contains at least $n^\beta/\log n$ different words.**

In the formal statement, two specific postulates are adopted. Firstly, the words are understood as the nonterminal symbols of the shortest grammar-based encoding of the text. Secondly, the texts are assumed to be emitted by a nonergodic source, with the described facts being binary IID variables that are asymptotically predictable in a shift-invariant way.

The proof of the formal proposition applies several new tools. These are: a construction of universal grammar-based codes for which the differences of code lengths can be bounded easily, ergodic decomposition theorems for mutual information between the past and future of a stationary process, and a lemma that bounds differences of a sublinear function.

The linguistic relevance of presented modeling assumptions, theorems, definitions, and examples is discussed in parallel. While searching for concrete processes to which our proposition can be applied, we introduce several instances of strongly nonergodic processes. In particular, we define the subclass of accessible description processes, which formalizes the notion of texts that describe facts in a self-contained way.

Index Terms—**Zipf's law, universal source coding, grammar-based codes, smallest grammar problem, ergodic decomposition, excess entropy, nonergodic processes, language models, sublinear functions, asymptotically mean stationary processes, variable-length coding**

CONTENTS

The research reported in this work was supported in part by the Polish Ministry of Scientific Research and Information Technology, 1/P03A/045/28, the Australian Research Council, DP0210999, and the IST Programme of the European Community, under the PASCAL II Network of Excellence, IST-2002-506778. This publication reflects only the author's views. Some parts of the material in this paper were published in the conference communication [1] and in the technical report [2]. The remaining part of [2] was submitted for publication elsewhere.

Ł. Dębowski is with the Centrum Wiskunde & Informatica, Kruislaan 413, NL-1098 SJ Amsterdam, the Netherlands (e-mail: debowski@cwi.nl), on leave from the Institute of Computer Science, Polish Academy of Sciences, ul. Ordona 21, 01-237 Warszawa, Poland.

I. THE PROBLEM STATEMENT

> "If a Martian scientist sitting before his radio in Mars accidentally received from Earth the broadcast of an extensive speech [...], what criteria would he have to determine whether the reception represented the effect of animate process on Earth, or merely the latest thunderstorm on Earth? It seems that the only criteria would be the arrangement of occurrences of the elements, and the only clue to the animate origin would be this: the arrangement of the occurrences would be neither of rigidly fixed regularity such as frequently found in wave emissions of purely physical origin nor yet a completely random scattering of the same."
>
> G. K. Zipf [3, page 187]

The aim of this paper is to present a new explanation for the empirical distribution of words in natural language. To achieve this goal, we shall redress the problem of grammar-based compression [4], [5] and we will research information-theoretic properties of a subclass of strongly nonergodic stationary processes. Thus both information theorists and linguists may find this paper interesting.

From the empirical point of view, the distribution of words is quite well described by the celebrated Zipf-Mandelbrot law [3], [6], which states that the word frequency in a text is an inverse power of the word rank. Some effort in probability theory has been devoted to inferring this law for various idealized settings. The most famous one is the monkey-typing model. In this model, the consecutive characters of the text are modeled as IID variables assuming values of both letters and spaces and the Zipf-Mandelbrot law is obeyed for words meant as strings of letters delimited by spaces [6], [7], [8]. There were also some derivations of the Zipf-Mandelbrot law as a result of multiplicative processes [9], [10] or code length games [11].

The probabilistic explanations that have been found out so far may be considered unsatisfactory from the linguistic point of view. The main source of dissatisfaction is the intuition that fairly nothing is purely random or regular in human language, cf. also [12]. The explanation proposed in this paper, based on previous partial insights [13], [2], [14], [1], addresses some of such concerns. To the best of our knowledge, two modelling challenges will be taken into account for the first time:

(i) Words in the linguistic sense are some nonarbitrary constituents of texts since they can be delimited in the text even when the spaces are absent.
(ii) Texts in the linguistic sense refer to many facts unknown a priori to the reader but they usually do this in a consistent and repetitive way.

Rather than the original Zipf-Mandelbrot law, we shall consider its integrated version, usually called Herdan's or Heaps' law in the English literature. This law says that the number of distinct words observed in a text is proportional to a power of the text length [15], [16], [17], [18]. The claim can be inferred from the original law assuming certain regularity of text growth [19], [20].

Thus the interest of this paper will be focused on proving a proposition which can be simply expressed in the following very informal way, assuming thereafter $\beta \in (0, 1)$:

(I) If an n-letter long text describes n^β independent facts in a consistent way then the text contains at least $n^\beta / \log n$ different words.

Thesis (I) resonates with the ideas of semantic information developed by Bar-Hillel and Carnap [21] but it will be successfully formalized and proved here using concepts of the modern Shannon information theory, including newly derived results. To argue that texts in natural language can describe so many independent facts, we will present simple stochastic processes with an appealing linguistic interpretation.

So as to reduce thesis (I) to a provable statement, we will assume several specific modeling postulates, the plausibility of which is discussed below:

The definition of words in texts: Firstly, the set of words contained in a text will be understood as the set of letter strings that are repeated within the text *sufficiently many times*. Rough empirical correspondence between such letter chunks and words in the linguistic sense has been observed for texts in some natural languages [22], [23], [24], [25]. Developing the ideas of the cited authors, the letter chunks will be understood specifically as the distinct nonterminal symbols in the *shortest grammar-based encoding* of the text. On the other hand, simple string repeats can be shown to abound in the outputs of memoryless sources, cf. [26], [14].

Grammar-based codes [4], [5] are uniquely decodable codes which compress strings by transforming them first into special context-free grammars and then encoding the grammars as less redundant strings. An example of such a grammar is

$$\left\{ \begin{array}{l} A_1 \mapsto A_2 A_2 A_4 A_5 \textbf{dear\_children} A_5 A_3 \textbf{all.} \\ A_2 \mapsto A_3 \textbf{you} A_5 \\ A_3 \mapsto A_4 \textbf{\_to\_} \\ A_4 \mapsto \textbf{Good\_morning} \\ A_5 \mapsto \textbf{,\_} \end{array} \right\}. \quad (1)$$

If we start the derivation with symbol A_1 and follow the rewriting rules, we obtain a predecessor of the song *Happy Birthday to You*, the latter debatedly copyrighted.

In the compressions of longer texts, nonterminals A_i often correspond to words in the linguistic sense, especially if it is additionally required that the nonterminals were defined as strings of only terminal symbols [25]. Thus the number of distinct nonterminal symbols in a grammar-based compression, which equals 5 for example (1), will be henceforth called its vocabulary size. A lower bound for the vocabulary size of some specific grammar will be given in terms of the number of independent facts described by the compressed text. The suitable grammar minimizes certain natural grammar length function, which has not been considered in the information-theoretic literature [4], [5] but is close to the one used in the computational linguistic experiments [23], [25].[1]

The definition of facts described by texts: In the second turn, we have to make precise the notion of a collection of texts that describe random facts in a consistent way. Let $(X_i)_{i\in\mathbb{Z}}$ be a stochastic process on a probability space $(\Omega, \mathcal{J}, P)$, where variables $X_i : \Omega \to \mathbb{X}$ assume values from a fixed countable set $\mathbb{X}$ (called the alphabet). Notation $X_{m:n} := (X_i)_{m\le k\le n}$ will be used for strings of the variables (also called blocks).

Definition 1: A stochastic process $(X_i)_{i\in\mathbb{Z}}$ is called an *uncountable description process (UDP)* if there exists a binary process $(Z_k)_{k\in\mathbb{N}} \sim$ IID (i.e., independent identically distributed variables) with $P(Z_k = 0) = P(Z_k = 1) = 1/2$

[1] Notwithstanding the adopted definition of the vocabulary size, we are aware that the proportionality between the number of distinct nonterminals in the smallest grammar and the number of different words in the linguistic sense can be valid only approximately. Let us notice that hapaxes, i.e., words that appear just once in the text, cannot be recognized as nonterminals by a good grammar-based compressor. By Zipf's law, roughly every second distinct word is a hapax in middle-sized texts. Thus the number of distinct nonterminals and that of distinct words can be proportional for middle-sized texts.

Nonetheless, the situation gets more complicated for very short or very long texts, where the proportion of hapaxes varies [27], [28]. With the text length, this proportion decreases and there appear many repeatedly used multiword expressions, recognized as convenient nonterminals by a good compressor [23]. Moreover, the vocabulary growth with the text size depends sharply on whether the text was written by a single author or is a multiauthor collection. The proportion of hapaxes can be observed to ultimately decrease exponentially in the first case [28], whereas it seems to stay away constantly from zero in the second one [29], [28].

and there exist functions $(f_{nk})_{n,k\in\mathbb{N}} : \mathbb{X}^n \to \{0,1\}$ such that

$$\lim_{n\to\infty} P(f_{nk}(X_{t+1:t+n}) = Z_k) = 1 \qquad (2)$$

for all $t \in \mathbb{Z}$ [2].

The name "UDP" is motivated by the idea that the process $(X_i)_{i\in\mathbb{Z}}$ attempts to describe the value of variable $Y = \sum_{k\in\mathbb{N}} 2^{-k} Z_k$, which has an uncountable range. In a linguistic interpretation, variables Z_i are some facts described by texts, whereas $X_i : \Omega \to \mathbb{X}$ are the consecutive characters (letters or spaces) of an unknown text or rather of an infinite collection of unknown texts written in natural language. Sufficiently long strings $X_{m:n}$ form individual (finite) texts. In this case the alphabet $\mathbb{X}$ is finite. Such a model was considered in [30]. Another plausible interpretation is to imagine that X_i are consecutive words or sentences in the text collection. In that case, we can imagine that $\mathbb{X}$ is infinite and certain strings $X_{m:n}$ correspond to contiguous texts, as well.

Regardless of the nature of the individual X_i, the number of facts described by the text $X_{1:n}$ will be identified with the number of Z_i's that may be predicted with probability at least δ given $X_{1:n}$. That is, this number will be understood as the cardinality of set

$$U_\delta(n) := \{k \in \mathbb{N} : P(f_{nk}(X_{1:n}) = Z_k) \geq \delta\}, \qquad (3)$$

where $\delta > \frac{1}{2}$.

To illustrate how the abstract concept of a UDP matches some preconceptions about human language communication, let us consider the following example. Let the alphabet be $\mathbb{X} = \mathbb{N} \times \{0,1\}$ and let the process $(X_i)_{i\in\mathbb{Z}}$ have the form

$$X_i := (K_i, Z_{K_i}), \qquad (4)$$

where $(Z_k)_{k\in\mathbb{N}}$ and $(K_i)_{i\in\mathbb{Z}}$ are probabilistically independent whereas $(K_i)_{i\in\mathbb{Z}}$ is such an ergodic stationary process that $P(K_i = k) > 0$ for every natural number $k \in \mathbb{N}$. For such assumptions it will be demonstrated that variables (4) form a UDP. In particular, the cardinality of set $U_\delta(n)$ is of order n^β if we assume $(K_i)_{i\in\mathbb{Z}} \sim$ IID with $P(K_i = k) \propto k^{-1/\beta}$.

Variables $X_i := (K_i, Z_{K_i})$ can be given some formal semantic interpretation. Imagine that $(X_i)_{i\in\mathbb{Z}}$ is a sequence of consecutive statements extracted from a random collection of texts which describe some random state of affairs $(Z_k)_{k\in\mathbb{N}}$ *consistently*. Each statement of form $X_i = (k, z)$ asserts that the value of a random k-th bit of the state of affairs is z, i.e., it affirms that $Z_k = z$ in such way that both the bit address k and its value z can be identified. Logical consistency of the description is reflected in the following property: If two statements $X_i = (k, z)$ and $X_j = (k', z')$ happen to describe bits of the same address ($k = k'$) then they always assert the same bit value ($z = z'$).

Although the concept of a UDP is much more general than example (4), it formalizes an optimistic notion of human communication. For an infinite collection of texts $(X_i)_{i\in\mathbb{Z}}$ there is an infinite collection of independent elementary facts $(Z_k)_{k\in\mathbb{N}}$ which are unknown to the text reader but being referred to in the texts. There is a fixed method of interpreting finite texts to learn these facts, namely functions f_{nk} that represent human language competence. They allow readers to determine any fact Z_k with a growing certainty the more texts they read, regardless of their starting point.

One can also think that the shift-invariance of successful prediction (2) reflects the intuition that human language competence does not change over generations of readers but it may depend on the amount or the kind of read texts. Exposed to exactly the same collection of texts from their birth, two ideal readers would understand them in the same way (i.e., they would infer the same values of Z_k).[2]

Other modeling assumptions: Although example (4) clearly illustrates the linguistic relevance of certain UDP's, the stochastic processes for which proposition (I) will be established rigorously do not have the specific form (4). For technical reasons, the alphabet $\mathbb{X}$ will be assumed finite. Moreover, we shall assume that the probabilistic source which generates the texts is a *stationary finite-energy* process. Finite-energy processes are processes with exponentially dumped conditional block probabilities [34]. Such a condition is satisfied for processes dithered with an IID noise [34]—so it seems reasonable in the context of natural language modeling. Assuming stationarity and the finite alphabet for natural language models has much longer tradition in information theory [30]. It has been also often supposed that the generation of texts should be modeled by a nonergodic process [35, Section 6.4]. In fact, a stationary process is a UDP if and only if it has a nonatomic shift-invariant sub-σ-field [2]. Thus stationary UDP's are nonergodic. Informally, they can be called very strongly nonergodic.

The plain-word statement of thesis (I) conceals its linkage with information theory and its historical origin. For the stationary process $(X_i)_{i\in\mathbb{Z}}$ of discrete variables X_i, let us define the n-symbol *block entropy* $H(n) := H(X_{t+1:t+n}) = -\mathbf{E} \log P(X_{t+1:t+n})$, $\mathbf{E}$ being the expectation operator. Then denote the block mutual information as

$$E(n) := 2H(n) - H(2n) = I(X_{1:n}; X_{n+1:2n}), \qquad (5)$$

called the n-symbol *excess entropy* after [36].

An idea that $E(n)$ is of order n^β, with $\beta \approx 1/2$, was supposed by [37] for printed English text, cf. also [38], [39], [40], [41], [42], [36]. Hence we refer to this proposition as Hilberg's thesis (or Hilberg's law). Hilberg's thesis provided

[2] *(Remark added after the submission)* Such modeling cries, of course, for a concrete semantic interpretation of the elementary facts $(Z_k)_{k\in\mathbb{N}}$. One can think of the halting probability Ω, an incompressible infinite binary sequence which formally represents certain amount of timeless independent truths that mathematicians may pursue [31, Section 4], [32], [33, Section 3.6.2]. From the idealized reader's perspective, the binary expansion of Ω may look like a probabilistically random sequence $(Z_k)_{k\in\mathbb{N}}$. We doubt, however, that the bits of Ω could be guessed at the power-law rate by any mathematical text author, having no access to a supernatural power.

The facts that are repetitively described in the everyday language use seem to be of more accidental nature and easier to collect. It is, however, unfeasible to produce an effective decomposition of the physical and cultural enviroment which human beings do describe repeatedly into a sequence of truly independent bits and to assure that this sequence is infinite, cf. [31]. Since the existing world is unique and probabilities are mostly theoretical concepts, it is advisable to regard the main result of this paper as a shadow of some yet unknown statement in the algorithmic information theory. That hypothetical proposition could deal with individual real texts and different particular worlds (also fictitious or elusive) described in them.

a direct inspiration for our research of proposition (I), since this can be split into two more specific assertions:

(Ia) Consider a stationary uncountable description process $(X_i)_{i\in\mathbb{Z}}$ over a finite alphabet. If the cardinality of the set $U_\delta(n)$, defined in (3), is greater than $c_1 n^\beta$ then $E(n)$ is not less than $c_2 n^\beta$, for some positive c_1 and c_2.

(Ib) Consider a finite-energy stationary process $(X_i)_{i\in\mathbb{Z}}$ over a finite alphabet. If $E(n)$ is greater than $c_2 n^\beta$ then the shortest grammar-based compression of the block $X_{1:n}$ applies at least $c_3 n^\beta / \log n$ distinct nonterminal symbols on average, for some positive c_2 and c_3.

The exact statements are to be understood in an asymptotic sense, explained in the following section. A heuristic proof of proposition (Ib) was sketched in [13]. This paper furnishes the formal proof and develops a discussion of the logically earlier proposition (Ia), supplemented by the construction of some suitable processes.

Although the tools used to demonstrate propositions (Ia) and (Ib) are different, it is reasonable to consider these propositions jointly as a means to formalize and prove thesis (I). The reasons are following. Hilberg's thesis was formulated merely on base of the block entropy estimates for printed English published by Shannon [30]. It can be argued that Shannon's estimates are too crude to infer the asymptotic behavior of excess entropies $E(n)$, cf. [43]. Proposition (Ia) makes Hilberg's thesis more likely, regardless of the estimation difficulties. Conversely, (Ib) adds some empirical aspect to the rationalist statement (Ia). Discussing propositions (Ia) and (Ib) together may also provide more insight into stronger formalizations of thesis (I) than the main theorem of this manuscript.

The remainder of this paper is split into several semi-independent parts. An overview of the composition is given in the next section. Since the manuscript is multidisciplinary, we tried to keep it self-contained.

II. An overview of the tools and results

The central result of this article is Theorem 9 in Section V, which formalizes thesis (I). The exact phrasing of the theorem is not reproduced in advance since it depends on a longer construction of the following Section III, which covers a new class of grammar-based codes. Contrary to typical scheme of presentation, it is easier to firstly give a heuristic sketch of the proof, which will be done right now.

Let $H(n) := H(X_{t+1:t+n}) = -\mathbf{E} \log P(X_{t+1:t+n})$ be the n-symbol block entropy of a stationary process $(X_i)_{i\in\mathbb{Z}}$, where variables $X_i : \Omega \to \mathbb{X}$ assume values from the countable set $\mathbb{X}$. An important parameter of the process is its entropy rate

$$h := \inf_{n\in\mathbb{N}} H(n)/n = \lim_{n\to\infty} H(n)/n \tag{6}$$

Consider also the set of well-predictable facts $U_\delta(n)$, defined in equation (3). Let us define the block "pseudoentropy"

$$H^U(n) := hn + [\log 2 - \eta(\delta)] \cdot \operatorname{card} U_\delta(n), \tag{7}$$

where

$$\eta(p) := -p \log p - (1-p)\log(1-p). \tag{8}$$

is the entropy of binary distribution $(p, 1-p)$ and $\operatorname{card} U_\delta(n)$ denotes the cardinality of set $U_\delta(n)$.

Using some facts about the ergodic decomposition, to be derived in Subsections IV-B and IV-C, we can prove that $H(n) \ge H^U(n)$ and $\lim_n H^U(n)/n = h$ for a finite alphabet $\mathbb{X}$. Thus the excess-bounding Lemma 1 from Appendix I can be applied to function $G(n) = H(n) - H^U(n)$. In particular, we obtain

$$\liminf_{n\to\infty} \frac{\operatorname{card} U_\delta(n)}{n^\beta} > 0 \implies \limsup_{n\to\infty} \frac{E(n)}{n^\beta} > 0 \tag{9}$$

for the n-symbol excess entropy $E(n) = 2H(n) - H(2n)$, as an instance of implication (99).

Implication (9) formalizes proposition (Ia). The premise is true in particular for the uncountable description process (4) with $(K_i)_{i\in\mathbb{Z}} \sim$ IID and the marginal distribution $P(K_i = k) \propto k^{-1/\beta}$ (cf. Subsection VI-B). Although this process is over an infinite alphabet, the right-hand side of (9) holds as well (cf. Subsection VI-C).

In the following, let us consider proposition (Ib). Before focusing on grammar-based codes, we discuss a less specific case. Denote the set of nonempty strings as $\mathbb{X}^+ := \bigcup_{n\in\mathbb{N}} \mathbb{X}^n$ and the set of all strings as $\mathbb{X}^* := \mathbb{X}^+ \cup \{\lambda\}$, where λ is the empty string. Let $C : \mathbb{X}^+ \to \mathbb{Y}^+$ be a *uniquely decodable code* over an input alphabet $\mathbb{X}$ and a finite output alphabet $\mathbb{Y} = \{0, 1, ..., D_Y - 1\}$, i.e., its extension $C^* : (u_1, ..., u_k) \mapsto C(u_1)...C(u_k)$ into finite tuples of strings $u_i \in \mathbb{X}^*$ is an injection. Denote the expected length of code C as

$$H^C(n) := \mathbf{E}\, |C(X_{1:n})| \log D_Y. \tag{10}$$

For a uniquely decodable code, the coding inequality $H^C(n) \ge H(n)$ is satisfied [35] and thus the code will be called *universal* if its limiting compression rate $\lim_n H^C(n)/n$ equals the entropy rate, i.e., $\lim_n H^C(n)/n = h$ for any stationary process. There are no universal codes for an infinite input alphabet $\mathbb{X}$ [44], [45] but they exist for a finite $\mathbb{X}$ [46], [47], [4].

Let us observe that if the code C is universal then there holds an equality of rates

$$\lim_{n\to\infty} H^C(n)/n = \lim_{n\to\infty} H(n)/n = \lim_{n\to\infty} H^U(n)/n \tag{11}$$

and a transitive inequality

$$H^C(u) \ge H(n) \ge H^U(n). \tag{12}$$

Hence, as an instance of relations (98) and (99), implications

$$\liminf_{n\to\infty} \frac{E(n)}{n^\beta} > 0 \implies \limsup_{n\to\infty} \frac{E^C(n)}{n^\beta} > 0, \tag{13}$$

$$\liminf_{n\to\infty} \frac{\operatorname{card} U_\delta(n)}{n^\beta} > 0 \implies \limsup_{n\to\infty} \frac{E^C(n)}{n^\beta} > 0 \tag{14}$$

hold for the expected excess length of the code

$$\begin{aligned} E^C(n) &:= 2H^C(n) - H^C(2n) \\ &= \mathbf{E}\, [|C(X_{1:n})| + |C(X_{n+1:2n})| - |C(X_{1:2n})|] \log D_Y. \end{aligned} \tag{15}$$

The implication converse to (13) is not true, which follows from the negative result of [48], see Appendix III.

Whereas relation (9) rephrases thesis (Ia), implications (13) and (14) correspond in part to theses (Ib) and (I) respectively. The missing part of the correspondence is that (13) and (14) do not contain the specific bound for the vocabulary size of a grammar-based code.

By the second line of formula (15), the suitable completion of (13) and (14) can be given by a lower bound for the vocabulary size of the shortest grammar-based code C in terms of the code's excess length

$$|C(u)| + |C(v)| - |C(uv)| \,, \tag{16}$$

provided code C is universal. If one considers the code length $|C(u)|$ as an analogue of the algorithmic complexity of string u, cf. [5], then (16) is the analogue of algorithmic mutual information [49]. The technical details of bounding the vocabulary size in terms of the excess code length (16) are easily motivated by the following heuristic reasoning.

Grammar-based codes compress strings by transforming them first into special grammars, called admissible grammars [4], and then encoding the grammars back into strings according to a fixed simple schema. An *admissible* grammar is a context-free grammar which generates some singleton language $\{w\}$, $w \in \mathbb{X}^+$, and whose production rules do not have empty right-hand sides [4]. In such a grammar, there is one rule per nonterminal symbol and the nonterminals can be ordered so that the symbols are rewritten onto strings of strictly succeeding symbols [4]. Hence, an admissible grammar is given by its set of production rules

$$G = \left\{ \begin{array}{l} A_1 \to \alpha_1, \\ A_2 \to \alpha_2, \\ ..., \\ A_n \to \alpha_n \end{array} \right\}, \tag{17}$$

where A_1 is the start symbol, other A_i are secondary nonterminals, and the right-hand sides of rules satisfy $\alpha_i \in (\{A_{i+1}, A_{i+2}, ..., A_n\} \cup \mathbb{X})^+$. The *vocabulary size* of G, i.e., the number of used nonterminal symbols, will be written

$$\mathbf{V}[G] := \operatorname{card}\{A_1, A_2, ..., A_n\} = n.$$

On the other hand, the *Yang-Kieffer length* of grammar G is

$$|G| := \textstyle\sum_i |\alpha_i| \,, \tag{18}$$

where $|\alpha|$ is the length of $\alpha \in (\{A_1, A_2, ..., A_n\} \cup \mathbb{X})^*$ [4].

If a string w contains many repeated substrings then some grammar for w can "factor out" the repetitions and may be used to represent w concisely. The set of admissible grammars will be denoted as $\mathcal{G}$ while $\mathcal{G}(w) \subset \mathcal{G}$ will stand for the subset of admissible grammars which generate the language $\{w\}$, $w \in \mathbb{X}^+$. A function $\Gamma : \mathbb{X}^+ \to \mathcal{G}$ such that $\Gamma(w) \in \mathcal{G}(w)$ for all $w \in \mathbb{X}^+$ is called a *grammar transform* [4].

We may suppose naively that the length of the shortest grammar $|\Gamma(w)|$ for w is a sufficiently good approximation of the length of the shortest universal grammar-based code $|C(w)|$, cf. [5], [50]. Thus, we could obtain an upper bound for the excess code length (16), needed to establish (Ib), from a similar bound for the excess grammar length

$$|\Gamma(u)| + |\Gamma(v)| - |\Gamma(uv)| \,.$$

Indeed, there is a simple bound for the latter quantity in terms of the vocabulary size.

Theorem 1: Let Γ be a *minimal grammar transform*, i.e.,

$$|\Gamma(w)| = \min_{G \in \mathcal{G}(w)} |G| \tag{19}$$

and let $\mathbf{L}(w)$ be the maximal length of a (possibly overlapping) repeat in w, i.e.,

$$\mathbf{L}(w) := \max\{|s| : w = x_1 s y_1 = x_2 s y_2 \wedge x_1 \neq x_2\} \,, \tag{20}$$

where $s, x_i, y_i \in \mathbb{X}^*$. For any strings $w = uv, u, v \in \mathbb{X}^*$ we have

$$0 \le |\Gamma(u)| + |\Gamma(v)| - |\Gamma(w)| \le \mathbf{V}[\Gamma(w)]\mathbf{L}(w). \tag{21}$$

Proof: This result was noticed in part in [13, Theorem 3]. A brief justification is as follows. For any string $\alpha \in (\{A_2, A_3, ..., A_n\} \cup \mathbb{X})^*$, denote its *expansion* with respect to (17) as $\langle\alpha\rangle_G$, i.e., $\{\langle\alpha\rangle_G\}$ is the language generated by grammar (17) with $\alpha_1 = \alpha$ [5]. Let a minimal grammar for $w = uv$ be of the form

$$G = \left\{ \begin{array}{l} A_1 \to x_L x_M x_R, \\ A_2 \to \alpha_2, \\ ..., \\ A_n \to \alpha_n \end{array} \right\}.$$

We will split it into two separate grammars for u and v:

$$G_L = \left\{ \begin{array}{l} A_1 \to x_L y_L, \\ A_2 \to \alpha_2, \\ ..., \\ A_n \to \alpha_n \end{array} \right\}, \quad G_R = \left\{ \begin{array}{l} A_1 \to y_R x_R, \\ A_2 \to \alpha_2, \\ ..., \\ A_n \to \alpha_n \end{array} \right\},$$

where the string x_M of length $|x_M| \le 1$ at the boundary of the descriptions for u and w gets expanded into a string of terminal symbols $\langle x_M\rangle_G = y_L y_R \in \mathbb{X}^*$. Since $|\alpha_i| \le \mathbf{L}(w)$ and $|y_L y_R| \le \mathbf{L}(w)$ by minimality of G, we obtain

$$|\Gamma(u)| + |\Gamma(v)| \le |G_L| + |G_R| \le |G| + n \cdot \mathbf{L}(w).$$

Regrouping the terms yields the right inequality in (21). The proof of the left inequality applies grammar joining rather than splitting and can be found in [13]. ∎

Inequality (21) constitutes a nontrivial lower bound of the vocabulary size only if the maximal repeat length $\mathbf{L}(w)$ can be upper-bounded well enough. A logarithmic bound for the latter is the best what we may count on, $\mathbf{L}(w) = \Omega(\log |w|)$, and it actually holds for finite-energy processes almost surely [34] , i.e., $\mathbf{L}(X_{1:n}) = O(\log n)$ a.s., as well as in expectation. These results and the definition of finite-energy processes are detailed for reference in Appendix II.

Although inequalities (9), (13), (14), and (21) combined with Lemma 2 from Appendix II provide a heuristic rationale in favor of theses (Ia), (Ib), and (I), they do not constitute a rigorous proof. The flaw is that the minimal Yang-Kieffer length $|\Gamma(\cdot)|$ is a too crude approximation of a universal code length. For any uniquely decodable code C, we have $\lim_n \max_{w \in \mathbb{X}^n} |C(w)|/n \ge 1$ necessarily. On the other hand, a grammar transform Γ is called *asymptotically compact* if

$$\lim_{n\to\infty} \max_{w \in \mathbb{X}^n} |\Gamma(w)|/n = 0 \tag{22}$$

and for each grammar in $\Gamma(\mathbb{X}^+)$ each nonterminal has a different expansion. In particular, any minimal grammar transform (19) is asymptotically compact [4], [5].

In the following three sections we will reconstruct a rigorous proof of some version of proposition (I):

(i) In Section III, we will build a new class of universal grammar-based codes over a finite alphabet. Rather than applying the standard grammar-to-string encoder of Kieffer and Yang [4], these codes use a novel local encoder inspired by the simplistic code of Neuhoff and Shields [47] (Subsections III-A through III-D). Thus the codes satisfy an analogue of (21) with the excess code length (16) substituted for the excess grammar length.

(ii) Section IV is a study of nonergodic stationary processes. It provides the proofs of equality (11) and inequality (12) for uncountable description processes over a finite alphabet. Some useful preliminary facts to be introduced include elementary algebraic identities satisfied by excess entropy and the ergodic decomposition of this quantity (Subsections IV-A and IV-B, cf. also [14]).

(iii) Section V puts the results together. Thesis (I) is expressed as a formal statement, namely Theorem 9. Several ideas for formulating propositions that would be stronger than Theorem 9 are discussed immediately after its proof.

The issue of this paper, is in fact, is to both formalize and prove thesis (I). From this point of view, it is important to construct examples of stochastic processes which satisfy the assumption of Theorem 9 and to demonstrate that they are relevant for natural language modelling. With a partial success, these questions will be dealt with in Section VI.

The article is briefly concluded in Section VII. Four appendices in the following provide supplementary material. The excess-bounding lemma for sublinear nonnegative functions is exposed in Appendix I. Appendix II presents bounds for the lengths of longest repeats. Two results concerning the difference $E^C(n) - E(n)$ are derived in Appendix III. In Appendix IV, we discuss a peculiar behavior of the vocabulary size for the Yang-Kieffer codes based on irreducible grammar transforms.

III. Grammar-based codes

For the set of admissible grammars $\mathcal{G}$, a *grammar-based code* is a uniquely decodable code of form $C = B(\Gamma(\cdot)) : \mathbb{X}^+ \to \mathbb{Y}^+$, where $\Gamma : \mathbb{X}^+ \to \mathcal{G}$ is a (string-to-)grammar transform and $B : \mathcal{G} \to \mathbb{Y}^+$ is called a *grammar(-to-string) encoder* [4]. In principle, the grammar encoder should be chosen as sufficiently good for many different grammar transforms. To guarantee the existence of universal codes of form $C = B(\Gamma(\cdot))$, we shall assume further in this section that both input and output alphabets are finite, $\mathbb{X} = \{0, 1, ..., D_X - 1\}$ and $\mathbb{Y} = \{0, 1, ..., D_Y - 1\}$ in particular.

Indeed, there exists a grammar encoder $B_{\text{YK}} : \mathcal{G} \to \mathbb{Y}^+$ [4], called Yang-Kieffer encoder, such that

(i) set $B_{\text{YK}}(\mathcal{G})$ is prefix-free,
(ii) $|B_{\text{YK}}(G)| \le |G| \left(A + \log_{D_Y} |G|\right)$ for some $A > 0$,
(iii) $C = B_{\text{YK}}(\Gamma(\cdot))$ is a universal code for any asymptotically compact transform Γ.

Unfortunately, in the case of code $C = B_{\text{YK}}(\Gamma(\cdot))$, it is hard to compare the excess grammar length $|\Gamma(u)| + |\Gamma(v)| - |\Gamma(uv)|$ with the excess code length $|C(u)| + |C(v)| - |C(uv)|$. Thus we will consider another grammar encoder.

Let us notice that notation (17) can be reduced to

$$G = (\alpha_1, \alpha_2, ..., \alpha_n) \tag{23}$$

without any confusion. Subsequently, we will write (23) instead of (17). We will also define a grammar encoder that represents G as a string resembling list (23). This encoder yields universal codes given a simple condition (Theorem 3) and provides nearly a homomorphism between some operations on grammars and strings. Hence the universal codes satisfy an analogue of Theorem 1 as well (Theorem 4).

A. Local grammar encoders

The proof of inequality (21) sketched in Section II applies certain "cut-and-paste" operations on grammars. Besides the operations mentioned there, the following one was used in [13] to prove that the left-hand side of (21) is nonnegative:

Definition 2: $\oplus : \mathcal{G} \times \mathcal{G} \to \mathcal{G}$ is called *grammar joining* if

$$G_1 \in \mathcal{G}(w_1) \wedge G_2 \in \mathcal{G}(w_1) \implies G_1 \oplus G_2 \in \mathcal{G}(w_1 w_2).$$

It would be convenient to use a grammar joining $\oplus$ and an encoder $B : \mathcal{G} \to \mathbb{X}^+$ such that the edit distance between $B(G_1 \oplus G_2)$ and $B(G_1)B(G_2)$ be small. Without making the idea too precise, such joining and encoder will be called *adapted*.

The following example of mutually adapted joining $\oplus$ and encoder B will be used in the consecutive sections. Firstly, let us introduce a useful notation.

Definition 3: For any function $f : \mathbb{U} \to \mathbb{W}^*$, where concatenation on domains $\mathbb{U}^*$ and $\mathbb{W}^*$ is defined, denote its *extension* onto strings as

$$f^* : \mathbb{U}^* \ni x_1 x_2 ... x_m \mapsto f(x_1) f(x_2) ... f(x_m) \in \mathbb{W}^*. \tag{24}$$

Now for $G_i = (\alpha_{i1}, \alpha_{i2}, ..., \alpha_{in_i})$, $i = 1, 2$, define

$$G_1 \oplus G_2 := (A_2 A_{n_1+2}, H_1^*(\alpha_{11}), H_1^*(\alpha_{12}), ..., H_1^*(\alpha_{1n_1}), H_2^*(\alpha_{21}), H_2^*(\alpha_{22}), ..., H_2^*(\alpha_{2n_2})),$$

where $H_1(A_j) := A_{j+1}$ and $H_2(A_j) := A_{j+n_1+1}$ for nonterminals and $H_1(x) := H_2(x) := x$ for terminals $x \in \mathbb{X}$.

In the next construction, the set of natural numbers $\mathbb{N}$ is treated as a generic infinite countable alphabet with concatenation ab, addition $a + b$, and subtraction $a - b$.

Definition 4: $B : \mathcal{G} \to \mathbb{Y}^+$ is a *local grammar encoder* if

$$B(G) = B_{\text{S}}^*(B_{\text{N}}(G)), \tag{25}$$

where:

(i) the function $B_{\text{N}} : \mathcal{G} \to (\{0\} \cup \mathbb{N})^*$ encodes grammars as strings of natural numbers so that the encoding of a grammar $G = (\alpha_1, \alpha_2, ..., \alpha_n)$ is the string

$$B_{\text{N}}(G) := F_1^*(\alpha_1) D_X F_2^*(\alpha_2) D_X ... D_X F_n^*(\alpha_n) (D_X + 1),$$

which employs relative indexing $F_i(A_j) := D_X+1+j-i$ for nonterminals and the identity transformation $F_i(x) := x$ for terminals $x \in \mathbb{X} = \{0, 1, ..., D_X - 1\}$,

(ii) B_{S} is a function of form $B_{\text{S}} : \{0\} \cup \mathbb{N} \to \mathbb{Y}^*$ (for technical purpose of the next subsection, not necessarily an injection)—we will call B_{S} the natural number encoder.

Indeed, local encoders are adapted to the joining operation $\oplus$. For instance, if $B(G_i) = u_i B_{\text{S}}(D_X + 1)$ for some grammars G_i, $i = 1, 2$, then $B(G_1 \oplus G_2) = B_{\text{S}}(D_X + 2) B_{\text{S}}(D_X + 2 + \mathbf{V}[G_1]) B_{\text{S}}(D_X) u_1 B_{\text{S}}(D_X) u_2 B_{\text{S}}(D_X + 1)$.

There exist many prefix-free local encoders. Obviously, the set $B_{\text{N}}(\mathcal{G})$ is prefix-free itself. Therefore, the encoder (25) is prefix-free (and uniquely decodable) if B_{S} is also prefix-free, i.e., if B_{S} is an injection and set $B_{\text{S}}(\{0\} \cup \mathbb{N})$ is prefix-free.

B. Encoder-induced grammar lengths

Let us generalize the definition of the grammar length to include the notion of a universal code length as a special case.

Definition 5: For a grammar encoder $B : \mathcal{G} \to \mathbb{Y}^+$, the function $|B(\cdot)|$ will be called the *B-induced grammar length*. For example, Yang-Kieffer length $|\cdot|$ is B-induced for a local grammar encoder $B = B_{\text{S}}^*(B_{\text{N}}(\cdot))$, where

$$B_{\text{S}}(x) = \begin{cases} \lambda & \text{for } x \in \{D_X, D_X + 1\}, \\ 0 & \text{else.} \end{cases} \tag{26}$$

In the same spirit, we can extend the idea of the smallest grammar with respect to the Yang-Kieffer length, discussed in [5]. A subclass $\mathcal{J} \subset \mathcal{G}$ of admissible grammars will be called *sufficient* if there exists a grammar transform $\Gamma : \mathbb{X}^+ \to \mathcal{J}$, i.e., if $\mathcal{J} \cap \mathcal{G}(w) \neq \emptyset$ for all $w \in \mathbb{X}^+$. On the other hand, a grammar transform Γ will be called a $\mathcal{J}$-grammar transform if $\Gamma(\mathbb{X}^+) \subset \mathcal{J}$.

Definition 6: For an arbitrary grammar length function $\|\cdot\| : \mathcal{G} \to \{0\} \cup \mathbb{N}$, a $\mathcal{J}$-grammar transform Γ will be called *$(\|\cdot\|, \mathcal{J})$-minimal grammar transform* if $\|\Gamma(w)\| \le \|G\|$ for all $G \in \mathcal{G}(w) \cap \mathcal{J}$ and $w \in \mathbb{X}^+$.

Definition 7: The code $B(\Gamma(\cdot))$ will be called *$(B, \mathcal{J})$-minimal* if Γ is $(\|\cdot\|, \mathcal{J})$-minimal for the B-induced grammar length $\|\cdot\|$.

Definition 8: For a grammar length $\|\cdot\|$, the grammar subclasses $\mathcal{J}, \mathcal{K} \subset \mathcal{G}$ are called *$\|\cdot\|$-equivalent* if

$$\min_{G \in \mathcal{G}(w) \cap \mathcal{J}} \|G\| = \min_{G \in \mathcal{G}(w) \cap \mathcal{K}} \|G\| \qquad \text{for all } w \in \mathbb{X}^+.$$

C. Subclasses of grammars

In subsection III-E, we will bound the excess lengths of $(B, \mathcal{J})$-minimal codes, where B are local encoders and $\mathcal{J}$ are some sufficient subclasses. In subsection III-D, we will show that several of these codes are universal. Prior to this, let us introduce several subclasses of grammars $\mathcal{J} \neq \mathcal{G}$ for which: (i) our results hold, (ii) the computation of $(B, \mathcal{J})$-minimal codes may be easier than for $(B, \mathcal{G})$-minimal ones, and (iii) the interpretation of grammars' vocabulary size as the number of distinct words in a linguistic sense seems more plausible.

First, we will say that $(\alpha_1, \alpha_2, ..., \alpha_n)$ is a *flat grammar* if $\alpha_i \in \mathbb{X}^+$ for $i > 1$. The set of flat grammars will be denoted as $\mathcal{F}$. In particular, flat grammars were considered in the computational linguistic experiment by [25]. Next, symbol $\mathcal{D}_k \subset \mathcal{F}$ will denote the class of *k-block interleaved grammars*, i.e., flat grammars $(\alpha_1, \alpha_2, ..., \alpha_n)$, where $\alpha_i \in \mathbb{X}^k$ for $i > 1$. As a further subclass, $\mathcal{B}_k \subset \mathcal{D}_k$ will stand for the set of *k-block grammars*, i.e., k-block interleaved grammars $(uw, \alpha_2, ..., \alpha_n)$, where string $u \in (\{A_2, A_3, ..., A_n\})^*$ contains occurrences of all $A_2, A_3, ..., A_n$ and string $w \in \mathbb{X}^*$ has length $|w| < k$, cf. [47]. Of course, classes $\mathcal{B}_k$, $\mathcal{D}_k$, $\mathcal{B} := \bigcup_{k \ge 1} \mathcal{B}_k$, $\mathcal{D} := \bigcup_{k \ge 1} \mathcal{D}_k$, and $\mathcal{F}$ are sufficient.

On the other hand, grammar $(\alpha_1, \alpha_2, ..., \alpha_n)$ is called *irreducible* if

(i) each string α_i has a different expansion $\langle \alpha_i \rangle_G$ and satisfies $|\alpha_i| > 1$,

(ii) each secondary nonterminal appears in string $\alpha_1 \alpha_2 ... \alpha_n$ at least twice,

(iii) each pair of consecutive symbols in strings $\alpha_1, \alpha_2, ..., \alpha_n$ appears at most once at nonoverlapping positions [4].

The set of irreducible grammars will be denoted as $\mathcal{I}$.

Class $\mathcal{I}$ is important in the theory of grammar-based compression for two reasons. Firstly, any $\mathcal{I}$-grammar transform is asymptotically compact [4] so it yields a universal code when combined with the grammar encoder B_{YK}. Secondly, there is an $\mathcal{I}$-grammar transform which is $(|\cdot|, \mathcal{G})$-minimal.

Theorem 2: The classes $\mathcal{I}$ and $\mathcal{G}$ are $|\cdot|$-equivalent.

Proof: Starting with any grammar $G_1 \in \mathcal{G}(w)$, a grammar $G_2 \in \mathcal{I} \cap \mathcal{G}(w)$ can be constructed by applying a sequence of certain reduction rules until the local minimum of functional $2|\cdot| - \mathbf{V}[\cdot]$ is achieved [4]. In fact, the only reduction applicable to a grammar that minimizes $|\cdot|$ is the introduction of a new nonterminal denoting a pair of symbols which appears exactly twice on the right-hand side of the grammar, cf. section VI in [4]. This reduction conserves the Yang-Kieffer length. ∎

D. Universal codes for local encoders

The local encoders in our sense resemble the encoder B_{NS} considered by Neuhoff and Shields [47] as an encoder for the class of block grammars $\mathcal{B}$. The authors have established that any $(B_{\text{NS}}, \mathcal{B})$-minimal code is universal. The main difference between the encoder B_{NS} and a local encoder is that B_{NS} encodes a nonterminal A_i as a string of length $\lfloor \log_{D_Y} \mathbf{V}[G] \rfloor + 1$ whereas the local encoder uses a string of length $|B_{\text{S}}(D_X + i)|$. This is not a big difference so we can easily prove the following proposition using some results of [47].

Theorem 3: Let B_{S} be such a prefix-free natural number encoder that $|B_{\text{S}}(\cdot)|$ is growing and asymptotically optimal, i.e.,

$$\limsup_{n \to \infty} |B_{\text{S}}(n)| / \log_{D_Y} n = 1. \tag{27}$$

For any sufficient subclass of grammars $\mathcal{J} \supset \mathcal{B}$, every $(B_{\text{S}}^*(B_{\text{N}}(\cdot)), \mathcal{J})$-minimal code C is strongly universal, i.e.,

$$\lim_{n \to \infty} \mathbf{E} \left(\frac{|C(X_{1:n})| \log D_Y}{n} \right) = h \tag{28}$$

and

$$\limsup_{n\to\infty} \frac{|C(X_{1:n})| \log D_Y}{n} \le h \quad \text{a.s.} \tag{29}$$

for every stationary ergodic process $(X_k)_{k\in\mathbb{Z}}$.

Remark: Claims (28) and (29) can be generalized to stationary nonergodic processes as follows. Firstly, the strong ergodic decomposition theorem [51, a statement in the proof of Theorem 9.12] and inequality (29) imply

$$\limsup_{n\to\infty} \frac{|C(X_{1:n})| \log D_Y}{n} \le h_F \quad \text{a.s.} \tag{30}$$

for any stationary process $(X_k)_{k\in\mathbb{Z}}$, where h_F is the entropy rate of the process's random ergodic measure, viz. (51) and (50). Since $0 \le |C(X_{1:n})| \le Kn$ for a $K > 0$, inequality (30) implies equality (28) for any stationary process $(X_k)_{k\in\mathbb{Z}}$ by formula (53) and the inverse Fatou lemma, cf. [52]. Consecutively, the generalized (28) and (30) imply that we have in fact equality

$$\limsup_{n\to\infty} \frac{|C(X_{1:n})| \log D_Y}{n} = h_F \quad \text{a.s.} \tag{31}$$

Proof: Consider a sequence of $\mathcal{B}_k$-grammar transforms Γ_k. For an $\epsilon > 0$ and a stationary ergodic process $(X_k)_{k\in\mathbb{Z}}$ with entropy rate h, let $k(n)$ be the largest integer k satisfying $k2^{k(H+\epsilon)} \le n$. We have

$$\begin{aligned}\limsup_{n\to\infty} \max_{w\in\mathbb{X}^n} \frac{\log_{D_Y} \mathbf{V}[\Gamma_{k(n)}(w)]}{k(n)} &\le h + 2\epsilon, \\ \lim_{n\to\infty} \mathbf{E}\, \mathbf{V}[\Gamma_{k(n)}(X_{1:n})] \cdot k(n)/n &= 0, \\ \lim_{n\to\infty} \mathbf{V}[\Gamma_{k(n)}(X_{1:n})] \cdot k(n)/n &= 0 \quad \text{a.s., cf. [47].}\end{aligned}$$

Since $\lim_n k(n) = \infty$, a $(B, \mathcal{J})$-minimal code is universal if

$$|B(\Gamma_k(w))| \le \alpha k \mathbf{V}[\Gamma_k(w)] + \gamma(k) \frac{n}{k} \log_{D_Y} \mathbf{V}[\Gamma_k(w)],$$

where $\alpha > 0$ and $\lim_k \gamma(k) = 1$. In particular, this inequality holds for (25), (27), and growing $|B_{\mathrm{S}}(\cdot)|$. ■

The prefix-free natural number encoder B_{S} satisfying (27) can be chosen, e.g., as the Elias D_Y-ary representation $\omega : \{0\} \cup \mathbb{N} \to \mathbb{Y}^*$ [53], $|\omega(n)| = \ell(n)$, where

$$\ell(n) := \begin{cases} 1 & \text{if } n < D_Y, \\ \ell(\lfloor \log_{D_Y} n \rfloor) + \lfloor \log_{D_Y} n \rfloor + 1 & \text{if } n \ge D_Y. \end{cases}$$

E. Bounds for the vocabulary size

Let us derive the analogue of Theorem 1 for some minimal grammar-based codes that use the local grammar encoders. Firstly, the code lengths are almost subadditive. Secondly, the excess code lengths are dominated by the vocabulary size multiplied by the length of the longest repeat. The code universality is irrelevant for the proofs.

Definition 9: Consider a grammar $G = (\alpha_1, \alpha_2, ..., \alpha_n) \in \mathcal{G}(w)$. For $0 \le p, q \le |w|$ and $p + q = |w|$, let $u, v \in \mathbb{X}^*$ be the strings such that $p = |u|$, $q = |v|$ and $uv = w$. Then define the *left* and *right croppings* of G as

$$\begin{aligned}\mathbb{L}_p G &:= (x_L y_L, \alpha_2, ..., \alpha_n) \in \mathcal{G}(u), \\ \mathbb{R}_q G &:= (y_R x_R, \alpha_2, ..., \alpha_n) \in \mathcal{G}(v),\end{aligned}$$

where exactly one of the following conditions holds:

(i) $\alpha_1 = x_L x_R$ and $y_L y_R = \lambda$,

(ii) $\alpha_1 = x_L A_i x_R$ for some nonterminal A_i, $2 \le i \le n$, with expansion $\langle A_i \rangle_G = y_L y_R$.

Moreover, define the *flattening*

$$\mathbb{F}G := (\alpha_1, \langle\alpha_2\rangle_G, \langle\alpha_3\rangle_G, ..., \langle\alpha_n\rangle_G)$$

and the *secondary part*

$$\mathbb{S}G := (\lambda, \alpha_2, \alpha_3, ..., \alpha_n).$$

Theorem 4: Let B be the local encoder (25). Introduce constants

$$W_m := \max_{0\le n\le D_X+2+m} |B_{\mathrm{S}}(n)|. \tag{32}$$

Let Γ be a $(\|\cdot\|, \mathcal{J})$-minimal grammar transform for the B-induced grammar length $\|\cdot\|$. Consider the code $C = B(\Gamma(\cdot))$, strings $w, u, v \in \mathbb{X}^+$, and a grammar class $\mathcal{K}$ which is $\|\cdot\|$-equivalent to $\mathcal{J}$.

(i) If $G_1, G_2 \in \mathcal{J} \implies G_1 \oplus G_2 \in \mathcal{K}$ then

$$|C(u)| + |C(v)| - |C(uv)| \ge -3W_0 - W_{\mathbf{V}[\Gamma(u)]}. \tag{33}$$

(ii) If $G \in \mathcal{J} \implies \mathbb{L}_n G, \mathbb{R}_n G \in \mathcal{K}$ for all valid n then

$$|C(u)|, |C(v)| \le |C(uv)| + W_0 \mathbf{L}(uv), \tag{34}$$

$$|C(u)| + |C(v)| - |C(uv)| \le \|\mathbb{S}\Gamma(uv)\| + W_0 \mathbf{L}(uv). \tag{35}$$

(iii) If $G \in \mathcal{J} \implies \mathbb{F}G \in \mathcal{K}$ then

$$\|\mathbb{S}\Gamma(w)\| + W_0 \mathbf{L}(w) \le W_0 \mathbf{V}[\Gamma(w)](1 + \mathbf{L}(w)). \tag{36}$$

Remark: In particular, (33) holds for $\mathcal{J} = \mathcal{G}, \mathcal{I}$ while inequalities (34)–(36) hold for $\mathcal{J} = \mathcal{G}, \mathcal{I}, \mathcal{F}, \mathcal{D}, \mathcal{D}_k$. Moreover, (35) and (36) together imply bound

$$|C(u)| + |C(v)| - |C(uv)| \le W_0 \mathbf{V}[\Gamma(uv)](1 + \mathbf{L}(uv)), \tag{37}$$

which generalizes the inequality (21).

Proof:

(i) The result is implied by $\|\Gamma(uv)\| \le \|\Gamma(u) \oplus \Gamma(v)\|$ and

$$\|G_1 \oplus G_2\| \le \|G_1\| + \|G_2\| + |B_{\mathrm{S}}(D_X + 2 + \mathbf{V}[G_1])| + 3W_0,$$

where $G_1 = \Gamma(u)$ and $G_2 = \Gamma(v)$.

(ii) Set $p = |u|$, $q = |v|$, and $w = uv$. The inequalities follow from

$$\begin{aligned}\|\Gamma(w)\| + W_0 \mathbf{L}(w) &\ge \|\mathbb{L}_p \Gamma(w)\| \ge \|\Gamma(u)\|, \\ \|\Gamma(w)\| + W_0 \mathbf{L}(w) &\ge \|\mathbb{R}_q \Gamma(w)\| \ge \|\Gamma(v)\|,\end{aligned}$$

and

$$\|\mathbb{L}_p \Gamma(w)\| + \|\mathbb{R}_q \Gamma(w)\| \le \|\Gamma(w)\| + \|\mathbb{S}\Gamma(w)\| + W_0 \mathbf{L}(w).$$

(iii) The claim is entailed by $\|\mathbb{S}\Gamma(w)\| \le \|\mathbb{S}\mathbb{F}\Gamma(w)\|$ and $\|\mathbb{S}\mathbb{F}\Gamma(w)\| \le W_0 (\mathbf{V}[\Gamma(w)] - 1)(1 + \mathbf{L}(w)) + W_0$. ■

Although Theorems 1 and 4 are analogous, there is a huge qualitative difference between the codes based on irreducible grammars that apply the Yang-Kieffer encoder B_{YK} and the universal grammar-based codes that minimize the length induced by the local encoder. The vocabulary size of the former codes is lower-bounded also by the square root of the code length. Thus these codes appear to see more structure in IID strings than in data that exhibit a sequential order. In contrast, the newly considered codes discover much less structure in the IID case, cf. Appendix IV and the experiment [14].

IV. Stationary processes

In this section we explore stationary processes rather than codes. The goal is to prove equality (11) and inequality (12) for uncountable description processes (UDP's) over a finite alphabet. The proofs will be given in Subsection IV-C. In the preparatory Subsection IV-A, we shall discuss some elementary algebraic identities satisfied by excess entropy $E = \lim_n E(n)$, the limit of the n-symbol excess entropies $E(n)$. This is followed by an analysis of the ergodic decomposition of E and $E(n)$ in Subsection IV-B, which also provides some necessary material. A by-product of this decomposition is the proposition that the excess code lengths $E^C(n)$ are unbounded on all but countably many ergodic processes for every universal code, proved in Appendix III.

A. The limit of the n-symbol excess entropies

Let the alphabet $\mathbb{X}$ be a generic countable set again. Consider the sequence of the n-symbol excess entropies $E(n) = I(X_{-n+1:0}; X_{1:n})$ defined in (5). Since $E(n)$ cannot decrease with growing n, we may define the limiting value

$$E = \sup_{n\in\mathbb{N}} E(n) = \lim_{n\to\infty} E(n), \tag{38}$$

called simply *excess entropy* [36]. Although less attention was paid in information theory to E than to entropy rate (6), excess entropy satisfies a number of neat identities.

Denote the difference operator as $\Delta f(n) = f(n) - f(n-1)$ and assume $H(X_1) < \infty$. The first two differences of block entropy $H(n) := H(X_{1:n})$, with $H(0) := 0$, are conditional entropy $\Delta H(n) = H(X_n|X_{1:n-1})$ and minus conditional mutual information $\Delta^2 H(n) = -I(X_1; X_n|X_{2:n-1})$. Observe that $H(n), \Delta H(n), -\Delta^2 H(n) \in [0, \infty[$, whereas the entropy rate (6) satisfies equality $h = \lim_n \Delta H(n)$ for any countable alphabet (only the finite alphabet case was considered e.g. in [54, Section 2.9]). Hence, as it was derived in [36], we have

$$\begin{aligned} E &= -\textstyle\sum_{n=2}^{\infty}(n-1)\Delta^2 H(n) \\ &= \lim_{n\to\infty}[H(n) - n\Delta H(n)] \\ &= \lim_{n\to\infty}[H(n) - nh] \end{aligned} \tag{39}$$

since

$$E(n) = -\textstyle\sum_{k=2}^{2n} u_{nk}\Delta^2 H(k), \tag{40}$$

$$H(n) - n\Delta H(n) = -\textstyle\sum_{k=2}^{n}(k-1)\Delta^2 H(k), \tag{41}$$

where

$$u_{nk} = \begin{cases} k-1, & 2 \le k \le n, \\ 2n-k+1, & n+1 \le k \le 2n. \end{cases} \tag{42}$$

In view of (39), excess entropy equals the nonnegative deviation of block entropy from the asymptotic linear growth. Since $I(X_1; X_n|X_{2:n-1}) = 0$ for $n > k$ for a k-th order Markov process, E is finite in this case. Moreover, excess entropy is finite for finite-state sources, a.k.a. hidden Markov processes [55], [36], by the data-processing inequality $I(X_{1:n}; X_{n+1:2n}) \le I(X_{1:n}; Y_n) \le \sup_m H(Y_m) < \infty$, where Y_n is the hidden state at time n.[3] Whereas finite-state sources are state-of-the-art models in many applications, including computational linguistics [58], [24], Hilberg's observation of an empirical power law $E(n) \asymp \sqrt{n}$ indicates that a larger class of models may be worth considering.

B. Ergodic decomposition of excess entropy

In this subsection we will discuss a representation of excess entropy that pertains to the ergodic decomposition of a stationary process. This result will be utilized in the following subsection, which concerns the uncountable description processes specifically. We shall take for granted many facts that were mentioned and derived elsewhere. The measure-theoretic generalization of conditional mutual information [59], [60], [61], [2] is some tool we need to recall in the very beginning.

For probability space $(\Omega, \mathcal{J}, P)$, a *partition* of the algebra $\mathcal{J} \subset 2^\Omega$, being the domain of probability measure $P : \mathcal{J} \to \mathbb{R}$, is a finite set of events $\{B_j\}_{j=1}^J$ such that $B_i \cap B_j = \emptyset$ and $\bigcup_{j=1}^J B_j = \Omega$. Just like for discrete variables, define *mutual information* between partitions $\alpha = \{A_i\}_{i=1}^I$ and $\beta = \{B_j\}_{j=1}^J$ with respect to probability measure P as

$$I_P(\alpha;\beta) := \sum_{i=1}^{I}\sum_{j=1}^{J} P(A_i \cap B_j)\log\frac{P(A_i \cap B_j)}{P(A_i)P(B_j)}, \tag{43}$$

where $0 \log 0/x := 0$.

Let $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ be the *subalgebras* of *algebra* $\mathcal{J}$. That is, $\{\emptyset, \Omega\} \subset \mathcal{A}, \mathcal{B}, \mathcal{C} \subset \mathcal{J}$ as well as $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, and $\mathcal{J}$ are closed against operations $\cap$, $\cup$, and $\setminus$. Moreover let the random variable $P(A||\mathcal{C})$ be the conditional probability of event $A \in \mathcal{J}$ w.r.t. the smallest σ-field containing $\mathcal{C}$ [62, Section 33]. We may extend the concepts of *conditional mutual information*, *mutual information*, *conditional entropy*, and *entropy* respectively as

$$I(\mathcal{A};\mathcal{B}|\mathcal{C}) := \sup_{\alpha\subset\mathcal{A},\beta\subset\mathcal{B}} \mathbf{E}\, I_{P(\cdot||\mathcal{C})}(\alpha;\beta), \tag{44}$$

$$I(\mathcal{A};\mathcal{B}) := I(\mathcal{A};\mathcal{B}|\{\emptyset,\Omega\}), \tag{45}$$

$$H(\mathcal{A}|\mathcal{C}) := I(\mathcal{A};\mathcal{A}|\mathcal{C}), \tag{46}$$

$$H(\mathcal{A}) := I(\mathcal{A};\mathcal{A}|\{\emptyset,\Omega\}), \tag{47}$$

cf. [2], [60], [61], [63, Section 12]. These concepts generalize the definitions for random variables in a natural way. If we consider discrete random variables Y_i and the smallest subalgebras $\mathcal{A}_i \subset \mathcal{J}$ such that all events of the form $(Y_i = y_i)$ belong to $\mathcal{A}_i$ then $I(Y_1; Y_2|Y_3) = I(\mathcal{A}_1; \mathcal{A}_2|\mathcal{A}_3)$, $I(Y_1; Y_2) = I(\mathcal{A}_1; \mathcal{A}_2)$, $H(Y_1|Y_3) = H(\mathcal{A}_1|\mathcal{A}_3)$, and $H(Y_1) = I(\mathcal{A}_1)$. Quantities (44)–(47) also satisfy the additivity relation (chain rules) and enjoy certain continuity [60], [61], [2, Theorem 1].

Consider process $(X_k)_{k\in\mathbb{Z}}$ again, where variables $X_i : (\Omega, \mathcal{J}) \to (\mathbb{X}, \mathcal{X})$ provide a mapping between measurable

[3] By a similar reasoning, excess entropy is finite also for the Gaussian ARMA. Some disguised expressions for the excess entropy of Gaussian processes were evaluated in [56, Section 5.5], [57].

spaces, $\mathbb{X}$ being countable. The *completion* of algebra $\mathcal{A}$ is the smallest algebra $\mathcal{B} \supset \mathcal{A}$ that contains all elements of 2^Ω of outer P-measure 0 and is closed with respect to countable sums. The completions of the two subalgebras that contain events $(X_i = x_i)$ for all $i \le 0$ and for all $i \ge 1$ respectively will be denoted as $\mathcal{G}_{\le 0}, \mathcal{G}_{\ge 1} \subset \mathcal{J}$. Then, by continuity of mutual information [60], [61, Section 2.2],

$$E = I(\mathcal{G}_{\le 0}; \mathcal{G}_{\ge 1}). \tag{48}$$

On the other hand, by the chain rule for conditional mutual information [61, Section 3.6], [2, Theorem 1(vi)], we have

$$\begin{aligned} I(\mathcal{G}_{\le 0}; \mathcal{G}_{\ge 1}) &= I(\mathcal{G}_{\le 0}; \mathcal{F}) + I(\mathcal{G}_{\le 0}; \mathcal{G}_{\ge 1}|\mathcal{F}) \\ &= H(\mathcal{F}) + I(\mathcal{G}_{\le 0}; \mathcal{G}_{\ge 1}|\mathcal{F}) \end{aligned} \tag{49}$$

for any algebra $\mathcal{F} \subset \mathcal{G}_{\le 0} \cap \mathcal{G}_{\ge 1}$.

In fact, we may choose $\mathcal{F}$ to be the preimage of the process's shift-invariant algebra [2, Lemma 3]. To be precise, denote the product measurable space of doubly infinite sequences $(\mathbb{U}, \mathcal{U}) = \times_{k\in\mathbb{Z}} (\mathbb{X}, \mathcal{X})$. For the shift transformation $T : \mathbb{U} \ni (x_k)_{k\in\mathbb{Z}} \mapsto (x_{k+1})_{k\in\mathbb{Z}} \in \mathbb{U}$, where $x_k \in \mathbb{X}$, define the *invariant algebra* $\mathcal{I} := \{A \in \mathcal{U} : TA = A\}$. Now let

$$\mathcal{F} = (X_i)^{-1}_{i\in\mathbb{Z}}(\mathcal{I}) \subset \mathcal{J}.$$

For this algebra, conditional information $I(\mathcal{G}_{\le 0}; \mathcal{G}_{\ge 1}|\mathcal{F})$ can be interpreted in an interesting way. Let $(\mathbb{S}, \mathcal{S})$ be the measurable space of stationary probability measures on $(\mathbb{U}, \mathcal{U})$ (i.e., $\mu \circ T = \mu$ for $\mu \in \mathbb{S}$) and let $(\mathbb{E}, \mathcal{E}) \subset (\mathbb{S}, \mathcal{S})$ be the subspace of *ergodic* measures (i.e., $\mu(A) \in \{0, 1\}$ for $\mu \in \mathbb{E}$ and $A \in \mathcal{I}$). Precisely, $\mathcal{S}$ and $\mathcal{E}$ are defined as the smallest σ-fields containing all cylinder sets $\{\mu \in \mathbb{S} : \mu(A) \le r\}$ and $\{\mu \in \mathbb{E} : \mu(A) \le r\}$, $A \in \mathcal{U}$, $r \in \mathbb{R}$, respectively. Since $\mathcal{U}$ is countably generated, all respective singletons $\{\mu\}$ belong to $\mathcal{S}$ and $\mathcal{E}$. According to the strong ergodic decomposition theorem [51, Theorems 9.10-12] there exists an almost surely unique random variable $F : (\Omega, \mathcal{F}) \to (\mathbb{E}, \mathcal{E})$ such that

$$F(A) = P((X_i)_{i\in\mathbb{Z}} \in A||\mathcal{F}), \tag{50}$$

for all $A \in \mathcal{U}$. The variable F will be called here the random ergodic measure of $(X_i)_{i\in\mathbb{Z}}$. For every stationary process, the distribution $\nu(W) := P(F \in W)$, $W \in \mathcal{E}$, is given uniquely.

In the following it is convenient to consider information measures for the process $(X_k)_{k\in\mathbb{Z}}$ as functions of the process distribution. For an arbitrary distribution $\mu = P((X_k)_{k\in\mathbb{Z}} \in \cdot) \in \mathbb{S}$, we will consider the following parametrization:

$$H_\mu(n) := H(n), \qquad h_\mu := h, \tag{51}$$

$$E_\mu(n) := E(n), \qquad E_\mu := E. \tag{52}$$

Plugging F for μ and using equality $\mathbf{E}\, F(A) = P((X_i)_{i\in\mathbb{Z}} \in A)$, we obtain:

Theorem 5: For a stationary process over an alphabet $\mathbb{X}$,

$$h = \mathbf{E}\, h_F \qquad \text{if } \mathbb{X} \text{ is finite,} \tag{53}$$

$$E = H(\mathcal{F}) + \mathbf{E}\, E_F \qquad \text{if } \mathbb{X} \text{ is countable.} \tag{54}$$

Equality (54) follows from (49), whereas decomposition (53) for the entropy rate was derived by [64] . The exact proof that $\mathbf{E}\, E_F = I(\mathcal{G}_{\le 0}; \mathcal{G}_{\ge 1}|\mathcal{F})$ can be found in [2, Theorem 3].

There are also finitely-dimensional analogues of (53) and (54), which will be applied in the following section. To write them down, define the *triple mutual information* (TMI) as

$$I(X; Y; Z) := I(X; Z) + I(Y; Z) - I(X, Y; Z) \tag{55}$$

in the case of finite mutual information $I(X; Z)$, $I(Y; Z)$, and $I(X, Y; Z)$. If entropies $H(X)$, $H(Y)$, and $H(Z)$ are finite then the value of $I(X; Y; Z)$ does not depend on the argument permutations [54]. Anyway, we cannot use a construction analogous to (44) to extend the TMI to a permutation-independent function of arbitrary fields, since $I(X; Y; Z)$ is not necessarily positive and monotonic.

Theorem 6: For a stationary process over a countable alphabet,

$$H(n) = I(X_{1:n}; \mathcal{F}) + \mathbf{E}\, H_F(n), \tag{56}$$

$$E(n) = I(X_{1:n}; X_{1:2n}; \mathcal{F}) + \mathbf{E}\, E_F(n), \tag{57}$$

where the second formula holds if $H(n) < \infty$. The limiting behavior of the appearing quantities is following:

(i) In general,

$$\lim_{n\to\infty} I(X_{1:n}; \mathcal{F}) = H(\mathcal{F}), \tag{58}$$

$$\lim_{n\to\infty} \mathbf{E}\, E_F(n) = \mathbf{E}\, E_F. \tag{59}$$

(ii) If the alphabet $\mathbb{X}$ is finite,

$$\lim_{n\to\infty} \mathbf{E}\, H_F(n)/n = h \text{ with } \mathbf{E}\, H_F(n) \ge hn, \tag{60}$$

$$\lim_{n\to\infty} I(X_{1:n}; X_{1:2n}; \mathcal{F})/n = \lim_{n\to\infty} I(X_{1:n}; \mathcal{F})/n = 0. \tag{61}$$

Proof: We have $H(n) = H(X_{1:n})$ and $\mathbf{E}\, H_F(n) = H(X_{1:n}|\mathcal{F})$. Hence (56) follows by the chain rule $H(\mathcal{A}) = I(\mathcal{A}; \mathcal{B}) + H(\mathcal{A}|\mathcal{B})$ [61, Section 3.6], [2, Theorem 1(vi)]. In the following, (56) implies (57).

The convergence $\lim_n I(X_{1:n}; \mathcal{F}) = H(\mathcal{F})$ can be established by continuity of (conditional) mutual information [61, Section 2.2], [2, Theorem 1(iv-v)]. Same for $\lim_n \mathbf{E}\, E_F(n) = \mathbf{E}\, E_F$ since $\mathbf{E}\, E_F(n) = E(X_{1:n}|\mathcal{F})$ and $\mathbf{E}\, E_F = I(\mathcal{G}_{\le 0}; \mathcal{G}_{\ge 1}|\mathcal{F})$.

The equality and inequality in (60) can be obtained from (53) and the general property

$$h_F = \inf_{n\in\mathbb{N}} H_F(n)/n = \lim_{n\to\infty} H_F(n)/n$$

via the dominated convergence theorem. Equalities (61) follow from (56), (60), and the definition of the entropy rate (6). ∎

C. Uncountable description processes

According to Theorem 5, excess entropy E is infinite if either $\mathbf{E}\, E_F$ or $H(\mathcal{F})$ is infinite. In fact, there exist ergodic processes with $E = \mathbf{E}\, E_F = \infty$ even for the binary alphabet $\mathbb{X} = \{0, 1\}$ [65]. Necessarily, entropy $H(\mathcal{F})$ equals zero if and only if the process is ergodic [2, Theorem 1(vii) and Lemma 3]. On the other hand, equality $H(\mathcal{F}) = \infty$ holds in particular if the completion of $\mathcal{F}$ contains a nonatomic sub-σ-field [2, Theorem 1(viii)]. The latter case corresponds to the class of uncountable description processes, mentioned in Section I.

Theorem 7: A stationary process $(X_i)_{i\in\mathbb{Z}}$ is a UDP if and only if the completion of $\mathcal{F}$ contains a nonatomic sub-σ-field. Moreover, in the case of a UDP, all events $(Z_k = 0)$ and $(Z_k = 1)$ belong to the completion of $\mathcal{F}$ [2, Theorem 4].

Whereas $E = \infty$ for UDP's in general, the n-symbol excess entropy $E(n)$ can be bounded by the number of bits Z_k that are individually predictable given $X_{1:n}$ with sufficiently high probability.

Theorem 8: For a UDP $(X_i)_{i\in\mathbb{Z}}$ we have

$$I\left(X_{1:n};(Z_k)_{k\in\mathbb{N}}\right) \geq [\log 2 - \eta(\delta)] \cdot \operatorname{card} U_\delta(n), \tag{62}$$

where $\delta \in (1/2, 1)$ and $U_\delta(n)$ is defined in equation (3).

Proof: By continuity of mutual information [61, Section 2.2], [2, Theorem 1(iv-v)],

$$\begin{aligned} I\left(X_{1:n};(Z_k)_{k\in\mathbb{N}}\right) &= \lim_{n\to\infty} I(X_{1:n};Z_{1:k}) \\ &= \sum_{k=1}^{\infty} I(X_{1:n};Z_k|Z_{1:k-1}). \end{aligned} \tag{63}$$

On the other hand,

$$\begin{aligned} I(X_{1:n};Z_k|Z_{1:k-1}) &= H(Z_k|Z_{1:k-1}) - H(Z_k|X_{1:n},Z_{1:k-1}) \\ &\geq \log 2 - H(Z_k|f_{nk}(X_{1:n})) \\ &\geq \log 2 - \eta(P(f_{nk}(X_{1:n}) = Z_k)) \end{aligned}$$

by the Fano inequality $H(Y_1|Y_2) \leq \eta(P(Y_1 = Y_2))$ for a binary variable Y_1 [54, Theorem 2.47]. Restricting the summation in (63) to $k \in U_\delta(n)$ yields the claim. ∎

A corollary of this result is the inequality $H(n) \geq H^U(n)$ satisfied for the pseudoentropy defined in (7) for the alphabet $\mathbb{X}$ being finite. The derivation is as follows. Since all events $(Z_k = 0)$ and $(Z_k = 1)$ belong to the completion of algebra $\mathcal{F}$ by Theorem 7, we have $I(X_{1:n};\mathcal{F}) \geq I\left(X_{1:n};(Z_k)_{k\in\mathbb{N}}\right)$ by the data processing inequality [2, Theorem 1(iii)]. In the following, by Theorems 6 and 8 we obtain

$$\begin{aligned} H(n) &\geq hn + I(X_{1:n};\mathcal{F}) \\ &\geq hn + I\left(X_{1:n};(Z_k)_{k\in\mathbb{N}}\right) \geq H^U(n), \end{aligned} \tag{64}$$

where $\lim_n H^U(n)/n = h$.

V. The Main Result

Intermediate results of the preceding sections allow to compile the main theorem of this article. The proposition bounds the vocabulary size of a minimal grammar-based compression in terms of the number of independent facts predictable from the compressed string, provided the string was sampled from a finite-energy uncountable description process. This theorem can be called an effective formalization of thesis (I).

Theorem 9: Let $B : \mathcal{G} \to \mathbb{Y}^+$ be the local grammar encoder (25) for the output alphabet $\mathbb{Y} = \{0, 1, ..., D_Y - 1\}$ and such a prefix-free natural number encoder B_S that $|B_S(\cdot)|$ is a growing function and

$$\limsup_{n\to\infty} |B_S(n)|/\log_{D_Y} n = 1.$$

Consider also a sufficient subclass of admissible grammars $\mathcal{J}$ that contains all block grammars, i.e. $\mathcal{J} \supset \mathcal{B}$, and satisfies:

(i) $G \in \mathcal{J} \implies \mathbb{F}G \in \mathcal{J}$ and

(ii) $G \in \mathcal{J} \implies \mathbb{L}_n G, \mathbb{R}_n G \in \mathcal{J}$ for all valid n,

where operations $\mathbb{F}$, $\mathbb{L}_n$, and $\mathbb{R}_n$ are given in Definition 9.

On the other hand, let $(X_i)_{i\in\mathbb{Z}}$ be a stationary finite-energy uncountable description process over the input alphabet $\mathbb{X} = \{0, 1, ..., D_X - 1\}$. Assume that inequality

$$\liminf_{n\to\infty} \frac{\operatorname{card} U_\delta(n)}{n^\beta} > 0 \tag{65}$$

holds for the set of predictable facts

$$U_\delta(n) := \left\{k \in \mathbb{N} : P\left(f_{nk}\left(X_{1:n}\right) = Z_k\right) \geq \delta\right\},$$

where $\delta \in (1/2/, 1)$ and $\beta \in (0, 1)$.

Consider the vocabulary size $\mathbf{V}[\Gamma(X_{1:n})]$ of a $(|B(\cdot)|, \mathcal{J})$-minimal grammar transform $\Gamma : \mathbb{X}^+ \to \mathcal{G}$. The accepted hypotheses entail inequality

$$\limsup_{n\to\infty} \mathbf{E} \left(\frac{\mathbf{V}[\Gamma(X_{1:n})]}{n^\beta(\log n)^{-1}}\right)^p > 0, \quad p > 1. \tag{66}$$

Proof: Code $C = B(\Gamma(\cdot))$ is universal by Theorem 3. Hence by Theorems 6, 7, and 8, viz. the derivation (64), we have $H^C(u) \geq H(n) \geq H^U(n)$ and

$$\lim_{n\to\infty} H^C(n)/n = \lim_{n\to\infty} H(n)/n = \lim_{n\to\infty} H^U(n)/n \tag{67}$$

for the expected code length $H^C(n)$ defined in (10) and the pseudoentropy $H^U(n)$ defined in (7). As a result, implication

$$\liminf_{n\to\infty} \frac{\operatorname{card} U_\delta(n)}{n^\beta} > 0 \implies \limsup_{n\to\infty} \frac{E^C(n)}{n^\beta} > 0 \tag{68}$$

holds as an instance of (99).

Consider $p, q > 1$ such that $(p-1)(q-1) = 1$. Define variables

$$U_n := \mathbf{V}[\Gamma(X_{1:2n})] n^{-\beta} \log n, \tag{69}$$

$$T_n := (1 + \mathbf{L}(X_{1:2n}))(\log n)^{-1}. \tag{70}$$

Theorem 4(ii)–(iii) assures that $E^C(n)\, n^{-\beta} \leq W_0 \,\mathbf{E}\, U_n T_n$ for the constant W_0 defined in (32). By Hölder's inequality, we also have $\mathbf{E}\, U_n T_n \leq (\mathbf{E}\, U_n^p)^{1/p} (\mathbf{E}\, T_n^q)^{1/q}$. Since $\mathbf{E}\, T_n^q$ are bounded by inequality (102), we obtain

$$\limsup_{n\to\infty} \frac{E^C(n)}{n^\beta} > 0 \implies \limsup_{n\to\infty} \mathbf{E}\, U_n^p > 0. \tag{71}$$

The conclusion follows from the conjunction of propositions (65), (68), and (71). ∎

We have sought to formulate Theorem 9 with possibly generic assumptions since few fundamental principles have been recognized for probabilistic modeling of natural language [66], [58], [67], [68] and several different grammar-based codes have been considered in the computational linguistic research [22], [23], [24], [25], [14]. In particular, our theorem applies to the codes based on flat grammars ($\mathcal{J} = \mathcal{F}$). These codes yield compressions of texts in natural language which are close to their decompositions into words in the linguistic sense, according to the experiment by [25].

We conjecture that there exist many processes that satisfy the hypothesis of Theorem 9. In the next section, some simple examples of processes will be presented that satisfy several of the four conditions: (65), stationarity, finite energy, and finite alphabet. In particular, we conjecture that all conditions are satisfied for the process which can be obtained from of a process of form (4) by a stationary variable-length coding into the chosen finite alphabet, cf. [69, Example 6]. For technical complications, we have not completed the proof but we will make the construction and its motivation clear in Subsection VI-E.

Before we pass on to the examples of uncountable description processes, let us discuss several ideas for strengthening the proposition of Theorem 9. Striving for realism in language modeling, it is desirable to relax the assumption of stationarity and to demonstrate the strong law version of inequality (66), under only a little modification of the remaining assumptions. Removing the hypothesis of strict stationarity would also ease the construction of processes to which the theorem is applicable.

What conditions suffice for the strong law version of inequality (66)? The respective strong law proposition reads

$$\limsup_{n\to\infty} \frac{\mathbf{V}[\Gamma(X_{1:n})]}{n^\beta (\log n)^{-1}} > 0 \quad \text{a.s.} \tag{72}$$

Let us trace back some plausible conditions for (72).

Consider the code $C = B(\Gamma(\cdot))$ and an arbitrary stationary process $(X_i)_{i\in\mathbb{Z}}$. According to the Remark after Theorem 3, we have

$$\limsup_{n\to\infty} \frac{|C(X_{1:n})| \log D_Y}{n} = h_F \quad \text{a.s.,} \tag{73}$$

where h_F is the entropy of the process's random ergodic measure. On the other hand, the asymptotic equipartition theorem for nonergodic processes [63, Theorem 13.1], [70] asserts that

$$\lim_{n\to\infty} (-\log P(X_{1:n}))/n = h_F \quad \text{a.s.} \tag{74}$$

Recall also Barron's lemma [71, Theorem 3.1], [34, Lemma 1], which states that

$$|C(X_{1:n})| \log D_Y + \log P(X_{1:n}) \ge -2\log n \tag{75}$$

for all but finitely many $n \in \mathbb{N}$ almost surely.

The above three facts imply that function

$$G(n) := |C(X_{1:n})| \log D_Y + \log P(X_{1:n}) + 2\log n$$

almost surely satisfies $\limsup_k G(k)/k = 0$ and $G(n) \ge 0$ for all but finitely many $n \in \mathbb{N}$. Hence

$$\limsup_{n\to\infty} [2G(n) - G(2n)] \ge 0 \tag{76}$$

by the excess-bounding Lemma 1.

Assume now that the process $(X_i)_{i\in\mathbb{Z}}$ is a finite-energy process and satisfies almost surely

$$\liminf_{n\to\infty} \frac{1}{n^\beta} \log \frac{P(X_{1:2n})}{P(X_{1:n})P(X_{n+1:2n})} > 0, \tag{77}$$

$$\lim_{n\to\infty} \frac{1}{n^\beta} \log \frac{P(X_{1:n})}{P(X_{n+1:2n})} = 0, \tag{78}$$

$$\lim_{n\to\infty} \frac{|C(X_{1:n})| - |C(X_{n+1:2n})|}{n^\beta} = 0. \tag{79}$$

Hence (72) follows by (76), Theorem 4(ii)–(iii) and the almost sure claim of Theorem 2.

For which processes are relations (77)–(79) satisfied? It seems plausible that (77) may hold only if the ergodic components of process $(X_i)_{i\in\mathbb{Z}}$ are sufficiently similar to one another, like for process (4). Some potentially relevant subclass of UDP's that includes (4), called accessible description processes, will be introduced in Subsection VI-C. These processes are characterized by variables Z_k being conditionally independent given any finite block $X_{n:m}$, $n \le m$. We suppose that different ergodic components of an accessible description process are also sufficiently similar to one another so that (77) may follow if the process satisfies $\lim_n E(n)/n^\beta > 0$.

Establishing relations (78)–(79) for some large class of processes seems harder. Anyway, let us observe that it suffices to demonstrate $\sum_n P(|f(X_{1:n}) - g((X_i)_{i\in\mathbb{Z}})| \ge \epsilon n^{-\beta}) < \infty$ for some shift-invariant function g and $f(X_{1:k}) = |C(X_{1:n})|$ or $f(X_{1:k}) = P(X_{1:n})$ respectively. By the stationarity of P we may substitute $X_{k_n+1:k_n+n}$ for $X_{1:k}$ in the infinite sum, whence $\lim_n [f(X_{k_n+1:k_n+n}) - g((X_i)_{i\in\mathbb{Z}})]/n^\beta = 0$ follows almost surely for any k_n by the Borel-Cantelli lemma. Consecutively, we obtain (78)–(79).

May the stationarity assumption be relaxed? There is no reason to assume that stationary processes are the best models for texts in natural language. Let us recall that several facts in information theory have been generalized from the domain of stationary processes to the larger class of asymptotically mean stationary (AMS) processes [69], [72]. So is conclusion (72) true for all AMS finite-energy processes that satisfy (77)–(79). Indeed, the asymptotic equipartition (74) was proved in the AMS case explicitly [69, Theorem 8], (73) holds by the Lemma on page 969 of [69], whereas Barron's lemma (75) and Theorem 2 apply also to a nonstationary process $(X_i)_{i\in\mathbb{Z}}$. (The right-hand side of (73) and (74) in the AMS case equals the entropy of the random stationary ergodic measure whose expectation dominates as the distribution of $(X_i)_{i\in\mathbb{Z}}$.)

May $\liminf$ be substituted for $\limsup$ in (66) or (72)? Replacing the upper limit by the lower limit requires developing an alternative to Lemma 1. As discussed in Appendix I, the claim of Lemma 1 cannot be strengthened if its hypothesis is kept intact.

Does Theorem 9 hold for tractably computable codes? Computing Yang-Kieffer minimal grammars is known to be NP-hard [5]. We conjecture that computing $(|B(\cdot)|, \mathcal{G})$-minimal grammars has also the NP complexity. Although we sought to constrain the grammar minimization in Theorem 9 to smaller domains of grammars $\mathcal{J} \subset \mathcal{G}$, it is a question of future research to decide whether any of these $(|B(\cdot)|, \mathcal{J})$-minimal codes is tractably computable in a sufficiently good approximation. This problem is important since any experimental comparisons of the vocabulary sizes can be done only for efficiently computable grammar transforms.

It is natural to ask whether a converse of thesis (I) is true. Suppose that the minimal grammar-based compression of an n-letter long string applies m different nonterminals. Does it mean that the string describes roughly $m \log n$ independent facts in a consistent way? We deem that such a conclusion is not sound even if we accept the most permissive formal notion of what independent facts are. Counterexamples might be sought at several levels.

First of all, it is known that the redundancy $H^C(n) - H(n)$ of any universal code cannot be bounded by a sublinear function against all stationary ergodic processes [48]. Hence the difference $E^C(n) - E(n)$ cannot be universally bounded by n^β for any fixed $\beta \in (0, 1)$, cf. Appendix III.

The most natural reasoning is, however, simpler. Suppose that the vocabulary size grows like (72) only for processes for which inequality (65) is true. The idea of a counterexample to this proposition is obvious. If there is a UDP for which (72) holds then the same holds also for some ergodic component of the process and this component forms the counterexample by virtue of being ergodic.

A possible line of deprecating this counterexample is to reply that almost every ergodic component of a UDP does describe an infinite sequence of facts which is algorithmically random rather than IID in the probabilistic sense. Thus the appropriate formalization of thesis (I) should be rather an analogue of Theorem 9 for individual sequences (i.e., individual texts) within the algorithmic information theory [73], [74], [33], [49]. An analogue of Definition 1 should be provided for individual sequences as well. This program correlates with Kolmogorov's doubt expressed in [73] that a probabilistic model can be identified for texts in natural language.

We conjecture, however, that the converse of the hypothetical analogue of Theorem 9 for individual sequences is not true, either. It can be easily shown that the vocabulary size of Yang-Kieffer codes based on irreducible grammars grows as $\Omega(\sqrt{n/\log n})$ on an algorithmically random input of length n, cf. Appendix IV. Although the minimal grammar-based codes introduced in this paper seem to compress algorithmically random strings of a fixed length much better, they may asymptotically behave in a similar way.

VI. Examples of uncountable description processes

In this section we shall us examine uncountable description processes in more detail. We will show what kind of processes may satisfy the hypothesis of Theorem 9. According to Definition 1, an uncountable description process (UDP) is such a stochastic process $(X_i)_{i\in\mathbb{Z}}$ that there exist independent equidistributed binary variables Z_k, $k = 1, 2, 3, ...$, asymptotically predictable in a shift-invariant way given $(X_i)_{i\in\mathbb{Z}}$. That is,

$$\lim_{n\to\infty} P(f_{nk}(X_{t+1:t+n}) = Z_k) = 1 \tag{80}$$

for all lags t, all indices k, and certain functions f_{nk}.

From a linguistic point of view, the notion of a UDP captures the idea of an infinite collection of texts describing at random a persistent state of some random world. It is not important whether this world is real or imaginary. It only matters that the state of the world is infinitely complex and a priori unpredictable for the text reader but learnable from any sufficiently large amount of texts.

According to Theorem 7, a stationary process is an uncountable description process if and only if its shift-invariant σ-field contains a nonatomic sub-σ-field. In consequence, all stationary UDP's are nonergodic. Definition 1 and example (4) provide an insight into how linguistically motivated nonergodic processes can look, which seems more concrete than popular discussions of nonergodicity in language modeling so far [35], [75]. Nonetheless, nonergodic processes received much attention in information theory at a very fundamental level [76], [77], [78], [79], mostly from the viewpoint of the general ergodic decomposition theorem [80], [64], [81, Theorem I.4.10], [51, Theorem 9.10–12].

Moreover, UDP's are not the first specific subclass of nonergodic processes to be introduced. Conditionally IID processes are a likewise subclass which has been researched since long [82], [83], [84], [85], [86], [87], [88]. A stationary process is called conditionally IID if its random ergodic measure is supported on the set of IID process measures. The celebrated historical predecessor of the ergodic decomposition theorem, known as the de Finetti theorem, states that each exchangeable process is conditionally IID [87].

There exist uncountable description processes which are conditionally IID. Several of our examples will be of such a form. The conditionally IID processes do not seem, however, to provide reasonable models for natural language. Let us define a more plausible subclass of UDP's:

Definition 10: An uncountable description process $(X_i)_{i\in\mathbb{Z}}$ is called an *accessible description process* if the variables Z_k, $k \in \mathbb{N}$, are conditionally independent given any finite block $X_{n:m}$, $n \le m$.

This condition is satisfied for example (4). Using the accessibility condition allows in particular to compute an asymptotic expression for the n-symbol excess entropies $E(n)$ if $(K_i)_{i\in\mathbb{Z}} \sim$ IID and $P(K_i = k) \propto k^{-1/\beta}$.

In a linguistic interpretation, accessible description processes seem to correspond to collections of texts that avoid describing facts in a cryptic or hermetic way. When a fact can be inferred from such a text, the external knowledge of other facts does not improve the inference. The text is self-contained and there can be only little ambiguity of the knowledge conveyed by the text.

Compare it with a cryptic way of writing. Like in alchemical treaties, the writer can adopt some specific expressions that replace the words known to the reader. As a result, the knowledge conveyed in the text is inaccessible to the reader unless the reader knows the random key to the text. If the key exists, of course, viz. the famous cases of the Voynich manuscript or Codex Seraphinianus [89], [90].

Let us state explicitly that the conditional independence is assumed in Definition 10 only for the variables mentioned in Definition 1. For some UDP's, there may be more independent variables Z_k having property (80) than $(Z_k)_{k\in\mathbb{N}}$ but they need not be conditionally independent given blocks $X_{n:m}$. Cryptic and explicit ways of referring to the described world may also coexist in natural language. Each UDP is in a sense asymptotically accessible, i.e., variables Z_k are conditionally independent given the whole process $(X_i)_{i\in\mathbb{Z}}$.

A. A mixture of Bernoulli processes

Presenting an uncountable description process over a finite alphabet is easy if we do not require the power-law growth of the n-symbol excess entropies $E(n)$. We can simply consider a conditionally IID process which is an uncountable mixture of IID processes. Let $\left(X_i^{(p)}\right)_{i\in\mathbb{Z}}$ denote a Bernoulli process with parameter p, $0 \le p \le 1$, namely, a binary IID process having the marginal distribution

$$P\left(X_i^{(p)} = x\right) = \begin{cases} p, & x = 1, \\ 1-p, & x = 0. \end{cases}$$

Now we can construct a process $(X_i)_{i\in\mathbb{Z}}$ such that

$$P(X_{1:n} = x_{1:n}||Y) = P\left(X_{1:n}^{(Y)} = x_{1:n}\right) \tag{81}$$

for a real variable Y supported on $[0,1]$ with continuous distribution $P(Y \le p) = \Psi(p) := \int_0^p f(x)dx$.

By Theorem 7, $(X_i)_{i\in\mathbb{Z}}$ is an uncountable description process since Y equals $\lim_n n^{-1}\sum_{i=1}^n X_i$ almost surely and thus is measurable against the completed shift-invariant σ-field. The suitable Z_k satisfying (80) may be constructed as the consecutive digits of the binary expansion of the distribution function Ψ taken at the value of Y, $\Psi(Y) = \sum_{k=1}^\infty Z_k 2^{-k}$. If Y has the beta distribution with nonnegative rational parameters, $(X_i)_{i\in\mathbb{Z}}$ is the well-known Pólya urn process [84], [85], [88]. On the other hand, we have $Y = \Psi(Y) = \sum_{k=1}^\infty Z_k 2^{-k}$ provided Y is uniformly distributed on $[0,1]$.

Regardless of the distribution of Y, the pace of inferring the hidden values of Z_k does not come close to a power-law. Let us observe that blocks $X_{1:n}$ are conditionally independent from futures $(X_i)_{i>n}$ given the block sums $S_n := \sum_{i=1}^n X_i$. Hence in view of the information inequality for Markov chain $X_{1:n} \to S_n = f(X_{1:n}) \to (X_i)_{i>n} \to (Z_k)_{k\in\mathbb{N}}$, we have

$$\begin{aligned} I(X_{1:n}; (Z_k)_{k\in\mathbb{N}}) &= I(S_n; (Z_k)_{k\in\mathbb{N}}) \le H(S_n), \\ E(n) = I(X_{1:n}; X_{n+1:2n}) &= I(S_n; X_{n+1:2n}) \le H(S_n). \end{aligned}$$

Moreover, $H(S_n) \le \log(n+1)$ since S_n takes only $n+1$ distinct values. The equality arises for the uniformly distributed Y since then

$$P(S_n = s) = \int_0^1 \binom{n}{s} x^s (1-x)^{n-s} dx = \frac{1}{n+1}$$

for $s = 0, 1, ..., n$.

Similar behavior may be observed for conditionally IID processes over any other finite alphabet $\mathbb{X} = (0, 1, ..., D-1)$. The role of S_n is played by the block type, namely, the tuple of random variables $\mathbf{S}_n = (S_n^0, S_n^1, ..., S_n^{D-1})$, where S_n^i is the number of occurrences of value i in the block $X_{1:n}$. By the Markov chain $X_{1:n} \to \mathbf{S}_n \to (X_i)_{i>n}$, we have $E(n) = I(X_{1:n}; X_{n+1:2n}) \le H(\mathbf{S}_n) \le D\log(n+1)$. This bound can be related to the expression for the minimax regret in exponential families [91, Theorem 7.2], [92].

B. Processes over an infinite alphabet

The previous example looked like estimating an unknown real valued parameter p of the Bernoulli process in the ordinary setting of Bayesian statistics. When the alphabet $\mathbb{X}$ is infinite, a different type of stronger dependence can arise in conditionally independent UDP's. Although we can always clump variables Z_k together into a single uniformly distributed real variable $Y = \sum_{k=1}^\infty Z_k 2^{-k}$, it can be sometimes quite unnatural to think that there is any specific number of real parameters to be estimated.

For instance, let the alphabet be $\mathbb{X} = \mathbb{N} \times \{0,1\}$ and let $(X_i)_{i\in\mathbb{Z}}$ take form

$$X_i := (K_i, Z_{K_i}), \tag{82}$$

just as in the initial example (4). This process is conditionally IID if $(K_i)_{i\in\mathbb{Z}} \sim$ IID and $(Z_k)_{k\in\mathbb{N}}$ and $(K_i)_{i\in\mathbb{Z}}$ are independent. Additionally, (82) constitutes a UDP if $P(K_i = k) > 0$ for all $k \in \mathbb{N}$. The guessing functions f_{nk} need not be the best predictors so some suitable instance is

$$f_{nk}(x_{1:n}) = \begin{cases} 0 & \text{if } x_i = (k,0) \text{ for an } i \in \{1,...,n\}, \\ 1 & \text{if } x_i = (k,1) \text{ for an } i \in \{1,...,n\}, \\ 2 & \text{else.} \end{cases} \tag{83}$$

A particularly interesting case arises when variables K_i are zeta distributed, namely,

$$P(K_i = k) = k^{-1/\beta}/\zeta(1/\beta), \tag{84}$$

where $\beta \in (0,1)$ and $\zeta(x) = \sum_{k=1}^\infty k^{-x}$ is the zeta function. In this case, the n-symbol excess entropy $E(n)$ grows as a power-law. To ascertain the latter proposition, let $U_\delta(n)$ be the set of well predictable facts, defined in (3). In view of equality

$$\begin{aligned} P(f_{nk}(X_{1:n}) = Z_k) &= P(K_i = k \text{ for some } i \in \{1,...,n\}) \\ &= 1 - [1 - P(K_i = k)]^n, \end{aligned}$$

we have $k \in U_\delta(n)$ if and only if $P(K_i = k) \ge 1 - (1-\delta)^{1/n}$. This yields

$$U_\delta(n) \supset \left\{k \in \mathbb{N} : P(K_i = k) \ge -n^{-1}\log(1-\delta)\right\}$$

by inequality $1 - x^{1/n} \le -n^{-1}\log x$ for $x > 0$ and hence

$$\operatorname{card} U_\delta(n) \ge \left[\frac{n}{-\zeta(1/\beta)\log(1-\delta)}\right]^\beta.$$

Thus the power-law growth of $E(n)$ follows by (9). A more precise calculation of $E(n)$ is presented in the next subsection.

As indicated in Section I, example (82) may be generalized. Keeping intact the alphabet $\mathbb{X} = \mathbb{N} \times \{0,1\}$, independence $(Z_k)_{k\in\mathbb{N}} \perp\!\!\!\perp (K_i)_{i\in\mathbb{Z}}$, and guessing functions (83), one may admit $(K_i)_{i\in\mathbb{Z}}$ to be any ergodic stationary process assuming values in natural numbers so that $P(K_i = k) > 0$ for all $k \in \mathbb{N}$. It is easy to prove that such a process $(X_i)_{i\in\mathbb{Z}}$ is a UDP. The proof applies the ergodic theorem [81, Theorem I.3.1] and the fact that the almost sure convergence implies convergence in probability [93, page 188].

The logical interpretation of this process was also mentioned in Section I. Namely, $(X_i)_{i\in\mathbb{Z}}$ can be imagined as a sequence of consecutive logical statements extracted from a random collection of texts. Each statement of form $X_i = (k,z)$ explicitly asserts that z is the value of the k-th bit of some abstract state of affairs $(Z_k)_{k\in\mathbb{N}}$. Regardless of the choice of described facts, determined by process $(K_i)_{i\in\mathbb{Z}}$, the statements are always logically consistent. Namely, if two propositions $X_i = (k, z)$ and $X_j = (k', z')$ describe the same bits ($k = k'$) then they always report the same value ($z = z'$).

Notwithstanding this, the strict logical consistency of statements X_i is not needed to form a UDP. For example, let us extend the probability space with an ergodic process $(U_i)_{i\in\mathbb{Z}}$ such that $P(U_i = 1) > 1/2$ and $(U_i)_{i\in\mathbb{Z}} \perp\!\!\!\perp (Z_k)_{k\in\mathbb{N}}, (K_i)_{i\in\mathbb{Z}}$. For the processes $(Z_k)_{k\in\mathbb{N}}$ and $(K_i)_{i\in\mathbb{Z}}$ as previously, we set

$$X_i := (K_i, Z_{K_i} \oplus U_i), \tag{85}$$

where $\oplus$ is addition modulo 2 (XOR). Such variables $X_i = (k, z)$ can be interpreted as statements which describe $(Z_k)_{k\in\mathbb{N}}$ but are not necessarily true. Yet the noise in the statements can be filtered out by using guessing functions

$$f_{nk}(x_{1:n}) = \begin{cases} 0 & \text{if } N_{k0}(x_{1:n}) > N_{k1}(x_{1:n}), \\ 1 & \text{if } N_{k1}(x_{1:n}) > N_{k0}(x_{1:n}), \\ 2 & \text{else}, \end{cases} \tag{86}$$

where $N_{kz}(x_{1:n})$ is the number of $i \in \{1, ..., n\}$ such that $x_i = (k, z)$. Hence it can be shown easily that the process of (86) is also a UDP.

C. Accessible description processes

Both (82) and (85) form accessible description processes. This observation opens an easy path to compute the n-symbol excess entropies. Assume that $(X_i)_{i\in\mathbb{Z}}$ is an accessible description process and the block entropy $H(X_{1:n})$ is finite. By the chain rules,

$$\begin{aligned} E(n) &= I\left(X_{1:n}; X_{n+1:2n}; (Z_k)_{k\in\mathbb{N}}\right) \\ &\quad + I\left(X_{1:n}; X_{n+1:2n} | (Z_k)_{k\in\mathbb{N}}\right). \end{aligned}$$

The conditional mutual information is nonnegative, whereas the triple mutual information satisfies

$$\begin{aligned} & I\left(X_{1:n}; X_{n+1:2n}; (Z_k)_{k\in\mathbb{N}}\right) \\ & = 2I\left(X_{1:n}; (Z_k)_{k\in\mathbb{N}}\right) - I\left(X_{1:2n}; (Z_k)_{k\in\mathbb{N}}\right) \\ & = \textstyle\sum_{k=1}^\infty [2I(X_{1:n}; Z_k) - I(X_{1:2n}; Z_k)] \\ & = \textstyle\sum_{k=1}^\infty I(X_{1:n}; X_{n+1:2n}; Z_k) \end{aligned} \tag{87}$$

for conditionally independent variables Z_k since

$$\begin{aligned} I\left(X_{1:n}; (Z_k)_{k\in\mathbb{N}}\right) &= \sum_{k=1}^\infty I(X_{1:n}; Z_k | Z_{1:k-1}) \\ &= \sum_{k=1}^\infty [H(Z_k|Z_{1:k-1}) - H(Z_k|X_{1:n}, Z_{1:k-1})] \\ &= \sum_{k=1}^\infty [H(Z_k) - H(Z_k|X_{1:n})] = \sum_{k=1}^\infty I(X_{1:n}; Z_k). \end{aligned} \tag{88}$$

Theorem 10: Let $(K_i)_{i\in\mathbb{Z}}$ be IID variables with the distribution (84), $\beta \in (0, 1)$. The n-symbol excess entropies of the process $(X_i)_{i\in\mathbb{Z}}$ given by (82) obey the asymptotic law

$$\lim_{n\to\infty} \frac{E(n)}{n^\beta} = \frac{(2 - 2^\beta)\Gamma(1 - \beta)\log 2}{[\zeta(1/\beta)]^\beta}. \tag{89}$$

Proof: Since $I\left(X_{1:n}; X_{n+1:2n} | (Z_k)_{k\in\mathbb{N}}\right)$ vanishes, we may write $E(n) = \sum_{k=1}^\infty I(X_{1:n}; X_{n+1:2n}; Z_k)$ by formula (87). Noticing that

$$\begin{aligned} H(Z_k|X_{1:n}) &= (\log 2) \cdot P(K_i \neq k \text{ for all } i \in \{1, ..., n\}) \\ &\quad + 0 \cdot P(K_i = k \text{ for some } i \in \{1, ..., n\}), \\ I(X_{1:n}; Z_k) &= (\log 2) \cdot P(K_i = k \text{ for some } i \in \{1, ..., n\}) \\ &= (1 - [1 - P(K_i = k)]^n) \log 2, \end{aligned}$$

we obtain the triple mutual information

$$I(X_{1:n}; X_{n+1:2n}; Z_k) = (1 - [1 - P(K_i = k)]^n)^2 \log 2.$$

Hence $E(n)$ equals up to a small constant to the integral

$$\begin{aligned} &(\log 2) \int_1^\infty \left(1 - \left(1 - \frac{A}{k^{1/\beta}}\right)^n\right)^2 dk \\ &\quad = \beta[An]^\beta (\log 2) \int_{(1-A)^n}^1 (1 - u)^2 f_n(u) du, \end{aligned}$$

where $A := 1/\zeta(1/\beta)$ and

$$f_n(u) := u^{1/n-1}[n(1 - u^{1/n})]^{-(\beta+1)}.$$

By the de l'Hôpital rule,

$$\lim_{n\to\infty} f_n(u) = f(u) := u^{-1}(-\log u)^{-(\beta+1)}.$$

The product $(1 - u)^2 f(u)$ has a pole only at 0, where it integrates like $(-\log u)^{-(\beta+1)} d(-\log u)$. Hence, $(1-u)^2 f(u)$ is integrable and

$$\lim_{n\to\infty} \frac{E(n)}{n^\beta} = \frac{\beta \log 2}{[\zeta(1/\beta)]^\beta} \int_0^1 \frac{(1 - u)^2\, du}{u\,(-\log u)^{\beta+1}}$$

by the dominated convergence theorem. The further substitution $t = -\log u$ yields

$$\begin{aligned} \int_0^1 \frac{(1 - u)^2\, du}{u\,(-\log u)^{\beta+1}} &= \int_0^\infty (1 - e^{-t})^2\, t^{-\beta-1} dt \\ &= (2 - 2^\beta)\beta^{-1}\Gamma(1 - \beta), \end{aligned}$$

where integrals

$$\begin{aligned} &\textstyle\int_0^\infty (e^{-kt} - 1)t^{-\beta-1} dt = (e^{-kt} - 1)(-\beta^{-1})t^{-\beta}|_0^\infty \\ &\quad - \textstyle\int_0^\infty (-ke^{-kt})(-\beta^{-1})t^{-\beta} dt = -k^\beta \beta^{-1}\Gamma(1 - \beta) \end{aligned}$$

can be safely integrated by parts for the considered β. ■

D. Satisfying the finite-energy condition

The condition of finite energy (see Definition 13 in Appendix II), which appears in the hypothesis of the Theorem 9, can be satisfied easily for the case of infinite alphabet. Processes (82) and (85) are finite-energy processes if so is the fact selection process $(K_i)_{i\in\mathbb{Z}}$.

For example, consider (82). We have

$$\begin{aligned} & P\left((X_i = (k_i, z_i))_{i=1}^L\right) \\ & = P\left((K_i = k_i)_{i=1}^L\right) P\left((Z_{k_i} = z_i)_{i=1}^L \middle| (K_i = k_i)_{i=1}^L\right) \\ & = P\left((K_i = k_i)_{i=1}^L\right) P\left((Z_{k_i} = z_i)_{i=1}^L\right). \end{aligned}$$

Thus

$$\begin{aligned}&P\left((X_i=(k_i,z_i))_{i=n+1}^{n+m}\,|\,(X_i=(k_i,z_i))_{i=1}^{n}\right)\\&=P\left((K_i=k_i)_{i=n+1}^{n+m}\,|\,(K_i=k_i)_{i=1}^{n}\right)\\&\quad\cdot P\left((Z_{k_i}=z_i)_{i=n+1}^{n+m}\,|\,(Z_{k_i}=z_i)_{i=1}^{n}\right)\\&\le P\left((K_i=k_i)_{i=n+1}^{n+m}\,|\,(K_i=k_i)_{i=1}^{n}\right).\end{aligned}$$

Hence the property of finite energy for $(X_i)_{i\in\mathbb{Z}}$ follows from that for $(K_i)_{i\in\mathbb{Z}}$.

The same reasoning applies to (85) if $Z_{k_i}\oplus U_i$ is plugged in for Z_{k_i}. Of course, the fact selection $(K_i)_{i\in\mathbb{Z}}$ is a finite-energy process if variables K_i are IID and assume more than one value.

E. Coding into a finite alphabet

It is more difficult to construct a stationary finite-energy process that satisfies both (65) and the condition of finite alphabet. The straightforward idea is to apply a variable-length coding to the process $(X_i)_{i\in\mathbb{Z}}$ given by (82).

Let $\mathbb{Y}$ be some finite alphabet, whereas $\mathbb{X}$ is the infinite alphabet of $(X_i)_{i\in\mathbb{Z}}$. For a function $f:\mathbb{X}\to\mathbb{Y}^*$, which transforms elements of $\mathbb{X}$ into strings over $\mathbb{Y}$, let us generalize its extension (24) to doubly infinite sequences $\mathbf{x}=(x_i)_{i\in\mathbb{Z}}$, $x_i\in\mathbb{X}$. Namely, let

$$f^*(\mathbf{x})=\mathbf{y}=(y_i)_{i\in\mathbb{Z}}, \tag{90}$$

where for every m there exists k and l such that $y_1y_2...y_k=f^*(x_1x_2...x_m)$ and $y_{-l}y_{-l+1}...y_0=f^*(x_{-m}x_{-m+1}...x_0)$.

Let us introduce the “expanded” process

$$(Y_i)_{i\in\mathbb{Z}}:=f^*((X_i)_{i\in\mathbb{Z}}) \tag{91}$$

for the “shrunk” process $(X_i)_{i\in\mathbb{Z}}$. If $(X_i)_{i\in\mathbb{Z}}$ is stationary, $|f(X_i)|>1$, and the expected expansion rate $\mathbf{E}\,|f(X_i)|$ is finite then process $(Y_i)_{i\in\mathbb{Z}}$ is asymptotically mean stationary (AMS), i.e., there exists a stationary process $(\bar{Y}_i)_{i\in\mathbb{Z}}$ with the distribution

$$\bar{P}(\bar{Y}_{1:n}=y_{1:n})=\lim_{M\to\infty}\frac{1}{M}\sum_{i=1}^{M}P(Y_{i+1:i+n}=y_{1:n}), \tag{92}$$

called the stationary mean of $(Y_i)_{i\in\mathbb{Z}}$ [69, Example 6].

Another important ingredient of our construction is that the variable-length coding must be reversible and commute with the shift operation $T:(x_k)_{k\in\mathbb{Z}}\mapsto(x_{k+1})_{k\in\mathbb{Z}}$.

Definition 11: The mapping $\mathbf{y}=\pi(\mathbf{x})$ is called a *synchronizable injection* if (i) $\pi(\mathbf{x})=\pi(\mathbf{x}')$ implies $\mathbf{x}=\mathbf{x}'$ and (ii) $T^i\pi(\mathbf{x})=\pi(\mathbf{x}')$ for an $i\in\mathbb{Z}$ implies $T^j\mathbf{x}=\mathbf{x}'$ for a $j\in\mathbb{Z}$.

Theorem 11: Let π be a synchronizable injection. The shift-invariant σ-fields of AMS processes $(X_i)_{i\in\mathbb{Z}}$, $(Y_i)_{i\in\mathbb{Z}}$, and $(\bar{Y}_i)_{i\in\mathbb{Z}}$ linked by $(Y_i)_{i\in\mathbb{Z}}:=\pi((X_i)_{i\in\mathbb{Z}})$ and (92) are in one-to-one correspondence and the distributions of these processes are equal on the corresponding shift-invariant sets.

(The proof will be published elsewhere, cf. [69, Theorem 1].)

Consider a comma-separated code $f(x)=g(x)c$ for an injection $g:\mathbb{X}\to(\mathbb{Y}\setminus\{c\})^*$. It is easy to see that the extension f^* is a synchronizable injection. By Theorems 7 and 11, processes $(Y_i)_{i\in\mathbb{Z}}$ and $(\bar{Y}_i)_{i\in\mathbb{Z}}$ given by (91) and (92) are UDP's if and only if $(X_i)_{i\in\mathbb{Z}}$ is a UDP. It is plausible that other interesting properties of $(X_i)_{i\in\mathbb{Z}}$ are preserved in the process $(\bar{Y}_i)_{i\in\mathbb{Z}}$ if f is a reasonably efficient code. A logarithmic decrease in the growth of the number of well predictable facts may be, however, due to encoding each statement $x=(k,z)$ as a string of length $|f(k,z)|=\Omega(\log k)$. Let us express our hypothesis formally:

Conjecture 1: For $\mathbb{X}=\mathbb{N}\times\{0,1\}$ and $\mathbb{Y}=\{0,1,2\}$ consider the comma-separated code $f:\mathbb{X}\to\mathbb{Y}^*$ given by

$$f(k,z)=b(2k+z)2,$$

where $b:\mathbb{N}\to\{0,1\}^*$ is the binary representation of natural numbers. Let process $(\bar{Y}_i)_{i\in\mathbb{Z}}$ be constructed via (91) and (92) from the process $(X_i)_{i\in\mathbb{Z}}$ given by (82) for the IID variables $(K_i)_{i\in\mathbb{Z}}$ with the distribution (84), $\beta\in(0,1)$. Denote the set of facts asymptotically predictable from $(\bar{Y}_i)_{i\in\mathbb{Z}}$ as $(\bar{Z}_k)_{k\in\mathbb{N}}$.

We conjecture that $(\bar{Y}_i)_{i\in\mathbb{Z}}$ is such a stationary finite-energy process that

$$\liminf_{n\to\infty}\frac{\operatorname{card}\bar{U}_\delta(n)}{n^\beta\log n}>0 \tag{93}$$

holds for the set of well-predictable facts

$$\bar{U}_\delta(n):=\left\{k\in\mathbb{N}:\bar{P}\left(\bar{f}_{nk}\left(\bar{Y}_{1:n}\right)=\bar{Z}_k\right)\ge\delta\right\},$$

where $\bar{f}_{nk}$ are certain guessing functions and $\delta\in(1/2/,1)$.

The proof of this conjecture relies on developing some auxiliary results for the variable-length coding of AMS processes.

F. Randomized coding

The process constructed in Conjecture 1 as an example for Theorem 9 has a clear motivation. Using the comma-separated code makes this process similar to the monkey-typing model [6], [7], [8], mentioned in Section I, with the commas playing the role of spaces. If a universal grammar-based compressor is run on this process then, we suppose, the majority of nonterminals A_i will encode whole chunks $f(k,z)$. And this provides a heuristic explanation why the distribution of nonterminals A_i and the distribution of fact addresses K_i are close to each other.

Notwithstanding this somehow disappointing remark, Theorem 9 appears applicable to more complex processes, where the nonterminals do not correspond to strings encoding whole statements (k,z). To compare, one would hardly agree that single words of natural language form complete statements in general. Let us build a class of likewise processes.

The example from the previous section may be developed in an attempt to implement the following metaphor. Assume that the process $(X_i)_{i\in\mathbb{Z}}$ of form (82) provides an abstract semantic representation of a text in natural language. On the other hand, process $(Y_i)_{i\in\mathbb{Z}}$ will correspond to the form of the text that is actually relayed among the language users.

Although one can usually answer the question whether a particular fact is described in a particular text section without reading the whole text, there exist usually several manners of stating this fact in an understandable way. Whereas the former property can apparently be provided by any synchronizable injection from $(X_i)_{i\in\mathbb{Z}}$ to $(Y_i)_{i\in\mathbb{Z}}$, we need another construction to account for the latter phenomenon.

Denote the manner of expressing the statement X_i as a new random variable $W_i : \Omega \to \mathbb{W}$. Instead of using function $f : \mathbb{X} \to \mathbb{Y}^*$, we shall apply function $\tilde{f} : \mathbb{X} \times \mathbb{W} \to \mathbb{Y}^*$ and consecutively let

$$(Y_i)_{i\in\mathbb{Z}} := \tilde{f}^*((X_i, W_i)_{i\in\mathbb{Z}}). \tag{94}$$

Let us firstly consider a comma-separated code $\tilde{f}(x, w) = \tilde{g}(x, w)c$ for a function $\tilde{g} : \mathbb{X} \times \mathbb{W} \to (\mathbb{Y} \setminus \{c\})^*$. For the comprehensibility of the text $(Y_i)_{i\in\mathbb{Z}}$, it matters that the reader can identify the abstract statements $(X_i)_{i\in\mathbb{Z}}$ but recognizing the abstract principles of the composition $(W_i)_{i\in\mathbb{Z}}$ is not necessary. Thus $\tilde{g}(x, w) = \tilde{g}(x', w')$ should imply $x = x'$ but function $\tilde{g}$ need not be an injection.

For this degenerate comma-separated code, the number of facts that are predictable from a string $X_{1:n}$ is still not greater than the number of facts that can be predicted from the corresponding string $Y_{1:m}$, where $m = \sum_{i=1}^{n} \left|\tilde{f}(X_i, W_i)\right|$. We suppose that a similar statement is true for the stationary mean $(\bar{Y}_i)_{i\in\mathbb{Z}}$ under mild conditions. Thus the number of nonterminals in the minimal compression of $\bar{Y}_{1:m}$ is lower-bounded by the number of facts described in the corresponding section $X_{1:n}$. It may be possible to tamper simultaneously with the definitions of coding $\tilde{f}$ and of process $(W_i)_{i\in\mathbb{Z}}$ to make chunks $\tilde{f}(x, w)$ hardly ever repeated within $\bar{Y}_{1:m}$. Referring to the same fact in yet different words, isn't it what renowned novelists can do?

To support some of our claims, let us write down a simple general fact. For the discussed comma-separated code, mapping $\pi(\mathbf{x}) := f^*((x_i, w_i)_{i\in\mathbb{Z}})$ is a synchronizable injection for any fixed sequence $\mathbf{w} = (w_i)_{i\in\mathbb{Z}}$, $w_i \in \mathbb{W}$. A stronger property, defined below, holds for $\tilde{\pi}(\mathbf{x}, \mathbf{w}) := f^*((x_i, w_i)_{i\in\mathbb{Z}})$.

Definition 12: The mapping $\mathbf{y} = \tilde{\pi}(\mathbf{x}, \mathbf{w})$ is called *synchronizable with respect to* $\mathbf{x}$ if (i) $\tilde{\pi}(\mathbf{x}, \mathbf{w}) = \tilde{\pi}(\mathbf{x}', \mathbf{w}')$ implies $\mathbf{x} = \mathbf{x}'$ and (ii) $T^i\tilde{\pi}(\mathbf{x}, \mathbf{w}) = \tilde{\pi}(\mathbf{x}', \mathbf{w}')$ for an $i \in \mathbb{Z}$ implies $T^j\mathbf{x} = \mathbf{x}'$ for a $j \in \mathbb{Z}$.

Theorem 12: Let $\tilde{\pi}(\mathbf{x}, \mathbf{w})$ be synchronizable with respect to $\mathbf{x}$. Let $(X_i, W_i)_{i\in\mathbb{Z}}$, $(Y_i)_{i\in\mathbb{Z}}$, and $(\bar{Y}_i)_{i\in\mathbb{Z}}$ be AMS processes linked by $(Y_i)_{i\in\mathbb{Z}} := \tilde{\pi}((X_i)_{i\in\mathbb{Z}}, (W_i)_{i\in\mathbb{Z}})$ and (92). There exists an injection from the shift-invariant σ-field of process $(X_i)_{i\in\mathbb{Z}}$ to the shift-invariant σ-field of $(Y_i)_{i\in\mathbb{Z}}$ and $(\bar{Y}_i)_{i\in\mathbb{Z}}$ and the distributions of these processes are equal on the corresponding sets.
(The proof proceeds like the proof of Theorem 11.)

Thus the processes $(Y_i)_{i\in\mathbb{Z}}$ and $(\bar{Y}_i)_{i\in\mathbb{Z}}$ which were mentioned in the above statement are UDP's, by Theorem 7, if so is the process $(X_i)_{i\in\mathbb{Z}}$. Simultaneously, the process $(W_i)_{i\in\mathbb{Z}}$ may be entangled in an arbitrary way. The way of expressing a fact W_i may depend on the previous expressions $(W_j)_{j<i}$, the whole selection of facts $(K_i)_{i\in\mathbb{Z}}$ and the whole set of facts $(Z_k)_{k\in\mathbb{N}}$. Hence we think that Theorem 9 can be applied to an extremely large class of processes, which encompasses quite adequate probabilistic models of natural language.

G. *Challenges of practical modeling*

Although the constructions discussed in this paper cast light on the multilayered structure of randomness in texts, they are too idealized or too underspecified for a direct application in natural language processing (NLP), the branch of engineering that seeks to model human language understanding with computers. We feel that we should briefly address the challenges of statistical NLP [58], [67], [68] since this domain has attracted far more researchers than quantitative linguistics [66], the theoretical investigation of language data distributions.

Complex nonergodic models have not been much practiced in NLP, at least overtly, for several reasons. Firstly, several smoothing (i.e., parameter estimation) procedures which are used in NLP produce ergodic models automatically for any training data. This feature has to do with the safety requirement that the probabilities of all blocks $(X_{1:n} = x_{1:n})$ should be strictly positive [58]. When this constraint is imposed onto a k-th order Markov process over a finite alphabet, which is a typical model class in speech recognition [58] or part-of-speech tagging [67], [68], the Markov process must be ergodic. Our example (85) shows, however, that inequality $P(X_{1:n} = x_{1:n}) > 0$ is satisfied also by certain UDP's.

Another obstacle for practical language modeling in speech recognition is the necessity of simultaneously considering two kinds of nonergodicity. One has to model both what the speaker says (i.e., the semantic content and logical consistency of her utterances) and how she says (e.g. her individual features of pronunciation). Whereas accessible description processes, such as (82), may be suited to achieve the first goal, the second aim requires modeling similar to the "parametrically" nonergodic model (81).

VII. The outlook

This paper provides an explanation for the distribution of words in natural language as a joint effect of narrative repetitions in texts and randomness of the described world. Besides developing a class of less redundant grammar-based codes, an important development of this work is a formal model of the repetitive knowledge, as it can be conveyed by texts in natural language. This notion has been differentiated from the ordinary information, or randomness, present in texts. The repetitive knowledge is even more elusive than the ordinary randomness [31] but to certain degree it constitutes a plausible object of empirical investigation, via its links with excess entropy and efficient universal coding.

We have brought together several research lines in linguistics and information theory, using very different concepts and terms. Thus it is worth resuming our main result in plain words at the expense of certain simplification. There seems to be little consolation linguists can find in observing Zipf's law, yet nothing scary is there either that could shake their preconceptions about human communications. The Martian scientist, speculated by G. K. Zipf in the passage quoted in the introduction, may not infer that human texts convey any timeless knowledge just from counting chunks in them. Nonetheless, there holds a converse implication.

A sufficient explanation for Zipf's law can be provided by the notion that human utterances convey some general knowledge that is mostly logically consistent, a priori unknown but learnable, and repetitive but potentially infinite. Moreover, the number of distinct words, or rather of the chunks obtained by minimal universal grammar-based coding, provides an upper bound on the total amount of repetitive knowledge expressed in the text. It does not matter whether a part of this knowledge is specific or general or objective or subjective or discovered or created.

Hence we find interesting the large scale experiments where a short tail of the distribution of words can be observed [28]. These experiments indicate a limitation of the active vocabulary of a single person, as opposed to the vocabulary of a language. If a similar exponential tail were observed for the nonterminals of the minimal grammar-based code, it would imply a limitation on an individual person's memory. Observing a comparably little number of distinct nonterminals in the grammar-based compression of the Voynich manuscript could also corroborate the hoax hypothesis [89], [90].

It is harder to foresee what kind of mathematical research may be inspired by this paper, which adopts a singular bird's-eye view on information theory and related domains. At least, several open problems have been introduced formally in Section V and Subsection VI-E. Addressing these questions requires further research of asymptotically mean stationary processes and the shortest grammar problem. Although we kept our argumentation within the scope of the Shannon information theory, certain prospective problems for the algorithmic information theory have been pointed out as well.

Acknowledgment

The author wishes to thank Peter Harremoës, Peter Grünwald, Alfonso Martinez, and Jan Mielniczuk for discussing the composition of this paper. Special thanks are due to James Crutchfield, Arthur Ramer, and Peter Grünwald for arranging for the author's stays with the Santa Fe Institute, the University of New South Wales, and the Centrum Wiskunde & Informatica, where some constructions for this long-born work were completed. Finally, we thank cordially Prof. Emerit. Gabriel Altmann for many words of support during our investigations.

Appendix I
The excess-bounding lemma

This paper deals with bounds on sublinear parts of functions which grow asymptotically at the same linear rate. Thus the following lemma, observed in [13] in a less general form, is a convenient tool to use:

Lemma 1 (Excess-bounding lemma): Consider a function $G : \mathbb{N} \to \mathbb{R}$ such that $G(n) \geq 0$ for all but finitely many n and $\lim_k G(k)/k = 0$. For any $A \in \mathbb{N}$, we have

$$\limsup_{n\to\infty} \left[AG(n) - G(An)\right] \geq 0. \tag{95}$$

Proof: For $n, m \in \mathbb{N}$, we have identity

$$\sum_{k=0}^{m-1} \frac{AG(A^k n) - G(A^{k+1}n)}{A^{k+1}} = G(n) - n \cdot \frac{G(A^m n)}{A^m n}.$$

Hence $G(n) - n \lim_k G(k)/k \geq 0$ implies $\sum_{k=0}^{\infty}[AG(A^k n) - G(A^{k+1}n)]/A^{k+1} \geq 0$. Putting $n = A^p M$, we obtain

$$\sum_{k=p}^{\infty} \left[AG(A^k M) - G(A^{k+1}M)\right] \cdot \frac{1}{A^{k+1}} \geq 0 \tag{96}$$

for any $M \in \mathbb{N}$ and all but finitely many $p \in \mathbb{N}$. If (95) did not hold then we would have $AG(A^k M) - G(A^{k+1}M) < 0$ for all k greater than some p. This contradicts, however, (96). ∎

As a special case, consider functions $G_1(n) \geq G_2(n) \geq 0$ and their excess values $F_i(n) = 2G_i(n) - G_i(2n)$. If the functions have equal limits $\lim_n G_i(n)/n = g < \infty$ then

$$\limsup_{n\to\infty} \left[F_2(n) - F_1(n)\right] \geq 0 \tag{97}$$

follows from Lemma 1. By inequality $\limsup_n (a_n + b_n) \geq \limsup_n a_n + \liminf_n b_n$ for arbitrary two sequences (a_n) and (b_n), inequality (97) entails the implication

$$\liminf_{n\to\infty} \frac{F_1(n)}{n^\beta} > 0 \implies \limsup_{n\to\infty} \frac{F_2(n)}{n^\beta} > 0. \tag{98}$$

Moreover, if $G_1(n) = gn + \tilde{G}_1(n)$ then

$$\liminf_{n\to\infty} \frac{\tilde{G}_1(n)}{n^\beta} > 0 \implies \limsup_{n\to\infty} \frac{F_2(n)}{n^\beta} > 0. \tag{99}$$

The latter statement is less obvious so here comes a brief justification. The left-hand side implies that $G(n) = G_2(n) - gn - Bn^\beta \geq 0$ for all but finitely n and a certain $B > 0$. Then it suffices to apply Lemma 1 to obtain the right-hand side.

Inequality (97) is not so easy to refine, even given some additional assumptions on the functions $G_i(n)$ that we can easily assure in our applications. The lower limit $\liminf_n [F_2(n) - F_1(n)]$ cannot be bounded so easily. For example, let $G_1(n) = H(n)$ be the block entropy and $G_2(n) = H^C(n)$ be the expected code length (10). Then G_1 is nondecreasing and concave whereas it is reasonable to assume that G_2 is nondecreasing but only subadditive, i.e., $G_2(n+m) \leq G_2(n) + G_2(m)$. Consequently, F_1 is nondecreasing and tends to $\lim_n [G_1(n) - ng]$ while F_2 can oscillate between 0 and F_1 in the worst case, as the following proposition assures.

Theorem 13: For any nonnegative, increasing, and concave function G_1 such that $\lim_n G_1(n) = \infty$, there exists a nondecreasing and subadditive function $G_2 \geq G_1$ such that $\lim_n G_2(n)/n = \lim_n G_1(n)/n$ and $\liminf_n F_2(n) = 0$.

Proof: We will construct a G_2 which is constant and linear on alternating intervals. Since G_1 is unbounded, there exists an infinite sequence of arguments $(b_i)_{i\in\mathbb{N}}$, where $b_1 := 1$ and $b_{i+1} := \min\{n \in \mathbb{N} : G_1(n) \geq 2G_1(b_i)\}$. In the next step, let us define

$$G_2(n) := \min\{nG_1(b_i)/b_i, G_1(b_{i+1})\}$$

for $b_i \leq n \leq b_{i+1}$.

This construction satisfies the required properties for the following reasons. Since G_1 is subadditive by concavity, we have $G_1(b_{i+1}) \geq 2G_1(b_i) \geq G_1(2b_i)$. Hence $b_{i+1} \geq 2b_i$ since G_1 is increasing. Moreover,

$$F_2(b_i) = 2G_1(b_i) - \min\{2G_1(b_i), G_1(b_{i+1})\} = 0.$$

Thus $\liminf_n F_2(n) = 0$. Inequality $G_2 \geq G_1$ holds since G_1 is growing and concave. G_2 is subadditive since $G_2(n)/n$ does not increase with n [94, Theorem 7.2.4]. Finally, since G_2 and G_1 are both subadditive and are equal on the infinite sequence $(b_i)_{i\in\mathbb{N}}$ we have $\lim_n G_2(n)/n = \inf_n G_2(n)/n = \inf_n G_1(n)/n = \lim_n G_1(n)/n$ by the Fekete lemma. ■

APPENDIX II
BOUNDS FOR THE LONGEST REPEAT

Let us review several bounds for the maximal length of a repeat in a string, defined in (20). First of all, if the alphabet is a finite set $\mathbb{X} = \{0, 1, ..., D_X - 1\}$ then

$$\mathbf{L}(w) \geq \log_{D_X} |w| - \log_{D_X} \log_{D_X} |w| - 1 \qquad (100)$$

for any string $w \in \mathbb{X}^*$. This bound is justified by the observation that if w can be split into at least $D_X^n + 1$ substrings of length n then at least two substrings must be identical. The right-hand side of (100) equals one of possible n's.

Bounding the maximal repeat length above by a sublinear function is impossible with respect to certain classes of probability measures. For any function $g(n)$ with $\lim_n g(n)/n = 0$ there exists such a stationary process $(X_i)_{i\in\mathbb{Z}}$ that

$$\limsup_{n\to\infty} \mathbf{L}(X_{1:n})/g(n) \geq 1 \quad \text{a.s.}$$

(i.e., almost surely) [95]. Nevertheless, a strong upper bound exists for quite a large class of processes:

Definition 13: $(X_i)_{i\in\mathbb{Z}}$ is called a *finite-energy process* if

$$P(X_{n+1:n+m}||X_{1:n}) \leq Kc^m \quad \text{a.s.} \qquad (101)$$

for $n, m \in \mathbb{N}$ and certain constants $c < 1$ and K.

Lemma 2: Let $(X_i)_{i\in\mathbb{Z}}$ be a finite-energy process. We have

$$\sup_{n\in\mathbb{N}} \mathbf{E}\left(\frac{\mathbf{L}(X_{1:n})}{\log n}\right)^q < \infty, \quad q > 0, \qquad (102)$$

$$\limsup_{n\to\infty} \frac{\mathbf{L}(X_{1:n})}{\log n} \leq A \quad \text{a.s.} \qquad (103)$$

for a constant $A < \infty$.

Remark: Lemma 2 is true for any countable alphabet $\mathbb{X}$, also if $(X_i)_{i\in\mathbb{Z}}$ is not stationary. Finite energy-processes can be obtained by dithering ergodic processes with an IID noise [34].

Proof: The almost sure part was shown by [34, Theorem 2]. It remains to demonstrate the bound in expectation. Assume (101) and consider a $j > i \geq 0$. Applying the idea of [96], let us notice that

$$\begin{aligned} &P(X_{j+1:j+k} = X_{i+1:i+k}) \\ &= \sum_{w\in\mathbb{X}^j} P(X_{j+1:j+k} = X_{i+1:i+k}|X_{1:i} = w)P(X_{1:j} = w) \\ &\leq \sum_{w\in\mathbb{X}^j} Kc^k P(X_{1:j} = w) = Kc^k. \end{aligned}$$

Hence

$$\begin{aligned} P(\mathbf{L}(X_{1:n}) \geq k) &= P(\exists_{0\leq i<j\leq n-k} X_{j+1:j+k} = X_{i+1:i+k}) \\ &\leq \sum_{0\leq i<j\leq n-k} P(X_{j+1:j+k} = X_{i+1:i+k}) \\ &\leq \frac{(n-k)(n-k-1)}{2} Kc^k \leq \frac{n^2 Kc^k}{2}. \end{aligned}$$

This bound is nontrivial for $k > A := (2\log n + \log K - \log 2)/\log c^{-1}$. Consider a sufficiently large n so that $A \geq 1$. Inequality (102) follows from the series of inequalities

$$\begin{aligned} \mathbf{E}\,(\mathbf{L}(X_{1:n}))^q &\leq A^q + \sum_{k>A} k^q P(\mathbf{L}(X_{1:n}) \geq k) \\ &\leq \sum_{k=0}^{\infty} (k+A)^q c^k \leq A^q \sum_{k=0}^{\infty} (k+1)^q c^k, \end{aligned}$$

where $\sum_{k=0}^{\infty} (k+1)^q c^k < \infty$. ■

APPENDIX III
EXCESS LENGTHS OF UNIVERSAL CODES

Let $\mathbb{X} = \{0, 1, ..., D_X - 1\}$ and $\mathbb{Y} = \{0, 1, ..., D_Y - 1\}$ be the input and the output alphabets. Denote the expected length of code $C : \mathbb{X}^+ \to \mathbb{Y}^+$ as

$$H^C(n) := \mathbf{E}\,|C(X_{1:n})| \log D_Y$$

and its excess $E^C(n) := 2H^C(n) - H^C(2n)$ as previously.

Universal codes are those uniquely decodable codes that achieve the best possible compression rate $\lim_n H^C(n)/n = h$ on the average, whereas inequalities $H^C(n) \geq H(n) \geq hn$ hold in general. By Lemma 1, we obtain that

$$\limsup_{n\to\infty} \left[E^{C'}(n) - E^C(n)\right] \geq 0 \text{ if } H^{C'}(\cdot) \geq H^C(\cdot) \quad (104)$$

for any universal codes C and C'. Thus the search for the shortest codes reduces to the task of finding universal codes that enjoy the smallest excess code length.

By Lemma 1, the excess code length is bounded below by the n-symbol excess entropy,

$$\limsup_{n\to\infty} \left[E^C(n) - E(n)\right] \geq 0. \qquad (105)$$

This bound is not so strong in view of such a fact:

Theorem 14: Let $\beta \in (0, 1)$. The following statements are equivalent:

(i) $H^C(n) - H(n) \leq An^\beta$ holds for an $A > 0$ and all n.

(ii) $E^C(n) - E(n) \leq Bn^\beta$ holds for a $B > 0$ and all n.

Proof: If (i) holds then (ii) holds for $B = 2A$ since $H^C(n) - H(n) \geq 0$. Conversely, if (ii) is true then

$$\begin{aligned} H^C(n) - H(n) &= \sum_{k=0}^{\infty} \frac{E^C(2^k n) - E(2^k n)}{2^{k+1}} \\ &\leq \sum_{k=0}^{\infty} Bn^\beta 2^{k(\beta-1)-1} \leq \frac{Bn^\beta}{2(1-2^{\beta-1})}. \end{aligned}$$

■

Fix some $\beta \in (0, 1)$. For each universal code there exists a stationary process for which statement (i) is not true [48]. Thus proposition (ii) is false in the same case, as well.

Using the ergodic decomposition of excess entropy, we can prove that $\sup_n E^C(n)$ is finite only for countably many ergodic processes. On the other hand, $E(n)$ is bounded by the finite excess entropy for uncountably many ergodic sources. For instance, $E < \infty$ for all irreducible Markov processes. Denote the expectation of the excess code length as

$$E^C_\mu(n) := \mathbf{E}\,(2\,|C(X_{1:n})| - |C(X_{1:2n})|) \log D_Y\,,$$

taken with respect to a measure $\mu = P((X_k)_{k\in\mathbb{Z}} \in \cdot) \in \mathbb{S}$.

Theorem 15: For a finite alphabet $\mathbb{X}$ and a universal code C, let $N^C(K)$ be the number of distinct ergodic measures $\mu \in \mathbb{E}$ such that $\limsup_n E^C_\mu(n) \le K$, $K \in \mathbb{R}$. We have

$$\log N^C(K) \le K$$

for $K \ge 0$ whereas $N^C(K) = 0$ for $K < 0$.

Proof: The case of $K < 0$ is directly captured by inequality (105). As for $K \ge 0$, let us firstly inspect the ergodic decomposition of the expected excess code length.

Consider the distribution of the random ergodic measure F. Using $\nu(W) := P(F \in W)$, $W \in \mathcal{E}$, equation (50) may be written as

$$P((X_i)_{i\in\mathbb{Z}} \in A) = \int \sigma(A) d\nu(\sigma).$$

The code lengths $|C(w)|$ for $w \in \mathbb{X}^n$ can be upper-bounded by a natural number if $\mathbb{X}$ is finite. Hence

$$E^C(n) = \mathbf{E}\, E^C_F(n) \tag{106}$$

follows by the disintegration formula

$$\int f((X_i)_{i\in\mathbb{Z}}) dP = \int f d\left(\int \sigma d\nu(\sigma)\right) = \int \left(\int f d\sigma\right) d\nu(\sigma)$$

for a bounded $\mathcal{U}$-measurable function f [62, Exercise 18.19].

By (54), (105), and (106) we obtain

$$\limsup_{n\to\infty} \mathbf{E}\, E^C_F(n) = \limsup_{n\to\infty} E^C(n) \ge H(\mathcal{F}) + \mathbf{E}\, E_F. \tag{107}$$

Inequality (107) will be used in the further reasoning.

Consider a natural number M such that $M \le N(K)$. Let $A \subset \mathbb{E}$ be a subset of M ergodic measures μ such that $\limsup_n E^C_\mu(n) \le K$. Let process $(X_i)_{i\in\mathbb{Z}}$ have distribution $P((X_i)_{i\in\mathbb{Z}} \in \cdot) = M^{-1}\sum_{\mu\in A}\mu$. By the uniqueness of its ergodic decomposition, the random ergodic measure F takes the value of each $\mu \in A$ with probability $1/M$. Hence $H(\mathcal{F}) = \log M$ by [2, Theorem 1(v) and Lemma 3].

Take some $\epsilon > 0$. Random variables $K + \epsilon - E^C_F(n)$, $n \in \mathbb{N}$, are almost surely nonnegative for all but finitely many n. Thus, by the Fatou lemma, $K + \epsilon - \mathbf{E}\,\limsup_n E^C_F(n) \le K + \epsilon - \limsup_n \mathbf{E}\, E^C_F(n)$. Hence from inequality (107) we obtain

$$\log M = H(\mathcal{F}) \le \limsup_{n\to\infty} \mathbf{E}\, E^C_F(n) \le \mathbf{E}\,\limsup_{n\to\infty} E^C_F(n) \le K.$$

Since this holds for any $M \le N(K)$, the claim follows. ■

Appendix IV
The vocabulary size of Y-K codes

There is a large qualitative difference between the grammar-based codes which were introduced in [4] and those that minimize the length induced by the local encoder B which satisfies condition (27). The empirical study of [14] compared the longest matching grammar transform (LMG) [4], [5], an irreducible transform which locally minimizes the Yang-Kieffer length $|\cdot|$, with a similar grammar transform called BLMG that locally minimizes the length $|B(\cdot)|$. It appeared that LMG and BLMG behave in a strikingly different way in terms of the vocabulary size.

The grammar transforms were applied to two novels in English and in Polish (abt. 6×10^5 characters) and their unigram approximations (roughly, random permutations of the texts). Paradoxically, the LMG discovered more structure in the unigram text (abt. 6×10^4 distinct nonterminals) than in the original data (abt. 3×10^4 nonterminals). For the BLMG, a difference of two orders was observed but in the opposite direction (about 1×10^2 nonterminals for the unigram data and 1×10^4 for the original).

As far as probabilistic modeling makes sense, the text in natural language has lower entropy rate and much higher excess entropy than its unigram approximation. Thus the vocabulary size of the BLMG seems proportional to the excess entropy of the source, in accordance with Theorem 4. In contrast, the vocabulary size of the LMG appears proportional rather to the entropy rate. Part (i) of the following proposition was mentioned by [14] as an explanation to this counterintuitive behavior of the longest matching transform:

Theorem 16: (i) If Γ is an $\mathcal{I}$-grammar transform then

$$\mathbf{V}[\Gamma(w)] > \sqrt{|\Gamma(w)|/2} - D_X - 1. \tag{108}$$

(ii) If Γ is an $\mathcal{F}\cap\mathcal{P}$-grammar transform then

$$\mathbf{V}[\Gamma(w)]\mathbf{L}(w) > \sqrt{|\Gamma(w)|/2} - D_X - 1. \tag{109}$$

Remark: The notations for grammar classes are as in Subsection III-C, except for $\mathcal{P}$, which stands for the set of partially irreducible grammars. Grammar $(\alpha_1, \alpha_2, ..., \alpha_n)$ is called *partially irreducible* if it satisfies conditions (i) and (ii) of irreducibility, as well as, each pair of consecutive symbols in string α_1 appears at most once at nonoverlapping positions. The LMG is an $\mathcal{I}$-grammar transform and there exists an $\mathcal{F}\cap\mathcal{P}$-grammar transform Γ which is a modification of the LMG. In order to compute the value of this transform for a string w, we start with the grammar $\{A_1 \to w\}$ and iteratively replace the longest repeated substrings u in the start symbol definition with the new nonterminals $A_i \to u$ until there is no repeat of length $|u| \ge 2$.

Proof: Write $G = \Gamma(w)$ and $V = \mathbf{V}[\Gamma(w)]$ for brevity. Notice that $x + a + 1 > \sqrt{y/2}$ follows from $(y - x)/2 \le (x + a)^2$ for $x, y, a \ge 0$.

(i) In this case, any pair of symbols occurs at most once at the every second position of all right-hand sides of G. Hence, $(|G| - V)/2 \le (V + D_X)^2$, which implies (108).

(ii) At the every second position of the start symbol definition in G, a pair of symbols can occur only once. Thus (109) follows by $[|G| - V\mathbf{L}(w)]/2 \le (V + D_X)^2 \le (V\mathbf{L}(w) + D_X)^2$.

■

Consider a stationary ergodic process $(X_i)_{i\in\mathbb{Z}}$ with an entropy rate h. For any $\mathcal{I}$-transform Γ the respective Yang-Kieffer code $C = B_{\text{YK}}(\Gamma(\cdot))$ is universal so

$$|\Gamma(X_{1:n})|\,(\text{const} + \log n) \ge hn - 2\log n$$

holds for all but finitely many n almost surely. This follows by the code construction, Barron's inequality and the asymptotic equipartition. Hence Theorem 16 implies

$$\mathbf{V}[\Gamma(X_{1:n})] \geq \sqrt{hn/\log n} + \text{const}$$

for all but finitely many n almost surely.

This reasoning cannot be transferred to the case of a $(B, \mathcal{J})$-minimal universal code, where B is a local encoder. In the grammars produced by this code, the substrings may appear more than twice and there is no fixed upper bound on the number of repeats.

References

[1] Ł. Dębowski, "On vocabulary size of grammar-based codes," in *2007 IEEE International Symposium on Information Theory. Nice, France, July 25-29*, 2007, pp. 91–95.

[2] ——, "Ergodic decomposition of excess entropy and conditional mutual information," 2006, IPI PAN Reports, Nr 993. Institute of Computer Science, Polish Academy of Sciences.

[3] G. K. Zipf, *The Psycho-Biology of Language: An Introduction to Dynamic Philology, 2nd ed.* The MIT Press, 1965.

[4] J. C. Kieffer and E. Yang, "Grammar-based codes: A new class of universal lossless source codes," *IEEE Trans. Inform. Theor.*, vol. 46, pp. 737–754, 2000.

[5] M. Charikar, E. Lehman, A. Lehman, D. Liu, R. Panigrahy, M. Prabhakaran, A. Sahai, and A. Shelat, "The smallest grammar problem," *IEEE Trans. Inform. Theor.*, vol. 51, pp. 2554–2576, 2005.

[6] B. Mandelbrot, "Structure formelle des textes et communication," *Word*, vol. 10, pp. 1–27, 1954.

[7] G. A. Miller, "Some effects of intermittent silence," *Amer. J. Psych.*, vol. 70, pp. 311–314, 1957.

[8] B. Conrad and M. Mitzenmacher, "Power laws for monkeys typing randomly: The case of unequal probabilities," *IEEE Trans. Inform. Theor.*, vol. 50, pp. 1403–1414, 2004.

[9] H. A. Simon, "On a class of skew distribution functions," *Biometrika*, vol. 42, pp. 425–440, 1955.

[10] R. Perline, "Strong, weak and false inverse power laws," *Statist. Sci.*, vol. 20, pp. 68–88, 2005.

[11] P. Harremoës and F. Topsøe, "Maximum entropy fundamentals," *Entropy*, vol. 3, pp. 191–226, 2001.

[12] A. Kilgarriff, "Language is never ever ever random," *Corpus Linguist. Linguist. Theor.*, vol. 1, pp. 263–276, 2005.

[13] Ł. Dębowski, "On Hilberg's law and its links with Guiraud's law," *J. Quantit. Linguist.*, vol. 13, pp. 81–109, 2006.

[14] ——, "Menzerath's law for the smallest grammars," in *Exact Methods in the Study of Language and Text*, P. Grzybek and R. Köhler, Eds. Mouton de Gruyter, 2007, pp. 77–85.

[15] W. Kuraszkiewicz and J. Łukaszewicz, "The number of different words as a function of text length," *Pamiętnik Literacki*, vol. 42(1), pp. 168–182, 1951, in Polish.

[16] P. Guiraud, *Les caractères statistiques du vocabulaire*. Paris: Presses Universitaires de France, 1954.

[17] G. Herdan, *Quantitative Linguistics*. Butterworths, 1964.

[18] H. S. Heaps, *Information Retrieval—Computational and Theoretical Aspects*. Academic Press, 1978.

[19] E. Khmaladze, "The statistical analysis of large number of rare events," 1988, Technical Report MS-R8804. Centrum voor Wiskunde en Informatica, Amsterdam.

[20] A. Kornai, "How many words are there?" *Glottometrics*, vol. 4, pp. 61–86, 2002.

[21] Y. Bar-Hillel and R. Carnap, "An outline of a theory of semantic information," in *Language and Information: Selected Essays on Their Theory and Application*. Addison-Wesley, 1964, pp. 221–274.

[22] J. G. Wolff, "Language acquisition and the discovery of phrase structure," *Lang. Speech*, vol. 23, pp. 255–269, 1980.

[23] C. G. de Marcken, "Unsupervised language acquisition," Ph.D. dissertation, Massachussetts Institute of Technology, 1996.

[24] C. G. Nevill-Manning, "Inferring sequential structure," Ph.D. dissertation, University of Waikato, 1996.

[25] C. Kit and Y. Wilks, "Unsupervised learning of word boundary with description length gain," in *Proceedings of the Computational Natural Language Learning ACL Workshop, Bergen*, M. Osborne and E. T. K. Sang, Eds., 1999, pp. 1–6.

[26] P. Jacquet, "Common words in two random strings," in *2007 IEEE International Symposium on Information Theory. Nice, France, July 25-29*, 2007, pp. 1481–1485.

[27] J. K. Orlov, "Linguostatistik: Aufstellung von Sprachnormen oder Analyse des Redeprozesses?" in *Sprache, Text, Kunst. Quantitative Analysen*, J. K. Orlov, M. G. Boroda, and I. Š. Nadarejšvili, Eds. Studienverlag Dr. N. Brockenmeyer, Bochum, 1982, pp. 1–55.

[28] M. A. Montemurro and D. H. Zanette, "New perspectives on Zipf's law in linguistics: from single texts to large corpora," *Glottometrics*, vol. 4, pp. 87–99, 2002.

[29] R. Ferrer i Cancho and R. V. Solé, "Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited," *J. Quantit. Linguist.*, vol. 8(3), pp. 165–173, 2001.

[30] C. Shannon, "Prediction and entropy of printed English," *Bell Syst. Tech. J.*, vol. 30, pp. 50–64, 1950.

[31] G. J. Chaitin, "Randomness and mathematical proof," *Sci. Am.*, vol. 232, no. 5, pp. 47–52, 1975.

[32] C. H. Bennett and M. Gardner, "The random number omega bids fair to hold the mysteries of the universe," *Sci. Am.*, vol. 241, pp. 20–34, 1979.

[33] M. Li and P. M. B. Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications, 2nd ed.* Springer, 1997.

[34] P. C. Shields, "String matching bounds via coding," *Ann. Probab.*, vol. 25, pp. 329–336, 1997.

[35] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley, 1991.

[36] J. P. Crutchfield and D. P. Feldman, "Regularities unseen, randomness observed: The entropy convergence hierarchy," *Chaos*, vol. 15, pp. 25–54, 2003.

[37] W. Hilberg, "Der bekannte Grenzwert der redundanzfreien Information in Texten — eine Fehlinterpretation der Shannonschen Experimente?" *Frequenz*, vol. 44, pp. 243–248, 1990.

[38] W. Ebeling and G. Nicolis, "Entropy of symbolic sequences: the role of correlations," *Europhys. Lett.*, vol. 14, pp. 191–196, 1991.

[39] ——, "Word frequency and entropy of symbolic sequences: a dynamical perspective," *Chaos Sol. Fract.*, vol. 2, pp. 635–650, 1992.

[40] W. Ebeling and T. Pöschel, "Entropy and long-range correlations in literary English," *Europhys. Lett.*, vol. 26, pp. 241–246, 1994.

[41] W. Bialek, I. Nemenman, and N. Tishby, "Predictability, complexity and learning," *Neural Comput.*, vol. 13, p. 2409, 2001.

[42] ——, "Complexity through nonextensivity," *Physica A*, vol. 302, pp. 89–99, 2001.

[43] I. Kontoyiannis, "The complexity and entropy of literary styles," 1996, NSF Technical Report no. 97, Department of Statistics, Stanford University.

[44] J. Kieffer, "A unified approach to weak universal source coding," *IEEE Trans. Inform. Theor.*, vol. 24, pp. 674–682, 1978.

[45] L. Györfi, I. Páli, and E. C. van der Meulen, "There is no universal source code for infinite alphabet," *IEEE Trans. Inform. Theor.*, vol. 40, pp. 267–271, 1994.

[46] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression," *IEEE Trans. Inform. Theor.*, vol. 23, pp. 337–343, 1977.

[47] D. Neuhoff and P. C. Shields, "Simplistic universal coding," *IEEE Trans. Inform. Theor.*, vol. IT-44, pp. 778–781, 1998.

[48] P. C. Shields, "Universal redundancy rates don't exist," *IEEE Trans. Inform. Theor.*, vol. IT-39, pp. 520–524, 1993.

[49] P. D. Grunwald and P. M. B. Vitanyi, "Kolmogorov complexity and information theory," *J. Log. Lang. Inform.*, vol. 12, pp. 497–529, 2003.

[50] W. Rytter, "Application of Lempel-Ziv factorization to the approximation of grammar-based compression," *Theor. Comput. Sci.*, vol. 302, pp. 211–222, 2003.

[51] O. Kallenberg, *Foundations of Modern Probability*. Springer, 1997.

[52] T. Weissman, "Not all universal source codes are pointwise universal," 2004, http://www.stanford.edu/~tsachy/interest.htm.

[53] P. Elias, "Universal codeword sets and representations for the integers," *IEEE Trans. Inform. Theor.*, vol. 21, pp. 194–203, 1975.

[54] R. W. Yeung, *First Course in Information Theory*. Kluwer Academic Publishers, 2002.

[55] Y. Ephraim and N. Merhav, "Hidden Markov processes," *IEEE Trans. Inform. Theor.*, vol. 48, pp. 1518–1569, 2002.

[56] U. Grenander and G. Szegő, *Toeplitz Forms and Their Applications*. Berkeley: University of California Press, 1958.

[57] P. D. Finch, "On the covariance determinants of autoregressive and moving average models," *Biometrika*, vol. 47, pp. 194–211, 1960.

[58] F. Jelinek, *Statistical Methods for Speech Recognition*. The MIT Press, 1997.

[59] I. M. Gelfand, A. N. Kolmogorov, and A. M. Yaglom, "Towards the general definition of the amount of information," *Dokl. Akad. Nauk SSSR*, vol. 111, pp. 745–748, 1956, in Russian.

[60] R. L. Dobrushin, "A general formulation of the fundamental Shannon theorems in information theory," *Usp. Matem. Nauk*, vol. 14(6), pp. 3–104, 1959, in Russian.

[61] M. S. Pinsker, *Information and Information Stability of Random Variables and Processes*. Holden-Day, 1964.

[62] P. Billingsley, *Probability and Measure*. Wiley, 1979.

[63] ——, *Ergodic Theory and Information*. Wiley, 1965.

[64] R. M. Gray and L. D. Davisson, "The ergodic decomposition of stationary discrete random processses," *IEEE Trans. Inform. Theor.*, vol. 20, pp. 625–636, 1974.

[65] R. C. Bradley, "On the strong mixing and weak Bernoulli conditions," *Z. Wahrsch. verw. Geb.*, vol. 50, pp. 49–54, 1980.

[66] R. Köhler, G. Altmann, and R. G. Piotrowski, Eds., *Quantitative Linguistik. Ein internationales Handbuch / Quantitative Linguistics. An International Handbook*. Walter de Gruyter, 2005.

[67] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999.

[68] D. Jurafsky and J. H. Martin, *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, 2000.

[69] R. M. Gray and J. C. Kieffer, "Asymptotically mean stationary measures," *Ann. Probab.*, vol. 8, pp. 962–973, 1980.

[70] A. R. Barron, "The strong ergodic theorem for densities: Generalized Shannon-McMillan-Breiman theorem," *Ann. Probab.*, vol. 13, pp. 1292–1303, 1985.

[71] ——, "Logically smooth density estimation," Ph.D. dissertation, Stanford University, 1985.

[72] R. Fontana, R. Gray, and J. Kieffer, "Asymptotically mean stationary channels," *IEEE Trans. Inform. Theor.*, vol. 27, pp. 308–316, 1981.

[73] A. N. Kolmogorov, "Three approaches to the quantitative definition of information," *Probl. Inform. Transm.*, vol. 1(1), pp. 1–7, 1965.

[74] G. J. Chaitin, "A theory of program size formally identical to information theory," *J. ACM*, vol. 22, pp. 329–340, 1975.

[75] T. C. Bell, J. G. Cleary, and I. H. Witten, *Text Compression*. Prentice Hall, 1990.

[76] R. M. Gray and L. D. Davisson, "Source coding theorems without the ergodic assumption," *IEEE Trans. Inform. Theor.*, vol. 20, pp. 502–516, 1974.

[77] J. C. Kieffer, "On the optimum average distortion attainable by a fixed rate coding of a nonergodic source," *IEEE Trans. Inform. Theor.*, vol. 21, pp. 190–193, 1975.

[78] P. C. Shields, D. L. Neuhoff, L. D. Davisson, and F. Ledrappier, "The distortion-rate function for nonergodic sources," *Ann. Probab.*, vol. 138–143, 1978.

[79] J. C. Kieffer, "On coding a stationary process to achieve a given marginal distribution," *Ann. Probab.*, vol. 8, pp. 131–141, 1980.

[80] V. A. Rokhlin, "On the fundamental ideas of measure theory," *Matem. Sborn.*, vol. 25, pp. 107–150, 1949, in Russian.

[81] P. C. Shields, *The Ergodic Theory of Discrete Sample Paths*. American Mathematical Society, 1996.

[82] B. de Finetti, "Funzione caratteristica di un fenomeno aleatorio," *Att. Real. Acad. Naz. Linc. Ser. 6*, vol. 4, pp. 251–299, 1931.

[83] E. Hewitt and L. J. Savage, "Symmetric measures on Cartesian products," *Trans. Amer. Math. Soc.*, vol. 80, pp. 470–501, 1955.

[84] D. Blackwell and D. Kendall, "The Martin boundary for Pólya's urn scheme, and an application to stochastic population growth," vol. 1, pp. 284–296, 1964.

[85] D. A. Freedman, "Bernard Friedman's urn," *Ann. Math. Statist.*, vol. 36, pp. 956–970, 1965.

[86] R. L. Dykstra, J. E. Hewett, and W. A. Thompson, "Events which are almost independent," *Ann. Statist.*, vol. 1, pp. 674–681, 1973.

[87] J. F. C. Kingman, "Uses of exchangeability," vol. 6, pp. 183–197, 1978.

[88] B. M. Hill, D. Lane, and W. Sudderth, "Exchangeable urn processes," *Ann. Probab.*, vol. 15, pp. 1586–1592, 1987.

[89] G. Landini, "Evidence of linguistic structure in the Voynich manuscript using spectral analysis," *Cryptologia*, vol. 25, pp. 275–295, 2001.

[90] A. Schinner, "The Voynich manuscript: Evidence of the hoax hypothesis," *Cryptologia*, vol. 31, pp. 95–107, 2007.

[91] P. D. Grünwald, *The Minimum Description Length Principle*. The MIT Press, 2007.

[92] A. Barron, J. Rissanen, and B. Yu, "The minimum description length principle in coding and modeling," *IEEE Trans. Inform. Theor.*, vol. 44, pp. 2743–2760, 1998.

[93] A. N. Shiryaev, *Probability. Second Edition*. Springer, 1996.

[94] E. Hille and R. S. Phillips, *Functional analysis and semigroups*. American Mathematical Society, 1957.

[95] P. C. Shields, "String matching: The ergodic case," *Ann. Probab.*, vol. 20, pp. 1199–1203, 1992.

[96] I. Kontoyiannis and Y. Suhov, "Prefixes and the entropy rate for long-range sources," in *Probability Statistics and Optimization: A Tribute to Peter Whittle*, F. Kelly, Ed. Wiley, 1994, pp. 89–98.