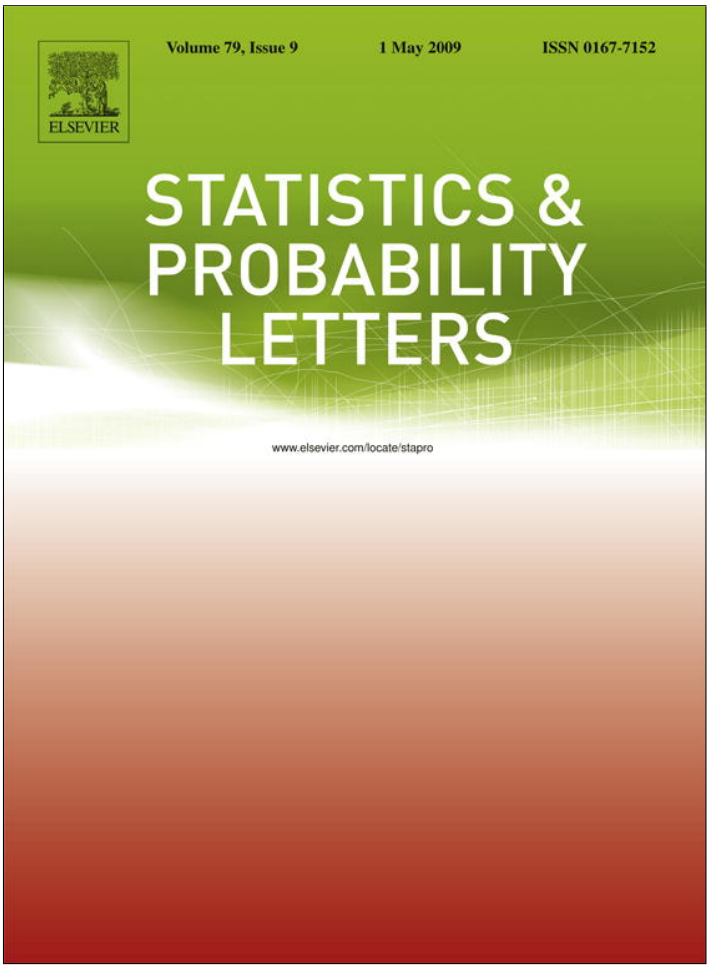

Statistics and Probability Letters 79 (2009) 1260–1268

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

A general definition of conditional information and its application to ergodic decomposition☆

Łukasz Dębowski *

Centrum Wiskunde & Informatica, Science Park 123, NL-1098 XG Amsterdam, Netherlands

ARTICLE INFO

Article history:
Received 15 January 2008
Received in revised form 17 June 2008
Accepted 13 January 2009
Available online 29 January 2009

ABSTRACT

We discuss a simple definition of conditional mutual information (CMI) for fields and σ-fields. The new definition is applicable also in nonregular cases, unlike the well-known but more restricted definition of CMI by Dobrushin. Certain properties of the two notions of CMI and their equivalence for countably generated σ-fields are established. We also consider an application, which concerns the ergodic decomposition of mutual information for stationary processes. In this case, CMI is tightly linked, via additivity of information, with entropy defined as self-information. Thus we reconsider the latter concept in some detail.

1. Introduction

The extension of entropy and related information measures into functionals of arbitrary algebras of events is some useful abstract tool in information theory (Gelfand et al., 1956; Dobrushin, 1959; Pinsker, 1964). This extension allows us to handle entropy and information not only for discrete and continuous variables simultaneously but also for the tail and invariant σ-fields of stochastic processes.

Unfortunately, the extension that is provided in the existing literature is neither fully general nor the simplest possible, see Dobrushin (1959, Section 2) and Pinsker (1964, Chapters 1–3) for detailed accounts. The aim of this paper is to show a simpler path to generalizing several information measures, including conditional Kullback–Leibler divergence.

For probability space $(\Omega, \mathcal{J}, P)$ let $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ be subfields of $\mathcal{J}$. Fields are set algebras closed under finite operations, whereas σ-fields are assumed to be closed also under denumerable sums and products. A field is called finite if it has finitely many elements. The smallest (finite) field containing partition $\{B_j\}_{j=1}^J$ of Ω, where $B_i \in \mathcal{J}$, will be denoted by $[B_1, \ldots, B_J]$. For any finite field $\mathcal{B}$ there is a unique partition $\{B_j\}_{j=1}^J$ such that $\mathcal{B} = [B_1, \ldots, B_J]$. Thus we can define four Shannon information measures for three finite fields $\mathcal{A} = [A_1, \ldots, A_I]$, $\mathcal{B} = [B_1, \ldots, B_J]$, and $\mathcal{C} = [C_1, \ldots, C_K]$:

- entropy $H(\mathcal{A}) := H_P(\mathcal{A}) := -\sum_{i=1}^I P(A_i) \log P(A_i)$,
- mutual information

$$I(\mathcal{A}; \mathcal{B}) := I_P(\mathcal{A}; \mathcal{B}) := \sum_{i=1}^I \sum_{j=1}^J P(A_i \cap B_j) \log \frac{P(A_i \cap B_j)}{P(A_i)P(B_j)},$$

☆ The work was partially supported by the Polish Ministry of Scientific Research and Information Technology, grant no. 1/P03A/045/28, and the IST Programme of the European Community, under the PASCAL II Network of Excellence, IST-2002-506778. This publication reflects only the author's views.

* Tel.: +31 20 592 4193.
E-mail address: debowski@cwi.nl.

0167-7152/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.spl.2009.01.016

- conditional entropy $H(\mathcal{A}|\mathcal{C}) := \sum_{k=1}^{K} P(C_k) H_{P(\cdot|C_k)}(\mathcal{A})$, and
- conditional mutual information $I(\mathcal{A};\mathcal{B}|\mathcal{C}) := \sum_{k=1}^{K} P(C_k) I_{P(\cdot|C_k)}(\mathcal{A};\mathcal{B})$,

where the algebraic relation $0 \log 0 = 0$ is assumed.

The above formulae mirror standard definitions for finite-valued random variables (e.g., Cover and Thomas, 1991, Eqs. 2.1, 2.10, 2.28, 2.60). If field $\mathcal{A}_i$ is the smallest field with respect to which variable Y_i is measurable, then one puts $I(Y_1;Y_2|Y_3) := I(\mathcal{A}_1;\mathcal{A}_2|\mathcal{A}_3)$, $I(Y_1;Y_2) := I(\mathcal{A}_1;\mathcal{A}_2)$, $H(Y_1|Y_2) := H(\mathcal{A}_1|\mathcal{A}_2)$, and $H(Y_1) := H(\mathcal{A}_1)$. Similar conventions are followed for other random variables once the information measures are extended to infinite fields (Pinsker, 1964, Translator's Remarks to Chapter 1).

It is easy to notice that $\eta(\mathcal{A}) \geq \eta(\mathcal{A}')$ for $\mathcal{A} \supset \mathcal{A}'$ in each case of $\eta(\mathcal{A}) = H(\mathcal{A}), H(\mathcal{A}|\mathcal{C}), I(\mathcal{A};\mathcal{B}), I(\mathcal{A};\mathcal{B}|\mathcal{C})$. Hence for finite $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ we have

$$H(\mathcal{A}) = \sup H(\mathcal{A}'), \qquad I(\mathcal{A};\mathcal{B}) = \sup I(\mathcal{A}';\mathcal{B}'), \tag{1}$$

$$H(\mathcal{A}|\mathcal{C}) = \sup H(\mathcal{A}'|\mathcal{C}), \qquad I(\mathcal{A};\mathcal{B}|\mathcal{C}) = \sup I(\mathcal{A}';\mathcal{B}'|\mathcal{C}), \tag{2}$$

where the supremum is taken over finite fields $\mathcal{A}' \subset \mathcal{A}$ and $\mathcal{B}' \subset \mathcal{B}$. The above equalities can also be used as *definitions* for infinite $\mathcal{A}$ and $\mathcal{B}$. Indeed, formulae (1) were discussed as definitions by Gelfand et al. (1956) and Pinsker (1964).[1]

Denote the expectation of the random variable Y as $\mathbf{E}\,Y = \int Y \mathrm{d}P$. To resolve the problem of generalizing conditional information measures to infinite $\mathcal{C}$, it suffices to observe that for finite $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ we have also

$$H(\mathcal{A}|\mathcal{C}) = \mathbf{E}\, H(\mathcal{A} \parallel \mathcal{C}), \qquad I(\mathcal{A};\mathcal{B}|\mathcal{C}) = \mathbf{E}\, I(\mathcal{A};\mathcal{B} \parallel \mathcal{C}), \tag{3}$$

where $H(\mathcal{A} \parallel \mathcal{C}) := H_{P(\cdot\parallel\mathcal{C})}(\mathcal{A})$ and $I(\mathcal{A};\mathcal{B} \parallel \mathcal{C}) := I_{P(\cdot\parallel\mathcal{C})}(\mathcal{A};\mathcal{B})$ are random variables and $P(A \parallel \mathcal{C})$ is the conditional probability of event $A \in \mathcal{J}$ w.r.t. the smallest σ-field containing $\mathcal{C}$ (cf. e.g. Billingsley, 1979, Section 33). Expressions (3) remain sound for any field $\mathcal{C}$. Thus we can generalize conditional information measures first to arbitrary $\mathcal{C}$ via (3) and then to arbitrary $\mathcal{A}$ and $\mathcal{B}$ via (2).

Whereas the left expression in (3) is well known (Billingsley, 1965, Section 12), the analogical approach seems to have never been investigated in depth for conditional mutual information. A rather cumbersome expression has been generally adopted instead. The motivation came from the equality

$$I(\mathcal{A};\mathcal{B}) = \tilde{I}(\mathcal{A};\mathcal{B}) := \begin{cases} \int \log \dfrac{\mathrm{d}P_{\mathcal{A}\mathcal{B}}}{\mathrm{d}P_{\mathcal{A}\times\mathcal{B}}} \mathrm{d}P_{\mathcal{A}\mathcal{B}} & P_{\mathcal{A}\mathcal{B}} \ll P_{\mathcal{A}\times\mathcal{B}}, \\ \infty & \text{else}, \end{cases} \tag{4}$$

(Gelfand et al., 1956, Theorem 4; Dobrushin, 1959, Section 2), where the "diagonal" measure $P_{\mathcal{A}\mathcal{B}}(A \times B) := P(A \cap B)$ and the product measure $P_{\mathcal{A}\times\mathcal{B}}(A \times B) := P(A)P(B)$ are defined as measures on product σ-field $\mathcal{A} \otimes \mathcal{B}$ via their unique extension from Cartesian product $\mathcal{A} \times \mathcal{B}$.

By analogy to (4), Dobrushin (1959, Eqs. 2.7.10–10'), followed by Pinsker (1964, Section 3.1), defined conditional mutual information

$$\tilde{I}(\mathcal{A};\mathcal{B}|\mathcal{C}) := \begin{cases} \int \log \dfrac{\mathrm{d}P_{\mathcal{A}\mathcal{B}\mathcal{C}}}{\mathrm{d}P_{\mathcal{A}\times\mathcal{B}|\mathcal{C}}} \mathrm{d}P_{\mathcal{A}\mathcal{B}\mathcal{C}} & P_{\mathcal{A}\mathcal{B}\mathcal{C}} \ll P_{\mathcal{A}\times\mathcal{B}|\mathcal{C}}, \\ \infty & \text{else}, \end{cases} \tag{5}$$

where $P_{\mathcal{A}\mathcal{B}\mathcal{C}}$ and $P_{\mathcal{A}\times\mathcal{B}|\mathcal{C}}$ are measures on $\mathcal{A} \otimes \mathcal{B} \otimes \mathcal{C}$ given by $P_{\mathcal{A}\mathcal{B}\mathcal{C}}(A \times B \times C) := P(A \cap B \cap C)$ and

$$P_{\mathcal{A}\times\mathcal{B}|\mathcal{C}}(A \times B \times C) := \int_C P(A \parallel \mathcal{C}) P(B \parallel \mathcal{C}) \mathrm{d}P. \tag{6}$$

Measure $P_{\mathcal{A}\times\mathcal{B}|\mathcal{C}}$ exists and hence expression (5) is valid if conditional probability $\{P(E \parallel \mathcal{C})\}_{E\in\mathcal{A}}$ is regular (Swart, 1996). Thus expressions (4) and (5) open way to simple algebraic expressions for information measures of Gaussian variables (Pinsker, 1964, Chapters 9–11; Cover and Thomas, 1991, Chapter 9). Nonetheless, expression (5) does not make sense in certain other cases, when the function $P_{\mathcal{A}\times\mathcal{B}|\mathcal{C}}$ on the Cartesian product $\mathcal{A} \times \mathcal{B} \times \mathcal{C}$ fails to be even finitely additive (Sazonov, 1964). With regard to these questions see also the Translator's remarks to the Chapter 3 of Pinsker (1964).[2]

In this paper we will pursue the properties and applications of conditional information defined via (2) and (3). In Section 2, we will show that this simpler definition is equivalent to (5) in the case of countably generated fields. Although the new concept can be applied to any probability space, its general algebraic properties can be established more easily than for

[1] This approach cannot be used to generalize non-Shannon information measures, such as triple mutual information, since they are not monotonic in general (Yeung, 2002, Chapter 6 on I-measure). Some generalization of the I-measure to σ-fields might be useful, however.

[2] The issue that $P_{\mathcal{A}\times\mathcal{B}|\mathcal{C}}$ need not be a measure seems to be first raised in the literature by A. Feinstein, the translator of Pinsker (1964). R. L. Dobrushin forwarded his question to V. V. Sazonov, who produced a counterexample in his 1964 paper. In the footnote on page 55 of Pinsker (1964), Feinstein mentions that $P_{\mathcal{A}\times\mathcal{B}|\mathcal{C}}$ can fail to be a measure but gives no reference to Sazonov, whose article was published in the same year. A very similar counterexample was given by Swart (1996), who was unaware of Sazonov's construction.

the old one. An application will be presented in Section 3. The example concerns the ergodic decomposition of mutual information between the past and future of a countably generated stationary process. Since the application is focused on the additivity relation $I(\mathcal{A};\mathcal{B}) = H(\mathcal{C}) + I(\mathcal{A};\mathcal{B}|\mathcal{C})$ for $\mathcal{C} \subset \mathcal{A} \cap \mathcal{B}$, we will reconsider some properties of self-information $H(\mathcal{C}) := I(\mathcal{C};\mathcal{C})$ in Section 4.

The presented application features regular conditional probabilities. Thus using $I(\mathcal{A};\mathcal{B})$ and $I(\mathcal{A};\mathcal{B}|\mathcal{C})$ rather than $\tilde{I}(\mathcal{A};\mathcal{B})$ and $\tilde{I}(\mathcal{A};\mathcal{B}|\mathcal{C})$ seems just a matter of taste. We feel, however, that the new definition of CMI is more natural and useful for the following reasons: (i) We avoid discussing whether $P_{\mathcal{A}\mathcal{B}\mathcal{C}}$ is dominated by $P_{\mathcal{A}\times\mathcal{B}|\mathcal{C}}$ and consider one Radon–Nikodym derivative less. (ii) We obtain in a rigorous way a more general additivity relation (chain rule) than established so far. (iii) The new definition explicitly stimulates thinking about information in terms of sets of events rather than in terms of random variables and densities.

These theoretical advantages are useful. The general additivity allows us to prove an impossibility result in coding theory mentioned in Section 3. Thinking in terms of σ-fields helps us to demonstrate an elementary characterization of some strongly nonergodic processes in Section 4. We hope that our paper provides a motivated and compact introduction to four generalized Shannon information measures.

2. Properties of conditional information

Let $\mathcal{A} \vee \mathcal{B}$ denote the intersection of all fields that contain $\mathcal{A}$ and $\mathcal{B}$. The newly proposed definition reads:

Definition 1. For finite fields $\mathcal{A}'$ and $\mathcal{B}'$ on the event space Ω and a probability measure P on $\mathcal{A}' \vee \mathcal{B}'$, let mutual information be

$$I_P(\mathcal{A}';\mathcal{B}') := \sum_{i=1}^{I}\sum_{j=1}^{J} P(A_i \cap B_j) \log \frac{P(A_i \cap B_j)}{P(A_i)P(B_j)},$$

where $\{A_i\}_{i=1}^{I}$ and $\{B_j\}_{j=1}^{J}$ are the partitions of Ω that satisfy $\mathcal{A}' = [A_1, \ldots, A_I]$ and $\mathcal{B}' = [B_1, \ldots, B_J]$.

Next, consider a probability space $(\Omega, \mathcal{J}, P)$. For an arbitrary field $\mathcal{C}$ and finite fields $\mathcal{A}'$ and $\mathcal{B}'$, where $\mathcal{A}', \mathcal{B}', \mathcal{C} \subset \mathcal{J}$, we define pointwise conditional mutual information

$$I(\mathcal{A}';\mathcal{B}' \parallel \mathcal{C}) := I_{P(\cdot\parallel\mathcal{C})}(\mathcal{A}';\mathcal{B}'),$$

where $P(E \parallel \mathcal{C})$ is the conditional probability of event $E \in \mathcal{J}$ w.r.t. the smallest σ-field containing $\mathcal{C}$.

The (average) conditional mutual information (or shortly CMI) between arbitrary fields $\mathcal{A}$ and $\mathcal{B}$ given a field $\mathcal{C}$ is defined as

$$I(\mathcal{A};\mathcal{B}|\mathcal{C}) := \sup \mathbf{E}\, I(\mathcal{A}';\mathcal{B}' \parallel \mathcal{C}), \tag{7}$$

where the supremum is taken over all finite fields $\mathcal{A}' \subset \mathcal{A}$ and $\mathcal{B}' \subset \mathcal{B}$.

For this definition and the other information measures discussed in the Introduction, we also have identities $I(\mathcal{A}_1;\mathcal{A}_2) = I(\mathcal{A}_1;\mathcal{A}_2|\{\emptyset,\Omega\})$, $H(\mathcal{A}_1|\mathcal{A}_2) = I(\mathcal{A}_1;\mathcal{A}_1|\mathcal{A}_2)$, and $H(\mathcal{A}_1) = I(\mathcal{A}_1;\mathcal{A}_1)$ like in the case of finite fields.

The expression on the right-hand side of (7) is meaningful for all $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ since conditional probabilities $P(\cdot \parallel \mathcal{C})$ are $\mathcal{J}$-measurable. No problems arise when the conditional probability is not regular (cf. Seidenfeld et al., 2001, Corollary 1) since the conditional distribution $(P(E \parallel \mathcal{C}))_{E\in\mathcal{E}}$ restricted to a finite field $\mathcal{E}$ is almost surely a probability measure (Billingsley, 1979, Theorem 33.2).

Although CMI has usually been discussed for σ-fields, the new definition makes sense also for fields. This point of view is convenient to prove continuity. We will write $\mathcal{B}_n \uparrow \mathcal{B}$ for a sequence $(\mathcal{B}_n)_{n\in\mathbb{N}}$ of fields such that $\mathcal{B}_1 \subset \mathcal{B}_2 \subset \cdots \subset \mathcal{B}$ and $\bigcup_{n\in\mathbb{N}} \mathcal{B}_n = \mathcal{B}$. ($\mathcal{B}$ need not be a σ-field.)

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

(i) $I(\mathcal{A};\mathcal{B}|\mathcal{C}) = I(\mathcal{B};\mathcal{A}|\mathcal{C})$;
(ii) $I(\mathcal{A};\mathcal{B}|\mathcal{C}) \geq 0$ *with the equality if and only if* $P(A \cap B \parallel \mathcal{C}) = P(A \parallel \mathcal{C})P(B \parallel \mathcal{C})$ *almost surely for all* $A \in \mathcal{A}$ *and* $B \in \mathcal{B}$;
(iii) $I(\mathcal{A};\mathcal{B}|\mathcal{C}) \leq \min(H(\mathcal{A}|\mathcal{C}), H(\mathcal{B}|\mathcal{C}))$;
(iv) $I(\mathcal{A};\mathcal{B}_1|\mathcal{C}) \leq I(\mathcal{A};\mathcal{B}_2|\mathcal{C})$ *if* $\mathcal{B}_1 \subset \mathcal{B}_2$;
(v) $I(\mathcal{A};\mathcal{B}_n|\mathcal{C}) \uparrow I(\mathcal{A};\mathcal{B}|\mathcal{C})$ *for* $\mathcal{B}_n \uparrow \mathcal{B}$.

Remark. Properties (i) and (ii) were established for definition (5) by Pinsker (1964) in Section 3.2, whereas (iv) and (v) are analogues of his Theorem 3.10.1.

Proof. Properties (i), (ii), (iii), and (iv) follow directly from the same properties for finite fields (Cover and Thomas, 1991, Eqs. 2.46, 2.91, 2.40, 2.122). Property (v) holds since every partition of $\mathcal{B} = \bigcup_{n\in\mathbb{N}} \mathcal{B}_n$ is a partition of $\mathcal{B}_m$ for almost all m. □

An important property of definition (7) is that the value of CMI does not change when the fields are extended to complete σ-fields (or any intermediate fields). A field is called *complete* if it contains all sets of outer P-measure 0. Let $\sigma(\mathcal{A})$ denote the intersection of all complete σ-fields containing $\mathcal{A}$. The unique extension of measure P from $\mathcal{J}$ to $\sigma(\mathcal{J})$ will be written as P, as well.

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 a finite field $\mathcal{C}_n \subset \mathcal{C}$ satisfy*

$$\{\omega \in \Omega : (i-1)/n < P(E \parallel \mathcal{C}) \leq i/n\} \in \mathcal{C}_n \quad \text{for } i = 1, \ldots, n \quad \text{and} \quad E \in \mathcal{A} \vee \mathcal{B}. \tag{8}$$

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

Remark. Such finite fields $\mathcal{C}_n$ exist since $P(E \parallel \mathcal{C})$ are $\mathcal{C}$-measurable.

Proof. Condition (8) implies $|P(E \parallel \mathcal{C}_n) - P(E \parallel \mathcal{C})| \leq 1/n$ almost surely. Thus

$$\lim_{n\to\infty} I(\mathcal{A}; \mathcal{B} \parallel \mathcal{C}_n) = I(\mathcal{A}; \mathcal{B} \parallel \mathcal{C}) \quad \text{almost surely} \tag{9}$$

by the continuity of $I_P(\mathcal{A}; \mathcal{B})$ as a function of P (Yeung, 2002, Section 2.3). For $\mathcal{A} = [A_1, \ldots, A_I]$ and $\mathcal{B} = \left[B_1, \ldots, B_J\right]$, we also have $I(\mathcal{A}; \mathcal{B}|\mathcal{C}_n) = \int I(\mathcal{A}; \mathcal{B} \parallel \mathcal{C}_n)\mathrm{d}P$, $I(\mathcal{A}; \mathcal{B}|\mathcal{C}) = \int I(\mathcal{A}; \mathcal{B} \parallel \mathcal{C})\mathrm{d}P$ and $0 \leq I(\mathcal{A}; \mathcal{B} \parallel \mathcal{C}) \leq \log \min\{I, J\}$ almost surely. Hence the thesis follows from (9) by the Lebesgue dominated convergence theorem. □

With Lemma 1, we can demonstrate a proposition, the first part of which has been mentioned.

Theorem 2. *Let $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, and $\mathcal{D}$ be subfields of $\mathcal{J}$.*

(i) $I(\mathcal{A}; \mathcal{B}|\mathcal{C}) = I(\mathcal{A}; \sigma(\mathcal{B})|\mathcal{C})$
and $I(\mathcal{A}; \mathcal{B}|\mathcal{C}) = I(\mathcal{A}; \mathcal{B}|\sigma(\mathcal{C}))$;
(ii) $I(\mathcal{A}; \mathcal{B} \vee \mathcal{C}|\mathcal{D}) = I(\mathcal{A}; \mathcal{C}|\mathcal{D}) + I(\mathcal{A}; \mathcal{B}|\mathcal{C} \vee \mathcal{D})$.

Remark. The analogue of (i) for $I(\mathcal{A}; \cdot)$ was proved by Dobrushin (1959, Section 2.2). Additivity (ii), often called the chain rule, is well known for finite-valued variables. For example, it implies $H(X) = I(X; Y) + H(X|Y)$. The analogue of (ii) for the other definition of CMI was also treated by Dobrushin (1959, Eqs. 2.7.1 and 2.7.9) for $\mathcal{D} = \{\emptyset, \Omega\}$ and by Pinsker (1964, Theorem 3.6.2 and Eq. 3.6.6) for a general $\mathcal{D}$. The assertion made by Pinsker covered all cases of measure dominance and singularity but assumed implicitly that the conditional product measures exist. After a discussion with Dobrushin, the translator of Pinsker's book showed in his remarks to Chapter 3 that the special case (11) holds if $P_{\mathcal{ABC}} \ll P_{\mathcal{A}\times(\mathcal{BC})}$. This assumption implies also that $P_{\mathcal{A}\times\mathcal{B}|\mathcal{C}}$ exists, $P_{\mathcal{ABC}} \ll P_{\mathcal{A}\times\mathcal{B}|\mathcal{C}}$, and $P_{\mathcal{AC}} \ll P_{\mathcal{A}\times\mathcal{C}}$. By the way, there are misprints in Eqs. 3.6.1–3 of Pinsker (1964), which correspond to (11) with $I(\mathcal{B}; \mathcal{C})$ substituted for $I(\mathcal{A}; \mathcal{C})$.

In the following proofs, we use symmetric difference $A \triangle B := A \setminus B \cap B \setminus A$.

Proof. (i) 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 $P(E \parallel \mathcal{C}) = P(E \parallel \sigma(\mathcal{C}))$. It remains to justify $I(\mathcal{A}; \mathcal{B}|\mathcal{C}) = I(\mathcal{A}; \sigma(\mathcal{B})|\mathcal{C})$. We will adapt the proof for case $\mathcal{C} = \{\emptyset, \Omega\}$ given by Dobrushin (1959, Section 2.2).

Fix a finite field $\mathcal{A}_1$ and $\epsilon > 0$. Consider $\sigma_0(\mathcal{B}) \supset \mathcal{B}$ defined as the intersection of *all* σ-fields containing $\mathcal{B}$ (not necessarily complete ones). According to Dobrushin (1959, Eq. 2.2.10), 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 a mapping $f : \mathcal{B}_2 \to \mathcal{B}_2'$ such that $P(B \triangle f(B)) = 0$ for all $B \in \mathcal{B}_2$.)

Now let us extend this result to $\mathcal{C} \neq \{\emptyset, \Omega\}$. Consider a finite field $\mathcal{C}_n \subset \mathcal{C}$ satisfying (8). By Dobrushin's result, for almost every $\omega \in \Omega$ there exists a finite field $\mathcal{B}_\omega \subset \mathcal{B}$ such that $I(\mathcal{A}_1; \mathcal{B}_\omega \parallel \mathcal{C}_n)(\omega) \geq I(\mathcal{A}_1; \mathcal{B}_2 \parallel \mathcal{C}_n)(\omega) - \epsilon$. For some version of conditional probability and $\mathcal{B}_\omega$, random variable $\omega \mapsto \mathcal{B}_\omega$ is $\mathcal{C}_n$-measurable and then $\mathcal{B}_1 := \bigvee_{\omega\in\Omega} \mathcal{B}_\omega$ is a finite field with $\mathcal{B}_1 \subset \mathcal{B}$. By Theorem 2(iv), $\mathcal{B}_1$ satisfies $I(\mathcal{A}_1; \mathcal{B}_1 \parallel \mathcal{C}_n) \geq I(\mathcal{A}_1; \mathcal{B}_\omega \parallel \mathcal{C}_n) \geq I(\mathcal{A}_1; \mathcal{B}_2 \parallel \mathcal{C}_n) - \epsilon$ for almost every ω and thus $I(\mathcal{A}_1; \mathcal{B}_1|\mathcal{C}_n) \geq I(\mathcal{A}_1; \mathcal{B}_2|\mathcal{C}_n) - \epsilon$.

Recall that $\lim_n I(\mathcal{A}_1; \mathcal{B}|\mathcal{C}_n) = I(\mathcal{A}_1; \mathcal{B}|\mathcal{C})$ by Lemma 1. Thus we have

$$\forall_{\delta>0}\, \forall_{\mathcal{B}_2\subset\sigma(\mathcal{B})}\, \exists_{\mathcal{B}_1\subset\mathcal{B}}\, I(\mathcal{A}_1; \mathcal{B}_1|\mathcal{C}) \geq I(\mathcal{A}_1; \mathcal{B}_2|\mathcal{C}) - \delta, \tag{10}$$

where $\mathcal{B}_1$ and $\mathcal{B}_2$ are assumed to be finite fields. For arbitrary δ and $\mathcal{B}_2$, a suitable $\mathcal{B}_1$ is given by the construction in the previous paragraph for a sufficiently large n and a sufficiently small ϵ. Equality $I(\mathcal{A}; \mathcal{B}|\mathcal{C}) = I(\mathcal{A}; \sigma(\mathcal{B})|\mathcal{C})$ follows from (10) and the inequality $I(\mathcal{A}; \mathcal{B}|\mathcal{C}) \leq I(\mathcal{A}; \sigma(\mathcal{B})|\mathcal{C})$.

(ii) Let $\mathcal{A}$ and $\mathcal{B}$ be finite fields and let $\mathcal{C}$ be any field. Subsequently, let $\mathcal{C}_n \subset \mathcal{C}$ be finite fields satisfying $I(\mathcal{A}; \mathcal{B} \vee \mathcal{C}) - I(\mathcal{A}; \mathcal{B} \vee \mathcal{C}_n) \leq 1/n, I(\mathcal{A}; \mathcal{C}) - I(\mathcal{A}; \mathcal{C}_n) \leq 1/n$, and (8). The latter requirement implies $\lim_n I(\mathcal{A}; \mathcal{B}|\mathcal{C}_n) = I(\mathcal{A}; \mathcal{B}|\mathcal{C})$. Thus, the well-known equalities $I(\mathcal{A}; \mathcal{B} \vee \mathcal{C}_n) = I(\mathcal{A}; \mathcal{C}_n) + I(\mathcal{A}; \mathcal{B}|\mathcal{C}_n)$ for finite $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}_n$ (Cover and Thomas, 1991, Eq. 2.60) imply

$$I(\mathcal{A}; \mathcal{B} \vee \mathcal{C}) = I(\mathcal{A}; \mathcal{C}) + I(\mathcal{A}; \mathcal{B}|\mathcal{C}). \tag{11}$$

By Theorems 1(v) and 2(i), we may extend (11) to any $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$. Assume finite $\mathcal{A}$ again. By (11) we also have

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

where all expressions are finite. Having established the claim for finite $\mathcal{A}$, we generalize it to infinite $\mathcal{A}$, using Theorems 1(v) and 2(i) again. □

Theorems 1(v) and 2(i) conjoined with the following lemma allow us to prove easily the partial equivalence of the two definitions of CMI.

Lemma 2. *Consider σ-fields $\mathcal{A}_n \uparrow \mathcal{A}'$, $\mathcal{A} = \sigma(\mathcal{A}')$, $\mathcal{B}_n \uparrow \mathcal{B}'$, $\mathcal{B} = \sigma(\mathcal{B}')$, and $\mathcal{C}$. If there exists measure $P_{\mathcal{A}\times\mathcal{B}|\mathcal{C}}$ then*

$$\tilde{I}(\mathcal{A};\mathcal{B}|\mathcal{C}) = \lim_{n\to\infty} \tilde{I}(\mathcal{A}_n;\mathcal{B}_n|\mathcal{C}). \tag{12}$$

Proof. Denote $S = P_{\mathcal{A}\times\mathcal{B}|\mathcal{C}} + P_{\mathcal{A}\mathcal{B}\mathcal{C}}$. By the existence of $P_{\mathcal{A}\times\mathcal{B}|\mathcal{C}}$, measure $P_{\mathcal{F}\times\mathcal{G}|\mathcal{C}}$ exists also for $\mathcal{F} \subset \mathcal{A}$ and $\mathcal{G} \subset \mathcal{B}$. Both cases of (5) can be written as

$$\tilde{I}(\mathcal{F};\mathcal{G}|\mathcal{C}) = \int \kappa\left(\mathrm{d}P_{\mathcal{F}\mathcal{G}\mathcal{C}}/\mathrm{d}S\right)\mathrm{d}S,$$

where $\kappa(x) := x\log x - x\log(1-x) - 2x + 1$. We have the martingale convergence $\lim_n \mathrm{d}P_{\mathcal{A}_n\mathcal{B}_n\mathcal{C}}/\mathrm{d}S = \mathrm{d}P_{\mathcal{A}\mathcal{B}\mathcal{C}}/\mathrm{d}S$ S-almost surely. Since function κ is continuous and nonnegative, we have $\tilde{I}(\mathcal{A};\mathcal{B}|\mathcal{C}) \le \liminf_n \tilde{I}(\mathcal{A}_n;\mathcal{B}_n|\mathcal{C})$ by the Fatou lemma. On the other hand, κ is convex so $\tilde{I}(\mathcal{A}_n;\mathcal{B}_n|\mathcal{C}) \le \tilde{I}(\mathcal{A};\mathcal{B}|\mathcal{C})$ by the Jensen inequality. Thus (12) must be satisfied. □

Theorem 3. *Let $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ be subfields of $\mathcal{J}$, where $\mathcal{A}$ and $\mathcal{B}$ are countably generated, i.e., $\mathcal{A} = \sigma(\mathcal{A}')$ and $\mathcal{B} = \sigma(\mathcal{B}')$ for some countable fields $\mathcal{A}'$ and $\mathcal{B}'$. Then we have*

$$\tilde{I}(\mathcal{A};\mathcal{B} \parallel \mathcal{C}) = I(\mathcal{A};\mathcal{B} \parallel \mathcal{C}). \tag{13}$$

Proof. Let us notice that both sides of (13) equal $\int I(\mathcal{A};\mathcal{B} \parallel \mathcal{C})\mathrm{d}P$ when $\mathcal{A}$ and $\mathcal{B}$ are finite. Thus the continuity properties expressed in Theorems 1(v) and 2(i) and Lemma 2 imply that (13) holds also when $\mathcal{A}$ and $\mathcal{B}$ are countably generated. □

3. An application to ergodic decomposition

As an example, we will apply the machinery developed in Section 2 to the ergodic decomposition of a stationary process. Consider a process $(X_k)_{k\in\mathbb{Z}}$ on $(\Omega, \mathcal{J}, P)$, where $X_i : (\Omega, \mathcal{J}) \to (\mathbb{X}, \mathcal{X})$. Set $\mathcal{G}_{m:n} \subset \mathcal{J}$ as the smallest σ-fields against which blocks $X_{m:n} := (X_k)_{m\le k\le n}$ are measurable, assuming $\mathcal{G}_i := \mathcal{G}_{i:i}$. 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 σ-fields. For any field $\mathcal{F} \subset \sigma(\mathcal{G}_{-\infty}) \cap \sigma(\mathcal{G}_{\infty})$, we have

$$H(\mathcal{G}_1|\mathcal{G}_{-\infty:0}) = H(\mathcal{G}_1|\mathcal{G}_{-\infty:0} \vee \mathcal{F}), \tag{14}$$

$$\begin{aligned} I(\mathcal{G}_{-\infty:0};\mathcal{G}_{1:\infty}) &= I(\mathcal{G}_{-\infty:0};\mathcal{G}_{1:\infty} \vee \mathcal{F}) \\ &= I(\mathcal{G}_{-\infty:0};\mathcal{F}) + I(\mathcal{G}_{-\infty:0};\mathcal{G}_{1:\infty}|\mathcal{F}) \\ &= H(\mathcal{F}) + I(\mathcal{G}_{-\infty:0};\mathcal{G}_{1:\infty}|\mathcal{F}) \end{aligned} \tag{15}$$

in view of Theorems 1(iii–iv) and 2(i–ii).

Assume that $(X_k)_{k\in\mathbb{Z}}$ is stationary. Then

$$E := I(\mathcal{G}_{-\infty:0};\mathcal{G}_{1:\infty}) = \lim_{n\to\infty} I(X_{-n:0};X_{1:n}) \tag{16}$$

is called excess entropy (Crutchfield and Feldman, 2003), cf. Theorems 1(iv) and 2(i). Moreover, if the variable range $\mathbb{X}$ is finite then $H(\mathcal{G}_1|\mathcal{G}_{-\infty:0})$ equals entropy rate

$$h := \lim_{n\to\infty} H(X_1|X_{-n:0}) = \lim_{n\to\infty} H(X_{1:n})/n, \tag{17}$$

cf. Yeung (2002, Section 2.9) and Theorems 1(iv) and 8(iii) in the next section. We shall interpret the right-hand sides of Eqs. (14) and (15) likewise using ergodic decomposition.

Consider the measurable space of doubly-infinite sequences $(\mathbb{U}, \mathcal{U}) = \times_{k\in\mathbb{Z}}(\mathbb{X}, \mathcal{X})$, 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 σ-field $\mathcal{I} := \{A \in \mathcal{U} : TA = A\}$. Let $(\mathbb{S}, \mathcal{S})$ be the measurable space of stationary probability measures on $(\mathbb{U}, \mathcal{U})$ (i.e., $\mu\circ T = \mu$ for $\mu \in \mathbb{S}$) and let $(\mathbb{E}, \mathcal{E}) \subset (\mathbb{S}, \mathcal{S})$ be the subspace of ergodic measures (i.e., $\mu(A) \in \{0, 1\}$ for $\mu \in \mathbb{E}$ and $A \in \mathcal{I}$). Precisely, $\mathcal{S}$ and $\mathcal{E}$ are defined as the smallest σ-fields containing all cylinder sets $\{\mu \in \mathbb{S} : \mu(A) \le r\}$ and $\{\mu \in \mathbb{E} : \mu(A) \le r\}$, $A \in \mathcal{U}$, $r \in \mathbb{R}$, respectively. Since $\mathcal{U}$ is countably generated, all respective singletons $\{\mu\}$ belong to $\mathcal{S}$ and $\mathcal{E}$. The ergodic decomposition theorem can be stated as follows:

Theorem 4. *Consider a stationary measure $\mu \in \mathbb{S}$.*

(i) *(Shields, 1996, Theorem I.4.10; Kallenberg, 1997, Theorem 9.10) There exists a version of conditional distribution $\mu(\cdot \parallel \mathcal{I}) : \mathcal{U} \times \mathbb{U} \to \mathbb{R}$ such that $\mu(\cdot \parallel \mathcal{I})(u) \in \mathbb{E}$ for all $u \in \mathbb{U}$.*

(ii) *(Kallenberg, 1997, Theorem 9.12) Measure*

$$\nu(W) := \mu(\{u \in \mathbb{U} : \mu(\cdot \parallel \mathcal{I})(u) \in W\}), \quad W \in \mathcal{E},$$

is the only measure on $\mathcal{E}$ that satisfies

$$\mu = \int \mu(\cdot \parallel \mathcal{I})\mathrm{d}\mu = \int \sigma(\cdot)\mathrm{d}\nu(\sigma), \quad \sigma \in \mathbb{E}. \tag{18}$$

It is convenient to leave the space of doubly-infinite sequences and apply Theorem 4 to the countably generated process $(X_k)_{k\in\mathbb{Z}}$ with distribution $\mu = P((X_k)_{k\in\mathbb{Z}} \in \cdot) \in \mathbb{S}$, on a possibly richer space $(\Omega, \mathcal{J}, P)$. Set $\mathcal{G}_I := (X_k)^{-1}_{k\in\mathbb{Z}}(\mathcal{I})$ and define the random ergodic measure

$$F := \mu(\cdot \parallel \mathcal{I})((X_k)_{k\in\mathbb{Z}}).$$

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

The following lemma asserts that $\mathcal{F}$ is a field that we need.

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

This is a simple fact in ergodic theory. Since we have not come across an explicit proof of the lemma, we sketch it for completeness.

Proof. By Theorem 4(ii) and $\mathcal{I}$-measurability of $\mu(A \parallel \mathcal{I})$ 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, $\mu(A \parallel \mathcal{I}) = I_A$ μ-almost surely for any $A \in \mathcal{I}$ so, by Theorem 4(ii), $(X_k)^{-1}_{k\in\mathbb{Z}}(A)$ is an element of the smallest complete σ-field w.r.t. which $F(A)$ is measurable. Hence $\mathcal{G}_I \subset \sigma(\mathcal{F})$.

Let $A \in \mathcal{U}_- := (X_k)_{k\in\mathbb{Z}}(\mathcal{G}_{-\infty:0})$. By the ergodic theorem (e.g. Shields, 1996, Theorem I.3.1), variable $F(A)$ is $\sigma(\mathcal{G}_{-\infty})$-measurable. This result may be extended to any $A \in \mathcal{U}$ using the stationarity assumption and approximation theorems (Billingsley, 1979, Theorem 11.4 and 13.4). Thus $\mathcal{F} \subset \sigma(\mathcal{G}_{-\infty})$ and, by analogy, $\mathcal{F} \subset \sigma(\mathcal{G}_\infty)$. □

It is convenient to consider information measures for the subfields of $\mathcal{G}_{-\infty:\infty}$ as functions of the process distribution. For an arbitrary distribution $\mu = P((X_k)_{k\in\mathbb{Z}} \in \cdot) \in \mathbb{S}$, notice that $P(A) = \mu((X_k)_{k\in\mathbb{Z}}(A))$ for any $A \in \mathcal{G}_{-\infty:\infty}$. Thus we may introduce an explicit parametrization $I_\mu(\mathcal{A}, \mathcal{B}) := I(\mathcal{A}, \mathcal{B})$ for $\mathcal{A}, \mathcal{B} \subset \mathcal{G}_{-\infty:\infty}$, $h_\mu := h$, and $E_\mu := E$.

Let us substitute the random ergodic measure F is for μ. Since $F(A)$ equals $P((X_k)_{k\in\mathbb{Z}} \in A \parallel \mathcal{F})$ almost surely then $I_F(\mathcal{A}; \mathcal{B})$ is measurable for finite fields $\mathcal{A}$ and $\mathcal{B}$ and

$$\mathbf{E}\, I_F(\mathcal{A}; \mathcal{B}) = I(\mathcal{A}; \mathcal{B}|\mathcal{F}). \tag{19}$$

By the monotone convergence theorem and by Theorems 1(v) and 2(i), Eq. (19) may be generalized to any countably generated σ-fields $\mathcal{A}$ and $\mathcal{B}$. Hence there follows an ergodic decomposition of entropy rate and excess entropy:

Theorem 5. *For a countably generated stationary process* $(X_k)_{k\in\mathbb{Z}}$,

$$h = \mathbf{E}\, h_F \quad \text{if the variable range } \mathbb{X} \text{ is finite}, \tag{20}$$

$$E = H(\mathcal{F}) + \mathbf{E}\, E_F. \tag{21}$$

Proof. Variables h_F and E_F are measurable since they are limits of measurable variables by (16) and (17). Eq. (20), proved also by Gray and Davisson (1974, Theorem 5.1), can be established in the following way. For D being the cardinality of the range of $\mathbb{X}$, set $K := \log D$ so that $K - H(X_1) \geq 0$. By the monotone convergence theorem and (14),

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

Hence Eq. (20) follows. On the other hand, Eq. (21) follows directly from Lemma 3, (15), and (19) for $\mathcal{A} = \mathcal{G}_{-\infty:0}$ and $\mathcal{B} = \mathcal{G}_{1:\infty}$. □

Establishing the general additivity (11) has some application in coding theory. Namely, the simultaneous presence of E, $H(\mathcal{F})$, and $\mathbf{E}\, E_F$ in formula (21) is crucial to obtain such an impossibility result:

Theorem 6. *Let* $C : \mathbb{X}^+ \to \mathbb{X}^+$ *be a* uniquely decodable code *over a finite alphabet* $\mathbb{X} = \{0, 1, \ldots, D-1\}$, *i.e., its extension* $C^* : (u_1, \ldots, u_k) \mapsto C(u_1)\cdots C(u_k)$ *into finite tuples of strings* $u_i \in \mathbb{X}^*$ *is an injection. For the code length* $|C(\cdot)|$ *consider the normalized expectation of its excess*

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

taken with respect to a stationary measure $\mu = P((X_k)_{k\in\mathbb{Z}} \in \cdot) \in \mathbb{S}$. *Let* $N^C(K)$ *be the number of distinct ergodic measures* $\mu \in \mathbb{E}$ *such that* $\limsup_n E^C_\mu(n) \leq K$, $K \in \mathbb{R}$. *If the code is universal, i.e.,* $\lim_n n^{-1}\mathbf{E}\,|C(X_{1:n})|\log D = h$, *then*

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

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

Theorem 6 states that there cannot be too good codes among the asymptotically optimal ones. Our proof relies on additional lemmas and will be published elsewhere.

4. Entropy as self-information

Eq. (15) illustrates that the concept of entropy as self-information $H(\mathcal{A}) := I(\mathcal{A};\mathcal{A})$ arises naturally when the additivity of conditional information is considered. For a real variable Y, however, $H(Y)$ should not be confused with the differential entropy defined $h(Y) = -\int p(y)\log p(y)\mathrm{d}\lambda(y)$, where λ is the Lebesgue measure and $p = \mathrm{d}P(Y \in \cdot)/\mathrm{d}\lambda$. Although the appropriate difference of differential entropies for two real variables equals mutual information by equality (4), usually $h(Y) \neq H(Y)$. For instance, $h(Y) < \infty$ for a Gaussian variable Y (Cover and Thomas, 1991, Theorem 9.4.1). In the same case, $H(Y) = \infty$ according to a known result, stated here in a slightly stronger form.

Theorem 7. *$H(\mathcal{A}) = \infty$ unless $\mathcal{A}$ is purely atomic.*

Remark. A less formal proof of a weaker statement is given by Pinsker (1964, Section 2.4), viz. the Translator's Remarks on pp. 25–27. We say that a field $\mathcal{B}$ is *purely atomic* if there exists an atom $E \subset B$ for every $B \in \mathcal{B}$ such that $P(B) > 0$. On the other hand, $\mathcal{B}$ is called *nonatomic* if it has no atoms. Set E is called an atom with respect to $\mathcal{B}$ and P if $E \in \mathcal{B}$, $P(E) > 0$, and for every $F \in \mathcal{B}$ we have $P(E \cap F) = 0$ or $P(E \setminus F) = 0$.

Proof. Any measure P on $\mathcal{A}$ can be written as the sum of a purely atomic measure and a nonatomic measure, supported on disjoint sets $\Omega_a, \Omega_n \in \mathcal{A}$ respectively (Johnson, 1970, Theorem 2.1). Moreover, Ω_n can be partitioned into sets $A_1, A_2, \ldots, A_k \in \mathcal{A}$ such that $P(A_i) = P(\Omega_n)/k$ for each $k \in \mathbb{N}$ (cf. Billingsley, 1979, Exercise 2.17(d)). Hence $H(\mathcal{A}) \geq H([\Omega_a, A_1, \ldots, A_k]) = -P(\Omega_a)\log P(\Omega_a) - \sum_i P(A_i)\log P(A_i) \geq P(\Omega_n)\log k$. If $\mathcal{A}$ is not purely atomic then $P(\Omega_n) > 0$ and thus $H(\mathcal{A}) = \infty$.—This proof is due to Richard Bradley, private communication. □

Theorem 7 corresponds to a clear intuition, namely that the binary expansion of a random real variable $Y = \sum_{k=1}^{\infty} 2^{-k} Z_k$, uniformly distributed on $[0, 1]$, is a sequence of independent uniformly distributed random binary digits Z_k. Hence we obtain that $H(Y) = \sum_{k=1}^{\infty} H(Z_k|Z_{1:k-1}) = \sum_{k=1}^{\infty} H(Z_k) = \sum_{k=1}^{\infty} \log 2 = \infty$ by additivity and continuity of conditional information.

Treating a continuous real variable as a sequence of independent bits is very natural when the probability space is generated by a discrete stochastic process. In the following final example, the term 'fair-coin process' will stand for a binary process $(Z_k)_{k\in\mathbb{N}} \sim \mathrm{IID}$ with $P(Z_k = 0) = P(Z_k = 1) = 1/2$.

Definition 2. A process $(X_i)_{i\in\mathbb{Z}}$ is called an *uncountable description process (UDP)* if there exist functions $(f_{nk})_{n,k\in\mathbb{N}}$ and a fair-coin process $(Z_k)_{k\in\mathbb{N}}$ such that $\lim_n P(f_{nk}(X_{p+1:p+n}) = Z_k) = 1$ for all $p \in \mathbb{Z}$.

For instance, let $X_i := (K_i, Z_{K_i})$ assume values in $\mathbb{N} \times \{0, 1\}$, where variables $(Z_k)_{k\in\mathbb{N}}$ are probabilistically independent from $(K_i)_{i\in\mathbb{Z}} \sim \mathrm{IID}$ and $P(K_i = k) > 0$ for all $k \in \mathbb{N}$. If we let

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

then $P(f_{nk}(X_{p+1:p+n}) = Z_k) = 1 - [1 - P(K_i = k)]^n$. Thus $(X_i)_{i\in\mathbb{Z}}$ is a UDP.

It seems intuitive that $\lim_n I(X_{-n:0}; X_{1:n}) = \infty$ for any UDP since an infinite sequence of bits $(Z_k)_{k\in\mathbb{N}}$ can be learned given either the past or the future of $(X_i)_{i\in\mathbb{Z}}$. The proof of this proposition that we give below uses the generalized Shannon information measures and connects Definition 2 with nonatomicity of a shift-invariant sub-σ-field.

Let us recompile an entropic analogue of Theorem 1. By symmetry to $\mathcal{B}_n \uparrow \mathcal{B}$, we shall use notation $\mathcal{B}_n \downarrow \mathcal{B}$ for $\mathcal{B}_1 \supset \mathcal{B}_2 \supset \cdots \supset \mathcal{B}$ and $\bigcap_{n\in\mathbb{N}} \mathcal{B}_n = \mathcal{B}$.

Theorem 8. *Let $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{B}_n$ be subfields of $\mathcal{J}$.*

(i) *$H(\mathcal{A}) = 0$ if and only if $\mathcal{A}$ is* trivial *, i.e, if $P(A) \in \{0, 1\}$ for all $A \in \mathcal{A}$;*
(ii) *$H(\mathcal{A}|\mathcal{B}_1) \geq H(\mathcal{A}|\mathcal{B}_2)$ if $\mathcal{B}_1 \subset \mathcal{B}_2$;*
(iii) *$H(\mathcal{A}|\mathcal{B}_n) \downarrow H(\mathcal{A}|\mathcal{B})$ for $\mathcal{B}_n \uparrow \mathcal{B}$ and finite $\mathcal{A}$;*
(iv) *$H(\mathcal{A}|\mathcal{B}_n) \uparrow H(\mathcal{A}|\mathcal{B})$ for $\mathcal{B}_n \downarrow \mathcal{B}$;*
(v) *$H(\mathcal{A}|\mathcal{B}) = 0$ if and only if $\mathcal{A} \subset \sigma(\mathcal{B})$.*

Proof. Property (i) follows trivially from the analogical property for finite fields. Property (ii) was proved by Billingsley (1965, Identity $(C3)$ in Section 12) for finite $\mathcal{A}$ and it can be extended to infinite $\mathcal{A}$ immediately, as well.

Whereas property (iii) was proved by Billingsley (1965, Theorem 12.1) using the martingale and dominated convergence theorems, (iv) can be established for finite $\mathcal{A}$ likewise through the martingale convergence in the opposite direction (Doob, 1953, Chapter 8, Theorem 4.3). In the following, (iv) may be generalized to infinite $\mathcal{A}$ by noticing that there always exist such finite fields $\mathcal{A}_n \uparrow \mathcal{A}' \subset \mathcal{A}$ that $H(\mathcal{A}_n|\mathcal{B}_n) \uparrow H(\mathcal{A}|\mathcal{B})$ and $H(\mathcal{A}_n|\mathcal{B}_n) \leq H(\mathcal{A}|\mathcal{B}_n) \leq H(\mathcal{A}|\mathcal{B})$.

It remains to prove (v). Equality $H(\mathcal{A}|\mathcal{B}) = 0$ is equivalent to $P(A \parallel \mathcal{B}) \in \{0, 1\}$ almost surely for all $A \in \mathcal{A}$. On the other hand, it is straightforward that $P(A \parallel \mathcal{B}) \in \{0, 1\}$ holds if and only if $A \in \sigma(\mathcal{B})$. Firstly, notice that $P(A \parallel \mathcal{B})$ for $A \in \sigma(\mathcal{B})$ equals almost surely the indicator function of set A. To prove the converse, construct set $B := \{\omega \in \Omega : P(A \parallel \mathcal{B}) = 1\} \in \mathcal{B}$. By the definition of conditional probability and that of B, probabilities $P(A)$, $P(A \cap B)$, and $P(B)$ equal all $\int_B P(A \parallel \mathcal{B})\mathrm{d}P$. Thus $P(A \triangle B) = 0$ and hence $A \in \sigma(\mathcal{B})$. □

Via the properties (iii) and (v), we can link the convergence of finitely-valued random variables with inclusion of fields:

Lemma 4. *Let X be a finite-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 finite-valued variables X_n;*
(ii) $\lim_n H(X|\mathcal{Y}_n) = 0$;
(iii) $H(X|\mathcal{Y}) = 0$;
(iv) *X is $\sigma(\mathcal{Y})$-measurable;*

Remark. The assumption that X assumes finitely many values is important. Consider an X that takes values in natural numbers and has $H(X) = \infty$. Let $Y_k = 1$ for $X \geq k$ and $Y_k = 0$ else. We have $H(X|Y_{1:n}) = \infty$ since $H(X) = H(X|Y_{1:n}) + H(Y_{1:n})$ and $H(Y_{1:n}) \leq n \log 2$. Nevertheless, $H(X|(Y_n)_{n\in\mathbb{N}}) = 0$.

Proof. Statements (ii) and (iii) are equivalent by Theorem 8(iii), whereas (iii) and (iv) are equivalent by Theorem 8(v). It remains to prove that (i) is equivalent to (ii). Without loss of generality, let X assume values in $\{1, 2, \ldots, N\}$.

It is obvious that condition (ii) follows from (i) by the Fano inequality $H(X|\mathcal{Y}_n) \leq H(X|X_n) \leq \eta(P(X = X_n)) + [1 - P(X = X_n)] \log(N - 1)$ (Yeung, 2002, Theorem 2.47), where η is given by

$$\eta(p) = -p \log p - (1 - p) \log(1 - p), \quad p \in (0, 1)$$

and $\eta(0) = \eta(1) = 0$ to assure continuity. To prove the converse, define the value of random variable X_n as the smallest x such that $P(X = x \parallel \mathcal{Y}_n) \geq P(X = x' \parallel \mathcal{Y}_n)$ for $x' = 1, 2, \ldots, N$. We have $P(X = X_n \parallel \mathcal{Y}_n) \geq 1/N$. By concavity of η,

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

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

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

Thus (ii) implies (i). □

Hence uncountable description processes enjoy such a characterization:

Theorem 9. *Let $\mathcal{F}$ be the shift-invariant σ-field defined in Section 3. A stationary process $(X_i)_{i\in\mathbb{Z}}$ is a UDP if and only if $\sigma(\mathcal{F})$ contains a nonatomic sub-σ-field. Moreover, in the case of a UDP, variables Z_k are $\sigma(\mathcal{F})$-measurable.*

Proof. Assume first that $(X_i)_{i\in\mathbb{Z}}$ is a UDP. By Lemma 4, each variable Z_k is $\sigma(\mathcal{G}_{\infty:\infty})$-measurable and thus there exists a function g_k measurable $\mathcal{U}$ such that $g_k((X_k)_{k\in\mathbb{Z}}) = Z_k$ almost surely. Consider the distribution $\mu = P((X_k)_{k\in\mathbb{Z}} \in \cdot)$ and functions $g_{nk}((x_k)_{k\in\mathbb{Z}}) = f_{nk}(x_{1:n})$. By the definition of a UDP, $\lim_n \mu(T^i g_{nk} = g_k) = 1$, $i \in \mathbb{Z}$, and hence $\lim_n \mu(g_{nk} = T^{-i} g_k) = 1$ by stationarity of $(X_i)_{i\in\mathbb{Z}}$. The latter implies $g_k = T^{-i} g_k$ μ-almost everywhere and thus Z_k are $\sigma(\mathcal{F})$-measurable for all k. Construct the $\sigma(\mathcal{F})$-measurable variable $Y = \sum_{k\in\mathbb{N}} 2^{-k} Z_k$. The distribution of Y is Lebesgue measure on $[0, 1]$. The Lebesgue measure is nonatomic so $\sigma(\mathcal{F})$ contains a nonatomic sub-σ-field.

As for the converse, take $(X_i)_{i\in\mathbb{Z}}$ with a nonatomic $\mathcal{F}_0 \subset \sigma(\mathcal{F})$. 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 $P(B) = x$. 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 λ, $A_{wa} \subset A_w$, 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 $(Z_k)_{k\in\mathbb{N}}$ is a fair-coin process. By Lemma 3, Z_k are also $\sigma(\mathcal{G}_{1:\infty})$-measurable. Hence, by Lemma 4, $\lim_n P(f_{nk}(X_{1:n}) = Z_k) = 1$ for some functions f_{nk}. Finally, stationarity of $(X_i)_{i\in\mathbb{Z}}$ and $\sigma(\mathcal{F})$-measurability of Z_k imply that the probabilities $P(f_{nk}(X_{p+1:p+n}) = Z_k)$ do not depend on p. So $(X_i)_{i\in\mathbb{Z}}$ is a UDP. □

By Theorems 1(iv), 7 and 9, we have $H(\mathcal{F}) = \infty$ for every UDP. As a consequence, the excess entropy is $E = I(\mathcal{G}_{-\infty:0}; \mathcal{G}_{1:\infty}) \geq H(\mathcal{F}) = \infty$. The proof of Theorem 9 may be easily adjusted to show directly that $E = \infty$ also in the nonstationary case. Uncountable description processes can be contrasted with ergodic processes, which satisfy $H(\mathcal{F}) = 0$ by Theorem 8(i).

Acknowledgements

We express our thanks to Jan Mielniczuk, Richard Bradley, Peter Harremoës, and Ronald de Wolf. We are also grateful to the anonymous reviewer for detailed remarks and for letting us know about Sazonov's paper. The manuscript was written in part during the author's work at the Institute of Computer Science, Polish Academy of Sciences.

References

Billingsley, P., 1965. Ergodic Theory and Information. Wiley.
Billingsley, P., 1979. Probability and Measure. Wiley.
Cover, T.M., Thomas, J.A., 1991. Elements of Information Theory. Wiley.

Crutchfield, J.P., Feldman, D.P., 2003. Regularities unseen, randomness observed: The entropy convergence hierarchy. Chaos 15, 25–54.
Dobrushin, R.L., 1959. A general formulation of the fundamental Shannon theorems in information theory. Usp. Mat. Nauk 14 (6), 3–104 (in Russian).
Doob, J.L., 1953. Stochastic Processes. Wiley.
Gelfand, I.M., Kolmogorov, A.N., Yaglom, A.M., 1956. Towards the general definition of the amount of information. Dokl. Akad. Nauk SSSR 111, 745–748 (in Russian).
Gray, R.M., Davisson, L.D., 1974. The ergodic decomposition of stationary discrete random processes. IEEE Trans. Inform. Theory 20, 625–636.
Johnson, R.A., 1970. Atomic and nonatomic measures. Proc. Amer. Math. Soc. 25, 650–655.
Kallenberg, O., 1997. Foundations of Modern Probability. Springer.
Pinsker, M.S., 1964. Information and Information Stability of Random Variables and Processes. Holden-Day.
Sazonov, V.V., 1964. On a question of R.L. Dobrušin. Teor. Veroyatnost. i Primenen. 9 (1), 180–181 (in Russian).
Seidenfeld, T., Schervish, M.J., Kadane, J.B., 2001. Improper regular conditional distributions. Ann. Probab. 29, 1612–1624.
Shields, P.C., 1996. The Ergodic Theory of Discrete Sample Paths. American Mathematical Society.
Swart, J.M., 1996. A conditional product measure theorem. Statist. Probab. Lett. 28, 131–135.
Yeung, R.W., 2002. First Course in Information Theory. Kluwer Academic Publishers.