

# Ergodic decomposition of excess entropy and conditional mutual information\*

Łukasz Dębowski

ldebowsk@ipipan.waw.pl

*Polish Academy of Sciences, Institute of Computer Science*

*ul. Orłona 21, 01-237 Warszawa, Poland*

## Abstract

The article discusses excess entropy defined as mutual information between the past and future of a stationary process. The central result is an ergodic decomposition: Excess entropy is the sum of self-information of shift-invariant  $\sigma$ -field and the average of excess entropies for the ergodic components of the process. The result is derived using generalized conditional mutual information for fields of events, developed in the paper anew. Some corollary of the ergodic decomposition is that excess entropy is infinite for the class of processes with uncountably many ergodic components, called here uncountable description processes (UDP's). UDP's can be defined without the use of measure theory and the article argues for their potential utility in linguistics. Moreover, it is shown that finite-order excess entropies (some approximations of excess entropy) are dominated by the expected excess lengths of any universal code. Hence, universal codes may be used for rough estimation of excess entropy. Nevertheless, the excess code lengths diverge to infinity for almost every process with zero excess entropy, which is another corollary of the ergodic decomposition.

**Key words:** conditional mutual information for  $\sigma$ -fields, ergodic decomposition, excess entropy, uncountable description processes, halting probability, universal coding, algorithmic mutual information

**MSC 2000:** 94A17, 94A29, 28D20

## 1 Introduction

In this article we will show utility of some measure-theoretic tools in the study of sublinear effects in entropy and universal coding of stationary processes. The tools that we use are based on the concept of mutual information for set algebras, developed by Gelfand et al. [28]. The mathematical problems that we seek to solve arose from the celebrated application of information theory to linguistics [46], reconsidered in the context of grammar-based compression [39, 12] and a specific hypothesis about long-memory effects in human texts [36]. Our work can be perceived as a part of the larger interest in algorithmic mutual information treated as a building block for estimation methods of probabilistic mutual information [11, 42, 41, 34, 40, 33].

---

\*The work was partially supported by Polish Ministry of Scientific Research and Information Technology, grant no. 1/P03A/045/28 (2005–2006). To be printed as a manuscript in ICS PAS Reports, no. 993, 2006.

The basic concepts that we investigate are excess entropy and its approximations. Let  $(X_k)_{k \in \mathbb{Z}}$  be a stationary process. Let  $X_{m:n} := (X_k)_{m \leq k \leq n}$  denote the blocks of variables. Finite-order excess entropies  $E(n)$  are defined (cf. e.g. [15]) as mutual information between the adjacent finite blocks of equal length,

$$E(n) := I(X_{-n+1:0}; X_{1:n}).$$

Excess entropy  $E := \lim_n E(n)$  can be equivalently defined as mutual information between the  $\sigma$ -fields of half-infinite past and future [28]. There exists, however, yet another expression. For any random variable  $Y$ , define entropy  $\bar{H}(Y) := I(Y; Y)$  as self-information. In particular, if block entropy  $H(n) := \bar{H}(X_{1:n})$  is finite then  $E(n) = 2H(n) - H(2n)$  and excess entropy is the deviation of block entropy from the linear growth,

$$E = \lim_{n \rightarrow \infty} [H(n) - hn], \tag{1}$$

where  $h$  is entropy rate, i.e.,  $h := \lim_n H(n)/n$  [15]. Formula (1) motivates the name of quantity  $E$ .

Excess entropy has been deemed to be some measure of “long memory” or “stochastic complexity” [4], especially for discrete processes. Quantity  $E$  is finite for Gaussian ARMA processes [27] and finite-state sources (i.e. hidden Markov processes) [15] but it may be infinite even for  $X_i$  having finite range [9, 29, 3]. The recent interdisciplinary interest in excess entropy for discrete processes [23, 24, 25, 4, 5, 15, 17, 19] was stimulated by a linguistic hypothesis. Namely, Hilberg [36] reanalyzed Shannon’s data [46] concerning the guessing estimates of entropy rate for English and supposed that proportionality  $E(n) \asymp \sqrt{n}$  holds for such stochastic models of text written in a human language that  $X_i$  take the values of the consecutive letters of a text.

Considering Hilberg’s hypothesis from a more abstract point, we can ask three kinds of questions:

- (i) It has been hypothesized (cf. e.g. [14, Section 6.4]) that stochastic models of texts written in natural language should be nonergodic. (The property of nonergodicity seemingly allows to perform successful authorship attribution for texts basing on their simple statistics, see e.g. [49].) Are there some relationships between (non)ergodicity of the process and the infinite value of  $E$  or the divergence rate for  $E(n)$ ?
- (ii) Does there exist a process  $(X_k)_{k \in \mathbb{Z}}$  with  $E(n) \asymp n^\beta$ , where  $\beta \in (0, 1)$ ? Can we present a large class of example processes with  $E(n) \asymp n^\beta$  or  $E = \infty$ ?
- (iii) Is it possible to estimate finite-order excess entropies  $E(n)$  efficiently? Can we bound them in terms of the lengths of universal codes? Which universal codes are suitable for such estimation and which are not?

In this paper, we will sketch partial answers for these questions, mostly using algebraic properties of conditional mutual information for set algebras [28, 20].

The central result of the article is the ergodic decomposition of excess entropy accompanied by corollaries. The well known basic ergodic decomposition theorem states that any stationary measure can be represented uniquely as an expectation of a random ergodic measure (see e.g. [38] for a very formal treatment). Additionally Gray and Davisson [30] proved that the entropy rate of a stationary measure is the average of the respective entropy rates for the ergodic components of the process. We will justify another, somewhat counterintuitive

result. By additivity of conditional mutual information, the difference between the excess entropy and the average of excess entropies for the ergodic components is not necessarily zero. The difference equals self-information of the invariant  $\sigma$ -field.

The concept of self-information for nonatomic fields has been considered useless. For example, we can read in [13, page 848]:

Although the general notion of mutual information provides a definition of entropy, in general the entropy of a continuous variable is infinite and this sheds no light on information content or coding.

We will try to argue that this opinion is unjust. Self-information of the invariant  $\sigma$ -field is infinite if the process has an uncountable number of ergodic components. Hence, excess entropy is infinite in this case, as well. We will construct elementarily the class of processes, called here uncountable description processes (UDP's), which have such uncountable ergodic decomposition. We suppose that UDP's might find an application in text statistics so we think that they deserve thorough investigation.

The ergodic decomposition of excess entropy affects also the problem of universal coding. We will define expected excess lengths of codes analogously to  $E(n)$ . If the code is weakly universal, i.e., the expectation of its rate equals entropy rate then the expected excess lengths of the code are greater than  $E(n)$  for infinitely many  $n$ . Simply speaking, we can use any kind of universal code to obtain a rough upper bound for finite-order excess entropies. This result is proved using a simple trick without measure theory. When we combine it with ergodic decomposition, we obtain some negative result: The expected excess lengths of the code diverge to infinity for every weakly universal code and almost every ergodic process. Excess code lengths cannot be too good estimates of  $E(n)$  since the gaps between these two kinds of quantities can diverge for any kind of universal code. The divergence of the excess code length for (algorithmically) random sequences is known when the code is prefix Kolmogorov complexity [41, Figure 3.3]. We prove here the general result with no reference to theoretical computer science.

The composition of the paper is following. In section 2, we develop an elementary and convenient definition of generalized conditional mutual information. In section 3, we prove the ergodic decomposition of excess entropy. In section 4, we present the class of uncountable description processes. In section 5, we discuss excess lengths of universal codes. Finally, a conclusion is offered in section 6.

## 2 Information theory for fields of events

A convenient prerequisite to discuss excess entropy is the extension of conditional mutual information (CMI) into a functional of arbitrary fields of events (algebras of sets). Apart from a general definition of CMI, we will need several algebraic identities holding for the generalized CMI. In fact, Gelfand, Kolmogorov, and Yaglom developed an extension of (unconditional) mutual information and presented some of its properties in their seminal short communication [28]. Subsequently they also attempted at generalizing CMI and their research was detailed by Dobrushin [20]. As for the other sources, there exist seemingly only fragmentary discussions of the results, concerning unconditional mutual information [13] and conditional entropy [6].

Unfortunately, even the presentation in [20] is not sufficient for our needs. Some important identities are lacking and there are also some problems concerning the definition.

Namely, Dobrushin [20, Eq. (2.7.7)] followed the definition of CMI using the concept of conditional product measure. On the other hand, Swart [48] proved recently that the existence of regular conditional probability implies the existence of conditional product measure but, simultaneously, he presented an example of such probability space that conditional product measure does not exist.

In this article, we will use CMI for fields only as a natural and elegant tool, so let us omit the lengthy discussion of conditional product measure. We will develop all necessary results basing on a new definition of CMI, which is simpler, more general, and easier to manipulate. We hope that our presentation will popularize this concept as it deserves.

For probability space  $(\Omega, \mathcal{J}, P)$  let  $\mathcal{A}$ ,  $\mathcal{B}$ , and  $\mathcal{C}$  be the subfields of  $\mathcal{J}$ . For simplicity, we assume hereinafter that  $(\Omega, \mathcal{J}, P)$  is complete, i.e.,  $\mathcal{J}$  contains all subsets of  $P$ -measure 0 sets. Let  $P(A|\mathcal{C})$  be the conditional probability of  $A \in \mathcal{J}$  w.r.t. to the smallest  $\sigma$ -field containing  $\mathcal{C}$  (cf. e.g. [7, Section 33]). We define conditional mutual information between  $\mathcal{A}$  and  $\mathcal{B}$  w.r.t.  $\mathcal{C}$  as

$$I(\mathcal{A}; \mathcal{B}|\mathcal{C}) := \sup \int \left[ \sum_{i=1}^I \sum_{j=1}^J P(A_i \cap B_j|\mathcal{C}) \log \frac{P(A_i \cap B_j|\mathcal{C})}{P(A_i|\mathcal{C})P(B_j|\mathcal{C})} \right] dP, \quad (2)$$

where the supremum is taken over finite partitions  $\{A_i\}_{i=1}^I$  and  $\{B_j\}_{j=1}^J$  of  $\Omega$  where  $A_i \in \mathcal{A}$  and  $B_j \in \mathcal{B}$ . It is straightforward that the expression on the right-hand side of (2) is well defined for all  $\mathcal{A}$ ,  $\mathcal{B}$ , and  $\mathcal{C}$  since conditional probabilities  $P(\cdot|\mathcal{C})$  are  $\mathcal{J}$ -measurable [7, Section 33]. There are no problems even when the regular conditional probability does not exist (c.f. e.g. [7, Exercise 33.13] and [45, Corollary 1]) since conditional probability distribution  $(P(E|\mathcal{C}))_{E \in \mathcal{E}}$  restricted to a finite field  $\mathcal{E}$  is almost surely a probability measure [7, Theorem 33.2].

Fix some subfields  $\mathcal{A}_i \subset \mathcal{J}$ ,  $i = 1, 2, 3, \dots$ , and random variables  $Y_i$  such that each  $\mathcal{A}_i$  is the smallest field with respect to which  $Y_i$  is measurable. Analogically to the case when  $Y_i$  are discrete (see e.g. [51]), we define Shannon information measures: conditional mutual information  $I(Y_1; Y_2|Y_3) := I(\mathcal{A}_1; \mathcal{A}_2|\mathcal{A}_3)$ , mutual information  $I(Y_1; Y_2) := I(\mathcal{A}_1; \mathcal{A}_2) := I(\mathcal{A}_1; \mathcal{A}_2|\{\emptyset, \Omega\})$ , conditional entropy  $\bar{H}(Y_1|Y_2) := \bar{H}(\mathcal{A}_1|\mathcal{A}_2) := I(\mathcal{A}_1; \mathcal{A}_1|\mathcal{A}_2)$ , and entropy  $\bar{H}(Y_1) := \bar{H}(\mathcal{A}_1) := I(\mathcal{A}_1; \mathcal{A}_1)$ . It is straightforward that  $I(\mathcal{A}_1; \mathcal{A}_2)$  and  $\bar{H}(\mathcal{A}_1|\mathcal{A}_2)$  reduce to the same expressions as those given in [28] and [6, Section 12]. We write  $\bar{H}$  instead of  $H$  so that  $\bar{H}(\cdot)$  be not confused with block entropy and especially with differential entropy. These functions do differ, e.g., for Gaussian variables differential entropy is finite [14, Theorem 9.4.1] while  $\bar{H}(\cdot)$  is infinite. No confusion arises about mutual information, see [28, Theorem 4] or [13, Eqs. (2.7) and (2.8)].

CMI satisfies a number of algebraic identities, which generalize the well known identities of the discrete case. Adopt conventions as follows. Let  $A \ominus B = A \setminus B \cap B \setminus A$  stand for the symmetric difference of sets. For fields  $\mathcal{A}$  and  $\mathcal{B}$ ,  $\mathcal{A} \wedge \mathcal{B}$  is the intersection of all fields containing  $\mathcal{A}$  and  $\mathcal{B}$ . We write  $\mathcal{B}_n \uparrow \mathcal{B}$  for a sequence  $(\mathcal{B}_n)_{n \in \mathbb{N}}$  of fields such that  $\mathcal{B}_1 \subset \mathcal{B}_2 \subset \dots \subset \mathcal{B}$  and  $\bigcup_{n \in \mathbb{N}} \mathcal{B}_n = \mathcal{B}$ . We say that (i) field  $\mathcal{A}$  is *trivial* when  $P(A) \in \{0, 1\}$  for all  $A \in \mathcal{A}$ , (ii) field  $\mathcal{B}$  is *nonatomic* when  $P(B_1) > 0$  for  $B_1 \in \mathcal{B}$  implies that there exists  $B_2 \in \mathcal{B}$  such that  $B_2 \subset B_1$  and  $0 < P(B_2) < P(B_1)$ , (iii) a field is *finite* if it has finitely many elements, and (iv) a field is *complete* if it contains all sets of  $P$ -measure 0. Let  $\sigma_0(\mathcal{A})$  be the intersection of all  $\sigma$ -fields containing  $\mathcal{A}$ . We also introduce notation  $\sigma(\mathcal{A})$  for the intersection of all *complete*  $\sigma$ -fields containing  $\mathcal{A}$ . Finally, let  $[B_1, \dots, B_n]$  be the smallest (finite) field containing partition  $\{B_j\}_{j=1}^n$  with  $B_i \in \mathcal{J}$ .

**Lemma 1** *Let  $\mathcal{A}$  and  $\mathcal{B}$  be finite fields and let  $\mathcal{C}$  be any field. For each  $n \in \mathbb{N}$ , let field*

$\mathcal{C}_n \subset \mathcal{C}$  satisfy

$$\{\omega \in \Omega : (i-1)/n < P(E|\mathcal{C}) \leq i/n\} \in \mathcal{C}_n \text{ for } i = 1, \dots, n \text{ and } E \in \mathcal{A} \wedge \mathcal{B}. \quad (3)$$

Then  $\lim_n I(\mathcal{A}; \mathcal{B}|\mathcal{C}_n) = I(\mathcal{A}; \mathcal{B}|\mathcal{C})$ .

Notice that finite fields  $\mathcal{C}_n$  satisfying (3) may be constructed for any  $\sigma$ -field  $\mathcal{C}$  since conditional probabilities  $P(E|\mathcal{C})$  are  $\mathcal{C}$ -measurable functions.

**Proof:** For finite  $\mathcal{A} = [A_1, \dots, A_I]$  and  $\mathcal{B} = [B_1, \dots, B_J]$ , we obtain

$$I(\mathcal{A}; \mathcal{B}|\mathcal{C}) = \int I(\mathcal{A}; \mathcal{B}|\mathcal{C}) dP \quad (4)$$

using pointwise conditional mutual information

$$I(\mathcal{A}; \mathcal{B}|\mathcal{C}) := \sum_{i=1}^I \sum_{j=1}^J P(A_i \cap B_j|\mathcal{C}) \log \frac{P(A_i \cap B_j|\mathcal{C})}{P(A_i|\mathcal{C})P(B_j|\mathcal{C})}. \quad (5)$$

$I(\mathcal{A}; \mathcal{B}|\mathcal{C})$  is almost surely positive and less than  $I(\mathcal{A}; \mathcal{A}|\mathcal{C}) \leq \log I$ . On the other hand, condition (3) implies  $|P(E|\mathcal{C}_n) - P(E|\mathcal{C})| \leq 1/n$  a.s. Hence, the thesis follows by the Lebesgue dominated convergence theorem and a convergence result in [51, Section 2.3].  $\square$

**Theorem 1** *Let  $\mathcal{A}$ ,  $\mathcal{B}$ ,  $\mathcal{B}_n$ ,  $\mathcal{C}$ , and  $\mathcal{D}$  be subfields of  $\mathcal{J}$ .*

- (i)  $I(\mathcal{A}; \mathcal{B}|\mathcal{C}) \geq 0$ , with equality if and only if  $\mathcal{A} \perp_{\mathcal{C}} \mathcal{B}$ ;
- (ii)  $I(\mathcal{A}; \mathcal{B}|\mathcal{C}) \leq \min(\bar{H}(\mathcal{A}|\mathcal{C}), \bar{H}(\mathcal{B}|\mathcal{C}))$ ;
- (iii)  $I(\mathcal{A}; \mathcal{B}_1|\mathcal{C}) \leq I(\mathcal{A}; \mathcal{B}_2|\mathcal{C})$  if  $\mathcal{B}_1 \subset \mathcal{B}_2$ ;
- (iv)  $I(\mathcal{A}; \mathcal{B}_n|\mathcal{C}) \uparrow I(\mathcal{A}; \mathcal{B}|\mathcal{C})$  for  $\mathcal{B}_n \uparrow \mathcal{B}$ ;
- (v)  $I(\mathcal{A}; \mathcal{B}|\mathcal{C}) = I(\mathcal{A}; \sigma(\mathcal{B})|\mathcal{C})$  and  $I(\mathcal{A}; \mathcal{B}|\mathcal{C}) = I(\mathcal{A}; \mathcal{B}|\sigma(\mathcal{C}))$ ;
- (vi)  $I(\mathcal{A}; \mathcal{B} \wedge \mathcal{C}|\mathcal{D}) = I(\mathcal{A}; \mathcal{C}|\mathcal{D}) + I(\mathcal{A}; \mathcal{B}|\mathcal{C} \wedge \mathcal{D})$ ;
- (vii)  $\bar{H}(\mathcal{A}) = 0$  if and only if  $\mathcal{A}$  is trivial;
- (viii)  $\bar{H}(\mathcal{A}) = \infty$  if  $\mathcal{A}$  is a nonatomic  $\sigma$ -field;
- (ix)  $\bar{H}(\mathcal{A}|\mathcal{B}_1) \geq \bar{H}(\mathcal{A}|\mathcal{B}_2)$  if  $\mathcal{B}_1 \subset \mathcal{B}_2$ ;
- (x)  $\bar{H}(\mathcal{A}|\mathcal{B}_n) \downarrow \bar{H}(\mathcal{A}|\mathcal{B})$  for  $\mathcal{B}_n \uparrow \mathcal{B}$  and finite  $\mathcal{A}$ ;
- (xi) if  $\mathcal{B}$  is complete then  $\bar{H}(\mathcal{A}|\mathcal{B}) = 0$  if and only if  $\mathcal{A} \subset \mathcal{B}$ .

In the case of finite fields, properties (i), (ii), (iii), (vi), (vii), (ix), and (xi) reduce to well known identities for finitely-valued variables (cf. e.g. [14, 51]). For example, (vi) implies  $\bar{H}(X) = I(X; Y) + \bar{H}(X|Y)$ .

**Proof:** Properties (i), (ii), (iii), and (vii) follow trivially from the analogical properties for finite fields. Property (iv) holds since every finite partition of  $\mathcal{B} = \bigcup_{n \in \mathbb{N}} \mathcal{B}_n$  is a partition of  $\mathcal{B}_m$  for almost all  $m$ . Property (ix) follows by the analogous inequality for finite  $\mathcal{A}$ , cf. e.g. [6, Identity (C<sub>3</sub>) in Section 12]. Property (x) was proved verbatim in [6, Theorem 12.1]. Here are the proofs of the less obvious results:

- (v) Equality  $I(\mathcal{A}; \mathcal{B}|\mathcal{C}) = I(\mathcal{A}; \mathcal{B}|\sigma(\mathcal{C}))$  is obvious in view of the almost sure equality  $P(E|\mathcal{C}) = P(E|\sigma(\mathcal{C}))$ . It remains to justify  $I(\mathcal{A}; \mathcal{B}|\mathcal{C}) = I(\mathcal{A}; \sigma(\mathcal{B})|\mathcal{C})$ . Our proof is an adaptation of the proof in case of  $\mathcal{C} = \{\emptyset, \Omega\}$  given by [20, Section 2.2]. Fix a finite field  $\mathcal{A}_1$  and  $\epsilon > 0$ . Consider the smallest  $\sigma$ -field  $\sigma_0(\mathcal{B}) \supset \mathcal{B}$ . Dobrushin [20, Eq 2.2.10] proved that for any finite field  $\mathcal{B}_2 \subset \sigma_0(\mathcal{B})$  there exists a finite field  $\mathcal{B}_1 \subset \mathcal{B}$  such that  $I(\mathcal{A}_1; \mathcal{B}_1) \geq I(\mathcal{A}_1; \mathcal{B}_2) - \epsilon$ . In fact, the proposition remains true also for any  $\mathcal{B}_2 \subset \sigma(\mathcal{B})$ . (Since there exists a finite field  $\mathcal{B}'_2 \subset \sigma_0(\mathcal{B})$  and mapping  $f : \mathcal{B}_2 \rightarrow \mathcal{B}'_2$  such that  $P(B \ominus f(B)) = 0$  for all  $B \in \mathcal{B}_2$ .)

Now let us extend the result in [20] to conditional mutual information. First, define  $I(\mathcal{A}_1; \mathcal{B}|\mathcal{C})$  by (5). Consider a finite field  $\mathcal{C}_n \subset \mathcal{C}$  satisfying (3). By Dobrushin's result, for almost each  $\omega \in \Omega$  there exists a finite field  $\mathcal{B}_\omega \subset \mathcal{B}$  such that  $I(\mathcal{A}_1; \mathcal{B}_\omega|\mathcal{C}_n)(\omega) \geq I(\mathcal{A}_1; \mathcal{B}_2|\mathcal{C}_n)(\omega) - \epsilon$ . For some version of conditional probability and  $\mathcal{B}_\omega$ , random variable  $\omega \mapsto \mathcal{B}_\omega$  is  $\mathcal{C}_n$ -measurable and then  $\mathcal{B}_1 := \bigwedge_{\omega \in \Omega} \mathcal{B}_\omega$  is a finite field. By data-processing inequality (iii),  $\mathcal{B}_1$  satisfies  $I(\mathcal{A}_1; \mathcal{B}_1|\mathcal{C}_n) \geq I(\mathcal{A}_1; \mathcal{B}_\omega|\mathcal{C}_n) \geq I(\mathcal{A}_1; \mathcal{B}_2|\mathcal{C}_n) - \epsilon$  almost surely and thus  $I(\mathcal{A}_1; \mathcal{B}_1|\mathcal{C}_n) \geq I(\mathcal{A}_1; \mathcal{B}_2|\mathcal{C}_n) - \epsilon$ . Taking  $n$  sufficiently large and  $\epsilon$  sufficiently small, by  $\lim_n I(\mathcal{A}_1; \mathcal{B}|\mathcal{C}_n) = I(\mathcal{A}_1; \mathcal{B}|\mathcal{C})$ , we infer that for every  $\delta > 0$  and  $\mathcal{B}_2 \subset \sigma(\mathcal{B})$  there exists finite field  $\mathcal{B}_1 \subset \mathcal{B}$  such that  $I(\mathcal{A}_1; \mathcal{B}_1|\mathcal{C}) \geq I(\mathcal{A}_1; \mathcal{B}_2|\mathcal{C}) - \delta$ . Considering all possible  $\delta > 0$ ,  $\mathcal{A}_1 \subset \mathcal{A}$ , and  $\mathcal{B}_2 \subset \sigma(\mathcal{B})$ , we obtain  $I(\mathcal{A}; \mathcal{B}|\mathcal{C}) = I(\mathcal{A}; \sigma(\mathcal{B})|\mathcal{C})$ .

- (vi) Let  $\mathcal{A}$  and  $\mathcal{B}$  be finite fields and let  $\mathcal{C}$  be any field. Let  $\mathcal{C}_n \subset \mathcal{C}$  be finite fields satisfying  $I(\mathcal{A}; \mathcal{B} \wedge \mathcal{C}) - I(\mathcal{A}; \mathcal{B} \wedge \mathcal{C}_n) \leq 1/n$ ,  $I(\mathcal{A}; \mathcal{C}) - I(\mathcal{A}; \mathcal{C}_n) \leq 1/n$ , and (3). The latter requirement implies  $\lim_n I(\mathcal{A}; \mathcal{B}|\mathcal{C}_n) = I(\mathcal{A}; \mathcal{B}|\mathcal{C})$ . Thus, well known equalities  $I(\mathcal{A}; \mathcal{B} \wedge \mathcal{C}_n) = I(\mathcal{A}; \mathcal{C}_n) + I(\mathcal{A}; \mathcal{B}|\mathcal{C}_n)$  for finite  $\mathcal{A}$ ,  $\mathcal{B}$ , and  $\mathcal{C}_n$  [14, 51] imply

$$I(\mathcal{A}; \mathcal{B} \wedge \mathcal{C}) = I(\mathcal{A}; \mathcal{C}) + I(\mathcal{A}; \mathcal{B}|\mathcal{C}) \quad (6)$$

By (iv-v), we may extend (6) to any  $\mathcal{A}$ ,  $\mathcal{B}$ , and  $\mathcal{C}$ . Assume finite  $\mathcal{A}$  again. By (6) we also have

$$\begin{aligned} 0 &= - [I(\mathcal{A}; \mathcal{B} \wedge \mathcal{C} \wedge \mathcal{D}) - I(\mathcal{A}; \mathcal{D}) - I(\mathcal{A}; \mathcal{B} \wedge \mathcal{C}|\mathcal{D})] \\ &\quad + [I(\mathcal{A}; \mathcal{C} \wedge \mathcal{D}) - I(\mathcal{A}; \mathcal{D}) - I(\mathcal{A}; \mathcal{C}|\mathcal{D})] \\ &\quad + [I(\mathcal{A}; \mathcal{B} \wedge \mathcal{C} \wedge \mathcal{D}) - I(\mathcal{A}; \mathcal{C} \wedge \mathcal{D}) - I(\mathcal{A}; \mathcal{B}|\mathcal{C} \wedge \mathcal{D})] \\ &= I(\mathcal{A}; \mathcal{B} \wedge \mathcal{C}|\mathcal{D}) - I(\mathcal{A}; \mathcal{C}|\mathcal{D}) - I(\mathcal{A}; \mathcal{B}|\mathcal{C} \wedge \mathcal{D}), \end{aligned}$$

where all expressions are finite. Having established (vi) for finite  $\mathcal{A}$ , we generalize it to infinite  $\mathcal{A}$ , using (iv-v) again.

- (viii) It is known [7, Exercise 2.17(d)] that for any  $n \in \mathbb{N}$  nonatomic  $\sigma$ -field  $\mathcal{A}$  can be partitioned into sets  $A_1, A_2, \dots, A_n \in \mathcal{A}$  such that  $P(A_i) = 1/n$ . Hence,  $\bar{H}(\mathcal{A}) \geq \bar{H}([A_1, \dots, A_n]) = -\sum_i P(A_i) \log P(A_i) = \log n$ , and thus  $\bar{H}(\mathcal{A}) = \infty$ .
- (xi) Equality  $\bar{H}(\mathcal{A}|\mathcal{B}) = 0$  is equivalent to  $P(A|\mathcal{B}) \in \{0, 1\}$  almost surely for all  $A \in \mathcal{A}$ . Now we will prove that  $P(A|\mathcal{B}) \in \{0, 1\}$  holds if and only if  $A \in \mathcal{B}$ . First, for  $P(A|\mathcal{B}) \in \{0, 1\}$ , we construct set  $B := \{\omega \in \Omega : P(A|\mathcal{B}) = 1\} \in \mathcal{B}$ . By the definition of conditional probability [7, Section 33] and that of  $B$ , we have  $P(A) = \int P(A|\mathcal{G})dP = \int_B P(A|\mathcal{G})dP = P(A \cap B) = P(B)$ . Thus  $P(A \ominus B) = 0$  and hence  $A \in \mathcal{B}$  by completeness of  $\mathcal{B}$ . To prove the converse, notice that for  $A \in \mathcal{B}$  some version of  $P(A|\mathcal{B})(\omega)$  equals 1 for  $\omega \in A$  and 0 for  $\omega \notin A$  [7, Example 33.3].

□

If all terms in equation (6) are finite, then CMI defined in (2) coincides with CMI defined in [20, Eq. 2.7.10] by means of conditional product measure [20, Eq. 2.7.7]. For the latter concept of CMI, additivity (6) was established independently [20, Eq. 2.7.2]. The approach to  $I(\mathcal{A}; \mathcal{B}|\mathcal{C})$  presented in [20, Section 2.7] is restricted, however, to the case when the “diagonal” measure  $P^{(3)} : \sigma(\mathcal{A} \times \mathcal{B} \times \mathcal{C}) \ni E \mapsto P(\{\omega \in \Omega : (\omega, \omega, \omega) \in E\})$  is absolutely continuous w.r.t. some arbitrary product measure, cf. [17, Chapter 1]. Such approach is indeed very insightful for Gaussian processes, cf. [17, Chapters 8–10], and properties like (iv)–(vi) and (viii) can be proved independently using very different proofs [17, Theorems 1.17, 1.25, and 3.1]. Nevertheless, Theorem 1 and the definition given in (2) are much more concise and free of cumbersome assumptions than the previous approaches.

Despite the general continuity property (iv), there are examples of infinite  $\mathcal{A}$  such that  $\lim_n \bar{H}(\mathcal{A}|\mathcal{B}_n) \neq \bar{H}(\mathcal{A}|\mathcal{B})$  for  $\mathcal{B}_n \uparrow \mathcal{B}$ . Consider a discrete variable  $X$  taking values in natural numbers with  $\bar{H}(X) = \infty$ . Set  $Y_k = 1$  for  $X \geq k$  and  $Y_k = 0$  else. We have  $\bar{H}(X) = \bar{H}(X, Y_{1:n}) = \bar{H}(X|Y_{1:n}) + \bar{H}(Y_{1:n})$  and so  $\bar{H}(X|Y_{1:n}) = \infty$  since  $\bar{H}(Y_{1:n}) \leq n \log 2$ . Nevertheless,  $X$  is a measurable function of  $(Y_n)_{n \in \mathbb{N}}$  so  $\bar{H}(X|(Y_n)_{n \in \mathbb{N}}) = 0$  by property (xi).

On the other hand, the zero value of conditional entropy for finite fields can be related to convergence of finitely-valued variables. The main use of the following lemma is to switch conveniently between the language of random variables and the language of fields. We apply it in section 4.

**Lemma 2** *Let  $X$  be a finitely-valued variable. Consider fields  $\mathcal{Y}_n \uparrow \mathcal{Y}$ . The following statements are equivalent:*

- (i)  $\lim_n P(X = X_n) = 1$  for some  $\mathcal{Y}_n$ -measurable finitely-valued variables  $X_n$ ;
- (ii)  $\lim_n \bar{H}(X|\mathcal{Y}_n) = 0$ ;
- (iii)  $\bar{H}(X|\mathcal{Y}) = 0$ ;
- (iv)  $X$  is  $\sigma(\mathcal{Y})$ -measurable;

**Proof:** Statements (ii) and (iii) are equivalent by Theorem 1(ix), while (iii) and (iv) are equivalent by Theorem 1(xi). It remains to prove that (i) is equivalent to (ii). Without loss of generality, we shall assume that  $X$  takes values in  $\{1, 2, \dots, N\}$ . Let us also define  $\eta(p) := -p \log p - (1-p) \log(1-p)$  for  $p \in (0, 1)$  and  $\eta(p) := 0$  for  $p \in \{0, 1\}$ .

It is obvious that condition (ii) follows from (i) by Fano inequality  $\bar{H}(X|\mathcal{Y}_n) \leq \bar{H}(X|X_n) \leq \eta(P(X = X_n)) + [1 - P(X = X_n)] \log(N-1)$  [51, Theorem 2.47]. To prove the converse, define the value of random variable  $X_n$  as the smallest  $x$  such that  $P(X = x|\mathcal{Y}_n) \geq P(X = x'|\mathcal{Y}_n)$  for  $x' = 1, 2, \dots, N$ . We have  $P(X = X_n|\mathcal{Y}_n) \geq 1/N$ . By concavity of  $\eta$ ,

$$\eta(p) \geq \eta(q) \frac{1-p}{1-q} + \eta(1) \frac{p-q}{1-q} = \eta(q) \frac{1-p}{1-q}$$

for  $p \in [q, 1]$ . In particular,

$$\begin{aligned} \bar{H}(X|\mathcal{Y}_n) &= \bar{H}(X, X_n|\mathcal{Y}_n) \geq \int \eta(P(X = X_n|\mathcal{Y}_n)) dP \\ &\geq \int \frac{\eta(1/N)}{1-1/N} \cdot [1 - P(X = X_n|\mathcal{Y}_n)] dP = \frac{\eta(1/N)}{1-1/N} \cdot [1 - P(X = X_n)]. \end{aligned}$$

Hence (ii) implies (i). □

The concept of entropy and CMI for set algebras is slightly simpler than similar notions defined for densities of random variables. Working directly with subfields on  $(\Omega, \mathcal{J}, P)$ , we escape problems concerning measurability of random variables and their functions. These problems are artificial since CMI for random variables is invariant w.r.t. their measurable bijections. Speaking about information measures, we can totally abstract from the values of variables.

### 3 Excess entropy and ergodic decomposition

Consider a process  $(X_k)_{k \in \mathbb{Z}}$  on  $(\Omega, \mathcal{J}, P)$ , where  $X_i : (\Omega, \mathcal{J}) \rightarrow (\mathbb{X}, \mathcal{X})$ . Set  $\mathcal{G}_{m:n} \subset \mathcal{J}$  as the smallest  $\sigma$ -fields w.r.t. which blocks  $X_{m:n} := (X_k)_{m \leq k \leq n}$  are measurable. Analogously we also use  $\mathcal{G}_i := \mathcal{G}_{i:i}$ . Let  $\mathcal{G}_{-\infty} := \bigcap_{n < 0} \mathcal{G}_{-\infty:n}$  and  $\mathcal{G}_{\infty} := \bigcap_{n > 0} \mathcal{G}_{n:\infty}$  be the tail  $\sigma$ -fields of past and future. For any field  $\mathcal{F} \subset \sigma(\mathcal{G}_{-\infty}) \cap \sigma(\mathcal{G}_{\infty})$ , we have

$$\bar{H}(\mathcal{G}_1 | \mathcal{G}_{-\infty:0}) = \bar{H}(\mathcal{G}_1 | \mathcal{G}_{-\infty:0} \wedge \mathcal{F}), \quad (7)$$

$$\begin{aligned} I(\mathcal{G}_{-\infty:0}; \mathcal{G}_{1:\infty}) &= I(\mathcal{G}_{-\infty:0}; \mathcal{G}_{1:\infty} \wedge \mathcal{F}) = I(\mathcal{G}_{-\infty:0}; \mathcal{F}) + I(\mathcal{G}_{-\infty:0}; \mathcal{G}_{1:\infty} | \mathcal{F}) \\ &= \bar{H}(\mathcal{F}) + I(\mathcal{G}_{-\infty:0}; \mathcal{G}_{1:\infty} | \mathcal{F}) \end{aligned} \quad (8)$$

in view of properties (ii), (iii), (v), and (vi) proved in Theorem 1.

Assume that  $(X_k)_{k \in \mathbb{Z}}$  is stationary. Then  $I(\mathcal{G}_{-\infty:0}; \mathcal{G}_{1:\infty})$  equals excess entropy  $E = \lim_n E(n) = \lim_n I(\mathcal{G}_{-n:0}; \mathcal{G}_{1:n})$  by Theorem 1(iv-v). Moreover, if the variable range  $\mathbb{X}$  is finite then  $\bar{H}(\mathcal{G}_1 | \mathcal{G}_{-\infty:0})$  is entropy rate  $h$  by Theorem 1(x) and equality

$$\lim_{n \rightarrow \infty} \bar{H}(X_1 | X_{-n:0}) = \lim_{n \rightarrow \infty} \bar{H}(X_{1:n})/n,$$

cf. e.g. [51, Section 2.9]. We shall interpret the right-hand sides of equations (7) and (8) likewise using ergodic decomposition.

Several statements of ergodic theorem and ergodic decomposition may be compiled into a single formal proposition. In the following,  $\sigma$ -field  $\mathcal{X}$  is called countably generated if  $\mathcal{X}$  is the smallest  $\sigma$ -field containing some countable field  $\mathcal{A}$ .

**Theorem 2** *Consider a product measurable space  $(\mathbb{U}, \mathcal{U}) = \times_{k \in \mathbb{Z}} (\mathbb{X}, \mathcal{X})$  of doubly infinite sequences, where  $\mathcal{X}$  is countably generated. For 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 invariant  $\sigma$ -field  $\mathcal{I} := \{A \in \mathcal{U} : TA = A\}$ . Let  $\mathbb{S}$  be the set of all stationary probability measures on  $\mathcal{U}$  (i.e.,  $\mu \circ T = \mu$  for all  $\mu \in \mathbb{S}$ ) and let  $\mathbb{E} \subset \mathbb{S}$  be the subset of ergodic measures (i.e.,  $\mu(A) \in \{0, 1\}$  for all  $\mu \in \mathbb{E}$  and  $A \in \mathcal{I}$ ). Let  $(\mathbb{S}, \mathcal{S})$  and  $(\mathbb{E}, \mathcal{E})$  be the measurable spaces of measures induced by  $(\mathbb{U}, \mathcal{U})$ .*

Take a stationary measure  $\mu \in \mathbb{S}$ .

(i) [47, Theorem I.3.1] For any  $A \in \mathcal{U}$ ,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} I_A \circ T^k = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} I_A \circ T^{-k} = \mu(A | \mathcal{I}) \quad \mu\text{-a.s.},$$

where  $I_A(u) := 1$  for  $u \in A$  and  $I_A(u) := 0$  for  $u \notin A$ .

(ii) [38, Theorem 9.10] There exists function  $\phi : \mathcal{U} \times \mathbb{U} \rightarrow \mathbb{R}$  such that  $\phi(A, \cdot) = \mu(A | \mathcal{I})$  for all  $A \in \mathcal{U}$   $\mu$ -almost surely and  $\phi(\cdot, u) \in \mathbb{E}$  for all  $u \in \mathbb{U}$ .

(iii) [47, Theorem I.4.10] We have

$$\mu(A) = \int \phi(A, \cdot) d\mu = \int \mu_E(A) d\nu(\mu_E) \quad \text{for all } A \in \mathcal{U}, \quad (9)$$

where  $\nu$  is a measure on  $\mathcal{E}$  defined by

$$\nu(W) = \mu(\{u \in \mathbb{U} : \phi(\cdot, u) \in W\}) \quad (10)$$

(iv) [38, Theorem 9.12] For any measure  $\nu$  on  $\mathcal{E}$  satisfying (9) we have

$$\mu_E(\{u \in \mathbb{U} : \phi(\cdot, u) = \mu_E\}) = 1 \quad \text{for } \nu\text{-almost all } \mu_E \in \mathbb{E} \quad (11)$$

and consequently  $\nu$  is given by (10).

Equation (11), omitted in weaker statements of ergodic decomposition, is crucial to prove the uniqueness of  $\nu$  in (iv) and it is very useful in the discussion of universal codes, cf. [30, 50, 17].

Stationary process  $(X_k)_{k \in \mathbb{Z}}$  is said to be countably generated if it is measurable w.r.t. a countably generated  $\sigma$ -field. Let us apply Theorem 2 to such process, assuming that it has distribution  $\mu = P((X_k)_{k \in \mathbb{Z}} \in \cdot) \in \mathbb{S}$ . Set  $\mathcal{G}_I := (X_k)_{k \in \mathbb{Z}}^{-1}(\mathcal{I})$  and define random ergodic measure

$$F := \phi(\cdot, (X_k)_{k \in \mathbb{Z}}).$$

The distribution of the latter is  $P(F \in W) = \nu(W)$ . Let  $\mathcal{F} \subset \mathcal{J}$  be the smallest  $\sigma$ -field such that  $F$  is  $\mathcal{F}$ -measurable.

**Lemma 3** We have  $\sigma(\mathcal{F}) = \sigma(\mathcal{G}_I) \subset \sigma(\mathcal{G}_{-\infty}) \cap \sigma(\mathcal{G}_{\infty})$ .

**Proof:** By Theorem 2(ii) and  $\mathcal{I}$ -measurability of  $\mu(A|\mathcal{I})$  for any  $A \in \mathcal{U}$ ,  $F(A)$  is  $\sigma(\mathcal{G}_I)$ -measurable. Hence  $\mathcal{F} \subset \sigma(\mathcal{G}_I)$ . On the other hand, for any  $A \in \mathcal{I}$ ,  $\mu(A|\mathcal{I}) = I_A$   $\mu$ -almost surely so, by Theorem 2(ii),  $(X_k)_{k \in \mathbb{Z}}^{-1}(A)$  is an element of the smallest complete  $\sigma$ -field w.r.t. which  $F(A)$  is measurable. Hence  $\mathcal{G}_I \subset \sigma(\mathcal{F})$ .

Fix any  $A \in \mathcal{U}_- := (X_k)_{k \in \mathbb{Z}}(\mathcal{G}_{-\infty:0})$ . By Theorem 2(i), variable  $F(A)$  is  $\sigma(\mathcal{G}_{-\infty:n})$ -measurable for all  $n < 0$ . Hence  $F(A)$  is  $\sigma(\mathcal{G}_{-\infty})$ -measurable. But for each  $B \in \mathcal{U}$  there exists a measurable real function  $f_B$  such that  $F(B) = f_B((F(A))_{A \in \mathcal{U}_-})$ . (This follows by stationarity assumption and approximation theorems [7, Theorem 11.4 and 13.4].) Hence  $F(B)$  is  $\sigma(\mathcal{G}_{-\infty})$ -measurable for all  $B \in \mathcal{U}$  and thus  $\mathcal{F} \subset \sigma(\mathcal{G}_{-\infty})$ . Analogously we prove that  $\mathcal{F} \subset \sigma(\mathcal{G}_{\infty})$ .  $\square$

By Theorem 1(vii) and Lemma 3, the process is ergodic if and only if  $\bar{H}(\mathcal{F}) = 0$ .

It is convenient to consider information measures for subfields of  $\mathcal{G}_{-\infty:\infty}$  as functions of the process distribution. For arbitrary distribution  $\mu = P((X_k)_{k \in \mathbb{Z}} \in \cdot) \in \mathbb{S}$ , introduce parameterization  $I_\mu(\mathcal{A}, \mathcal{B}) := I(\mathcal{A}, \mathcal{B})$  with  $\mathcal{A}, \mathcal{B} \subset \mathcal{G}_{-\infty:\infty}$ ,  $H_\mu(n) := H(n)$ ,  $h_\mu := h$ ,  $E_\mu := E$ , and  $E_\mu(n) := E(n)$ . By Lemma 3,  $F = P((X_k)_{k \in \mathbb{Z}} \in \cdot | \mathcal{F})$  almost surely. Hence, for finite fields  $\mathcal{A}$  and  $\mathcal{B}$  we have

$$\mathbf{E} I_F(\mathcal{A}; \mathcal{B}) = I(\mathcal{A}; \mathcal{B} | \mathcal{F}), \quad (12)$$

where  $\mathbf{E} Y$  is the expectation of random variable  $Y$ . By the monotone convergence theorem, cf. e.g. [7, Theorem 16.2], and by Theorem 1(iv-v), we can generalize (12) to the case when  $\mathcal{A}$  and  $\mathcal{B}$  are countable fields or countably generated  $\sigma$ -fields. Therefore we can easily justify the following ergodic decomposition of entropy rate and excess entropy.

**Theorem 3** For a countably generated stationary process  $(X_k)_{k \in \mathbb{Z}}$ ,

$$h = \mathbf{E} h_F \text{ if the variable range } \mathbb{X} \text{ is finite,} \quad (13)$$

$$E = \bar{H}(F) + \mathbf{E} E_F. \quad (14)$$

Decomposition (13) was already proved in [30]. We assort it with (14) so as to show that the latter one has a simpler proof.

**Proof:** Equation (14) follows directly from Lemma 3, (8), and (12) for  $\mathcal{A} = \mathcal{G}_{-\infty:0}$  and  $\mathcal{B} = \mathcal{G}_{1:\infty}$ . Now let us justify (13). For  $M$  being the cardinality of range  $\mathbb{X}$ , set  $K = \log M$  so that  $K - \bar{H}(X_1) \geq 0$ . By monotone convergence theorem and (7),

$$\begin{aligned} \mathbf{E} [K - h_F] &= \mathbf{E} \left[ K - \lim_{n \rightarrow \infty} \bar{H}_F(X_1 | X_{-n:0}) \right] = \lim_{n \rightarrow \infty} \mathbf{E} [K - \bar{H}_F(X_1 | X_{-n:0})] \\ &= \lim_{n \rightarrow \infty} [K - \bar{H}(\mathcal{G}_1 | \mathcal{G}_{-n:0} \wedge \mathcal{F})] = [K - \bar{H}(\mathcal{G}_1 | \mathcal{G}_{-\infty:0})] \\ &= K - h. \end{aligned}$$

Hence the thesis follows.  $\square$

## 4 Uncountable description processes

The ergodic decomposition of excess entropy proved in Theorem 3 may help to discuss the construction of processes with infinite excess entropy. It is obvious that equality  $E = \infty$  can be a result of  $\bar{H}(\mathcal{F}) = \infty$  or  $\mathbf{E} E_F = \infty$ . By Theorem 1(viii), the first equality holds in particular if the process is nonergodic and  $\mathcal{F}$  is nonatomic. Let us give a simple characterization of these processes. By Bernoulli process we understand the fair-coin binary process  $(Z_n)_{n \in \mathbb{N}} \sim \text{IID}$  with  $P(Z_k = 0) = P(Z_k = 1) = 1/2$ .

**Definition 1** Stationary process  $(X_i)_{i \in \mathbb{Z}}$  is called an uncountable description process (UDP) if there exist functions  $f_{nk}$  and Bernoulli process  $(Z_n)_{n \in \mathbb{N}}$  such that

$$(i) \ P(f_{nk}(X_{i+1:i+n}) = Z_k) \text{ does not depend on } i \in \mathbb{Z},$$

$$(ii) \ \lim_n P(f_{nk}(X_{1:n}) = Z_k) = 1.$$

**Theorem 4** Countably generated stationary process  $(X_i)_{i \in \mathbb{Z}}$  is UDP if and only if  $\mathcal{F}$  contains a nonatomic sub- $\sigma$ -field.

Hence  $E = \infty$  for every UDP.

**Proof:** Assume first that  $(X_i)_{i \in \mathbb{Z}}$  is UDP. By Lemma 2, (i) and (ii) imply that for each  $i \in \mathbb{Z}$  variable  $Z_k$  is  $\sigma(\mathcal{G}_{i:\infty})$ -measurable. Furthermore, by (i) and (ii), there exists a single measurable function  $f$  of the half-infinite future such that  $f(X_{i:\infty}) = Z_k$  for all  $i$  almost surely. Hence  $P(A) = 1$  for event  $A := \bigcap_{i \in \mathbb{Z}} (f(X_{i:\infty}) = f(X_{i+1:\infty}))$ . But the set of infinite sequences  $(X_i)_{i \in \mathbb{Z}}(A)$  is shift invariant so  $A \in \mathcal{G}_I \subset \sigma(\mathcal{F})$  according to Lemma 3. Therefore  $Z_k$  are  $\mathcal{F}$ -measurable for all  $k$ . Construct real variable  $Y = \sum_{n \in \mathbb{N}} 2^{-n} Z_n$ . The distribution of  $Y$  is Lebesgue measure on interval  $[0, 1]$ . Lebesgue measure is nonatomic [7, Exercise 2.17(a)] so  $\mathcal{F}$  contains a nonatomic sub- $\sigma$ -field by  $\mathcal{F}$ -measurability of  $Y$ .

As for the converse, take  $(X_i)_{i \in \mathbb{Z}}$  with nonatomic  $\mathcal{F}_0 \subset \mathcal{F}$ . For any  $A \in \mathcal{F}_0$  and  $x \in [0, P(A)]$  there exists  $B \in \mathcal{F}_0$  such that  $B \subset A$  and  $P(B) = x$  [7, Exercise 2.17(c)]. Obviously,

this property can be used to define a family of nested sets  $A_w \in \mathcal{F}_0$  indexed by binary words  $w \in \{0, 1\}^*$  such that  $A_\lambda = \Omega$  for the empty word  $\lambda$ ,  $A_{wa} \subset A_w$ , and  $P(A_{w0}) = P(A_{w1}) = P(A_w)/2$ . For each  $k \in \mathbb{N}$  define  $Z_k$  as the characteristic function of set  $B_k = \bigcup_{w \in \{0,1\}^k} A_{w0}$ . Sequence  $(Z_n)_{n \in \mathbb{N}}$  is a Bernoulli process. By Lemma 3,  $Z_k$  are  $\sigma(\mathcal{G}_{1:\infty})$ -measurable. Hence by Lemma 2,  $\lim_n P(f_{nk}(X_{1:n}) = Z_k) = 1$  for some functions  $f_{nk}$ . Stationarity of  $(X_i)_{i \in \mathbb{Z}}$  and  $\mathcal{F}$ -measurability of  $Z_k$  imply that probabilities  $P(f_{nk}(X_{i+1:i+n}) = Z_k)$  do not depend on  $i \in \mathbb{Z}$ . So  $(X_i)_{i \in \mathbb{Z}}$  is UDP.  $\square$

Definition 1 allows to construct examples of uncountable description processes without delving into measure theory. One of the simplest UDP's is the process of form

$$X_i := (N_i, Z_{N_i}), \quad (15)$$

where  $(Z_n)_{n \in \mathbb{N}} \perp\!\!\!\perp (N_i)_{i \in \mathbb{Z}}$  and  $(N_i)_{i \in \mathbb{Z}}$  is any ergodic stationary process assuming values in natural numbers so that  $P(N_i = n) > 0$  for all  $n \in \mathbb{N}$  (e.g. we may take  $(N_i)_{i \in \mathbb{Z}} \sim \text{IID}$ ). It is easy to prove that (15) is UDP, setting functions

$$f_{lk}(x_{1:l}) = \begin{cases} 0 & \text{if } N_{k0} > N_{k1}, \\ 1 & \text{if } N_{k1} > N_{k0}, \\ 2 & \text{else,} \end{cases} \quad (16)$$

where  $N_{nz}(x_{1:l})$  is the number of  $j \in \{1, \dots, l\}$  such that  $x_j = (n, z)$ . Then it suffices to apply the ergodic theorem, i.e. Theorem 2(i) here, and the fact that almost sure convergence implies convergence in probability (cf. e.g. [7, Theorem 25.2]).

Our motivation for name ‘‘uncountable description processes’’ was as follows: One can imagine that the realization of a UDP attempts at describing the state of a random object  $(Z_n)_{n \in \mathbb{N}}$  tossed ‘‘prior’’ to tossing the UDP (cf. a similar utterance in [30, page 625]). On the other hand, object  $(Z_n)_{n \in \mathbb{N}}$ , isomorphic to a continuous real random variable  $Y = \sum_{n \in \mathbb{N}} 2^{-n} Z_n$ , has an uncountable number of states.

We suppose that some UDP's may be useful for probabilistic modeling of texts written in natural language since the process consisting of  $X_i$  as in (15) can be given a quasi-linguistic interpretation. Idealizing  $(X_i)_{i \in \mathbb{N}}$  as a stationary process, imagine that  $(X_i)_{i \in \mathbb{N}}$  is a sequence of consecutive statements extracted from a collection of texts describing *coherently* some random state of affairs  $(Z_n)_{n \in \mathbb{N}}$ . Each statement of form  $X_i = (n, z)$  asserts that the value of a random  $n$ -th bit of the state of affairs is  $z$ , i.e., it affirms that  $Z_n = z$  in such way that both bit address  $n$  and its value  $z$  can be identified. For two statements  $X_i = (n, z)$  and  $X_j = (n', z')$ , it is not set in advance which bits they describe and what values they assert but if they describe bits of the same address ( $n = n'$ ) then they must assert the same bit value ( $z = z'$ ).

Keeping the interpretation of variables  $X_i$  as statements describing random bits of  $(Z_n)_{n \in \mathbb{N}}$ , we can prove that  $(X_i)_{i \in \mathbb{Z}}$  is UDP also in some cases when statements  $X_i = (n, z)$  are not necessarily true, i.e., when  $Z_n \neq z$  with a positive probability. For processes  $(Z_n)_{n \in \mathbb{N}}$  and  $(N_i)_{i \in \mathbb{Z}}$  as previously, extend the probability space with an ergodic process  $(U_i)_{i \in \mathbb{Z}}$  such that  $P(U_i = 1) > 1/2$  and  $(U_i)_{i \in \mathbb{Z}} \perp\!\!\!\perp (Z_n)_{n \in \mathbb{N}}, (N_i)_{i \in \mathbb{Z}}$ . We claim that  $(X_i)_{i \in \mathbb{Z}}$  is UDP for

$$X_i := \begin{cases} (N_i, Z_{N_i}) & \text{if } U_i = 1, \\ (N_i, 1 - Z_{N_i}) & \text{if } U_i = 0. \end{cases}$$

Analogously to the previous example, the proof can be easily established for functions  $f_{lk}$  like in (16).

The presented examples of UDP's cannot be called satisfactory models of human text generation. Nevertheless, the general notion of UDP seems to capture the informal idea of a collection of texts from which it is theoretically possible to extract an infinite number of independent statements about a state of affairs which is both random and persistent. Concerning the state of affairs described by a collection of fiction books, one can debate about the practical feasibility of its extraction, its size, and its ontological status. Yet there exist a very precise, infinite, (algorithmically) random, and objective (in particular, persistent) state of affairs which, in some naive imagination, seems to be continually described in the collection of texts treating on mathematics. It is Chaitin's constant  $\Omega$ , i.e., halting probability.

Exactly, halting probability with respect to a chosen universal prefix Turing machine is

$$\Omega = \sum_{x \in P} 2^{-|x|},$$

where  $|x|$  is the length of binary string  $x$  and  $P \subset \{0, 1\}^*$  is the subset of input binary strings for which the machine halts (cf. e.g. [11], [41, Section 3.6.2], [21]). Number  $\Omega \in (0, 1)$  is incompressible (algorithmically random) and encodes compactly all facts in axiomatic mathematical theories. Consider its binary expansion  $\Omega = \sum_{k \in \mathbb{N}} 2^{-k} \omega_k$ ,  $\omega_k \in \{0, 1\}$ . Given first  $n$  bits  $\omega_{1:n} := (\omega_k)_{1 \leq k \leq n}$  and a mathematical theory which can be written down in less than  $n$  bits, it is possible to compute for every statement in the theory whether it is true, false, or independent of the axioms of the theory [41, page 218]. Conversely, we might naively suppose that the more mathematical texts we scan the more bits of  $\Omega$  we may be able to determine.

By analogy to  $(X_i)_{i \in \mathbb{Z}}$  given by (15), construct process  $(Y_i)_{i \in \mathbb{Z}}$  with

$$Y_i := (N_i, \omega_{N_i}).$$

For the same ergodic process  $(N_i)_{i \in \mathbb{Z}}$ ,  $(X_i)_{i \in \mathbb{Z}}$  is UDP and  $(Y_i)_{i \in \mathbb{Z}}$  is ergodic. Processes  $(X_i)_{i \in \mathbb{Z}}$  and  $(Y_i)_{i \in \mathbb{Z}}$  are very different but in view of Martin-Löf randomness of  $\Omega$  [41, Exercise 3.6.8], we suppose that certain laws of randomness which hold for  $(X_i)_{i \in \mathbb{Z}}$  hold also for  $(Y_i)_{i \in \mathbb{Z}}$ . A typical realization of process  $(Y_i)_{i \in \mathbb{Z}}$  may be practically impossible to distinguish from a typical realization of process  $(X_i)_{i \in \mathbb{Z}}$ . In particular, asymptotic behavior of recursive (effectively computable) codes for these two processes may be the same. We hope that this hypothesis could be formalized, generalized, and proved so that one could use some UDP's conveniently in the analysis of universal source coding for certain ergodic processes.

## 5 Excess lengths of universal codes

It is known that entropy rate of an ergodic finitely-valued process can be estimated via the length of a universal code for the process realization [16, 31]. (This is not true for some infinitely-valued processes [35].) One can ask if excess entropy can be estimated via universal codes in a similar manner.

Some partial positive answer is known for recursive probability distributions (i.e. distributions which are effectively computable). For stationary process  $(X_i)_{i \in \mathbb{Z}}$  define probability distribution of form  $\mathbf{P} : \mathbb{X}^* \ni w \mapsto P(X_{i:i+|w|-1} = w)$ . Let  $K(w)$  be algorithmic prefix complexity of string  $w \in \mathbb{X}^*$  [41, Section 3.1, page 194]. Analogically, if  $\mathbf{P}$  is a recursive real function [41, Section 1.7.3, page 53] then let  $K(\mathbf{P})$  be algorithmic prefix complexity of  $\mathbf{P}$ . Algorithmic prefix complexity of the restriction of  $\mathbf{P}$  to subdomain  $\mathbb{X}^n$  is less than

$K(\mathbf{P}) + K(n) + O(1)$ , where  $K(n)$  is algorithmic prefix complexity of natural number  $n \in \mathbb{N}$ . Hence, we have

$$0 \leq \mathbf{E} K(X_{1:n}) - H(n) \leq K(\mathbf{P}) + K(n) + O(1) \quad (17)$$

in view of [41, Theorem 8.1.1]. For strings  $u, v \in \mathbb{X}^*$  define algorithmic mutual information

$$I(u : v) := K(u) + K(v) - K(uv),$$

cf. [41, Definition 3.9.1]. By (17) there is

$$\lim_{n \rightarrow \infty} \mathbf{E} K(X_{1:n})/n = h, \quad (18)$$

$$\lim_{n \rightarrow \infty} \frac{\mathbf{E} I(X_{1:n} : X_{n+1:2n}) - E(n)}{\log^\alpha n} = 0, \quad \alpha > 1, \quad (19)$$

because of inequality  $K(n) + O(1) \leq O(1) \cdot \log n$ , cf. [34, Theorem 5.3]. That is, the function computing the shortest program which generates the argument is a weakly universal code w.r.t. the class of recursive probability distributions. Moreover, we can bound finite-order excess entropies  $E(n)$ , up to a logarithmically growing factor, by the expectations of algorithmic mutual information  $I(X_{1:n} : X_{n+1:2n})$ .

In applications of probability modeling such as constructing probabilistic theories of physical or linguistic reality, we cannot confine to using mere effectively computable distributions. There is no guarantee that the adequate description of the real world does not transcend our finite models of computation. Neither it is possible to use algorithmic complexity, which is not a recursive function, to estimate experimentally finite-order excess entropies predicted by some probabilistic theories or hypotheses. Thus, it is of great scientific interest to answer if (19) can be generalized to any stationary probability distribution with the length of some *efficiently* computable code substituted for  $K(\cdot)$ . By efficiently computable codes we do not mean only codes with  $O(n \log n)$ -time complexity, useful in practical data compression. For verifying hypotheses on excess entropy of human texts, we would be satisfied to use a much slower computable code, say with  $O(n^2)$ -time complexity, if only we had a guarantee, in form of a theorem, that an analog of (19) is satisfied even in the worst case.

Surely there exist codes, called weakly universal codes, which satisfy the analog of (18) for every stationary process. Some instances of universal codes are Lempel-Ziv code [52, 53, 14] and a large class of Yang-Kieffer codes based on irreducible grammar transforms [39]. Usually their universality has been asserted in form of a strong law of large numbers [14] but using (11) one can deduce the appropriate statement for the expectations, given in (20) [50, 17]. Do these codes satisfy the analog of (19)? We suppose that they do not. Since the analysis of concrete codes is very hard, we will present here only some introductory results on a hierarchy of universal codes. One of the results makes use of the ergodic decomposition of excess entropy proved in Theorem 3.

First, let us fix the terms. Consider injection  $C : \mathbb{X}^* \rightarrow \{0, 1, \dots, D\}^*$ , where  $\mathbb{X}$  is finite. We will call it a code. Define normalized code length  $K^C(w) := |C(w)| \log D$ , where  $|\cdot|$  is the usual length of string, i.e., the number of atomic symbols in it. Moreover, denote the expected length of the code as

$$H^C(n) := \mathbf{E} K^C(X_{1:n}).$$

Code  $C$  is called a weakly universal code if

$$\lim_{n \rightarrow \infty} H^C(n)/n = h \quad (20)$$

for any stationary distribution  $P((X_k)_{k \in \mathbb{Z}} \in \cdot) \in \mathbb{S}$ .

For a pair of strings  $u, v \in \mathbb{X}^*$  we shall introduce an “approximation” of algorithmic mutual information based on code  $C$ , i.e.,

$$I^C(u : v) := K^C(u) + K^C(v) - K^C(uv).$$

By analogy to finite-order excess entropy, define the expected excess length of code  $C$  as

$$E^C(n) := \mathbf{E} I^C(X_{1:n} : X_{n+1:2n}) = 2H^C(n) - H^C(2n).$$

Additionally, we will use explicit parameterization  $E_\mu^C(n) := E^C(n)$  for an arbitrary distribution  $\mu = P((X_k)_{k \in \mathbb{Z}} \in \cdot) \in \mathbb{S}$ .

A simple algebraic identity proves that  $E^C(n)$  dominates finite-order excess entropy  $E(n)$  for infinitely many  $n$ . A similar inequality arises for two codes, where the length of the first one is almost always bigger than the length of the latter one.

**Theorem 5 (cf. a less general statement in [19])** *For any function  $f : \mathbb{N} \rightarrow \mathbb{R}$  such that  $f(n) \geq 0$  for almost all  $n$  and  $\lim_k f(k)/k = 0$ , we have*

$$\limsup_{n \rightarrow \infty} [2f(n) - f(2n)] \geq 0. \quad (21)$$

**Proof:** We have identity

$$\sum_{k=0}^{m-1} [2f(2^k n) - f(2^{k+1} n)] \cdot \frac{1}{2^{k+1}} = f(n) - n \cdot \frac{f(2^m n)}{2^m n}, \quad n, m \in \mathbb{N}.$$

Hence  $f(n) - n \lim_k f(k)/k \geq 0$  implies  $\sum_{k=0}^{\infty} [2f(2^k n) - f(2^{k+1} n)]/2^{k+1} \geq 0$ . Putting  $n = 2^p M$ , we obtain

$$\sum_{k=p}^{\infty} [2f(2^k M) - f(2^{k+1} M)] \cdot \frac{1}{2^{k+1}} \geq 0, \quad M \in \mathbb{N}, \text{ almost all } p \in \mathbb{N}. \quad (22)$$

If (21) did not hold then there would be  $2f(2^k M) - f(2^{k+1} M) < 0$  for all  $k \geq p$  and some  $p$ . This stays, however, in contradiction with (22).  $\square$

**Corollary:** *For any weakly universal codes  $C$  and  $D$  we have*

$$\limsup_{n \rightarrow \infty} [E^C(n) - E^D(n)] \geq 0 \text{ if } H^C(n) \geq H^D(n) \text{ for almost all } n, \quad (23)$$

$$\limsup_{n \rightarrow \infty} [E^C(n) - E(n)] \geq 0. \quad (24)$$

*Relation (24) follows by channel inequality  $H^C(n) \geq H(n)$  (cf. e.g. [14, Theorem 5.3.1]).*

In view of Theorem 5, we can propose a heuristic rule for stating empirically driven conjectures about source coding: The smaller are the excess lengths of a code in some particular cases then the lower its redundancy may be in general. We suppose that in practical tests on finite individual strings, scaling of the excess lengths of a code can be a much more sensitive and reliable indicator of its asymptotic quality than the best achieved compression rate. Furthermore, we find it useful to introduce the following notation:

**Definition 2** A weakly universal code  $C$  is said to have order  $O(f(n))$  if

$$\lim_n \left| [E_\mu^C(n) - E_\mu(n)] / f(n) \right| < \infty \text{ for all } \mu \in \mathbb{S}.$$

For every ergodic process we can approximate each  $E(n)$ ,  $n \in \mathbb{N}$ , given the process realization. This is a straightforward corollary of the ergodic theorem. Nevertheless there exists no code of order  $O(1)$ . That result is a consequence of Theorems 3 and 5. Let  $(X_i)_{i \in \mathbb{Z}}$  be an arbitrary stationary process. We have disintegration formula

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

[7, Exercise 18.19], where  $f$  is a bounded  $\mathcal{U}$ -measurable function and  $\nu = P(F \in \cdot)$  is the distribution of the random ergodic measure  $F$  of the process, discussed in section 3. Letting  $f((X_i)_{i \in \mathbb{Z}}) := I^C(X_{1:n} : X_{n+1:2n})$  we obtain  $E^C(n) = \mathbf{E} E_F^C(n)$ . Hence formulae (24) and (14) imply

$$\limsup_{n \rightarrow \infty} \mathbf{E} E_F^C(n) = \limsup_{n \rightarrow \infty} E^C(n) \geq E = \bar{H}(F) + \mathbf{E} E_F. \quad (25)$$

It is the presence of  $\bar{H}(F)$  in (25) which implies the nonexistence of  $O(1)$ -codes.

**Theorem 6** For a weakly universal code  $C$ , let  $N(K)$  be the number of distinct stationary ergodic measures  $\mu$  such that  $\limsup_n E_\mu^C(n) \leq K$ . We have

$$\log N(K) \leq K.$$

**Proof:** Consider any  $M \in \mathbb{N}$  such that  $M \leq N(K)$ . Let  $A \subset \mathbb{E}$  be a subset of some  $M$  ergodic measures  $\mu$  such that  $\limsup_n E_\mu^C(n) \leq K$ . Construct a probability space with process  $(X_i)_{i \in \mathbb{Z}}$  such that  $P((X_i)_{i \in \mathbb{Z}} \in \cdot) = M^{-1} \sum_{\mu \in A} \mu$ . By the uniqueness of ergodic decomposition, random ergodic measure  $F$  takes the value of each  $\mu \in A$  with equal probability. Hence  $\bar{H}(F) = \log M$ . Set some  $\epsilon > 0$ . Random variables  $K + \epsilon - E_F^C(n)$ ,  $n \in \mathbb{N}$ , are almost surely nonnegative for almost all  $n$ . Thus by Fatou's lemma,  $K + \epsilon - \mathbf{E} \limsup_n E_F^C(n) \leq K + \epsilon - \limsup_n \mathbf{E} E_F^C(n)$ . Hence from inequality (25) we obtain

$$\log M = \bar{H}(F) \leq \bar{H}(F) + \mathbf{E} E_F \leq \limsup_{n \rightarrow \infty} \mathbf{E} E_F^C(n) \leq \mathbf{E} \limsup_{n \rightarrow \infty} E_F^C(n) \leq K.$$

We considered any finite  $M \leq N(K)$  so the displayed inequality implies the thesis.  $\square$

The set of ergodic measures with finite excess entropy is uncountable. This set includes for instance the measures of all hidden Markov processes (finite-state sources) [26]. On the other hand, by Theorem 6, there is only a countable number of measures  $\mu$  with finite  $\limsup_n E_\mu^C(n)$ . The gap between  $E^C(n)$  and  $E(n)$  can diverge for any kind of universal code. Some analog of this phenomenon is superlinear growth of prefix algorithmic complexity  $K(X_{1:n})$  for an IID process  $(X_i)_{i \in \mathbb{Z}}$  (cf. [1] or [41, Exercise 3.6.3(c)]). The novelty of Theorem 6 is that it concerns all universal codes (also those not effectively computable) and its proof uses tools from measure theory rather than theoretical computer science.

Although there are no codes of order  $O(1)$ , we may try to seek for weakly universal codes of order which is only slightly higher. Constructing a code with order  $O(\log^\alpha n)$  like in (19) may be very hard if not impossible. Despite some analogies between the problem of universal coding and the problem of the smallest grammar for a string [12], recently discovered

efficient algorithms constructing  $O(\log^\alpha n)$ -approximations of the smallest grammars [12, 44] probably cannot be converted into Yang-Kieffer codes of order  $O(\log^\alpha n)$  via the grammar encoder from the proof of Theorem 6 in [39].

We would be satisfied, however, to know much worse universal codes. For instance, in order to verify Hilberg hypothesis [36], which states that  $E(n) \asymp \sqrt{n}$  for human language production, we need an efficiently computable code of order  $O(n^\beta)$  where  $\beta$  is much less than  $1/2$ . Otherwise we may be hardly able to distinguish between the realizations of processes with  $E(n) \asymp \sqrt{n}$  and of processes with  $E(n) = O(1)$ . In fact, in view of [18, Eqs. (7) and (8)] we suppose that  $E^C(n) = \Omega(\sqrt{n})$  for every IID process with entropy rate  $h > 0$  and every Yang-Kieffer code  $C$  based on whatever irreducible grammar transform [39, Section 3.2] and the grammar encoder by [39]. In [18] we proposed an ad hoc grammar-based code which seemingly does not exhibit this property but, frankly speaking, we do not know any codes having *established* order  $O(n^\beta)$ .

## 6 Conclusion

We hope that our few examples show convincingly that conditional mutual information for  $\sigma$ -fields is a valuable tool in the analysis of excess entropy and excess lengths of universal codes. We would like to remark that the results reported in this paper originated as a part of our doctoral dissertation [17] inspired by and inspiring our parallel research in quantitative linguistics [19, 18]. For this article, we excerpted only the results connected to the ergodic decomposition of excess entropy. We bypassed about a half of our survey, concerning properties of excess entropy for Gaussian processes, effective construction of processes with  $E(n) \asymp n^\beta$ , and existence of ergodic process with infinite excess entropy. We would like to mention some of these omitted topics here to put our article in context.

First, the concept of excess entropy is not restricted to discrete processes. It is particularly insightful to discuss it for Gaussian processes. In such case we can bring together results in information theory and in time series analysis for weakly stationary processes. Assume that  $(X_i)_{i \in \mathbb{Z}}$  is a stationary zero-mean complex-valued Gaussian process. We have  $I(X_1; X_n) = -\log(1 - |\rho(n-1)|^2)$  and  $I(X_1; X_2 | X_{2:n-1}) = -\log(1 - |\alpha(n-1)|^2)$ , where  $\rho(i-j) = \mathbf{E}[X_i^* X_j]$  is autocorrelation (ACF) and  $\alpha(\cdot)$  is partial autocorrelation (PACF) [22]. Define  $E$  and  $E(n)$  as in section 1. Through identities as in [15], we obtain

$$E(n) = \sum_{k=2}^n (k-1) I(X_1; X_k | X_{2:k-1}) + \sum_{k=n+1}^{2n} (2n-k+1) I(X_1; X_k | X_{2:k-1}).$$

Hence, finiteness of excess entropy  $E$  can be simply expressed in terms of asymptotics of PACF rather than ACF. Namely  $E < \infty$  if and only if  $\sum_k k |\alpha(k)|^2 < \infty$ , where  $|\alpha(\cdot)| < 1$ . In fact, Grenander and Szegö [32, section 5.5] gave an integral expression for  $E$  in terms of the moving-average representation for the purely nondeterministic process [10, Theorem 5.7.1]. On the other hand, a Gaussian process has  $E < \infty$  only if it is completely nondeterministic [8]. Thus there are no nonergodic Gaussian processes with finite excess entropy [17, Theorem 10.4]. There are neither nonergodic Gaussian processes satisfying a similar condition, namely  $\sum_k |\alpha(k)| < \infty$ , where  $|\alpha(\cdot)| < 1$  [17, Theorem 11.17].

It is well known that there is a one-to-one correspondence between the ACF and the PACF [22]. Moreover, parameterization of the process in terms of PACF is unconstrained, i.e., the sole condition on  $\alpha(k)$  is  $|\alpha(k)| \leq 1$  and otherwise  $\alpha(k)$  are free to vary independently of one another [43]. Thus it is easy to prove the existence of Gaussian processes with  $E(n) \asymp n^\beta$ .

There is no comparably simple parameterization of discrete processes so the construction of discrete processes with  $E(n) \asymp n^\beta$  other than UDP's seems much harder.

In view of the ergodic decomposition of excess entropy, it is natural to ask about links between excess entropy and ergodicity. Although there are no nonergodic Gaussian processes with  $E < \infty$ , it is obvious from (14) that there exist both ergodic and nonergodic discrete processes with finite  $E$  (consider finite mixtures of ergodic finite-state source measures) as well as nonergodic measures with infinite  $E$  (consider some infinite mixtures or the examples given by [29, 3]). It is harder to construct ergodic processes with  $E = \infty$  but they exist and include some Gaussian processes (such as ARIMA(0,  $d$ , 0) [37] and probably also fractional Gaussian noise [2]), some denumerable one-class Markov chains [17, Example 2.13], and even some finitely-valued processes [9].

The above examples show that there is no link between ergodicity and the value of  $E$  for discrete processes. Nevertheless, one can ask about the links between ergodicity and the divergence rate for  $E(n)$ . Unlike the case of Gaussian processes, there is no fact of form: If  $E(n)$  grows sufficiently slow then the process must be ergodic. A simple counterexample are finite mixtures of ergodic finite-state source measures, having finite  $E$ . Thus, we should rather ask about the links between the divergence rate for  $E(n)$  and the question whether the process is UDP. We are almost certain that there are UDP's like (15) with arbitrarily slowly diverging  $E(n)$  for sufficiently fast decaying  $P(N_i > n)$  although no UDP can have  $E(n) = O(1)$ . From the perspective of potential linguistic applications, we wonder, however, if there is a general *converse* fact of form: If  $E(n) \asymp n^\beta$  for some sufficiently large  $\beta$  then the process must be UDP.

**Acknowledgments.** We would like to thank to Jan Mielniczuk for discussion and remarks.

## References

- [1] J. M. Barzdins and Freivalds. On the prediction of general recursive functions. *Soviet Mathematics. Doklady*, 13:1251–1254, 1972.
- [2] J. Beran. *Statistics for Long-Memory Processes*. New York: Chapman & Hall, 1994.
- [3] V. Berthé. Conditional entropy of some automatic sequences. *Journal of Physics A*, 27: 7993–8006, 1994.
- [4] W. Bialek, I. Nemenman, and N. Tishby. Complexity through nonextensivity. *Physica A*, 302:89–99, 2001.
- [5] W. Bialek, I. Nemenman, and N. Tishby. Predictability, complexity and learning. *Neural Computation*, 13:2409, 2001.
- [6] P. Billingsley. *Ergodic Theory and Information*. New York: Wiley, 1965.
- [7] P. Billingsley. *Probability and Measure*. New York: Wiley, 1979.
- [8] P. Bloomfield, N. P. Jewell, and E. Hayashi. Characterizations of completely nondeterministic stochastic process. *Pacific Journal of Mathematics*, 107:307–317, 1983.
- [9] R. C. Bradley. On the strong mixing and weak Bernoulli conditions. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 50:49–54, 1980.

- [10] P. J. Brockwell and R. A. Davis. *Time Series: Theory and Methods*. New York: Springer, 1987.
- [11] G. J. Chaitin. A theory of program size formally identical to information theory. *Journal of the ACM*, 22:329–340, 1975.
- [12] M. Charikar, E. Lehman, A. Lehman, D. Liu, R. Panigrahy, M. Prabhakaran, A. Sahai, and A. Shelat. The smallest grammar problem. *IEEE Transactions on Information Theory*, 51:2554–2576, 2005.
- [13] T. M. Cover, P. Gacs, and R. M. Gray. Kolmogorov’s contributions to information theory and algorithmic complexity. *Annals of Probability*, 17:840–865, 1989.
- [14] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. New York: Wiley, 1991.
- [15] J. P. Crutchfield and D. P. Feldman. Regularities unseen, randomness observed: The entropy convergence hierarchy. *Chaos*, 15:25–54, 2003.
- [16] L. D. Davisson. Universal noiseless coding. *IEEE Transactions on Information Theory*, 19:783–795, 1973.
- [17] Ł. Dębowski. *Excess entropy for stochastic processes over various alphabets*. PhD thesis, Institute of Computer Science, Polish Academy of Sciences, 2005. In Polish.
- [18] Ł. Dębowski. Menzerath’s law for the smallest grammars. In R. Köhler and P. Grzybek, editors, *The Exact Science of Language and Text*. 2006. To appear.
- [19] Ł. Dębowski. On Hilberg’s law and its links with Guiraud’s law. *Journal of Quantitative Linguistics*, 13:81–109, 2006.
- [20] R. L. Dobrushin. A general formulation of the fundamental Shannon theorems in information theory. *Uspekhi Matematicheskikh Nauk*, 14(6):3–104, 1959. In Russian.
- [21] R. G. Downey. Some recent progress in algorithmic randomness. In J. Fiala, V. Koubek, and J. Kratochvíl, editors, *Mathematical Foundations of Computer Science 2004, 29th International Symposium.*, pages 42–83. Springer, 2004.
- [22] J. Durbin. The fitting of time series models. *Review of the International Statistical Institute*, 28:233–244, 1960.
- [23] W. Ebeling and G. Nicolis. Entropy of symbolic sequences: the role of correlations. *Europhysics Letters*, 14:191–196, 1991.
- [24] W. Ebeling and G. Nicolis. Word frequency and entropy of symbolic sequences: a dynamical perspective. *Chaos, Solitons and Fractals*, 2:635–650, 1992.
- [25] W. Ebeling and T. Pöschel. Entropy and long-range correlations in literary English. *Europhysics Letters*, 26:241–246, 1994.
- [26] Y. Ephraim and N. Merhav. Hidden Markov processes. *IEEE Transactions on Information Theory*, 48:1518–1569, 2002.

- [27] P. D. Finch. On the covariance determinants of autoregressive and moving average models. *Biometrika*, 47:194–211, 1960.
- [28] I. M. Gelfand, A. N. Kolmogorov, and A. M. Yaglom. Towards the general definition of the amount of information. *Doklady Akademii Nauk SSSR*, 111:745–748, 1956. In Russian.
- [29] T. Gramss. Entropy of the symbolic sequence for critical circle maps. *Physical Review E*, 50:2616–2620, 1994.
- [30] R. M. Gray and L. D. Davisson. The ergodic decomposition of stationary discrete random processes. *IEEE Transactions on Information Theory*, 20:625–636, 1974.
- [31] R. M. Gray and L. D. Davisson. Source coding theorems without the ergodic assumption. *IEEE Transactions on Information Theory*, 20:502–516, 1974.
- [32] U. Grenander and G. Szegő. *Toeplitz Forms and Their Applications*. Berkeley: University of California Press, 1958.
- [33] P. Grünwald and P. Vitányi. Shannon information and Kolmogorov complexity. <http://www.arxiv.org/abs/cs.IT/0410002>, 2004.
- [34] P. D. Grünwald and P. M. B. Vitányi. Kolmogorov complexity and information theory (with an interpretation in terms of questions and answers). *Journal of Logic, Language, and Information*, 12:497–529, 2003.
- [35] L. Györfi, I. Páli, and E. C. van der Meulen. There is no universal source code for infinite alphabet. *IEEE Transactions on Information Theory*, 40:267–271, 1994.
- [36] W. Hilberg. Der bekannte Grenzwert der redundanzfreien Information in Texten — eine Fehlinterpretation der Shannonschen Experimente? *Frequenz*, 44:243–248, 1990.
- [37] J. R. M. Hosking. Fractional differencing. *Biometrika*, 68:165–176, 1981.
- [38] O. Kallenberg. *Foundations of Modern Probability*. New York: Springer, 1997.
- [39] J. C. Kieffer and E. Yang. Grammar-based codes: A new class of universal lossless source codes. *IEEE Transactions on Information Theory*, 46:737–754, 2000.
- [40] M. Li, X. Chen, X. Li, B. Ma, and P. M. B. Vitányi. The similarity metric. *IEEE Transactions on Information Theory*, 50:3250–3264, 2004.
- [41] M. Li and P. M. B. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications, 2nd ed.* New York: Springer, 1997.
- [42] A. Milosavljević. Discovering dependencies via algorithmic mutual information: A case study in DNA sequence comparisons. *Machine Learning*, 21:35–50, 1995.
- [43] F. L. Ramsey. Characterization of the partial autocorrelation function. *The Annals of Statistics*, 2:1296–1301, 1974.
- [44] W. Rytter. Application of Lempel-Ziv factorization to the approximation of grammar-based compression. *Theoretical Computer Science*, 302:211–222, 2003.

- [45] T. Seidenfeld, M. J. Schervish, and J. B. Kadane. Improper regular conditional distributions. *The Annals of Probability*, 29:1612–1624, 2001.
- [46] C. Shannon. Prediction and entropy of printed English. *Bell System Technical Journal*, 30:50–64, 1950.
- [47] P. C. Shields. *The Ergodic Theory of Discrete Time Series*. Providence: American Mathematical Society, 1996.
- [48] J. M. Swart. A conditional product measure theorem. *Statistics & Probability Letters*, 28:131–135, 1996.
- [49] H. van Halteren, R. H. Baayen, F. Tweedie, M. Haverkort, and A. Neijt. New machine learning methods demonstrate the existence of a human stylome. *Journal of Quantitative Linguistics*, 12:65–77, 2005.
- [50] T. Weissman. Not all universal source codes are pointwise universal. <http://www.stanford.edu/~tsachy/interest.htm>, 2004.
- [51] R. W. Yeung. *First Course in Information Theory*. Dordrecht: Kluwer Academic Publishers, 2002.
- [52] J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23:337–343, 1977.
- [53] J. Ziv and A. Lempel. Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory*, 24:530–536, 1978.