On vocabulary size of grammar-based codes

Łukasz Dębowski

Institute of Computer Science
Polish Academy of Sciences
01-237 Warszawa, Poland
Email: ldebowsk@ipipan.waw.pl

Abstract— We discuss inequalities holding between the vocabulary size, i.e., the number of distinct nonterminal symbols in a grammar-based compression for a string, and the excess length of the respective universal code, i.e., the code-based analog of algorithmic mutual information. The aim is to strengthen inequalities which were discussed in a weaker form in linguistics but shed some light on redundancy of efficiently computable codes. The main contribution of the paper is a construction of universal grammar-based codes for which the excess lengths can be bounded easily.

I. INTRODUCTION

In recent years some interest in the theory of universal coding has focused on detecting hierarchical structure in compressed data. In fact, a large class of universal grammar-based codes can be constructed [1]. The codes compress strings by transforming them first into special context-free grammars [2] and then encoding the grammars into less redundant strings.

This article presents several bounds for the vocabulary size, i.e., the number of distinct nonterminal symbols in a grammar-based compression for a string. The bounds are linked with investigation of code redundancy so we shall begin our discussion with a motivating comment.

Let $X_{m:n} := (X_k)_{m \leq k \leq n}$ be the blocks of finitely-valued variables $X_i : \Omega \to \mathbb{X} = \{0, 1, ..., D-1\}$ drawn from stationary process $(X_k)_{k \in \mathbb{Z}}$ on $(\Omega, \mathcal{J}, P)$. Assuming expectation operator $\mathbf{E}$, define n-symbol block entropy $H(n) := -\mathbf{E} \log P(X_{1:n} = \cdot)$ and excess entropy $E(n) := I(X_{1:n}; X_{n+1:2n}) = 2H(n) - H(2n)$, being mutual information between adjacent blocks [3].

On the other hand, let $C : \mathbb{X}^+ \to \mathbb{X}^+$ be a uniquely decodable code. For normalized code length $K^C(w) := |C(w)| \log D$, define $I^C(u : v) := K^C(u) + K^C(v) - K^C(uv)$ as an analog of algorithmic mutual information, cf. [4]. We will denote the expected length of the code as $H^C(n) := \mathbf{E} K^C(X_{1:n})$ and its expected excess length as $E^C(n) := \mathbf{E} I^C(X_{1:n} : X_{n+1:2n}) = 2H^C(n) - H^C(2n)$.

For a uniquely decodable code, noiseless coding inequality $H^C(n) \geq H(n)$ is satisfied and the code is called universal if compression rate $\lim_n H^C(n)/n$ equals entropy rate $h := \lim_n H(n)/n$ for any stationary distribution $P((X_k)_{k \in \mathbb{Z}} \in \cdot)$. Additionally, for any universal codes C and C' we have

$$\limsup_{n \to \infty} \left[E^C(n) - E(n) \right] \geq 0, \tag{1}$$

$$\limsup_{n \to \infty} \left[E^C(n) - E^{C'}(n) \right] \geq 0 \text{ if } H^C(\cdot) \geq H^{C'}(\cdot), \tag{2}$$

cf. [5], [6]. Thus the search for codes having the lowest redundancy can be restated as the task of finding universal codes with the smallest excess code lengths $I^C(\cdot : \cdot)$.

The specific aim of the present note is to justify links between the vocabulary size and the excess code length for certain universal grammar-based codes. A weaker form of this connection was mentioned in the context of following linguistic investigations, cf. [5], [7]:

(i) Majority of words in a natural language text can be identified as frequently repeated strings of letters. Grammar-based codes can be used to detect these repeats. Distinct words of the text happen to get represented as distinct nonterminal symbols in an approximately smallest context-free grammar for the text [8], [9]. The number of different "significantly" often repeated substrings in a typical text can be 100 times greater than in a comparable realization of a memoryless source [7].

(ii) There is a hypothesis that excess entropy of a random natural language text (imagined as a stationary stochastic process where variables X_i are consecutive letters of the text) obeys $E(n) \asymp \sqrt{n}$ rather than $E(n) = 0$ as a memoryless source [10] (cf. [6] for a connection of such an effect with nonergodicity). We asked whether the power-law growth of $E(n)$ can be linked with the known empirical power-law growth of the number of distinct words in a text against the text length [11].

In view of observation (i), our question in (ii) could be restated as: Are excess entropy $E(n)$ and the expected vocabulary size of some minimal code for string $X_{1:2n}$ approximately equal for every stationary process? Trying to answer the question, we derived inequality (1) in [5] and sought for further links between the excess code length and the vocabulary size. The result of [5] concerning the latter is encouraging but too weak. It relates the vocabulary size of the smallest grammar in the sense of [2] to the Yang-Kieffer excess grammar length rather than to the excess length of an actual universal code.

In this article, we will strengthen the connection. We will prove that excess code length $I^C(u : v)$ of some grammar-based code C is dominated by the product of the length of the longest repeated substring in string $w := uv$ and the vocabulary size of the code for w. To get this inequality, it suffices that C be the shortest code in an algebraically closed subclass of codes using a special grammar-to-string encoder. There exist universal codes satisfying this requirement.

Besides the main dominance result, we will justify an

inequality in the opposite direction. As an additional outcome, we will show that the vocabulary size of any irreducible grammar for string w cannot be less than the square root of the grammar length, cf. [7], [1]. This pair of inequalities might be used to lower-bound the redundancy of codes based on irreducible grammars.

The exposition is following. Section II reviews grammar-based coding. We construct local grammar-to-string encoders (II-A) and define minimal codes (II-B) with respect to some classes of grammars (II-C). Subsection II-D justifies universality of certain minimal codes which use local encoders. Section III presents the upper (III-A) and the lower (III-B) bounds for the excess lengths of a minimal code expressed in terms of its vocabulary size. Section IV resumes the article.

II. GRAMMAR-BASED CODING REVISITED

Grammar-based compression is founded on the following concept. An *admissible* grammar is a context free-grammar which generates singleton language $\{w\}$, $w \in \mathbb{X}^+$, and whose production rules do not have empty right-hand sides [1]. 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 [1].

Hence, an admissible grammar is given by its set of production rules $\{A_1 \to \alpha_1, A_2 \to \alpha_2, ..., A_n \to \alpha_n\}$, where A_1 is the start symbol, other A_i are secondary nonterminals, and the right-hand sides of rules satisfy $\alpha_i \in (\{A_{i+1}, A_{i+2}, ..., A_n\} \cup \mathbb{X})^+$. Since the grammar can be recovered also from sequence

$$G = (\alpha_1, \alpha_2, ..., \alpha_n), \tag{3}$$

we will call G simply the grammar. Its *vocabulary size*, i.e., the number of used nonterminal symbols, will be written

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

Let $\mathbb{X}^* = \mathbb{X}^+ \cup \{\lambda\}$, where λ is the empty word. For any string $\alpha \in (\{A_2, A_3, ..., A_n\} \cup \mathbb{X})^*$, we denote its *expansion* with respect to $G = (\alpha_1, \alpha_2, ..., \alpha_n)$ as $\langle \alpha \rangle_G$ [2], i.e., $\{\langle \alpha \rangle_G\}$ is the language generated by grammar $(\alpha, \alpha_2, \alpha_3, ..., \alpha_n)$. The set of admissible grammars will be denoted as $\mathcal{G}$ and $\mathcal{G}(w)$ will be the subset of admissible grammars which generate language $\{w\}$, $w \in \mathbb{X}^+$. 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* [1].

If string w contains many repeated substrings then some grammar in $\mathcal{G}(w)$ can "factor out" the repetitions and may be used to represent w concisely. It is not straightforward, however, how to quantify the size of a grammar. In [1] the length of grammar $G = (\alpha_1, \alpha_2, ..., \alpha_{\mathbf{V}[G]})$ was defined as

$$|G| := \sum_i |\alpha_i|, \tag{4}$$

where $|\alpha|$ is the length of $\alpha \in (\{A_1, A_2, ..., A_n\} \cup \mathbb{X})^*$. Function (4) will be called Yang-Kieffer length.

For a grammar transform, ratio $|\Gamma(w)| / |w|$ can be quite a biased measure of string compressibility. Precisely, transform Γ is called *asymptotically compact* if

$$\lim_{n \to \infty} \max_{w \in \mathbb{X}^n} |\Gamma(w)| / n = 0 \tag{5}$$

and for each grammar in $\Gamma(\mathbb{X}^+)$ each nonterminal has a different expansion. There is plenty of such transforms [1], [2].

The compression given by (5) is apparent. Consider *grammar-based codes*, i.e., uniquely decodable codes $C = B(\Gamma(\cdot)) : \mathbb{X}^+ \to \mathbb{X}^+$, where $\Gamma : \mathbb{X}^+ \to \mathcal{G}$ is a grammar transform and $B : \mathcal{G} \to \mathbb{X}^+$ is called a *grammar encoder* [1]. We have $\lim_n \max_{w \in \mathbb{X}^n} |C(w)| / n \geq 1$ necessarily. Nevertheless, there exists a grammar encoder $B_{\mathrm{YK}} : \mathcal{G} \to \mathbb{X}^+$ [1] such that

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

A. Local grammar encoders

It is hard to analyze the excess lengths of grammar-based codes which use B_{YK} given by [1] as their grammar-to-string encoder. We will define a more convenient encoder. The encoder will represent a grammar as a string resembling list (3) but, simultaneously, it will constitute nearly a homomorphism between some operations on grammars and strings.

Definition 1: $\oplus : \mathcal{G} \times \mathcal{G} \to \mathcal{G}$ 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 such grammar joining $\oplus$ and encoder $B : \mathcal{G} \to \mathbb{X}^+$ 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 encoders will be used in the next sections. For any function $f : \mathbb{U} \to \mathbb{W}$ of symbols, where concatenation on domains $\mathbb{U}^*$ and $\mathbb{W}^*$ is defined, denote its extension onto strings as $f^* : \mathbb{U}^* \ni x_1 x_2 ... x_m \mapsto f(x_1) f(x_2) ... f(x_m) \in \mathbb{W}^*$. For grammars $G_i = (\alpha_{i1}, \alpha_{i2}, ..., \alpha_{in_i})$, $i = 1, 2$, define joining

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

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 $x \in \mathbb{X}$.

Definition 2: $B : \mathcal{G} \to \mathbb{X}^+$ is a *local grammar encoder* if

$$B(G) = B_{\mathrm{S}}^*(B_{\mathrm{N}}(G)), \tag{6}$$

where:

(i) function $B_{\mathrm{N}} : \mathcal{G} \to (\{0\} \cup \mathbb{N})^*$ encodes grammars as strings of natural numbers so that the encoding of grammar $G = (\alpha_1, \alpha_2, ..., \alpha_n)$ is string $B_{\mathrm{N}}(G) = F_1^*(\alpha_1) D F_2^*(\alpha_2) D ... D F_n^*(\alpha_n)(D+1)$, which employs relative indexing $F_i(A_j) := D+1+j-i$ for nonterminals and identity transformation $F_i(x) := x$ for terminals $x \in \mathbb{X} = \{0, 1, ..., D-1\}$,
(ii) B_{S} is any function of form $B_{\mathrm{S}} : \{0\} \cup \mathbb{N} \to \mathbb{X}^+$ (for technical purposes, not necessarily an injection)—we will call B_{S} the natural number encoder.

Indeed, local encoders are adapted to joining operation $\oplus$. For instance, if $B(G_i) = u_i B_{\mathrm{S}}(D+1)$ for some grammars

G_i, $i = 1, 2$, then $B(G_1 \oplus G_2) = B_S(D + 2)B_S(D + 2 + \mathbf{V}[G_1])B_S(D)u_1 B_S(D)u_2 B_S(D + 1)$.

There exist many prefix-free local encoders. Obviously, set $B_N(\mathcal{G})$ itself is prefix-free. Therefore, encoder (6) is prefix-free (and uniquely decodable) if B_S is also prefix-free, i.e., if B_S is an injection and set $B_S(\{0\} \cup \mathbb{N})$ is prefix-free.

B. Encoder-induced grammar lengths

Let us generalize the concept of grammar length.

Definition 3: For a grammar encoder B, 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 $B = B_S^*(B_N(\cdot))$, where

$$B_S(D) = \lambda \text{ and } B_S(x) \in \mathbb{X} \text{ for } x \neq D. \qquad (7)$$

In the same spirit, we can extend the idea of the smallest grammar with respect to the Yang-Kieffer length, discussed in [2]. 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 $w \in \mathbb{X}^+$. Conversely, we will call grammar transform Γ a *$\mathcal{J}$-grammar transform* if $\Gamma(\mathbb{X}^+) \subset \mathcal{J}$.

Definition 4: For grammar length $\|\cdot\|$, $\mathcal{J}$-grammar transform Γ will be called *$(\|\cdot\|, \mathcal{J})$-minimal grammar transform* if $\|\Gamma(w)\| \leq \|G\|$ for all $G \in \mathcal{G}(w) \cap \mathcal{J}$ and $w \in \mathbb{X}^+$.

Definition 5: Code $B(\Gamma(\cdot))$ will be called *$(B, \mathcal{J})$-minimal* if Γ is $(\|\cdot\|, \mathcal{J})$-minimal for a B-induced grammar length $\|\cdot\|$.

Definition 6: For a grammar length $\|\cdot\|$, grammar sub-classes $\mathcal{J}, \mathcal{K} \subset \mathcal{G}$ 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. Classes of grammars

In section III, we will bound the excess lengths for $(B, \mathcal{J})$-minimal codes, where B are local encoders and $\mathcal{J}$ are some sufficient subclasses. In subsection II-D, we will show that several of these codes are universal. Prior to this, we have to define some necessary subclasses of grammars.

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 $\mathcal{F}$. 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 $i > 1$. On the other hand, $\mathcal{B}_k \subset \mathcal{D}_k$ 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. [12]. Of course, classes $\mathcal{B}_k$, $\mathcal{D}_k$, $\mathcal{B} := \bigcup_{k \geq 1} \mathcal{B}_k$, $\mathcal{D} := \bigcup_{k \geq 1} \mathcal{D}_k$, and $\mathcal{F}$ are sufficient.

Next, 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 [1].

The set of irreducible grammars will be denoted as $\mathcal{I}$. Any $\mathcal{I}$-grammar transform is asymptotically compact [1] so it yields a universal code when combined with grammar encoder B_{YK}.

Starting with any grammar $G_1 \in \mathcal{G}(w)$, one can construct an irreducible grammar $G_2 \in \mathcal{G}(w)$ by applying a sequence of certain reduction rules until the local minimum of functional $2 |\cdot| - \mathbf{V}[\cdot]$ is achieved [1]. This leads to the following lemma.

Lemma 1: Classes $\mathcal{I}$ and $\mathcal{G}$ are $|\cdot|$-equivalent.
Proof: The only reduction rule applicable to a grammar minimizing $|\cdot|$ is the introduction of a new nonterminal standing for a pair of symbols which appears exactly twice on the right-hand side of the grammar, cf. section VI in [1]. This reduction conserves Yang-Kieffer length. ∎

Additionally, we will say that grammar $(\alpha_1, \alpha_2, ..., \alpha_n)$ is *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. Let $\mathcal{P}$ stand for the set of partially irreducible grammars. Of course, $\mathcal{I} \subset \mathcal{P} \subset \mathcal{G}$ and $\mathcal{P}$ is sufficient.

Although $\mathcal{F} \cap \mathcal{P}$ and $\mathcal{F}$ are not $|\cdot|$-equivalent, class $\mathcal{F} \cap \mathcal{P}$ is sufficient and relates to $\mathcal{F}$ partially like $\mathcal{I}$ relates to $\mathcal{G}$. Some $\mathcal{F} \cap \mathcal{P}$-grammar transform Γ is a modification of the longest matching $\mathcal{I}$-grammar transform [1], [2]. In order to compute $\Gamma(w)$, we start with grammar $\{A_1 \to w\}$ and we replace iteratively the longest repeated substrings u in the start symbol definition with new nonterminals $A_i \to u$ until there is no repeat of length $|u| \geq 2$. $\Gamma(w)$ is the modified grammar.

D. Universal codes for local encoders

Neuhoff and Shields proved that any $(B_{NS}, \mathcal{B})$-minimal code is universal for some encoder B_{NS} and the class of block grammars $\mathcal{B}$ [12]. Encoder B_{NS} resembles a local encoder. The main difference is encoding nonterminals A_i as strings of length $\lfloor \log_D \mathbf{V}[G] \rfloor + 1$ rather than strings of length $|B_S(D + i)|$. Therefore we can establish the following proposition.

Theorem 1: Let B_S be such a prefix-free natural number encoder that $|B_S(\cdot)|$ is growing and

$$\limsup_{n \to \infty} |B_S(n)| / \log_D n = 1. \qquad (8)$$

Then for any sufficient subclass of grammars $\mathcal{J} \supset \mathcal{B}$, every $(B_S^*(B_N(\cdot)), \mathcal{J})$-minimal code C is universal, that is, $\lim_n H^C(n)/n = h$ and $\limsup_n K^C(X_{1:n})/n \leq h$ almost surely for every stationary process $(X_k)_{k \in \mathbb{Z}}$.

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

$$\limsup_{n \to \infty} \max_{w \in \mathbb{X}^n} \frac{\log_D \mathbf{V}[\Gamma_{k(n)}(w)]}{k(n)} \leq h + 2\epsilon,$$

$$\lim_{n \to \infty} \mathbf{E} \, \mathbf{V}[\Gamma_{k(n)}(X_{1:n})] \cdot k(n)/n = 0,$$

$$\lim_{n \to \infty} \mathbf{V}[\Gamma_{k(n)}(X_{1:n})] \cdot k(n)/n = 0 \text{ almost surely, cf. [12].}$$

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

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

where $\alpha > 0$ and $\lim_k \gamma(k) = 1$. In particular, this inequality holds for (6), (8), and growing $|B_{\mathrm{S}}(\cdot)|$. ∎

Prefix-free natural number encoder B_{S} satisfying (8) can be chosen, e.g., as the D-ary representation $\omega : \mathbb{N} \to \mathbb{X}^*$ [13], $|\omega(n)| = \ell(n)$, where

$$\ell(n) := \begin{cases} 1 & \text{if } n < D, \\ \ell(\lfloor \log_D n \rfloor) + \lfloor \log_D n \rfloor + 1 & \text{if } n \geq D. \end{cases}$$

Alternatively, we can use the D-ary representation $\delta : \mathbb{N} \to \mathbb{X}^*$ [13], $|\delta(n)| = 1 + 2 \lfloor \log_D(1 + \lfloor \log_D n \rfloor) \rfloor + \lfloor \log_D n \rfloor$.

III. BOUNDS INVOLVING THE VOCABULARY SIZE

We will derive several inequalities for the vocabulary size of certain minimal grammar-based codes. Strictly speaking, code universality is irrelevant for the proofs. It is important, however, that the codes use the local grammar encoders.

A. Upper bounds for the excess lengths

We will begin with defining several operations on grammars. For strings $u, v \in \mathbb{X}^*$ and $w = uv$, we define the *left and right croppings* of grammar $G = (\alpha_1, \alpha_2, ..., \alpha_n) \in \mathcal{G}(w)$ as

$$\mathbb{L}_u G := (x_L y_L, \alpha_2, ..., \alpha_n) \in \mathcal{G}(u),$$
$$\mathbb{R}_v G := (y_R x_R, \alpha_2, ..., \alpha_n) \in \mathcal{G}(v),$$

where exactly one of the following conditions holds:

(i) $\alpha_1 = x_L x_R$, while $y_L = y_R = \lambda$ are empty words,
(ii) $\alpha_1 = x_L A_i x_R$ for some nonterminal A_i, $2 \leq i \leq n$, with expansion $\langle A_i \rangle_G = y_L y_R$.

Next, for grammar $G = (\alpha_1, \alpha_2, ..., \alpha_n)$, define its *flattening* $\mathbb{F}G := (\alpha_1, \langle \alpha_2 \rangle_G, \langle \alpha_3 \rangle_G, ..., \langle \alpha_n \rangle_G)$. The secondary part of the grammar will be denoted as $\mathbb{S}G := (\lambda, \alpha_2, \alpha_3, ..., \alpha_n)$. Additionally, we will use a notation for the maximal length of a nonoverlapping repeat in string $w \in \mathbb{X}^*$, i.e.,

$$\mathbf{L}(w) := \max_{u,x,y,z \in \mathbb{X}^*: w = xuyuz} |u|.$$

Now we can generalize Theorem 3 from [5]. We will show that the lengths of some minimal codes are almost subadditive. Moreover, the excess lengths are dominated by the vocabulary size multiplied by the length of the longest repeat.

Theorem 2: Let B be local encoder (6). Introduce constants

$$W_m := \max_{0 \leq n \leq D+2+m} |B_{\mathrm{S}}(n)|.$$

Let $\|\cdot\|$ be the B-induced grammar length, Γ be a $(\|\cdot\|, \mathcal{J})$-minimal grammar transform, and $C = B(\Gamma(\cdot))$ be the respective grammar-based code. Set $w = uv$ for $u, v \in \mathbb{X}^+$ and consider 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

$$\frac{I^C(u:v)}{\log D} \geq -3W_0 - W_{\mathbf{V}[\Gamma(u)]}. \tag{9}$$

(ii) If $G \in \mathcal{J} \implies \mathbb{L}_u G, \mathbb{R}_v G \in \mathcal{K}$ provided $\mathbb{L}_u G$ and $\mathbb{R}_v G$ are defined then

$$\frac{K^C(w) - K^C(u)}{\log D} \geq -W_0 \mathbf{L}(w), \tag{10}$$

$$\frac{K^C(w) - K^C(v)}{\log D} \geq -W_0 \mathbf{L}(w), \tag{11}$$

$$\frac{I^C(u:v)}{\log D} \leq \|\mathbb{S}\Gamma(w)\| + W_0 \mathbf{L}(w). \tag{12}$$

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

$$\|\mathbb{S}\Gamma(w)\| + W_0 \mathbf{L}(w) \leq W_0 \mathbf{V}[\Gamma(w)](1 + \mathbf{L}(w)). \tag{13}$$

Remark 1: In particular, (9) holds for $\mathcal{J} = \mathcal{G}, \mathcal{P}, \mathcal{I}$ while inequalities (10)–(13) hold for $\mathcal{J} = \mathcal{G}, \mathcal{P}, \mathcal{I}, \mathcal{F}, \mathcal{D}, \mathcal{D}_k$. Moreover, (12) and (13) imply together bound

$$\frac{I^C(u:v)}{\log D} \leq W_0 \mathbf{V}[\Gamma(w)](1 + \mathbf{L}(w)), \tag{14}$$

which we have mentioned in the introduction.

Remark 2: Theorem 3 in [5] is a restriction of Theorem 2 to B_{S} given by (7) and $\|\cdot\|$ equal to Yang-Kieffer length $|\cdot|$.

Proof:

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

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

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

(ii) The inequalities follow from

$$\|\Gamma(w)\| + W_0 \mathbf{L}(w) \geq \|\mathbb{L}_u \Gamma(w)\| \geq \|\Gamma(u)\|,$$
$$\|\Gamma(w)\| + W_0 \mathbf{L}(w) \geq \|\mathbb{R}_v \Gamma(w)\| \geq \|\Gamma(v)\|,$$

and

$$\|\mathbb{L}_u \Gamma(w)\| + \|\mathbb{R}_v \Gamma(w)\| \leq \|\Gamma(w)\| + \|\mathbb{S}\Gamma(w)\| + W_0 \mathbf{L}(w).$$

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

B. Lower bounds for the excess lengths

For Yang-Kieffer length function, the excess lengths can be lower-bounded by another vocabulary-related quantity. For grammars $G_i = (\alpha_{i1}, \alpha_{i2}, ..., \alpha_{in_i})$, $i = 1, 2$, denote the number of their common nonterminal expansions as

$$\mathbf{V}[G_1; G_2] := \mathrm{card} \bigcap_{i=1,2} \left\{ \langle \alpha_{i2} \rangle_{G_i}, \langle \alpha_{i3} \rangle_{G_i}, ..., \langle \alpha_{in_i} \rangle_{G_i} \right\}.$$

Moreover, introduce a different grammar joining $\otimes$. For grammars $G_i = (\alpha_{i1}, \alpha_{i2}, ..., \alpha_{in_i})$, $i = 1, 2$, define

$$G_1 \otimes G_2 := (\alpha_{11}\alpha_{21}, Q_1^*(\alpha_{12}), ..., Q_1^*(\alpha_{1n_1}),$$
$$Q_2^*(\alpha_{22}), ..., Q_2^*(\alpha_{2n_2})),$$

where $Q_1(A_j) := A_j$ and $Q_2(A_j) := A_{j+n_1-1}$ for nonterminals and $Q_1(x) := Q_2(x) := x$ for terminals $x \in \mathbb{X}$.

Recall Grammar Reduction Rule 5 from [1], which deletes useless nonterminals from the grammar and, for all nonterminals sharing the same expansion, substitutes one of them. Let $\mathbb{I}G$ be the result of applying the rule to grammar G. We have

$$|\mathbb{I}(G_1 \otimes G_2)| \leq |G_1 \otimes G_2| - \mathbf{V}[G_1; G_2] \min_{ij} |\alpha_{ij}|$$
$$\leq |G_1 \otimes G_2| - \mathbf{V}[G_1; G_2]$$

if the production rules of grammars G_1 and G_2 do not have empty right-hand sides (i.e., if the grammars are admissible, in particular). This entails the following proposition.

Theorem 3: Let Γ be a $(|\cdot|, \mathcal{J})$-minimal grammar transform, where the production rules of grammars in $\mathcal{J}$ have nonempty right-hand sides. Let grammar class $\mathcal{K}$ be $|\cdot|$-equivalent to $\mathcal{J}$. If $G_1, G_2 \in \mathcal{J} \implies G_1 \otimes G_2 \in \mathcal{K}$ then

$$|\Gamma(u)| + |\Gamma(v)| - |\Gamma(uv)| \geq \mathbf{V}[\Gamma(u); \Gamma(v)]. \quad (15)$$

Remark: In particular, (15) holds for $\mathcal{J} = \mathcal{G}, \mathcal{P}, \mathcal{I}, \mathcal{F}, \mathcal{D}_k$.

Proof: Let $G_1 = \Gamma(u)$ and $G_2 = \Gamma(v)$. We have

$$|\mathbb{I}(G_1 \otimes G_2)| + \mathbf{V}[G_1; G_2] \leq |G_1 \otimes G_2| \leq |G_1| + |G_2|$$

and the result follows from $|\Gamma(uv)| \leq |\mathbb{I}(\Gamma(u) \otimes \Gamma(v))|$. ∎

The next two propositions suggest that the size of common vocabulary $\mathbf{V}[\Gamma(u); \Gamma(v)]$ for irreducible grammar transforms may grow quite fast with the length of strings u and v.

Theorem 4 (mentioned in [7]): For the $\mathcal{I}$-grammar transform Γ, we have $\mathbf{V}[\Gamma(w)] > \sqrt{|\Gamma(w)|/2} - (D - 1)$.

Proof: In the right-hand sides of $G = \Gamma(w)$, each concatenation of two symbols occurs at most once at nonoverlapping positions. Hence, $(|G| - \mathbf{V}[G])/2 \leq (\mathbf{V}[G] + D - 1)^2$. In particular $(\mathbf{V}[G] + D - 1)^2 > |G|/2 - (\mathbf{V}[G] + D - 1)$, which implies the thesis. ∎

Theorem 5: For the $\mathcal{F} \cap \mathcal{P}$-grammar transform Γ, inequality $\mathbf{V}[\Gamma(w)]\mathbf{L}(w) > \sqrt{|\Gamma(w)|/2} - (D - 1)$ is satisfied.

Proof: In the start symbol definition of $\Gamma(w)$, each concatenation of two symbols occurs at most once at nonoverlapping places. This yields $(|\Gamma(w)| - \mathbf{V}[\Gamma(w)]\mathbf{L}(w))/2 \leq (\mathbf{V}[\Gamma(w)] + D - 1)^2 \leq (\mathbf{V}[\Gamma(w)]\mathbf{L}(w) + D - 1)^2$. Hence the thesis follows. ∎

IV. CONCLUSION

Discussing certain minimal universal grammar-based codes, we have shown that the vocabulary size is greater than the excess code length divided by the length of the longest repeated substring $\mathbf{L}(\cdot)$. Recall that $\mathbf{L}(X_{1:n})$ for a block of n symbols drawn from an arbitrary stationary stochastic process cannot be upper-bounded by a universal function $o(n)$ [14]. Nevertheless, $\mathbf{L}(X_{1:n}) = O(\log n)$ almost surely if $(X_i)_{i \in \mathbb{Z}}$ is a finite-energy process [15]. Hence, an extended Hilberg hypothesis [10], stating that a good model for texts in natural languages is a finite-energy process with excess entropy $E(n) \asymp \sqrt{n}$, is consistent with some observations asserting that vocabulary size for certain text compressions is $\Omega(\sqrt{n}/\log n)$ where n is the text length [16, Figure 3.12 (b), p. 69].

While some premises appealing to ergodic decomposition make Hilberg's hypothesis plausible even without the evidence of grammar-based compression [6], there remains an important theoretical problem. Can the vocabulary size or the excess length of a grammar-based code be used to estimate excess entropy accurately? Inequality (1) giving the lower bounds, the upper bounds for difference $E^C(n) - E(n)$ are less recognized. Although $\left|E^C(n) - E(n)\right| = O(\log n)$ when the length of code C equals prefix algorithmic complexity and

block distribution $\mathbf{P}(w) = P(X_{1:|w|} = w)$ is a recursively computable function [6], [4], some results in ergodic theory indicate that there is no universal bound for $\left|E^C(n) - E(n)\right|$ in the class of stationary processes [6], [17].

Simpler arguments could be used to infer that difference $E^C(n) - E(n)$ is large for certain codes and stochastic processes. Consider compressing a memoryless source with entropy rate $h > 0$. We have $E(n) = 0$. On the other hand, let code C be formed by a local encoder satisfying (8) and an irreducible transform Γ. Then $E^C(n) = \Omega(\sqrt{hn/\log n})$ would be implied by Theorems 3 and 4 if relation $\mathbf{V}[\Gamma(X_{1:n}); \Gamma(X_{n+1:2n})] \asymp \mathbf{V}[\Gamma(X_{1:n})]$ held.

Let us notice that the bound for $E^C(n)$ conjectured for memoryless sources and irreducible grammar-based codes is almost the same as the inequality established for general minimal codes and sources with $E(n) \asymp \sqrt{n}$. Although there are huge differences between the statistics of grammar-based compressions for texts in natural language and memoryless sources [7], the theoretical understanding of them is lacking.

REFERENCES

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

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

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

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

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

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

[7] ——, "Menzerath's law for the smallest grammars," in *Viribus Quantitatis. The Exact Science of Language and Text*, R. Köhler and P. Grzybek, Eds. Berlin: de Gruyter, 2006, pp. 77–85.

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

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

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

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

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

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

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

[15] ——, "String matching bounds via coding," *Ann. Probab.*, vol. 25, pp. 329–336, 1997.

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

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