entropy

MDPI

Article

Approximating Information Measures for Fields

Łukasz Dębowski

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

Received: 20 November 2019; Accepted: 8 January 2020; Published: 9 January 2020

Abstract: We supply corrected proofs of the invariance of completion and the chain rule for the Shannon information measures of arbitrary fields, as stated by Dębowski in 2009. Our corrected proofs rest on a number of auxiliary approximation results for Shannon information measures, which may be of an independent interest. As also discussed briefly in this article, the generalized calculus of Shannon information measures for fields, including the invariance of completion and the chain rule, is useful in particular for studying the ergodic decomposition of stationary processes and its links with statistical modeling of natural language.

Keywords: Shannon information measures; fields; invariance of completion; chain rule

MSC: 94A17

1. Introduction

As it was noticed by Dębowski [1–3], a generalized calculus of Shannon information measures for arbitrary fields—initiated by Gelfand et al. [4] and later developed by Dobrushin [5], Pinsker [6], and Wyner [7]—is useful in particular for studying the ergodic decomposition of stationary processes and its links with statistical modeling of natural language. Fulfilling this need, Dębowski [1] has developed the calculus of Shannon information measures for arbitrary fields, relaxing the requirement of regular conditional probability, assumed implicitly by Dobrushin [5] and Pinsker [6]. He has done it unaware of the classical paper by Wyner [7], which pursued exactly the same idea, with some differences due to an independent interest.

Compared to exposition [7], the added value of the paper [1] was considering continuity and invariance of Shannon information measures with respect to completion of fields. Unfortunately, the proof of Theorem 2 in [1] establishing this invariance and the generalized chain rule contains some mistakes and gaps, which we have discovered recently. For this reason, in this article, we would like to provide a correction and a few new auxiliary results which may be of an independent interest. In this way, we will complete the full generalization of Shannon information measures and their properties, which was developed step-by-step by Gelfand et al. [4], Dobrushin [5], Pinsker [6], Wyner [7], and Dębowski [1]. By the way, we will also rediscuss the linguistic motivations of our results.

The preliminaries are as follows. Fix a probability space $(\Omega, \mathcal{J}, P)$. Fields are set algebras closed under finite Boolean operations, whereas σ-fields are assumed to be closed also under countable unions and products. A field is called finite if it has finitely many elements. A finite partition is a finite collection of events $\left\{B_j\right\}_{j=1}^{J} \subset \mathcal{J}$ which are disjoint and whose union equals Ω. The definition proposed by Wyner [7] and Dębowski [1] independently reads as follows:

Definition 1. *For finite partitions $\alpha = \{A_i\}_{i=1}^{I}$ and $\beta = \{B_j\}_{j=1}^{J}$ and a probability measure P, the entropy and mutual information are defined as*

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

Subsequently, for an arbitrary field $\mathcal{C}$ and finite partitions α and β, we define the pointwise conditional entropy and mutual information as

$$H_P(\alpha||\mathcal{C}) := H_{P(\cdot|\mathcal{C})}(\alpha), \qquad I_P(\alpha;\beta||\mathcal{C}) := I_{P(\cdot|\mathcal{C})}(\alpha;\beta), \tag{2}$$

where $P(E|\mathcal{C})$ is the conditional probability of event $E \in \mathcal{J}$ with respect to the smallest complete σ-field containing $\mathcal{C}$. Subsequently, for arbitrary fields $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$, the (average) conditional entropy and mutual information are defined as

$$H_P(\mathcal{A}|\mathcal{C}) := \sup_{\alpha\subset\mathcal{A}} \mathbf{E}_P H_P(\alpha||\mathcal{C}), \qquad I_P(\mathcal{A};\mathcal{B}|\mathcal{C}) := \sup_{\alpha\subset\mathcal{A},\beta\subset\mathcal{B}} \mathbf{E}_P I(\alpha;\beta||\mathcal{C}), \tag{3}$$

where the supremum is taken over all finite subpartitions and $\mathbf{E}_P X := \int X dP$ is the expectation. Finally, we define the unconditional entropy $H_P(\mathcal{A}) := H_P(\mathcal{A}|\{\emptyset,\Omega\})$ and mutual information $I_P(\mathcal{A};\mathcal{B}) := I_P(\mathcal{A};\mathcal{B}|\{\emptyset,\Omega\})$, as it is generally done in information theory. When the probability measure P is clear from the context, we omit subscript P from all above notations.

Although the above measures, called Shannon information measures, have usually been discussed for σ-fields, the defining equations (3) also make sense for fields. We observe a number of identities, such as $H(\mathcal{A}) = I(\mathcal{A};\mathcal{A})$ and $H(\mathcal{A}|\mathcal{C}) = I(\mathcal{A};\mathcal{A}|\mathcal{C})$. It is important to stress that Definition 1, in contrast to the earlier expositions by Dobrushin [5] and Pinsker [6], is simpler—as it applies one Radon–Nikodym derivative less—and does not require regular conditional probability, i.e., it does not demand that conditional distribution $(P(E|\mathcal{C}))_{E\in\mathcal{J}}$ be a probability measure almost surely. In fact, the expressions on the right-hand sides of the equations in (3) are defined for all $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$. No problems arise when conditional probability is not regular since conditional distribution $(P(E|\mathcal{C}))_{E\in\mathcal{E}}$ restricted to a finite field $\mathcal{E}$ is a probability measure almost surely [8] (Theorem 33.2).

We should admit that in the context of statistical language modeling, the respective probability space is countably generated so regular conditional probability is guaranteed to exist. Thus, for linguistic applications, one might think that expositions [5,6] are sufficient, although for a didactic reason, the approaches proposed by Wyner [7] and Dębowski [1] lead to a simpler and more general calculus of Shannon information measures. Yet, there is a more important reason for Definition 1. Namely, to discuss the ergodic decomposition of entropy rate and excess entropy—some highly relevant results for statistical language modeling, developed in [1] and to be briefly recalled in Section 3—we need the invariance of Shannon information measures with respect to completion of fields. But within the framework of Dobrushin [5] and Pinsker [6], such invariance of completion does not hold for strongly nonergodic processes, which seem to arise quite naturally in statistical modeling of natural language [1–3]. Thus, the approach proposed by Wyner [7] and Dębowski [1] is in fact indispensable.

Thus, let us inspect the problem of invariance of Shannon information measures with respect to completion of fields. A σ-field is called complete, with respect to a given probability measure P, if it contains all sets of outer P-measure 0. Let $\sigma(\mathcal{A})$ denote the intersection of all complete σ-fields containing class $\mathcal{A}$, i.e., $\sigma(\mathcal{A})$ is the completion of the generated σ-field. Let $\mathcal{A} \wedge \mathcal{B}$ denote the intersection of all fields that contain $\mathcal{A}$ and $\mathcal{B}$. Assuming Definition 1, the following statement has been claimed true by Dębowski [1] (Theorem 2):

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

1. $I(\mathcal{A};\mathcal{B}|\mathcal{C}) = I(\mathcal{A};\sigma(\mathcal{B})|\mathcal{C}) = I(\mathcal{A};\mathcal{B}|\sigma(\mathcal{C}))$ *(invariance of completion);*
2. $I(\mathcal{A};\mathcal{B}\wedge\mathcal{C}|\mathcal{D}) = I(\mathcal{A};\mathcal{B}|\mathcal{D}) + I(\mathcal{A};\mathcal{C}|\mathcal{B}\wedge\mathcal{D})$ *(chain rule).*

The property stated in Theorem 1. 1 will be referred to as the invariance of completion. It was not discussed by Wyner [7]. The property stated in Theorem 1. 2 is usually referred to as the chain rule or the polymatroid identity. It was proved independently by Wyner [7].

As we have mentioned, the invariance of completion is crucial to prove the ergodic decomposition of the entropy rate and excess entropy of stationary processes. But the proof of the invariance of completion given by Dębowski [1] contains a mistake in the order of quantifiers, and the respective proof of the chain rule is too laconic and contains a gap. For this reason, we would like to supplement the corrected proofs in this article. As we have mentioned, the chain rule was proved by Wyner [7], using an approximation result by Dobrushin [5] and Pinsker [6]. For completeness, we would like to provide a different proof of this approximation result—which follows easily from the invariance of completion—and to supply proofs of both parts of Theorem 1.

The corrected proofs of Theorem 1, to be presented in Section 2, are much longer than the original proofs by Dębowski [1]. In particular, for the sake of proving Theorem 1, we will discuss a few other approximation results, which seem to be of an independent interest. To provide more context for our statements, in Section 3, we will also recall the ergodic decomposition of excess entropy and its application to statistical language modeling.

2. Proofs

Let us write $\mathcal{B}_n \uparrow \mathcal{B}$ for a sequence $(\mathcal{B}_n)_{n\in\mathbb{N}}$ of fields such that $\mathcal{B}_1 \subset \mathcal{B}_2 \subset \cdots \subset \mathcal{B} = \bigcup_{n\in\mathbb{N}} \mathcal{B}_n$. ($\mathcal{B}$ need not be a σ-field.) Our proof of Theorem 1 will rest on a few approximation results and this statement by Dębowski [1] (Theorem 1):

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

1. $I(\mathcal{A};\mathcal{B}|\mathcal{C}) = I(\mathcal{B};\mathcal{A}|\mathcal{C})$*;*
2. $I(\mathcal{A};\mathcal{B}|\mathcal{C}) \geq 0$ *with the equality if and only if* $P(A\cap B|\mathcal{C}) = P(A|\mathcal{C})P(B|\mathcal{C})$ *almost surely for all* $A \in \mathcal{A}$ *and* $B \in \mathcal{B}$*;*
3. $I(\mathcal{A};\mathcal{B}|\mathcal{C}) \leq \min(H(\mathcal{A}|\mathcal{C}), H(\mathcal{B}|\mathcal{C}))$*;*
4. $I(\mathcal{A};\mathcal{B}_1|\mathcal{C}) \leq I(\mathcal{A};\mathcal{B}_2|\mathcal{C})$ *if* $\mathcal{B}_1 \subset \mathcal{B}_2$*;*
5. $I(\mathcal{A};\mathcal{B}_n|\mathcal{C}) \uparrow I(\mathcal{A};\mathcal{B}|\mathcal{C})$ *for* $\mathcal{B}_n \uparrow \mathcal{B}$*.*

Let $A^c = \Omega \setminus A$. Subsequently, let us denote the symmetric difference

$$A \triangle B := (A\setminus B)\cup(B\setminus A) = (A\cup B)\setminus(A\cap B). \tag{4}$$

Symmetric difference satisfies the following identities, which will be used:

$$A^c \triangle B^c = A \triangle B, \tag{5}$$
$$A\triangle B \subset (A\triangle C)\cup(C\triangle B), \tag{6}$$
$$(A\setminus C)\triangle B \subset (A\triangle B)\cup(C\cap B), \tag{7}$$
$$\left(\bigcup_{i\in C} A_i\right) \triangle \left(\bigcup_{i\in C} B_i\right) \subset \bigcup_{i\in C}(A_i\triangle B_i). \tag{8}$$

Moreover, we will apply the Bonferroni inequalities

$$0 \leq \sum_{1\leq i\leq n} P(A_i) - P\left(\bigcup_{1\leq i\leq n} A_i\right) \leq \sum_{1\leq i<j\leq n} P(A_i\cap A_j) \tag{9}$$

and inequality $P(A) \leq P(B) + P(A \triangle B)$.

In the following, we will derive the necessary approximation results. Our point of departure is the following folklore fact.

Theorem 3 (approximation of σ-fields)**.** *For any field $\mathcal{K}$ and any event $G \in \sigma(\mathcal{K})$, there is a sequence of events $K_1, K_2, \cdots \in \mathcal{K}$ such that*

$$\lim_{n\to\infty} P(G \triangle K_n) = 0. \tag{10}$$

Proof. Denote the class of sets G that satisfy (10) as $\mathcal{G}$. It is sufficient to show that $\mathcal{G}$ is a complete σ-field that contains the field $\mathcal{K}$. Clearly, all $G \in \mathcal{K}$ satisfy (10) so $\mathcal{G} \supset \mathcal{K}$. Now, we verify the conditions for $\mathcal{G}$ to be a σ-field.

1. We have $\Omega \in \mathcal{K}$. Hence, $\Omega \in \mathcal{G}$.
2. For $A \in \mathcal{G}$, consider $K_1, K_2, \cdots \in \mathcal{K}$ such that $\lim_{n\to\infty} P(A \triangle K_n) = 0$. Then, $A \triangle K_n = A^c \triangle K_n^c$, where $K_1^c, K_2^c, \cdots \in \mathcal{K}$. Hence, $A^c \in \mathcal{G}$.
3. For $A_1, A_2, \cdots \in \mathcal{G}$, consider events $K_i^n \in \mathcal{K}$ such that $P(A_i \triangle K_i^n) \leq 2^{-n}$. Then,

$$P\left(\left(\bigcap_{i=1}^{n} A_i\right) \triangle \left(\bigcap_{i=1}^{n} K_i^{i+n}\right)\right) \leq \sum_{i=1}^{n} P(A_i \triangle K_i^{i+n}) \leq 2^{-n}. \tag{11}$$

Moreover,

$$P\left(\left(\bigcap_{i=1}^{\infty} A_i\right) \triangle \left(\bigcap_{i=1}^{n} A_i\right)\right) = P\left(\bigcap_{i=1}^{n} A_i\right) - P\left(\bigcap_{i=1}^{\infty} A_i\right). \tag{12}$$

Hence,

$$\begin{aligned} & P\left(\left(\bigcap_{i=1}^{\infty} A_i\right) \triangle \left(\bigcap_{i=1}^{n} K_i^{i+n}\right)\right) \\ & \leq P\left(\left(\bigcap_{i=1}^{\infty} A_i\right) \triangle \left(\bigcap_{i=1}^{n} A_i\right)\right) + P\left(\left(\bigcap_{i=1}^{n} A_i\right) \triangle \left(\bigcap_{i=1}^{n} K_i^{i+n}\right)\right) \\ & \leq P\left(\bigcap_{i=1}^{n} A_i\right) - P\left(\bigcap_{i=1}^{\infty} A_i\right) - 2^{-n}, \end{aligned} \tag{13}$$

which tends to 0 for n going to infinity. Since $\bigcap_{i=1}^{n} K_i^{i+n} \in \mathcal{K}$, we thus obtain that $\bigcap_{i=1}^{\infty} A_i \in \mathcal{G}$.

Completeness of σ-field $\mathcal{G}$ is straightforward since, for any $A \in \mathcal{G}$ and $P(A \triangle A') = 0$, we obtain $A' \in \mathcal{G}$ using the same sequence of approximating events in field $\mathcal{K}$ as for event A. □

The second approximation result is the following bound:

Theorem 4 (continuity of entropy)**.** *Fix an $\epsilon \in (0, e^{-1}]$ and a field $\mathcal{C}$. For finite partitions $\alpha = \{A_i\}_{i=1}^{I}$ and $\alpha' = \{A_i'\}_{i=1}^{I}$ such that $P(A_i \triangle A_i') \leq \epsilon$ for all $i \in \{1, \ldots, I\}$, we have*

$$\left|H(\alpha|\mathcal{C}) - H(\alpha'|\mathcal{C})\right| \leq I\sqrt{\epsilon}\log\frac{I}{\sqrt{\epsilon}}. \tag{14}$$

Proof. We have the expectation $\int P(A_i \triangle A_i'|\mathcal{C})dP = P(A_i \triangle A_i') \leq \epsilon$. Hence, by the Markov inequality we obtain

$$P(P(A_i \triangle A_i'|\mathcal{C}) \geq \sqrt{\epsilon}) \leq \sqrt{\epsilon}. \tag{15}$$

Denote

$$B = \left(P(A_i \triangle A_i' | \mathcal{C}) < \sqrt{\epsilon}\right) \text{ for all } i \in \{1, \dots, I\}\right). \tag{16}$$

From the Bonferroni inequality, we obtain $P(B^c) \leq I\sqrt{\epsilon}$. Subsequently, we observe that $|H(\alpha||\mathcal{C}) - H(\alpha'||\mathcal{C})| \leq \log I$ holds almost surely. Hence,

$$\begin{aligned} \left|H(\alpha|\mathcal{C}) - H(\alpha'|\mathcal{C})\right| &= \left|\int \left[H(\alpha|\mathcal{C}) - H(\alpha'|\mathcal{C})\right] dP\right| \\ &\leq P(B^c) \log I + \int_B \left|H(\alpha||\mathcal{C}) - H(\alpha'||\mathcal{C})\right| dP \\ &\leq I\sqrt{\epsilon} \log I + \int_B \left|H(\alpha||\mathcal{C}) - H(\alpha'||\mathcal{C})\right| dP. \end{aligned} \tag{17}$$

Function $-x \log x$ is subadditive and increasing for $x \in (0, e^{-1}]$. In particular, we have $|(x+y)\log(x+y) - x \log x| \leq -y \log y$ for $x, y \geq 0$. Thus, on the event B we obtain

$$\begin{aligned} \left|H(\alpha||\mathcal{C}) - H(\alpha'||\mathcal{C})\right| &= \left|\sum_{i=1}^{I} P(A_i'|\mathcal{C}) \log P(A_i'|\mathcal{C}) - \sum_{i=1}^{I} P(A_i|\mathcal{C}) \log P(A_i|\mathcal{C})\right| \\ &\leq -\sum_{i=1}^{I} \left|P(A_i|\mathcal{C}) - P(A_i'|\mathcal{C})\right| \log \left|P(A_i|\mathcal{C}) - P(A_i'|\mathcal{C})\right| \\ &\leq -\sum_{i=1}^{I} P(A_i \triangle A_i'|\mathcal{C}) \log P(A_i \triangle A_i'|\mathcal{C}) \\ &\leq -I\sqrt{\epsilon} \log \sqrt{\epsilon} \end{aligned} \tag{18}$$

Plugging (18) into (17) yields the claim. □

Now, we can prove the invariance of completion. Note that

$$I(\alpha; \beta|\mathcal{C}) = H(\alpha|\mathcal{C}) + H(\beta|\mathcal{C}) - H(\alpha \wedge \beta|\mathcal{C}). \tag{19}$$

Proof of Theorem 1. 1 (invariance of completion): Consider some measurable fields $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$. We are going to demonstrate

$$I(\mathcal{A}; \mathcal{B}|\mathcal{C}) = I(\mathcal{A}; \sigma(\mathcal{B})|\mathcal{C}) = I(\mathcal{A}; \mathcal{B}|\sigma(\mathcal{C})). \tag{20}$$

Equality $I(\mathcal{A}; \mathcal{B}|\mathcal{C}) = I(\mathcal{A}; \mathcal{B}|\sigma(\mathcal{C}))$ is straightforward since $P(A|\mathcal{C}) = P(A|\sigma(\mathcal{C}))$ almost surely for all $A \in \mathcal{J}$. It remains to prove $I(\mathcal{A}; \mathcal{B}|\mathcal{C}) = I(\mathcal{A}; \sigma(\mathcal{B})|\mathcal{C})$. For this goal, it suffices to show that for any $\epsilon > 0$ and any finite partitions $\alpha \subset \mathcal{A}$ and $\beta' \subset \sigma(\mathcal{B})$ there exists a finite partition $\beta \subset \mathcal{B}$ such that

$$\left|I(\alpha; \beta|\mathcal{C}) - I(\alpha; \beta'|\mathcal{C})\right| < \epsilon. \tag{21}$$

Fix then some $\epsilon > 0$ and finite partitions $\alpha := \{A_i\}_{i=1}^{I} \subset \mathcal{A}$ and $\beta' := \left\{B_j'\right\}_{j=1}^{J} \subset \sigma(\mathcal{B})$. Invoking Theorem 3, we know that for each $\eta > 0$ there exists a class of sets $\{C_j\}_{j=1}^{J} \subset \mathcal{B}$ which need not be a partition, such that

$$P(C_j \triangle B_j') \leq \eta \tag{22}$$

for all $j \in \{1, \dots, J\}$. Let us put $B_{J+1}' := \emptyset$ and let us construct sets $D_0 := \emptyset$ and $D_j := \bigcup_{k=1}^{j} C_k$ for $j \in \{1, \dots, J\}$. Subsequently, we put $B_j := C_j \setminus D_{j-1}$ for $j \in \{1, \dots, J\}$ and $B_{J+1} := \Omega \setminus D_J$. In this way, we obtain a partition $\beta := \{B_j\}_{j=1}^{J+1} \subset \mathcal{B}$.

The next step of the proof is showing an analogue of bound (22) for partitions β and β'. To begin, for $j \in \{1, \dots, J\}$, we have

$$\begin{aligned} P(B_j \triangle B'_j) &= P((C_j \setminus D_{j-1}) \triangle B'_j) \le P(C_j \triangle B'_j) + P(D_{j-1} \cap B'_j) \\ &\le \eta + \sum_{k=1}^{j-1} P(C_k \cap B'_j) \\ &\le \eta + \sum_{k=1}^{j-1} \left[P(B'_k \cap B'_j) + P((C_k \cap B'_j) \triangle (B'_k \cap B'_j)) \right] \\ &\le \eta + \sum_{k=1}^{j-1} \left[0 + P(C_k \triangle B'_k) \right] \le j\eta. \end{aligned} \tag{23}$$

Now, we observe for $j, k \in \{1, \dots, J\}$ and $j \ne k$ that

$$P(C_j) \ge P(B'_j) - P(C_j \triangle B'_j) \ge P(B'_j) - \eta \tag{24}$$

$$\begin{aligned} P(C_j \cap C_k) &\le P(B'_j \cap B'_k) + P((C_j \cap C_k) \triangle (B'_j \cap B'_k)) \\ &\le 0 + P(C_j \triangle B'_j) + P(C_k \triangle B'_k) \le 2\eta. \end{aligned} \tag{25}$$

Hence, by the Bonferroni inequality we derive

$$\begin{aligned} P(B_{J+1} \triangle B'_{J+1}) &= P((\Omega \setminus D_J) \triangle \emptyset) = P(\Omega \setminus D_J) = 1 - P(D_J) \\ &\le 1 - \sum_{1 \le j \le J} P(C_j) + \sum_{1 \le j < k \le J} P(C_j \cap C_k) \\ &\le 1 - \sum_{1 \le j \le J} P(B'_j) + J\eta + \sum_{1 \le j < k \le J} 2\eta = J^2 \eta. \end{aligned} \tag{26}$$

Resuming our bounds, we obtain

$$P((A_i \cap B_j) \triangle (A_i \cap B'_j)) \le P(B_j \triangle B'_j) \le J^2 \eta \tag{27}$$

for all $i \in \{1, \dots, I\}$ and $j \in \{1, \dots, J+1\}$. Then, invoking Theorem 4 yields

$$\begin{aligned} \left| I(\alpha; \beta | \mathcal{C}) - I(\alpha; \beta' | \mathcal{C}) \right| &\le \left| H(\alpha \wedge \beta | \mathcal{C}) - H(\alpha \wedge \beta' | \mathcal{C}) \right| + \left| H(\beta | \mathcal{C}) - H(\beta' | \mathcal{C}) \right| \\ &\le I(J+1)\sqrt{J^2 \eta} \log \frac{I(J+1)}{\sqrt{J^2 \eta}} + (J+1)\sqrt{J^2 \eta} \log \frac{J+1}{\sqrt{J^2 \eta}}. \end{aligned} \tag{28}$$

Taking η sufficiently small, we obtain (21), which is the desired claim. □

Some consequence of the above result is this approximation result proved by Dobrushin [5] and Pinsker [6] and used by Wyner [7] to demonstrate the chain rule. Applying the invariance of completion, we supply a different proof than Dobrushin [5] and Pinsker [6].

Theorem 5 (split of join)**.** *Let $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, and $\mathcal{D}$ be subfields of $\mathcal{J}$. We have*

$$I(\mathcal{A}; \mathcal{B} \wedge \mathcal{C} | \mathcal{D}) = \sup_{\alpha \subset \mathcal{A}, \beta \subset \mathcal{B}, \gamma \subset \mathcal{C}} \boldsymbol{E}\, I(\alpha; \beta \wedge \gamma || \mathcal{D}), \tag{29}$$

where the supremum is taken over all finite subpartitions.

Proof. Define class

$$\mathcal{E} := \bigcup_{\beta\subset\mathcal{B},\gamma\subset\mathcal{C}} \sigma(\beta\wedge\gamma). \tag{30}$$

It can be easily verified that $\mathcal{E}$ is a field such that $\sigma(\mathcal{E}) = \sigma(\mathcal{B}\wedge\mathcal{C})$. Thus, for all finite partitions $\beta\subset\mathcal{B}$ and $\gamma\subset\mathcal{C}$ we have $\beta\wedge\gamma\subset\mathcal{E}$. Moreover, by definition of $\mathcal{E}$, for each finite partition $\varepsilon\subset\mathcal{E}$ there exists finite partitions $\beta\subset\mathcal{B}$ and $\gamma\subset\mathcal{C}$ such that partition $\beta\wedge\gamma$ is finer than ε. Hence, by Theorem 2.4, we obtain in this case,

$$\mathbf{E}\, I(\alpha;\varepsilon||\mathcal{D}) \le \mathbf{E}\, I(\alpha;\beta\wedge\gamma||\mathcal{D}) \le I(\alpha;\mathcal{E}|\mathcal{D}). \tag{31}$$

In consequence, by Theorem 1. 1, we obtain the claim

$$\begin{aligned} I(\mathcal{A};\mathcal{B}\wedge\mathcal{C}|\mathcal{D}) = I(\mathcal{A};\mathcal{E}|\mathcal{D}) &= \sup_{\alpha\subset\mathcal{A},\varepsilon\subset\mathcal{E}} \mathbf{E}\, I(\alpha;\varepsilon||\mathcal{D}) \\ &= \sup_{\alpha\subset\mathcal{A},\beta\subset\mathcal{B},\gamma\subset\mathcal{C}} \mathbf{E}\, I(\alpha;\beta\wedge\gamma||\mathcal{D}). \end{aligned} \tag{32}$$

□

The final approximation result which we need to prove the chain rule is as follows:

Theorem 6 (convergence of conditioning)**.** *Let $\alpha = \{A_i\}_{i=1}^I$ be a finite partition and let $\mathcal{C}$ be a field. For each $\epsilon > 0$, there exists a finite partition $\gamma' \subset \sigma(\mathcal{C})$ such that for any partition $\gamma\subset\sigma(\mathcal{C})$ finer than γ' we have*

$$|H(\alpha|\mathcal{C}) - H(\alpha|\gamma)| \le \epsilon. \tag{33}$$

Proof. Fix an $\epsilon > 0$. For each $n\in\mathbb{N}$ and $A\in\mathcal{J}$, partition

$$\gamma_A := \{((k-1)/n < P(A|\mathcal{C}) \le k/n) : k\in\{0,1,\dots,n\}\} \tag{34}$$

is finite and belongs to $\sigma(\mathcal{C})$. If we consider partition $\gamma' := \bigwedge_{i=1}^I \gamma_{A_i}$, it remains finite and still satisfies $\gamma'\subset\sigma(\mathcal{C})$. Let a partition $\gamma\subset\sigma(\mathcal{C})$ be finer than γ'. Then,

$$|P(A_i|\mathcal{C}) - P(A_i|\gamma)| \le 1/n \tag{35}$$

almost surely for all $i\in\{1,\dots,I\}$. We also observe

$$|H(\alpha|\mathcal{C}) - H(\alpha|\gamma)| \le \int |H(\alpha||\mathcal{C}) - H(\alpha||\gamma)|\, dP. \tag{36}$$

We recall that function $-x\log x$ is subadditive and increasing for $x\in(0,e^{-1}]$. In particular, we have $|(x+y)\log(x+y) - x\log x| \le -y\log y$ for $x,y\ge 0$. Hence, for $n\ge e$ we obtain almost surely

$$\begin{aligned} |H(\alpha||\mathcal{C}) - H(\alpha||\gamma)| &= \left| \sum_{i=1}^I P(A_i|\mathcal{C})\log P(A_i|\mathcal{C}) - \sum_{i=1}^I P(A_i|\gamma)\log P(A_i|\gamma) \right| \\ &\le -\sum_{i=1}^I |P(A_i|\mathcal{C}) - P(A_i|\gamma)| \log |P(A_i|\mathcal{C}) - P(A_i|\gamma)| \\ &\le \frac{I\log n}{n}. \end{aligned} \tag{37}$$

Taking n so large that $n^{-1} I \log n \le \epsilon$ yields the claim. □

Taking the above into account, we can demonstrate the chain rule. Our proof essentially follows the ideas of Wyner [7], except for invoking Theorem 6.

Proof of Theorem 1. 2 (chain rule): Let $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, and $\mathcal{D}$ be arbitrary fields, and let α, β, γ, and δ be finite partitions. The point of our departure is the chain rule for finite partitions [9] (Equation 2.60)

$$I(\alpha;\beta \wedge \gamma) = I(\alpha;\beta) + I(\alpha;\gamma|\beta). \tag{38}$$

By Definition 1 and Theorems 1. 1, 5, and 6, conditional mutual information $I(\mathcal{A};\mathcal{B}|\mathcal{C})$ can be approximated by $I(\alpha;\beta|\gamma)$, where we take appropriate limits of refined finite partitions with a certain care.

In particular, by Theorems 1. 1, 5, and 6, taking sufficiently fine finite partitions of arbitrary fields $\mathcal{B}$ and $\mathcal{C}$, the chain rule (38) for finite partitions implies

$$I(\alpha;\mathcal{B} \wedge \mathcal{C}) = I(\alpha;\mathcal{B}) + I(\alpha;\mathcal{C}|\mathcal{B}), \tag{39}$$

where all expressions are finite. Hence, we also obtain

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

where all expressions are finite. Having established the above claim for a finite partition α, we generalize it to

$$I(\mathcal{A};\mathcal{B}\wedge\mathcal{C}|\mathcal{D}) = I(\mathcal{A};\mathcal{B}|\mathcal{D}) + I(\mathcal{A};\mathcal{C}|\mathcal{B}\wedge\mathcal{D}) \tag{40}$$

for an arbitrary field $\mathcal{A}$, taking its appropriately fine finite partitions. □

3. Applications

This section borrows its statements largely from Dębowski [1–3] and is provided only to sketch some context for our research and justify its applicability to statistical language modeling. Let $(X_i)_{i\in\mathbb{Z}}$ be a two-sided infinite stationary process over a countable alphabet $\mathbb{X}$ on a probability space $(\mathbb{X}^{\mathbb{Z}}, \mathcal{X}^{\mathbb{Z}}, P)$, where $X_k((\omega_i)_{i\in\mathbb{Z}}) := \omega_k$. We denote random blocks $X_j^k := (X_i)_{j\le i\le k}$ and complete σ-fields $\mathcal{G}_j^k := \sigma(X_j^k)$ generated by them. By the generalized calculus of Shannon information measures, i.e., Theorems 1 and 2, we can define the entropy rate h_P and the excess entropy E_P of process $(X_i)_{i\in\mathbb{Z}}$ as

$$h_P := \lim_{n\to\infty} H_P(\mathcal{G}_0|\mathcal{G}_{-n}^{-1}) = H_P(\mathcal{G}_0|\mathcal{G}_{-\infty}^{-1}) \text{ if } \mathbb{X} \text{ is finite,} \tag{41}$$

$$E_P := \lim_{n\to\infty} I_P(\mathcal{G}_{-n}^{-1};\mathcal{G}_0^{n-1}) = I_P(\mathcal{G}_{-\infty}^{-1};\mathcal{G}_0^{\infty}), \tag{42}$$

see [10] for more background.

Let $T((\omega_i)_{i\in\mathbb{Z}}) := (\omega_{i+1})_{i\in\mathbb{Z}}$ be the shift operation and let $\mathcal{I} := \left\{A \in \mathcal{X}^{\mathbb{Z}} : T^{-1}(A) = A\right\}$ be the invariant σ-field. By the Birkhoff ergodic theorem [11], we have $\sigma(\mathcal{I}) \subset \sigma(\mathcal{G}_{-\infty}) \cap \sigma(\mathcal{G}_{\infty})$ for the tail σ-fields $\mathcal{G}_{-\infty} := \bigcap_{n=1}^{\infty} \mathcal{G}_{-\infty}^{-n}$ and $\mathcal{G}_{\infty} := \bigcap_{n=1}^{\infty} \mathcal{G}_n^{\infty}$. Hence, by Theorems 1 and 2 we further obtain expressions

$$h_P = H_P(\mathcal{G}_0|\mathcal{G}_{-\infty}^{-1}) = H_P(\mathcal{G}_0|\mathcal{G}_{-\infty}^{-1} \wedge \mathcal{I}) \text{ if } \mathbb{X} \text{ is finite,} \tag{43}$$

$$E_P = I_P(\mathcal{G}_{-\infty}^{-1};\mathcal{G}_0^{\infty}) = H_P(\mathcal{I}) + I_P(\mathcal{G}_{-\infty}^{-1};\mathcal{G}_0^{\infty}|\mathcal{I}). \tag{44}$$

Denoting the conditional probability $F(A) := P(A|\mathcal{I})$, which is a random stationary ergodic measure by the ergodic decomposition theorem [12], we notice that $H_P(\mathcal{G}_0|\mathcal{G}_{-\infty}^{-1} \wedge \mathcal{I}) = \mathbf{E}\,_P H_F(\mathcal{G}_0|\mathcal{G}_{-\infty}^{-1})$ and $I_P(\mathcal{G}_{-\infty}^{-1};\mathcal{G}_0^{\infty}|\mathcal{I}) = \mathbf{E}\,_P I_F(\mathcal{G}_{-\infty}^{-1};\mathcal{G}_0^{\infty})$, and consequently we obtain the ergodic decomposition of the entropy rate and excess entropy, which reads

$$h_P = \mathbf{E}\,_P h_F \text{ if } \mathbb{X} \text{ is finite,} \tag{45}$$

$$E_P = H_P(\mathcal{I}) + \mathbf{E}\,_P E_F. \tag{46}$$

Formulae (45) and (46) were derived by Gray and Davisson [13] and Dębowski [1] respectively. The ergodic decomposition of the entropy rate (45) states that a stationary process is asymptotically deterministic, i.e., $h_P = 0$, if and only if almost all its ergodic components are asymptotically deterministic, i.e., $h_F = 0$ almost surely. In contrast, the ergodic decomposition of the excess entropy (46) states that a stationary process is infinitary, i.e., $E_P = \infty$, if some of its ergodic components are infinitary, i.e., $E_F = \infty$ with a nonzero probability, or if $H_P(\mathcal{I}) = \infty$, i.e., if the process is strongly nonergodic in particular, see [14,15].

The linguistic interpretation of the above results is as follows. There is a hypothesis by Hilberg [16] that the excess entropy of natural language is infinite. This hypothesis can be partly confirmed by the original estimates of conditional entropy by Shannon [17], by the power-law decay of the estimates of the entropy rate given by the PPM compression algorithm [18], by the approximately power-law growth of vocabulary called Heaps' or Herdan's law [2,3,19,20], and by some other experiments applying neural statistical language models [21,22]. In parallel, Dębowski [1–3] supposed that the very large excess entropy in natural language may be caused by the fact that texts in natural language describe some relatively slowly evolving and very complex reality. Indeed, it can be mathematically proved that if the abstract reality described by random texts is unchangeable and infinitely complex, then the resulting stochastic process is strongly nonergodic, i.e., $H_P(\mathcal{I}) = \infty$ in particular [1–3]. Consequently, its excess entropy is infinite by formula (46). We suppose that a similar mechanism may work for natural language, see [23–26] for further examples of abstract stochastic mechanisms leading to infinitary processes.

Funding: This research received no external funding.

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

References

1. Dębowski, Ł. A general definition of conditional information and its application to ergodic decomposition. *Stat. Probab. Lett.* **2009**, *79*, 1260–1268. [CrossRef]
2. Dębowski, Ł. On the Vocabulary of Grammar-Based Codes and the Logical Consistency of Texts. *IEEE Trans. Inf. Theory* **2011**, *57*, 4589–4599. [CrossRef]
3. Dębowski, Ł. Is Natural Language a Perigraphic Process? The Theorem about Facts and Words Revisited. *Entropy* **2018**, *20*, 85. [CrossRef]
4. Gelfand, I.M.; Kolmogorov, A.N.; Yaglom, A.M. Towards the general definition of the amount of information. *Dokl. Akad. Nauk. SSSR* **1956**, *111*, 745–748. (In Russian)
5. Dobrushin, R.L. A general formulation of the fundamental Shannon theorems in information theory. *Uspekhi Mat. Nauk.* **1959**, *14*, 3–104. (In Russian)
6. Pinsker, M.S. *Information and Information Stability of Random Variables and Processes*; Holden-Day: San Francisco, CA, USA, 1964.
7. Wyner, A.D. A definition of conditional mutual information for arbitrary ensembles. *Inf. Control.* **1978**, *38*, 51–59. [CrossRef]
8. Billingsley, P. *Probability and Measure*; John Wiley: New York, NY, USA, 1979.
9. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; John Wiley: New York, NY, USA, 1991.
10. Crutchfield, J.P.; Feldman, D.P. Regularities unseen, randomness observed: The entropy convergence hierarchy. *Chaos* **2003**, *15*, 25–54. [CrossRef]

11. Birkhoff, G.D. Proof of the ergodic theorem. *Proc. Natl. Acad. Sci. USA* **1932**, *17*, 656–660. [CrossRef]
12. Rokhlin, V.A. On the fundamental ideas of measure theory. *Am. Math. Soc. Transl. Ser. 1* **1962**, *10*, 1–54.
13. Gray, R.M.; Davisson, L.D. The ergodic decomposition of stationary discrete random processses. *IEEE Trans. Inf. Theory* **1974**, *20*, 625–636. [CrossRef]
14. Löhr, W. Properties of the Statistical Complexity Functional and Partially Deterministic HMMs. *Entropy* **2009**, *11*, 385–401. [CrossRef]
15. Crutchfield, J.P.; Marzen, S. Signatures of infinity: Nonergodicity and resource scaling in prediction, complexity, and learning. *Phys. Rev. E* **2015**, *91*, 050106. [CrossRef]
16. Hilberg, W. Der bekannte Grenzwert der redundanzfreien Information in Texten—eine Fehlinterpretation der Shannonschen Experimente? *Frequenz* **1990**, *44*, 243–248. [CrossRef]
17. Shannon, C. Prediction and entropy of printed English. *Bell Syst. Tech. J.* **1951**, *30*, 50–64. [CrossRef]
18. Takahira, R.; Tanaka-Ishii, K.; Dębowski, Ł. Entropy Rate Estimates for Natural Language—A New Extrapolation of Compressed Large-Scale Corpora. *Entropy* **2016**, *18*, 364. [CrossRef]
19. Herdan, G. *Quantitative Linguistics*; Butterworths: London, UK, 1964.
20. Heaps, H.S. *Information Retrieval—Computational and Theoretical Aspects*; Academic Press: New York, NY, USA, 1978.
21. Hahn, M.; Futrell, R. Estimating Predictive Rate-Distortion Curves via Neural Variational Inference. *Entropy* **2019**, *21*, 640. [CrossRef]
22. Braverman, M.; Chen, X.; Kakade, S.M.; Narasimhan, K.; Zhang, C.; Zhang, Y. Calibration, Entropy Rates, and Memory in Language Models. *arXiv* **2019**, arXiv:1906.05664.
23. Dębowski, Ł. Mixing, Ergodic, and Nonergodic Processes with Rapidly Growing Information between Blocks. *IEEE Trans. Inf. Theory* **2012**, *58*, 3392–3401. [CrossRef]
24. Dębowski, Ł. On Hidden Markov Processes with Infinite Excess Entropy. *J. Theor. Probab.* **2014**, *27*, 539–551. [CrossRef]
25. Travers, N.F.; Crutchfield, J.P. Infinite Excess Entropy Processes with Countable-State Generators. *Entropy* **2014**, *16*, 1396–1413. [CrossRef]
26. Dębowski, Ł. Maximal Repetition and Zero Entropy Rate. *IEEE Trans. Inf. Theory* **2018**, *64*, 2212–2219. [CrossRef]