

Article

Not peer-reviewed version

An Intrinsic Characterization of Shannon's and Rényi's Entropy

[Martin Schlather](#) * and [Carmen Ditscheid](#)

Posted Date: 19 November 2024

doi: 10.20944/preprints202411.1373.v1

Keywords: chain rule; characterization; min-entropy; Rényi entropy; Shannon entropy



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Article

An Intrinsic Characterization of Shannon's and Rényi's Entropy

Martin Schlather * and Carmen Ditscheid

University of Mannheim; Institute of Mathematics

* Correspondence: martin.schlather@uni-mannheim.de; Tel.: +49-621-181-2552

Abstract: All characterizations of Shannon's entropy include the so-called chain rule, a formula on a hierarchically structured probability distribution, which is based on at least two elementary distributions. We show that the chain rule can be split into two natural components, the well-known additivity of the entropy in case of cross-products and a variant of the chain rule that involves only a single elementary distribution. The latter is given as a proportionality relation and hence allows a vague interpretation as self-similarity, hence intrinsic property of the Shannon entropy. A similar characterization is given also for the Rényi entropy and the min-entropy.

Keywords: chain rule; characterization; min-entropy; Rényi entropy; Shannon entropy

0. Introduction

The Shannon entropy H has been fully characterized by different sets of conditions [1–4]. The conditions themselves have been investigated deeply by [5,6], for instance. All characterizations use variants of the so-called chain rule, which can be summarized as follows. Let $p = \{p_1, \dots, p_n\}$ and $q = \{q_1, \dots, q_m\}$ be discrete probability distributions, and

$$p_{k,q} = \{p_1, \dots, p_{k-1}, p_k q_1, \dots, p_k q_m, p_{k+1}, \dots, p_n\}. \quad (1)$$

The chain rule states that

$$H(p_{k,q}) = H(p) + p_k H(q). \quad (2)$$

The construction in Equation (1) can be iterated, i.e., p is extended at several positions k_i with different distributions $q^{(i)}$. In case of two positions, this is

$$p_{k_1,q^{(1)};k_2,q^{(2)}} := (p_{k_2,q^{(2)}})_{k_1,q^{(1)}},$$

where $k_1 < k_2$. The most elementary version of the chain rule is due to [2], where q in Equation (1) is restricted to a Bernoulli distribution. [7] rely on the full version of Equation (1), where q is arbitrary. Since the full version allows an immediate interpretation in various areas of application, they consider Equation (1) as preferable from a didactical point of view. Often, the completely iterated version,

$$H(p_{1,q^{(1)};\dots;n,q^{(n)}}) = H(p) + \sum_{k=1}^n p_k H(q^{(k)}), \quad (3)$$

is taken [4], since this version allows the practical interpretation that first an alphabet α_k is chosen with probability p_k and subsequently a letter within α_k is chosen according to $q^{(k)}$.

[8] give the first algebraic approach, namely in terms of information loss. The latter is defined as $F_{p,q} = H(p) - H(q)$, where the probability distribution q is a function of p , hence $F_{p,q}$ quantifies the change in entropy of the transformation. The information loss $F_{p,q}$ is uniquely characterized by a number of properties, including $F_{p,q}$ being a convex linear map, i.e., $H(\lambda p \cup (1 - \lambda)q) = \lambda H(p) + (1 - \lambda)H(q)$, which can be considered as a restatement of Faddeev's chain rule.

Common to all versions of the chain rule is that they describe situations where canonically at least two different distributions, p and q , are involved. [Definition 3.1.1] [6] introduce an alternative notation of Faddeev's condition, which formally considers only one distribution but nevertheless relies on computing the Shannon entropy of two distributions, namely p and a Bernoulli distribution, to state the chain rule.

The Rényi entropy generalizes the Shannon entropy and can be fully characterized via quasi-arithmetic means [Section 5] [5]. Another approach to characterize the Rényi entropy relies on a transform. The order- α information energy S with Onicescu's information energy as a special case [9] is, up to a multiplicative constant, the sum of the α -powers of the probabilities. [10] gives a full characterization by means of a chain rule that is similar to Faddeev's variant,

$$S(p_{k,q}) = S(p) - p_1^\alpha (\text{const} - S(q)), \quad (4)$$

where q is a Bernoulli distribution and the constant depends on α only. [10] also generalizes this equation by introducing weighted sums. [Theorem 4.5.1] [4] characterizes the information energy by a formula that is analogous to the fully iterated version of the chain rule (3).

[11] deals with a rather general approach to the entropy, which is based on the additivity assumption of the entropy for independent systems. Under certain conditions, called scale invariance there, the entropy is unique. This result, however, is not applicable here, since it is closely related to scalar real numbers and not to probability distributions. Yet, the characterizations given here are in this spirit. To be specific, let $q = p$ in Equation (1). Then, we get

$$p_{k,p} = \{p_1, \dots, p_{k-1}, p_k p_1, \dots, p_k p_n, p_{k+1}, \dots, p_n\}, \quad (5)$$

and hence a proportionality relation for the Shannon entropy by Equation (2),

$$H(p_{k,p}) = (1 + p_k)H(p). \quad (6)$$

Theorem 1 in Section 1 states that the chain rule in Khinchin's (1953) characterization [1,2] can be split into two canonical components, the additivity property and the proportionality between $H(p_{k,p})$ and $H(p)$. The characterization of the Rényi entropy is in analogy to the characterization of the Shannon entropy, except that the functional Equation (4), i.e.,

$$S(p_{k,p}) = (1 + p_k^\alpha)S(p) - p_k^\alpha,$$

reflects the proportionality relation. A related characterization also exists for the min-entropy, cf. Theorem 3. All technical parts are postponed to Section 2. Section 3 finalizes with some comments on our approach.

1. Results

We denote by $\{.\}$ an ordered set und let

$$\begin{aligned} a\{p_1, p_2, \dots\} &= \{ap_1, ap_2, \dots\}, & a \geq 0, \\ \{p_1, p_2, \dots\} \cup \{q_1, q_2, \dots\} &= \{p_1, p_2, \dots, q_1, q_2, \dots\}. \end{aligned} \quad (7)$$

As we will address only the very first elements of an ordered set, the intuitive, but sloppy notation on the right hand side of Equation (7) is sufficient.

Let denote by P the set of all discrete probability distributions and by $P_n \subset P$ the set of distributions with at most n atomic events that have a probability greater than zero. We denote such a probability distribution by $\{p_1, \dots, p_n\} \in P_n$. The uniform distribution is denoted by $U_n = \{1/n, \dots, 1/n\}$.

Reordering the elements in Equation (5) will turn out to be very useful in the proofs. We define for $p \in P$

$$\begin{aligned} p_{\hat{k}} &= \{p_i : i \neq k\}, \\ \partial_k p &= p_k p \cup p_{\hat{k}}. \end{aligned} \quad (8)$$

For instance,

$$\partial_3 \{p_1, p_2, p_3\} = \{p_3 p_1, p_3 p_2, p_3 p_3, p_1, p_2\}.$$

We use the convention $0 \log 0 = 0$.

Theorem 1. Let $H : P \rightarrow [0, \infty]$ be a function. Then the following two assertions are equivalent:

1. H is the Shannon entropy,

$$H(p) = - \sum_{p_i \in p} p_i \log p_i,$$

up to a positive multiplicative constant.

2. H has the following properties:

- (a) H is a continuous function in the topology of convergence in distribution;
- (b) $H(\{p_1, \dots, p_n\}) = H(\{p_{\pi(1)}, \dots, p_{\pi(n)}\})$ for any permutation π and $\{p_1, \dots, p_n\} \in P_n, n \in \mathbb{N}$;
- (c) $H(\{p_1, \dots, p_n\}) = H(\{0, p_1, \dots, p_n\})$ for all $\{p_1, \dots, p_n\} \in P_n, n \in \mathbb{N}$;
- (d) $H(p) \leq H(U_n) < \infty$ for all $p \in P_n, n \in \mathbb{N}$;
- (e) $H(p) \in (0, \infty)$ if p is a non-degenerate geometric distribution;
- (f) $H(p \times q) = H(p) + H(q)$ for all $p, q \in P$;
- (g) some function $f : [0, 1] \rightarrow [0, \infty)$ exists such that for all $p \in P$ we have

$$H(\partial_1 p) = (1 + f(p_1))H(p). \quad (9)$$

Conds. 2a-2g can be regarded as interpretable, hence intrinsic properties, cf. [Section 1.2] [5]. For this reason we prefer Cond. 2d over a monotonicity assumption on H . Note that Conds. 2a-2d are precisely Khinchin's (1953) conditions as given in [2] except for replacing the chain rule by Conds. 2e-2g.

Remark 1. Note that f is not unique since it is undetermined at 1. A convenient choice is to take $f(1) = \lim_{a \rightarrow 1} f(a)$.

Remark 2. Our approach has a simple, but nice implication. Let $p, q \in P_n, n \in \mathbb{N}$. Then,

$$\text{KL}(\partial_1 p, \partial_1 q) = (1 + p_1) \text{KL}(p, q),$$

where KL denotes the Kullback-Leibler divergence.

In contrast to our characterization of the Shannon entropy, our analogous characterization of the Rényi entropy needs stronger assumptions on the unknown function f .

Definition 1 (Def 2.3.2 in [12]). A function $f : U \rightarrow \mathbb{R}, U \subset \mathbb{R}$, is called real analytic if for each $y \in U$ a neighborhood U_y of y exists, such that for all $x \in U_y$ there is an absolutely and uniformly convergent series representation of the form $f(x) = \sum_k a_k (x - y)^k$.

In the following, we call a function $f : [0, 1] \rightarrow [0, \infty)$ smooth if $\prod_{k=0}^{\infty} (1 + f(a^{2^k})) < \infty$ for all $a \in (0, 1)$ and if it can be extended to a real analytic function on a domain including $(0, 1]$.

Theorem 2. Let $H : P \rightarrow [0, \infty]$ be a function. Then, the following two assertions are equivalent:

1. H is the Rényi entropy,

$$H(p) = \frac{1}{1-\alpha} \log \sum_{p_i \in p} p_i^\alpha,$$

for some $\alpha \in (0, \infty) \setminus \{1\}$, up to a positive multiplicative constant.

2. H satisfies Conds. 2a-2f of Theorem 1. A value $\gamma \in \mathbb{R} \setminus \{0\}$ and a smooth function $f : [0, 1] \rightarrow [0, \infty)$ exist such that, for $S = e^{\gamma H}$, we have

$$S(\partial_1 p) = (1 + f(p_1))S(p) - f(p_1) \quad \forall p \in P. \quad (10)$$

Since Theorem 2 gives a characterization only up to a multiplicative constant, the missing constant is typically determined by the requirement that the Shannon entropy is canonically obtained as $\alpha \rightarrow 1$. The limit of the Rényi entropy as $\alpha \rightarrow \infty$ is called the min-entropy and will be considered next. Two major differences to the two previous theorems exist: First, the function f appears on the left hand side of the proportionality relation. Second, two characterizing equations are necessary. Unfortunately, the nice interpretability of Equation (9) reduces in Equation (10) and even further in Equations (11) and (12).

Theorem 3. Let $H : P \rightarrow [0, \infty]$ be a function. Then, the following two assertions are equivalent:

1. H is the min-entropy,

$$H(p) = -\log \max_{p_i \in P} p_i.$$

2. H satisfies Conds. 2a-2c, 2e and 2f of Theorem 1. Furthermore, $H(p) < \infty$ for $p \in P_n$, $n \in \mathbb{N}$. A smooth, monotone function $f : [0, 1] \rightarrow [0, \infty)$ exists, such that for $S = e^{-H}$ and all $p \in P \setminus \{\{1\}\}$, we have

$$f(p_1) \vee S(\partial_1 p) = S(p), \quad (11)$$

$$(1 + f(p_1))S\left(\frac{(\partial_1 p)_1}{1 - p_1^2}\right) = S\left(\frac{p_1}{1 - p_1}\right). \quad (12)$$

2. Proofs

2.1. Auxiliary Results

For the reader's convenience, we first repeat several known functional equations and properties.

Lemma 1. Let H be the Shannon entropy or the Rényi entropy. Then, $H(U_n)$ is maximal on P_n .

See, for instance, Lemma 2.2.4 and Remark 4.4.4 in [4] for a proof.

Lemma 2. Let $H : P \rightarrow [0, \infty]$ be a function fulfilling Conds. 2d and 2f of Theorem 1. Then, $H(U_n) = c \log n$ for all $n \in \mathbb{N}$ and some $c > 0$.

Proof. The inclusion $P_n \subset P_{n+1}$ and Cond. 2d immediately imply monotonicity, i.e.

$$H(U_n) \leq H(U_{n+1}) \quad \forall n \in \mathbb{N}.$$

By Cond. 2f and Implication 1 of Theorem V in [13] follows the claim.

□

Lemma 3 (Identity theorem, Corollary 1.2.7 in [14]). *If f and g are real analytic functions on an open interval $U \subset \mathbb{R}$ and if there is a sequence of distinct points x_1, x_2, \dots in U with $x_\infty = \lim_{n \rightarrow \infty} x_n \in U$ such that $f(x_n) = g(x_n)$ for all $n \in \mathbb{N}$, then*

$$f(x) = g(x) \quad \text{for all } x \in U.$$

The proof for the Shannon entropy needs the uniqueness of the solution of the functional Equation (13) below. We state two lemmas for the same purpose. The first needs the strong additional assumption that the solution is real analytic. The lemma itself remains unused, but the ideas behind the proof are used several times. The second lemma states the version used in Theorem 1; see Theorem 6.4.8 in [6] for a condition similar to Equation (16).

Lemma 4. *Let f be a real-valued function on a domain that includes $[0, 1)$. The following two assertions are equivalent:*

1. $f(a_0) = a_0$ for some $a_0 \in (0, 1)$, f is real analytic and

$$(1 - f(a)) \prod_{k=0}^{\infty} (1 + f(a^{2^k})) = 1 \quad \forall a \in (0, 1); \quad (13)$$

2. f is the identity.

Proof. By Equation 0.266 in [15], the identity satisfies Equation (13). Assume now that Equation (13) holds. Then,

$$f(a^2) = f^2(a),$$

since

$$\begin{aligned} 1 &= (1 - f(a))(1 + f(a)) \prod_{k=1}^{\infty} (1 + f(a^{2^k})) \\ &= (1 - f(a))(1 + f(a)) / (1 - f(a^2)). \end{aligned} \quad (14)$$

It follows that

$$f(a^{2^k}) = (f(a))^{2^k} \quad \forall a \in (0, 1), k \in \mathbb{Z}. \quad (15)$$

Since $a^{2^k} \rightarrow 0$ as $k \rightarrow \infty$, Lemma 3 yields that f is uniquely determined by the knowledge of a single value $f(a_0)$ with $a_0 \in (0, 1)$. \square

Lemma 5. *Let f be a real-valued function on $(0, 1)$. The following two assertions are equivalent:*

1. Equation (13) holds, f is continuous and for all $a \in (0, 1)$ we have

$$f(a) + f(1 - a) = 1; \quad (16)$$

2. f is the identity.

Proof. Equation (16) yields $f(1/2) = 1/2$. Equations (15) and (16) imply that $f((1 - a^{2^k})^{2^r}) = (1 - f^{2^k}(a))^{2^r}$ for all $a \in (0, 1)$ and $k, r \in \mathbb{Z}$. Let $T(A) = \{(1 - a^{2^k})^{2^r} : a \in A, r, k \in \mathbb{Z}\}$ for $A \subset (0, 1)$, $D_1 = T(\{1/2\})$, $D_{n+1} := D_n \cup T(D_n)$ and $D = \bigcup_n D_n$. It suffices to show that D is dense in $(0, 1)$.

Assume that it is not i.e., an interval $(a, b) \subset (0, 1)$ exists with $(a, b) \cap D = \emptyset$. Since $1/2 \in D$, we may assume without loss of generality, that $(a, b) \subset (1/2, 1)$. Then,

$$(a^2, b^2) \cap D = (1 - b^2, 1 - a^2) \cap D = \emptyset.$$

Consider the map f on the open intervals of $(0, 1)$,

$$f((a, b)) \mapsto \begin{cases} (a^2, b^2), & \text{if } a \geq 1/2, b^2 > 1/2 \\ (1 - b^2, 1 - a^2), & \text{if } a \geq 1/2, b^2 \leq 1/2, \\ (a, b) & a < 1/2 \end{cases}$$

Let $f^{(n)}$ be the n -times iterative application of f , i.e.,

$$f^{(n)} = f \circ \dots \circ f$$

and $(c, d) = f^{(n-1)}((a, b))$. Per assumption f maps $(a, b) \subset (1/2, 1)$ to either (a^2, b^2) or $(1 - b^2, 1 - a^2)$. Per construction, if $f^{(n)}((a, b)) = (1 - c^2, 1 - d^2)$, then $f^{(n+1)}((a, b)) \neq (1 - c^2, 1 - d^2)$. If eventually $f^{(n)}((a, b)) = f^{(n+1)}((a, b)) = (c^2, d^2)$, then it always holds $1/2 \in (c^2, d^2)$. Let $| (a, b) |$ be the length of an interval (a, b) , i.e., $b - a$. If $a \geq 1/2$, then

$$|f(a, b)| = b^2 - a^2 = (b + a)(b - a) > b - a.$$

Hence, $f^{(n+1)}((a, b))$ finally contains $1/2$ or its length is strictly monotoneously increasing. If the latter were true, the limit would exist as $|f| \leq 1/2$. In the limit we would have $b^2 - a^2 = b - a$, a contradiction to $b > a \geq 1/2$. \square

2.2. Properties of the Operator ∂_k

We will use the operator ∂_k , defined in Equation (8), iteratively. For instance,

$$\begin{aligned} \partial_2 \partial_k p &= (p_2 p_k)(p_k p \cup p_{\hat{k}}) \cup (p_k p \cup p_{\hat{k}})_2, \quad 1 \leq k \leq n, n \geq 2, \\ \partial_2 \partial_3 \{p_1, p_2, p_3\} &= \partial_2 \{p_3 p_1, p_3 p_2, p_3 p_3, p_1, p_2\} \\ &= \{p_3 p_2 p_3 p_1, p_3 p_2 p_3 p_2, p_3 p_2 p_3 p_3, p_3 p_2 p_1, p_3 p_2 p_2, \\ &\quad p_3 p_1, p_3 p_3, p_1, p_2\} \\ &= \{p_1 p_2 p_3^2, p_2^2 p_3^2, p_2 p_3^3, p_1 p_2 p_3, p_2^2 p_3, p_1 p_3, p_2^2, p_1, p_2\}. \end{aligned}$$

Note that the terms in general do not commute, e.g., $(p_k p \cup p_{\hat{k}})_{\hat{k}} \neq (p_{\hat{k}} \cup p_k p)_{\hat{k}}$ and $\partial_i \partial_k p \neq \partial_k \partial_i p$. The proofs of the theorems in Section 1 rely on the properties of various probability distributions. We denote by b_a the Bernoulli distribution, i.e.,

$$b_a = \{a, 1 - a\}, \quad a \in (0, 1),$$

and by g_a the geometric distribution, i.e.,

$$g_a = \{1 - a, (1 - a)a, (1 - a)a^2, \dots\}, \quad a \in (0, 1).$$

For $n \in \mathbb{N}$, let

$$g_{n,a} = \frac{1-a}{1-a^n} \{1, a, a^2, \dots, a^{n-1}\}, \quad a \in [0, 1),$$

$$G_{n,a} = \{a^n, 1 - a, (1 - a)a, \dots, (1 - a)a^{n-1}\}, \quad a \in [0, 1], \quad (17)$$

$$U_{k,n} = \{1 - k/n, \underbrace{1/n, \dots, 1/n}_{k \text{ times}}\}, \quad 0 \leq k < n. \quad (18)$$

The geometric distribution plays a dominant role in all of the subsequent proofs. Its important role has been known in information theory: it maximizes the Shannon entropy among discrete probability distributions given the mean [Theorem 5.8] [16] and minimizes the min-entropy under fixed variance, among all discrete log-concave random variables in \mathbb{Z} [Theorem 1] [17]. Let denote by \doteq equality up to inserting zeros and reordering.

Lemma 6. Let $p \in P$ and $\partial_1^\infty p := \lim_{n \rightarrow \infty} \partial_1^n p$. Then,

$$\partial_1^\infty p \doteq g_{p_1} \times \frac{p_{\hat{1}}}{1 - p_1}, \quad p_1 < 1, \quad (19)$$

$$\{1\} \doteq \{1\} \times \{1\}. \quad (20)$$

In particular, for $a \in (0, 1)$, $n \in \mathbb{N}$ and $0 \leq k < n$, we have

$$\partial_1^\infty b_a \doteq g_a, \quad (21)$$

$$\partial_1^\infty g_a \doteq g_{1-a} \times g_a, \quad (22)$$

$$\partial_1^\infty G_{n,a} \doteq g_{a^n} \times g_{n,a} \doteq g_a \doteq G_{\infty,a}, \quad (23)$$

$$\partial_2^\infty G_{n,a} \doteq g_{1-a} \times G_{n-1,a}, \quad (24)$$

$$\partial_1^\infty U_{k,n} \doteq g_{1-k/n} \times U_k, \quad (25)$$

$$\partial_2^\infty U_{k,n} \doteq g_{1/n} \times U_{k-1,n-1}, \quad 0 \leq k < n, \quad (26)$$

$$\partial_3^\infty \partial_2 b_a \doteq g_a \times b_a. \quad (27)$$

Furthermore, for $a \in (0, 1)$ and $n \in \mathbb{N}$,

$$\partial_1 G_{n,a} \doteq G_{2n,a}, \quad (28)$$

$$(G_{n,a})_{\hat{2}} = a G_{n-1,a}, \quad (29)$$

$$(1 - a^n) g_{n,a} = (G_{n,a})_{\hat{1}},$$

$$U_{n-1,n} = U_n.$$

Proof. We have

$$\partial_1^n p = \{p_1^{2^n}\} \cup \bigcup_{k \in \{2^n-1, \dots, 0\}} p_1^k p_{\hat{1}},$$

since this is true for $n = 1$ and

$$\begin{aligned} \partial_1^n p &= \partial_1(\partial_1^{n-1} p) = \partial_1 \left(\{p_1^{2^{n-1}}\} \cup \bigcup_{k \in \{2^{n-1}-1, \dots, 0\}} p_1^k p_{\hat{1}} \right) \\ &= p_1^{2^{n-1}} \left(\{p_1^{2^{n-1}}\} \cup \bigcup_{k \in \{2^{n-1}-1, \dots, 0\}} p_1^k p_{\hat{1}} \right) \cup \bigcup_{k \in \{2^{n-1}-1, \dots, 0\}} p_1^k p_{\hat{1}} \\ &= \{p_1^{2^n}\} \cup \bigcup_{k \in \{2^n-1, \dots, 2^{n-1}\}} p_1^k p_{\hat{1}} \cup \bigcup_{k \in \{2^{n-1}-1, \dots, 0\}} p_1^k p_{\hat{1}} \end{aligned}$$

by induction. Hence,

$$\partial_1^\infty p \doteq \bigcup_{k \in \{\dots, 1, 0\}} p_1^k p_{\hat{1}}.$$

Now, $p_1^k p_1 = ((1 - p_1)p_1^k) \cdot \frac{p_1}{1-p_1}$ for $p_1 < 1$, implying Equation (19). Further, we have

$$\partial_2 b_a = (1 - a)b_a \cup \{a\},$$

so that Equation (27) holds by Equation (19). The other equalities are immediate. \square

2.3. Proof of Theorem 1

Assume that H is the Shannon entropy. Then,

$$H(g_a) = - \sum_{k=0}^{\infty} (1-a)a^k \log((1-a)a^k) = -(a \log a + (1-a) \log(1-a))/(1-a) \in (0, \infty)$$

for $a \in (0, 1)$. All the other assertions in the second condition hold obviously, when choosing f to be the identity. Now, assume that the second condition holds. By Cond. 2a, Cond. 2b holds for all $p \in P$. Cond. 2f and Equation (20) yield $H(\{1\}) = 0$. It is easy to check that

$$H(\partial_1^n p) = H(p) \prod_{k=0}^{n-1} (1 + f(p_1^{2^k})).$$

Thus,

$$H(\partial_1^\infty p) = H(\lim_{n \rightarrow \infty} \partial_1^n p) = \lim_{n \rightarrow \infty} H(\partial_1^n p) = H(p)F(p_1) \quad (30)$$

with

$$F(a) = \prod_{k=0}^{\infty} (1 + f(a^{2^k})), \quad a \in (0, 1).$$

The function F is finite on $(0, 1)$ due to Cond. 2e, Cond. 2f, and Equations (22) and (30), i.e.,

$$H(g_a)F(1-a) = H(g_{1-a}) + H(g_a).$$

It follows

$$F(1-a) - 1 = \frac{H(g_{1-a})}{H(g_a)}, \quad a \in (0, 1).$$

Replacing a by $1-a$ in the equation above yields

$$(F(1-a) - 1)(F(a) - 1) = 1, \quad a \in (0, 1). \quad (31)$$

Now,

$$H(b_a)F(a) = H(g_a), \quad (32)$$

by Equations (21) and (30). Equations (27), (9) and (30) yield

$$H(b_a)(1 + f(1-a))F(a) = H(g_a) + H(b_a). \quad (33)$$

Since $H(b_a) > 0$ by Equation (32) and Cond. 2e, we immediately get that

$$f(1-a)F(a) = 1, \quad a \in (0, 1). \quad (34)$$

Equation (31) now writes $f(a) + f(1-a) = 1$. Plugging in $p = b_a$ into Equation (9) shows that f is continuous on $(0, 1)$. Equation (34) and Lemma 5 yield that $f(a) = a$ and $F(a) = 1/(1-a)$ for $a \in (0, 1)$. Equation (9) implies that $f(0) = 0$ so that f is the identity on $[0, 1)$. Lemma 2 implies that

$H(U_n) = c \log n$ for some $c > 0$. Henceforth, we may assume that $c = 1$. Cond. 2f and Equations (19) and (30) deliver

$$H(p)F(p_1) = H(g_{p_1}) + H\left(\frac{p_1}{1-p_1}\right), \quad p_1 < 1. \quad (35)$$

This implies

$$\log n \cdot F(1/n) = H(U_n)F(1/n) = H(g_{1/n}) + H(U_{n-1}) = H(g_{1/n}) + \log(n-1), \quad n \geq 2,$$

so that, by Equation (32),

$$H(b_{1/n}) = \log n - \frac{\log(n-1)}{F(1/n)} = -\frac{1}{n} \log \frac{1}{n} - \left(1 - \frac{1}{n}\right) \log \left(1 - \frac{1}{n}\right), \quad n \geq 2.$$

Due to the recursion Formula (35) and Equation (32), it suffices to show that the above formula extends to $H(b_a) = -a \log a - (1-a) \log(1-a)$, $a \in (0, 1)$, to finish the proof. By Equations (26) and (30), we get

$$H(U_{k,n})F(1/n) = H(g_{1/n}) + H(U_{k-1,n-1}),$$

so that

$$H(U_{k,n}) = \sum_{i=0}^{k-1} \frac{H(g_{1/(n-i)})}{\prod_{j=n-i}^n F(1/j)} = \frac{1}{n} \sum_{i=0}^{k-1} (n-i) H(b_{1/(n-i)})$$

by iteration and by Equation (32), respectively. On the other hand, by Equations (25) and (30),

$$H(U_{k,n})F(1 - k/n) = H(g_{1-k/n}) + \log(k),$$

so that

$$\begin{aligned} & H(g_{1-k/n}) + \log(k) \\ &= \frac{1}{k} \sum_{i=0}^{k-1} (n-i) \left(-\frac{1}{n-i} \log \frac{1}{n-i} - \left(1 - \frac{1}{n-i}\right) \log \left(1 - \frac{1}{n-i}\right) \right) \\ &= \frac{1}{k} \sum_{i=0}^{k-1} ((n-i) \log(n-i) - (n-i-1) \log(n-i-1)) \\ &= \frac{1}{k} (n \log n - (n-k) \log(n-k)) \end{aligned}$$

and by Equation (32)

$$H(b_{k/n}) = -\frac{k}{n} \log \frac{k}{n} - \left(1 - \frac{k}{n}\right) \log \left(1 - \frac{k}{n}\right), \quad n \geq 2, k < n.$$

Cond. 2a finalizes the proof.

2.4. Proof of Theorem 2

Assume that H is the Rényi entropy. Let

$$S(p) = e^{(1-\alpha)H(p)},$$

i.e., $S(p) = \|p\|_\alpha^\alpha$. Then,

$$S(\partial_1 p) = p_1^\alpha \sum_i p_i^\alpha + \sum_i p_i^\alpha - p_1^\alpha$$

so that $f(r) = r^\alpha$. Reversely, assume that the second condition holds. Since $H \geq 0$ and S is strictly monotone in H we can choose the sign of γ such that S is increasing in H and $S \geq 1$. Furthermore, S is continuous. If $S(p) = 1$, then $1 = S(\partial_1 p) = S(\partial_2 p) = \dots = S(\partial_1^\infty p)$. Hence, $S(b_a) > 1$ for $a \in (0, 1)$ as otherwise, by Equation (21), we would get a contradiction to Cond. 2e. It follows that $f(a) = (S(\partial_1 b_a) - S(b_a)) / (S(b_a) - 1)$ is continuous on $(0, 1)$. Equation (10) implies $f(0)(S(\{0\} \cup p) - 1) = 0$ for all $p \in P$, so that $f(0) = 0$. Cond. 2f and Equation (20) yield $H(\{1\}) = 0$ and thus $S(\{1\}) = 1$. Cond. 2f in Theorem 1 reads

$$S(p \times q) = S(p)S(q). \quad (36)$$

By Cond. 2a, Cond. 2b holds for all $p \in P$. It is easy to check by induction that Equation (10) implies

$$S(\partial_1^n p) = S(p) \prod_{k=0}^{n-1} (1 + f(p_1^{2^k})) - \sum_{\ell=0}^{n-1} f(p_1^{2^\ell}) \prod_{k=\ell+1}^{n-1} (1 + f(p_1^{2^k})). \quad (37)$$

Henceforth, we assume that $a \in (0, 1)$. Cond. 2e and Equation (22) yield

$$\infty > S(g_a)S(g_{1-a}) \leftarrow S(g_{1-a}) \prod_{k=0}^{n-1} (1 + f(a^{2^k})) - \sum_{\ell=0}^{n-1} f(a^{2^\ell}) \prod_{k=\ell+1}^{n-1} (1 + f(a^{2^k})) \quad (38)$$

as $n \rightarrow \infty$. Since $F(a) = \prod_{k=0}^{\infty} (1 + f(a^{2^k}))$ is assumed to be finite, the function G given by

$$G(a) = \sum_{\ell=0}^{\infty} f(a^{2^\ell}) F(a^{2^{\ell+1}})$$

is finite by Equation (38). Hence, by Equations (19), (36) and (37),

$$S(p)F(p_1) - G(p_1) = S(g_{p_1})S\left(\frac{p_1}{1-p_1}\right), \quad p_1 < 1. \quad (39)$$

Equation (39) implies that

$$S(g_a)F(1-a) - G(1-a) = S(g_{1-a})S(g_a),$$

so that $G(a) = (F(a) - S(g_a))S(g_{1-a})$ and Equation (39) reads

$$S(p)F(p_1) - (F(p_1) - S(g_{p_1}))S(g_{1-p_1}) = S(g_{p_1})S\left(\frac{p_1}{1-p_1}\right), \quad p_1 < 1. \quad (40)$$

In particular,

$$S(b_a)F(a) - (F(a) - S(g_a))S(g_{1-a}) = S(g_a). \quad (41)$$

In analogy to Equation (33) we get from Equation (27) that

$$[S(b_a)(1 + f(1-a)) - f(1-a)]F(a) - (F(a) - S(g_a))S(g_{1-a}) = S(g_a)S(b_a),$$

so that, by Equation (41),

$$[S(b_a)(1 + f(1-a)) - f(1-a)]F(a) - S(b_a)F(a) + S(g_a) = S(g_a)S(b_a),$$

i.e., as $S(b_a) > 1$,

$$f(1-a)F(a) = S(g_a). \quad (42)$$

Since $S(g_a) > 1$ for all $a \in (0, 1)$ it follows that $f(a), F(a) \neq 0$ for all $a \in (0, 1)$. Now, Equation (40) reads

$$S(p) - (1 - f(1 - p_1))S(g_{1-p_1}) = f(1 - p_1)S\left(\frac{p_1}{1 - p_1}\right). \quad (43)$$

Since

$$1 < S(b_a) = (1 - f(a))S(g_a) + f(a), \quad (44)$$

the function f cannot take the value 1 on $(0, 1)$. Due to the continuity of f we may summarize that

$$f(0) = 0, \quad f(a) \in (0, 1) \text{ for } a \in (0, 1).$$

Lemma 2 implies that $S(U_n) = n^c$ for some $c > 0$. Then, Equation (43) gives

$$(1 - f(1 - 1/n))S(g_{1-1/n}) = n^c - f(1 - 1/n)(n - 1)^c.$$

Furthermore, taking the idea from Equation (25),

$$S(U_{k,n}) = (1 - f(k/n))S(g_{k/n}) + f(k/n)k^c$$

and, cf. Equation (26),

$$\begin{aligned} S(U_{k,n}) &= (1 - f(1 - 1/n))S(g_{1-1/n}) + f(1 - 1/n)S(U_{k-1,n-1}) \\ &= n^c + f(1 - 1/n)[-(n - 1)^c + S(U_{k-1,n-1})] \\ &= n^c + f(1 - 1/n)f(1 - 1/(n - 1))[-(n - 2)^c + S(U_{k-2,n-2})] \\ &= n^c + (1 - (n - k)^c) \prod_{j=0}^{k-1} f(1 - 1/(n - j)), \end{aligned}$$

so that, for $0 \leq k < n$, we have

$$(1 - f(k/n))S(g_{k/n}) = n^c + (1 - (n - k)^c) \prod_{j=0}^{k-1} f(1 - 1/(n - j)) - f(k/n)k^c. \quad (45)$$

The continuity of S and Equations (43) and (42) imply that H is uniquely determined if f and c are unique. So, it remains to determine them. In line with the idea from Equation (24), Equations (43) and (29) yield

$$\begin{aligned} S(G_{n,a}) &= (1 - f(a))S(g_a) + f(a)S(G_{n-1,a}) \\ &= (1 - f(a))S(g_a) \sum_{k=0}^{n-1} f^k(a) + f^n(a) \\ &= f^n(a) + (1 - f^n(a))S(g_a). \end{aligned} \quad (46)$$

By Equations (28) and (10) we have

$$S(G_{2n,a}) = (1 + f(a^n))S(G_{n,a}) - f(a^n).$$

Equation (46) yields

$$f^{2n}(a) + (1 - f^{2n}(a))S(g_a) = (1 + f(a^n))(f^n(a) + (1 - f^n(a))S(g_a)) - f(a^n),$$

i.e.,

$$(f(a^n) - f^n(a))(1 - f^n(a))(S(g_a) - 1) = 0.$$

Since $S(g_a) > 1$ and $f < 1$ for $a \in (0, 1)$ we get for $n = 2$ that

$$f^2(a) = f(a^2) \quad \forall a \in (0, 1),$$

or, $f(a^{2^k}) = f^{2^k}(a)$, $k \in \mathbb{Z}$. Since $(a^{2^{-k}})_{k \in \mathbb{N}}$ converges to 1 and f is assumed to be smooth, Lemma 3 and the proof of Lemma 4 yield the uniqueness of f , and $f(a) = a^\alpha$ for $a \in [0, 1]$ and some $\alpha > 0$. Letting $n = 2k$, Equations (44) and (45) yield

$$\begin{aligned} S(b_{1/2}) &= (2k)^c + (1 - k^c) \prod_{j=0}^{k-1} f(1 - 1/(n - j)) - f(1/2)k^c + f(1/2) \\ &= (2k)^c + (1 - k^c)2^{-\alpha} - 2^{-\alpha}k^c + 2^{-\alpha}. \end{aligned}$$

Hence,

$$2^\alpha S(b_{1/2}) = (2^{\alpha+c} - 2)k^c + 2.$$

Since the left hand side is independent of k , the equation $c = 1 - \alpha$ must hold.

2.5. Proof of Theorem 3

Assume H is the min-entropy and let f be the identity. Then, Equations (11) and (12) obviously hold. Reversely, let the second condition of the theorem hold. Cond. (2f) and Equation (20) imply $H(\{1\}) = 0$ and thus $S(\{1\}) = 1$. Let

$$\hat{\partial}_i^n p = \frac{1}{1 - p_i^{2^n}} (\partial_i^n p)_{\hat{i}},$$

$\hat{\partial}_i^0 p = p_i / (1 - p_i)$ and $S = e^{-H}$. Equation (12) yields

$$S(\hat{\partial}_1^0 p) = S(\hat{\partial}_1^n p) \prod_{k=0}^{n-1} (1 + f(p_1^{2^k})) = S(g_{2^n, p_1}) S(\hat{\partial}_1^0 p) \prod_{k=0}^{n-1} (1 + f(p_1^{2^k})), \quad p_1 < 1, \quad (47)$$

by iteration. By Equation (12) follows $f(0) = 0$ and f has to be monotonously increasing. Assume $f(a) = 0$ for some $a \in (0, 1]$. Then, $f(p_1) = 0$ for all $0 \leq p_1 \leq a$ and, by Equation (11),

$$S(p) = S(\hat{\partial}_1 p) = S(\hat{\partial}_1^2 p) = \dots = S(\hat{\partial}_1^\infty p) = S(g_{p_1}) S(\hat{\partial}_1^0 p), \quad p_1 \leq a, p_1 < 1.$$

Letting $p = g_{1-a}$ we get $S(g_a) = 1$, which contradicts Cond. 2e. Now, Equation (47) yields

$$S(g_{2^n, a}) \prod_{k=0}^{n-1} (1 + f(a^{2^k})) = 1,$$

so that

$$S(g_a) = 1 / \prod_{k=0}^{\infty} (1 + f(a^{2^k})) \in (0, 1), \quad a \in (0, 1). \quad (48)$$

Henceforth, we assume that $a \in (0, 1)$. Since $\hat{\partial}_1 b_a = b_{1/(1+a)}$, we have, by Equation (12)

$$(1 + f(a)) S(b_{1/(1+a)}) = 1,$$

and

$$S(b_a) = \begin{cases} (1 + f((1-a)/a))^{-1}, & a \geq 1/2 \\ (1 + f(a/(1-a)))^{-1}, & a < 1/2 \end{cases}. \quad (49)$$

Equation (11) yields

$$\Phi(p_1) \vee S(\partial_1^\infty p) = S(p),$$

where the limit $\Phi(a) := \bigvee_{k=0}^\infty f(a^{2^k})$ exists since f is monotonously increasing and thus implies $\Phi = f$. Hence, by Equation (19),

$$S(p) = S(\partial_1^\infty p) \vee f(p_1) = S(g_{p_1})S(\hat{\partial}_1^0 p) \vee f(p_1), \quad (50)$$

in particular,

$$S(g_{1-a}) = S(g_a)S(g_{1-a}) \vee f(a). \quad (51)$$

Since $S(g_a) < 1$, it follows from Equation (51) that

$$S(g_a) = f(1-a). \quad (52)$$

Thus, the recursion formula (50) determines H uniquely as soon as f is unique. Equations (52) and (48) yield

$$f(1-a) \prod_{k=0}^\infty (1 + f(a^{2^k})) = 1,$$

so that in analogy to Equation (14)

$$\begin{aligned} (1 + f(a))f(1-a) &= (1 + f(a))f(1-a)f(1-a^2) \prod_{k=1}^\infty (1 + f(a^{2^k})) \\ &= f(1-a^2). \end{aligned} \quad (53)$$

Now, Equations (50), (23) and (52) yield

$$S(G_{n,a}) = S(\partial_1^\infty G_{n,a}) \vee f(a^n) = S(g_a) \vee f(a^n) = f(1-a) \vee f(a^n).$$

On the other hand, by Equation (29) and iterating Equation (50) follows

$$\begin{aligned} S(G_{n,a}) &= f(1-a) \vee S(g_{1-a})S(G_{n-1,a}) \\ &= f(1-a) \vee S^n(g_{1-a}) \\ &= f(1-a) \vee f^n(a). \end{aligned}$$

Hence,

$$f(a^n) = f^n(a), \quad \text{if } 1-a < a^n.$$

In particular,

$$f(a^2) = f^2(a) \quad \text{for } a > (\sqrt{5}-1)/2,$$

and $\lim_{a \rightarrow 1} f(a)$ is necessarily 1. We have

$$S(b_a) = S(\partial_2 b_a) \vee f(1-a) = (S(g_a)S(b_a) \vee f(a)) \vee f(1-a) = f(a) \vee f(1-a). \quad (54)$$

Equations (49) and (54) and the monotonicity of f entail

$$f(a)(1 + f(-1 + 1/a)) = 1, \quad a \geq 1/2.$$

In particular $f(1/2) = 1/2$. Equation (53) delivers $f(3/4) = 3/4 > (\sqrt{5} - 1)/2$. Hence, for $a = 3/4$, we have $f(a^{2^{-k}}) = a^{2^{-k}}$ and $a^{2^{-k}} \rightarrow 1$ as $k \rightarrow \infty$. Now, Lemma 3 yields the uniqueness of f .

3. Discussion & Conclusions

The characterization given here is not only novel with respect to splitting the chain rule into its natural components, it is also novel in the sense that only a rather weak relationship is required for the chain rule, i.e. the proportionality factor itself is not given, but arises naturally. Furthermore, the recursion formula has the advantage that it allows an interpretation as an intrinsic property, since only a single distribution is involved and since recursion formulae in general may allow additional practical interpretations such as self-similarity or scale invariance.

In comparison to Faddeev's approach, our proofs emphasize the role of the geometric distribution g_a , which arises as the infinite iteration of the Bernoulli distribution b_a , $a \in (0, 1)$, i.e. $\partial_1^\infty b_a = g_a$. The uniqueness proof for the Shannon entropy uses almost only the resulting proportionality relation of the geometric distribution, simplifying Faddeev's approach. Further, while Equation (2) is a rule, proportionality can be considered a law. This yields a deeper understanding of the Shannon entropy and may justify a broader application outside telecommunication.

The proofs rely strongly on distributions that generalize the uniform distribution and the geometric distribution, respectively. Possibly, these distributions play or will play a role also in other areas.

Although the characterization and the proofs of the Shannon and the Rényi entropy follow the same scheme, they show remarkable differences. For instance, the generalization $G_{n,a}$ of the geometric distribution, see Equation (17), does not play a role in case of the Shannon entropy. Also, the functional characterization in Equation (9) of the Shannon entropy is not the limit of the characterization in Equation (10) of the Rényi entropy. However, Equation (9) might be understood as the derivative of Equation (10) in the following sense, cf. [18]. Assume that f in Equation (10) is differentiable. Then,

$$\frac{d}{d\alpha} S(\partial_1 p) = \frac{d}{d\alpha} ((1 + f(p_1))S(p) - f(p_1)).$$

For $\alpha = 1$ we define $S(p) = \lim_{\alpha \rightarrow 1} \|p\|_\alpha^\alpha$. Then, we get

$$\frac{d}{d\alpha} S(\partial_1 p) = (1 + f(p_1)) \frac{d}{d\alpha} S(p).$$

Another interpretation of our results is that S is not a natural transformation of the Rényi entropy, in contrast to

$$T(p) = S(p) - S(\{1\}).$$

This definition bridges to the generalized information function [5, Section 6.2] and to the definite functions, cf. [19], Chapter 3, Corollary 3.3 and proof of Theorem 2.2, and Chapter 6, Example 5.16. Indeed, Equation (10) can be rewritten in full analogy to Equation (9), namely as

$$T(\partial_1 p) = (1 + f(p_1))T(p),$$

where $T(\partial_1 p) = S(\partial_1 p) - S(\{1\})$. Note that also the interpretation $T(\partial_1 p) = S(\partial_1 p) - S(\partial_1 \{1\})$ is possible.

Several technical assumptions had to be included in the characterization of the Rényi entropy, such as the condition that f shall be real analytic. Although we consider this fact as minor, we hope that these assumptions can be dropped in the future.

Author Contributions: Conceptualization and methodology, M.S.; investigation, validation and resources, C.D.; writing—original draft, M.S.; writing—review and editing, C.D.; all authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors thank Christopher Dörr for many hints and comments. They also thank Thomas Reichelt for valuable hints.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Khinchin, A. The concept of entropy in the theory of probability. *Uspekhi Matem. Nauk* **1953**, *8*, 3–20.
2. Faddeev, D. On the concept of entropy of a finite probabilistic scheme. *Uspekhi Matem. Nauk* **1956**, *11*, 227–231.
3. Rényi, A. On measures of entropy and information. *Fourth Berkeley Symposium on Mathematical Statistics and Probability: Contributions to the Theory of Statistics*. University of California Press, 1961, pp. 547–562.
4. Leinster, T. *Entropy and Diversity — The Axiomatic Approach.*; Cambridge, 2021.
5. Aczél, J.; Daróczy, Z. *On Measures of Information and Their Characterizations*; Academic Press, 1975.
6. Ebanks, B.; Sahoo, P.K.; Sander, W. *Characterization of information measures*; World Scientific, 1998.
7. Carcassi, G.; Aidala, C.; Barbour, J. Variability as a better characterization of Shannon entropy. *Europ. J. Phys.* **2021**, *42*, 045102.
8. Baez, J.; Fritz, T.; Leinster, T. A characterization of entropy in terms of information loss. *Entropy* **2011**, *13*, 1945–1957.
9. Onicescu, O. Energie informationnelle. *Sci. Paris* **1966**, *263*, 841–842.
10. Pardo, L. Order- α weighted information energy. *Information Sci.* **1986**, *40*, 155–164.
11. Schlather, M. An algebraic generalization of the entropy and its application to statistics **2024**. arxiv: 2404.05854.
12. Krantz, S. *Function Theory of Several Complex Variables*, 2nd ed.; AMS: Pacific Grove, Calif, 1992.
13. Erdős, P. On the distribution function of additive functions. *Ann. Math.* **1946**, *47*, 1–20.
14. Krantz, S.; Parks, H. *A Primer of Real Analytic Functions*; Birkhäuser Boston, 2002.
15. Gradshteyn, I.; Ryzhik, I. *Table of Integrals, Series, and Products*, 6th ed.; Academic Press: London, 2000. MR1773820.
16. Conrad, K. Probability Distributions and Maximum Entropy. Technical report, University of Connecticut. URL <https://kconrad.math.uconn.edu/blurbs/analysis/entropypost.pdf>. Accessed 14 Nov. 2024.
17. Jakimiuk, J.; Murawski, D.; Nayar, P.; Słobodianiuk, S. Log-concavity and discrete degrees of freedom. *Discr. Math.* **2024**, *347*, 114020.
18. Życzkowski, K. Rényi extrapolation of Shannon entropy. *Open Sys. & Inform. Dyn.* **2003**, *10*, 297–310.
19. Berg, C.; Christensen, J.P.R.; Ressel, P. *Harmonic Analysis on Semigroups. Theory of Positive Definite and Related Functions*; Springer: New York, 1984.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.