Article

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

Martin Schlather * and Carmen Ditscheid

Institute of Mathematics, University of Mannheim, 68131 Mannheim, Germany
* Correspondence: martin.schlather@uni-mannheim.de; Tel.: +49-621-181-2552

Abstract: All characterizations of the Shannon 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. Analogous characterizations are given for the Rényi entropy and its limits, the min-entropy and the Hartley entropy.

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

1. 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 $p = \{p_1, \ldots, p_n\}$ and $q = \{q_1, \ldots, q_m\}$ be discrete probability distributions, and

$$p_{k,q} = \{p_1, \ldots, p_{k-1}, p_k q_1, \ldots, p_k q_m, p_{k+1}, \ldots, p_n\}. \tag{1}$$

The chain rule states that

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

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 Faddeev [2], where q in Equation (1) is restricted to a Bernoulli distribution. Carcassi et al. [7] considered the full version of Equation (1), where q is arbitrary, as preferable from a didactical point of view, since this allows an immediate interpretation in various areas of application. Often, the completely iterated version,

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

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

Baez et al. [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 might 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. Ebanks et al. ([6], Definition 3.1.1) introduced 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 ([5], Section 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. Pardo [10] gives a full characterization by means of a chain rule that is similar to Faddeev's variant,

$$S(p_{k,q}) = S(p) + p_k^\alpha(S(q) - \text{const}), \tag{4}$$

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

The limits of the Rényi entropy as its parameter goes to 0 and ∞ are the Hartley entropy [11] and the min-entropy [12], respectively. Mathematical characterizations of these two entropies are rare, e.g., conical combinations of the Hartley entropy and the Shannon entropy are characterized in [13] as the only reasonable simultaneous solution to the chain rule and the subadditivity property. Ref. [13] also characterized the Hartley entropy as being insensitive. An advanced investigation of the min-entropy exists in form of the Fourier min-entropy [14].

Schlather [15] deals with a rather general approach to the entropy, which is axiomatically 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 obtain

$$p_{k,p} = \{p_1, \ldots, p_{k-1}, p_k p_1, \ldots, p_k p_n, p_{k+1}, \ldots, p_n\}, \tag{5}$$

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

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

Theorem 1 in Section 2 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. Related characterizations also exist for the min-entropy and the Hartley entropy, cf. Theorems 3 and 4. All technical parts are postponed to Section 3. There, the geometric distribution plays an overwhelming role. Section 4 illuminates practical and theoretical aspects of this paper. Section 5 finalizes with a digest of our approach.

2. Results

We denote by $\{.\}$ an ordered set of non-negative numbers and let

$$
\begin{aligned}
a\{p_1, p_2, \ldots\} &= \{ap_1, ap_2, \ldots\}, \qquad a \geq 0, \\
\{p_1, p_2, \ldots\} \cup \{q_1, q_2, \ldots\} &= \{p_1, p_2, \ldots, q_1, q_2, \ldots\}.
\end{aligned} \tag{6}
$$

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 (6) is sufficient.

Let us 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 $\{p_1, \ldots, p_n\} \in P_n$. The uniform distribution is denoted by $U_n = \{1/n, \ldots, 1/n\}$. Let $P_0 = \bigcup_n P_n$ be the set of all discrete probability distributions with finite support.

Reordering the elements in Equation (5) will turn out to be useful in the proofs. We define

$$
\begin{aligned}
p_{\hat{k}} &= \{p_i : i \neq k\}, \\
\partial_k p &= p_k p \cup p_{\hat{k}},
\end{aligned} \tag{7}
$$

where $p = \{p_1, p_2, \ldots\} \in P$. 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 $H : P \to [0, \infty]$ 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) *$H(\{p_1, \ldots, p_n\}) = H(\{p_{\pi(1)}, \ldots, p_{\pi(n)}\})$ for $\{p_1, \ldots, p_n\} \in P_0$ and any permutation π;*

   (c) *$H(\{p_1, \ldots, p_n\}) = H(\{0, p_1, \ldots, p_n\})$ for all $\{p_1, \ldots, p_n\} \in P_0$;*

   (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 nondegenerate geometric distribution;*

   (f) *$H(p \times q) = H(p) + H(q)$ for all $p, q \in P$;*

   (g) *a function $f : [0, 1] \to [0, \infty)$ exists such that*

$$
H(\partial_1 p) = (1 + f(p_1)) H(p), \qquad p \in P. \tag{8}
$$

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

Remark 1. *Note that, here and in the subsequent theorems, the function f is not unique, since it is undetermined at 1. A convenient choice is to take $f(1) = \lim_{a \to 1} f(a)$.*

Remark 2. *Our approach has a simple, but nice implication. Let $p, q \in P_0$. Then,*

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

where KL denotes the Kullback–Leibler divergence.

Theorem 2. *Let $H : P \to [0, \infty]$ 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 Conditions 2(a)–2(f) of Theorem 1. A value $\gamma \in \mathbb{R} \setminus \{0\}$ and a function $f : [0,1] \to [0, \infty)$ exist such that for $S = e^{\gamma H}$ we have*

$$S(\partial_1 p) = (1 + f(p_1)) S(p) - f(p_1), \qquad p \in P. \tag{9}$$

We denote the max-operator by $\vee$ and the min-operator by $\wedge$.

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

1. *H is the min-entropy,*

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

up to a positive multiplicative constant.
2. *H satisfies Conditions 2(a)–2(f) of Theorem 1. A function $f : [0,1] \to [0, \infty)$ exists, such that for $S = e^{-H}$, we have*

$$f(p_1) \vee S(\partial_1 p) = S(p), \qquad p \in P. \tag{10}$$

As the Hartley entropy H is discontinuous on P and $H(g_a) = \infty$, Conditions 2(a) and 2(e) are replaced by Equations (11) and (12), respectively.

Theorem 4. *Let $H : P \to [0, \infty]$ be a function. Then, the following two assertions are equivalent:*

1. *H is the Hartley entropy,*

$$H(p) = \log \sum_{p_i \in p} \mathbf{1}_{(0,1]}(p_i),$$

where $\mathbf{1}_{(0,1]}(p_i) = 1$ if $p_i \in (0,1]$, and 0 otherwise.
2. *H satisfies Conditions 2(b)–2(d) and 2(f) of Theorem 1 and $H(U_2) \in (0, \infty)$. Further, for $S = e^H$ holds*

$$\lim_{n \to \infty} \frac{S(\{p_1, \ldots, p_n\} \cup q \sum_{i=n+1}^{\infty} p_i)}{S(\{p_1, \ldots, p_n\} / \sum_{i=1}^{n} p_i)} \in \{1, \infty\}, \qquad \{p_1, p_2, \ldots\}, q \in P. \tag{11}$$

Moreover, $S(p) < \infty, p \in P$ with $p_1 < 1$ implies $S(p_{\hat{1}}/(1-p_1)) < \infty$. A function $f : [0,1] \to [0, \infty)$ exists such that Equation (9) holds true and

$$S\left(\frac{(\partial_1 p)_{\hat{1}}}{1 - p_1^2}\right) = (1 + f(p_1)) S\left(\frac{p_{\hat{1}}}{1 - p_1}\right), \qquad p \in P, p_1 < 1. \tag{12}$$

3. Proofs

We use the following abbreviations:

$$F_n(a) = \prod_{k=0}^{n-1} \left(1 + f(a^{2^k})\right), \qquad a \in [0,1],$$

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

3.1. Functional Equations

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

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

$$f(a) + f(1-a) = 1 \tag{13}$$
$$(1 - f(a))F(a) = 1; \tag{14}$$

2.   *f is continuous and, for all $a \in (0,1)$, Equation (13) holds and $f(a^2) = f^2(a)$;*
3.   *f is the identity.*

Proof. Let f be the identity. Equation 0.266 in [16] shows that Equation (14) holds true. The other properties are obvious.

Now, assume that the first condition applies. Then, the second condition holds true, since

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

from which the equation $f(a^2) = f^2(a)$, $a \in (0,1)$, follows.

We now show that the second condition implies that f is the identity. For an endomorphism h, we denote by $h^{(n)}$ the n-fold composition $h \circ \ldots \circ h$. The assumptions imply that $f((1 - a^{2^k})^{2^r}) = (1 - f^{2^k}(a))^{2^r}$ for all $a \in (0,1)$ and $k, r \in \mathbb{Z}$. Note that the set of maps $\{a \mapsto (1 - a^{2^k})^{2^r} : k, r \in \mathbb{Z}\}$ is identical to the set of its inverses. Equation (13) yields $f(1/2) = 1/2$. Let $T(A) = \{(1 - a^{2^k})^{2^r} : a \in A, r, k \in \mathbb{Z}\} \cup A$ for $A \subset (0,1)$, and $D = \bigcup_n T^{(n)}(\{1/2\})$. 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 $D \cap (a,b) = \emptyset$. Then, necessarily, also $D \cap \bigcup_n T^{(n)}((a,b)) = \emptyset$. Since $1/2 \in D$, we may assume, without loss of generality, that $(a,b) \subset (1/2, 1)$. Consider the map h defined on the open intervals of $(0,1)$,

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

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

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

Hence, $h^{(n)}((a,b))$ finally contains $1/2$ or its length is strictly monotonously increasing. If the latter were true, the limit would exist as $|h| \leq 1/2$. Close to the limit, we would have $|h^{(n+1)}((a,b)) - h^{(n)}((a,b))| = |d^2 - c^2 - (d - c)| = |(d - c)(d + c - 1)| < \epsilon$ for some $\epsilon > 0$, i.e., both boundaries of $h^{(n)}((a,b))$ must converge to $1/2$, contradicting the fact that an interval with upper bound close to $1/2$ is mapped to an interval with upper bound close to $3/4$.   □

Lemma 2. *Let $\Phi : [0,1] \to [0,1]$ be an endomorphism and $c > 0$. The following two assertions are equivalent:*

1. Φ *is continuous on* $(0, 1)$ *with values in* $(0, 1)$, $\Phi(1/2) > 0$ *and* $\Phi(1/n) \to 0 = \Phi(0)$ *as* $n \to \infty$. *For all* $n \in \mathbb{N}$ *and* $0 \leq k < n$, *we have*

$$\Phi(a^n) \vee \Phi(1-a) = \Phi^n(a) \vee \Phi(1-a), \qquad a \in (0,1),$$

$$\frac{\Phi(1-k/n)}{(n-k)^c} \vee \Phi(k/n) = \prod_{m=k+1}^{n} \Phi(1-1/m) \vee \bigvee_{\ell=k+1}^{n} \Phi(1/\ell) \prod_{m=\ell+1}^{n} \Phi(1-1/m);$$

2. $\Phi(a) = a^c$.

Proof. We assume that the first condition holds true. To begin with, we show that Φ is strictly increasing. For $j \in \mathbb{N}$, we have

$$\frac{\Phi(1-k/n)}{j^c(n-k)^c} \vee \Phi(k/n) = \prod_{m=jk+1}^{jn} \Phi(1-1/m) \vee \bigvee_{\ell=jk+1}^{jn} \Phi(1/\ell) \prod_{m=\ell+1}^{jn} \Phi(1-1/m).$$

Letting $j \to \infty$, we obtain

$$\Phi(k/n) = \lim_{j \to \infty} \prod_{m=jk+1}^{jn} \Phi(1-1/m),$$

so that $\Phi(k/n) \leq \Phi((k+1)/n)$ for all $0 < k < n-1$. Hence, Φ is increasing on $(0,1) \cap \mathbb{Q}$, thus increasing on $(0,1)$. In particular, $\Phi \geq \Phi(1/2) > 0$ on $[1/2, 1)$. Let $n/2 \leq k < n-1$ and $j \in \mathbb{N}$ such that $j^c(n-k-1)^c\Phi(k/n) > 1$. Then,

$$\begin{aligned}
\Phi((k+1)/n) &= \Phi(1-(k+1)/n)j^{-c}(n-k-1)^{-c} \vee \Phi((k+1)/n) \\
&= \prod_{m=j(k+1)+1}^{jn} \Phi(1-1/m) \\
&> \prod_{m=jk+1}^{jn} \Phi(1-1/m) = \Phi(k/n),
\end{aligned}$$

so that Φ is strictly increasing on $[1/2, 1)$. Therefore, $\Phi(a^n) = \Phi^n(a)$ for all $a \in (0,1)$ and $n \in \mathbb{N}$ with $1 - a < a^n$. Let $m, k \in \mathbb{N}$. Let ℓ be a multiple of k such that $(a \wedge a^{m/k}) + a^{1/\ell} - 1 > 0$. Then,

$$\Phi(a^{m/k}) = \Phi\left(\left(a^{1/\ell}\right)^{m\ell/k}\right) = \left(\Phi\left(a^{1/\ell}\right)\right)^{m\ell/k} = \left(\left(\Phi\left(a^{1/\ell}\right)\right)^\ell\right)^{m/k} = (\Phi(a))^{m/k}.$$

The continuity of Φ implies that $\Phi(a^x) = \Phi(a)^x$ for $a \in (0,1)$ and $x > 0$, so that $\Phi(a) = a^\gamma$ for a $\gamma \in \mathbb{R}$. Since $\Phi(1/n) \to 0$ and $\Phi(1/2) > 0$, we have $\gamma > 0$. Finally, for $k = 0$, we obtain from the assumptions that $\Phi(1)/n^c = 1/n^\gamma$, $n \in \mathbb{N}$, so that $\Phi(1) = 1$ and $c = \gamma$. $\square$

Finally, we restate two results from the literature for the reader's convenience.

Lemma 3. *Let* $H : P \to [0, \infty]$ *be a function fulfilling Conditions* 2(d) *and* 2(f) *of Theorem* 1. *Then,* $H(U_n) = c \log n$ *for all* $n \in \mathbb{N}$ *and some* $c \geq 0$.

Proof. The inclusion $P_n \subset P_{n+1}$ and Condition 2(d) immediately imply monotonicity, i.e., $H(U_n) \leq H(U_{n+1})$ for all $n \in \mathbb{N}$. By Condition 2(f) and Implication 1 of Theorem V in [17], the claim is followed. $\square$

Lemma 4. *Let* H *be the Shannon entropy, the Rényi entropy, the min-entropy, or the Hartley 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 proofs.

3.2. Properties of the Operator ∂_k

We will use the operator ∂_k, defined in Equation (7), iteratively. For instance,

$$
\begin{aligned}
\partial_2 \partial_k p &= (p_2 p_k)(p_k p \cup p_{\hat{k}}) \cup (p_k p \cup p_{\hat{k}})_{\hat{2}}, \\
\partial_2 \partial_3 \{p_1, p_2, p_3\} &= \partial_2 \{p_3 p_1, p_3 p_2, p_3 p_3, p_1, p_2\} \\
&= \{p_3 p_2 p_3 p_1, \; p_3 p_2 p_3 p_2, \; p_3 p_2 p_3 p_3, \; p_3 p_2 p_1, \; p_3 p_2 p_2, \\
&\qquad\quad p_3 p_1, \qquad\qquad\qquad\quad p_3 p_3, \qquad\quad p_1, \qquad p_2\} \\
&= \{p_1 p_2 p_3^2, \; p_2^2 p_3^2, \; p_2 p_3^3, \; p_1 p_2 p_3, \; p_2^2 p_3, \; p_1 p_3, \; p_3^2, \; p_1, \; p_2\}.
\end{aligned}
$$

The proofs of the theorems in Section 2 rely on the properties of various probability distributions. We denote by b_a the Bernoulli distribution, i.e.,

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

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

$$
g_a = (1-a)\{1, a, a^2, \ldots\}, \qquad a \in [0,1).
$$

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

$$
\begin{aligned}
g_{n,a} &= \frac{1-a}{1-a^n}\{1, a, a^2, \ldots, a^{n-1}\}, \quad a \in [0,1), \\
G_{n,a} &= \{a^n\} \cup (1-a)\{1, a, a^2, \ldots, a^{n-1}\}, \quad a \in [0,1], \quad (15) \\
U_{k,n} &= \{1 - k/n, \underbrace{1/n, \ldots, 1/n}_{k \text{ times}}\}, \qquad 0 \le k < n. \quad (16)
\end{aligned}
$$

Let $G_{0,a} = U_{0,n} = \{1\}$ and

$$
\begin{aligned}
\hat{\partial}_k^n p &= \frac{1}{1 - p_k^{2^n}} (\partial_k^n p)_{\hat{k}}, \qquad p_k < 1, n \in \mathbb{N}, \\
\hat{\partial}_k^0 p &= p_{\hat{k}}/(1 - p_k), \qquad p_k < 1.
\end{aligned}
$$

We denote by $\doteq$ equality up to inserting zeros and reordering.

Lemma 5. *Let $p \in P$ and $\partial_k^\infty p := \lim_{n \to \infty} \partial_k^n p$. Then,*

$$
\begin{aligned}
\partial_k^\infty p &\doteq g_{p_k} \times \hat{\partial}_k^0 p, \qquad p_k < 1, & (17) \\
\{1\} &\doteq \{1\} \times \{1\}, & (18) \\
\hat{\partial}_k^n p &\doteq g_{2^n, p_k} \times \hat{\partial}_k^0 p, \qquad p_k < 1. & (19)
\end{aligned}
$$

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

$$
\begin{aligned}
\partial_1^\infty b_a &\doteq g_a, & (20) \\
\partial_1^\infty g_a &\doteq g_{1-a} \times g_a, & (21) \\
\partial_1^\infty G_{n,a} &\doteq g_{a^n} \times g_{n,a} \doteq g_a \doteq G_{\infty,a}, & (22) \\
\partial_2^\infty G_{n,a} &\doteq g_{1-a} \times G_{n-1,a}, & (23) \\
\partial_1^\infty U_{k,n} &\doteq g_{1-k/n} \times U_k, \quad 1 \le k < n, & (24) \\
\partial_2^\infty U_{k,n} &\doteq g_{1/n} \times U_{k-1,n-1}, \quad 1 \le k < n, & (25) \\
\partial_3^\infty \partial_2 b_a &\doteq g_a \times b_a. & (26)
\end{aligned}
$$

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

$$\partial_1 G_{n,a} \doteq G_{2n,a}, \tag{27}$$

$$(G_{n,a})_{\hat{2}} = a G_{n-1,a}. \tag{28}$$

Proof. Without loss of generality, let $k = 1$ in Equation (17). We have

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

since this is true for $n = 1$ and

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

by induction. Hence,

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

Now, $p_1^\ell p_{\hat{1}} = (1-p_1)p_1^\ell \cdot \frac{p_{\hat{1}}}{1-p_1}$ for $p_1 < 1$, implying Equation (17). Further, we have $\partial_2 b_a = (1-a)b_a \cup \{a\}$, so that Equation (26) holds by Equation (17). The other equalities of the lemma are immediate. $\square$

Lemma 6. *Let $S : P \to [0,\infty]$ and $f : [0,1] \to [0,\infty)$, such that Equation (9) holds true. Then,*

$$S(\partial_1^n p) = (S(p) - 1)F_n(p_1) + 1, \qquad p \in P, n \in \mathbb{N}, \tag{29}$$

$$S(G_{2n,a}) = (1 + f(a^n))S(G_{n,a}) - f(a^n), \qquad n \in \mathbb{N}. \tag{30}$$

Proof. It is easy to check by induction that Equation (9) implies

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

Now,

$$1 + \sum_{i=1}^n c_i \prod_{j=i+1}^n (1 + c_j) = \prod_{j=1}^n (1 + c_j), \qquad c_j \in \mathbb{R}, n \in \mathbb{N},$$

by induction, so that the first assertion of the lemma holds. Equation (27) yields the second assertion. $\square$

3.3. 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 f is chosen to be the identity. Now, assume that the second condition applies. By Condition 2(a), Condition 2(b) holds for all $p \in P$. Condition (f) and Equation (18) yield $H(\{1\}) = 0$. It is easy to check that $H(\partial_1^n p) = H(p)F_n(p_1)$, so that

$$H(\partial_1^\infty p) = H(\lim_{n \to \infty} \partial_1^n p) = \lim_{n \to \infty} H(\partial_1^n p) = H(p)F(p_1). \tag{31}$$

The function F is finite due to Conditions 2(e) and 2(f) and Equations (21) and (31), i.e.,

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

It follows that

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

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

$$(F(1-a) - 1)(F(a) - 1) = 1, \qquad a \in (0,1). \tag{32}$$

Now,

$$H(b_a)F(a) = H(g_a) > 0, \tag{33}$$

by Equations (20) and (31) and Condition 2(e). Equations (26), (8), and (31) yield

$$H(b_a)(1 + f(1-a))F(a) = H(g_a) + H(b_a);$$

thus, $f(1-a)F(a) = 1$ for $a \in (0,1)$. Equation (32) now writes $f(a) + f(1-a) = 1$. Plugging in $p = b_a$ into Equation (8) shows that f is continuous on $(0,1)$. Now, Lemma 1 yields $f(a) = a$ and $F(a) = 1/(1-a)$ for $a \in (0,1)$. Equation (8) implies that $f(0) = 0$, so that f is the identity on $[0,1)$. Lemma 3 delivers $H(U_n) = c \log n$ for a $c \geq 0$. Condition 2(f), Equations (17) and (31) yield

$$H(p)F(p_1) = H(g_{p_1}) + H(\hat{\partial}_1^0 p), \quad p_1 < 1. \tag{34}$$

This implies that

$$c \log n \cdot F(1/n) = H(g_{1/n}) + c \log(n-1)$$

for $n \geq 2$, so that $c > 0$ and

$$c^{-1}H(g_{1/n}) = \frac{1}{1 - 1/n} \log n - \log(n-1) = -\frac{1/n}{1 - 1/n} \log \frac{1}{n} - \log\left(1 - \frac{1}{n}\right).$$

Henceforth, we may assume that $c = 1$. Due to the recursion Formula (34), it suffices to show that the above formula extends to $H(g_a) = -a(1-a)^{-1} \log a - \log(1-a)$, $a \in (0,1)$, to finish the proof. By Equations (25) and (31), we obtain

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

i.e.,

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

by iteration. On the other hand, by Equations (24) and (31),

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

so that

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

Condition 2(a), i.e., the continuity of H, finalizes the proof.

3.4. Proof of Theorem 2

Assume that H is the Rényi entropy. Let $S(p) = \exp((1-\alpha)H(p))$, 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 $H(\{1\}) = 0$ by Condition 2(f) and Equation (18), we have $S(\{1\}) = 1$. Furthermore, S is continuous. If $S(p) = 1$, then $1 = S(\partial_1 p) = S(\partial_2 p) = \ldots = S(\partial_1^\infty p)$ by Equation (9). Hence, $S(b_a) \neq 1$ for $a \in (0,1)$, as otherwise we would obtain a contradiction to Condition 2(e) by Equation (20). It follows that $f(a) = (S(\partial_1 b_a) - S(b_a))/(S(b_a) - 1)$ is continuous on $(0,1)$. Equation (9) implies that $f(0)(S(\{0\} \cup p) - 1) = 0$ for all $p \in P$, so that $f(0) = 0$. Condition 2(f) reads

$$
S(p \times q) = S(p)S(q). \tag{35}
$$

By Condition 2(a), Condition 2(b) holds for all $p \in P$. Henceforth, we assume that $a \in (0,1)$. Condition 2(e), Equations (29) and (21) yield

$$
\infty > S(g_a)S(g_{1-a}) \leftarrow (S(g_{1-a}) - 1)F_n(a) + 1,
$$

as $n \to \infty$. Hence, $F(a) = \lim_{n\to\infty} F(a)$ is finite. Now, by Equations (17), (35) and (29),

$$
(S(p) - 1)F(p_1) + 1 = S(g_{p_1})S(\hat\partial_1^0 p), \qquad p_1 < 1. \tag{36}
$$

In particular,

$$
(S(b_a) - 1)F(a) + 1 = S(g_a). \tag{37}
$$

We obtain by means of Equation (26) that

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

so that, by Equation (37),

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

i.e., as $S(b_a) \neq 1$,

$$
f(1-a)F(a) = S(g_a). \tag{38}
$$

Since $S(g_a) \neq 0$, it follows that $f(a), F(a) \neq 0$. Equations (36) and (38) yield for $p = g_{1-a}$ that $1 - F(a)^{-1} = (1 - f(1-a))S(g_{1-a})$, so that Equation (36) now reads

$$
S(p) - (1 - f(1-p_1))S(g_{1-p_1}) = f(1-p_1)S(\hat\partial_1^0 p), \tag{39}
$$

again by Equation (38). In line with the idea from Equation (23), Equations (39) and (28) yield

$$
\begin{aligned}
S(G_{n,a}) &= (1 - f(a))S(g_a) + f(a)S(G_{n-1,a}) \\
&= (1 - f(a))S(g_a)\sum_{k=0}^{n-1} f^k(a) + f^n(a) \\
&= f^n(a) + (1 - f^n(a))S(g_a).
\end{aligned}
\tag{40}
$$

Plugging in Equation (40) into Equation (30) 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.,

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

Assume that $f(a) = 1$. Then, by Equation (40), $1 = S(G_{n,a}) \to S(g_a)$ as $n \to \infty$. This contradicts $S(g_a) \neq 1$. Hence, we obtain for $n = 2$ that

$$
f^2(a) = f(a^2), \qquad a \in (0,1). \tag{41}
$$

Lemma 3 implies that $H(U_n) = \tilde{c}\log n$ for a $\tilde{c} \geq 0$; hence, $S(U_n) = n^c$ for $c = \gamma\tilde{c} \in \mathbb{R}$. By $S(U_2) = S(b_{1/2}) \neq 1$, it follows that $c \neq 0$. Equation (39) gives

$$
(1 - f(1 - 1/n))S(g_{1-1/n}) = n^c - f(1 - 1/n)(n-1)^c. \tag{42}
$$

Furthermore, taking the ideas from Equations (24) and (25), we obtain, by Equation (39),

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

and

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

by Equation (42) and iteration. Thus, for $0 < k < n$, we have

$$
(1 - f(k/n))S(g_{k/n}) = n^c + (1 - (n-k)^c)\frac{M_n}{M_{n-k}} - f(k/n)k^c, \tag{43}
$$

where

$$
M_1 = 1, \qquad M_n = \prod_{j=2}^{n} f(1 - 1/j), \quad n \geq 2.
$$

Note that $M_n > 0$ for all $n \in \mathbb{N}$ as a consequence of Equation (38). We obtain for $n = 2k$ in Equation (43) that

$$
(1 - f(1/2))S(g_{1/2}) = (2k)^c + (1 - k^c)\frac{M_{2k}}{M_k} - f(1/2)k^c. \tag{44}
$$

For $k = 1$, this is

$$
(1 - f(1/2))S(g_{1/2}) = 2^c - f(1/2). \tag{45}
$$

Plugging Equation (45) into Equation (44) we obtain $M_{2k}/M_k = 2^c - f(1/2)$ for $k > 1$. An immediate consequence of this constant ratio is that $M_{2k}/M_{2\ell} = M_k/M_\ell$ for $\ell, k > 1$.

Equation (43), the symmetry of the Bernoulli distribution, and the fact that, by Equation (39), $S(b_a) = (1 - f(a))S(g_a) + f(a)$, yield

$$(1 - (n - k)^c)\frac{M_n}{M_{n-k}} - f(k/n)(k^c - 1) = (1 - k^c)\frac{M_n}{M_k} - f(1 - k/n)((n - k)^c - 1),$$

i.e., for all $\ell \geq 0$ and $1 < k < n - 1$,

$$\begin{aligned}
f(k/n) &= \frac{M_{2^\ell n}}{M_{2^\ell k}} - N(2^\ell)\left[\frac{M_{2^\ell n}}{M_{2^\ell (n-k)}} - f(1 - k/n)\right] \\
&= \frac{M_n}{M_k} - N(2^\ell)\left[\frac{M_n}{M_{n-k}} - f(1 - k/n)\right]
\end{aligned} \tag{46}$$

with

$$N(j) = \frac{1 - j^c(n - k)^c}{1 - (jk)^c}.$$

In order for the right-hand side of Equation (46) to be independent of ℓ, we necessarily have $f(1 - k/n) = M_n/M_{n-k}$. Now, Equation (43) ensures that for all $\ell \geq 1$ and $1 < k < n - 1$, we have

$$\begin{aligned}
(1 - f(k/n))S(g_{k/n}) &= \ell^c n^c + (1 - \ell^c(n - k)^c)f(1 - k/n) - f(k/n)\ell^c k^c \\
&= f(1 - k/n) + k^c l^c((n/k)^c - (n/k - 1)^c f(1 - k/n) - f(k/n)).
\end{aligned}$$

In order for the right-hand side to be independent of ℓ, we necessarily have

$$(k/n)^c f(k/n) + (1 - k/n)^c f(1 - k/n) = 1.$$

Since f is continuous, we even have

$$a^c f(a) + (1 - a)^c f(1 - a) = 1, \qquad a \in (0, 1).$$

Let $\tilde{f}(a) = a^c f(a)$; then $\tilde{f}(a^2) = \tilde{f}^2(a)$ by Equation (41) and $\tilde{f}(a) + \tilde{f}(1 - a) = 1$. Lemma 1 yields that $\tilde{f}$ is the identity on $(0, 1)$; hence, $f(a) = a^\alpha$ with $\alpha = 1 - c$.

3.5. Proof of Theorem 3

Assume that H is the min-entropy and let f be the identity. Then, Equation (10) obviously holds true. Reversely, assume that the second condition of the theorem applies. Condition 2(f) and Equation (18) imply that $H(\{1\}) = 0$ and $S(\{1\}) = 1$. Equations (10) and (17) yield

$$S(p) = S(\partial_1^\infty p) \vee \Phi(p_1) = S(g_{p_1})S(\hat{\partial}_1^0 p) \vee \Phi(p_1), \qquad p_1 < 1, \tag{47}$$

where $\Phi(a) = \sup_{k=0}^\infty f(a^{2^k})$. Equations (10), (17), and (21) deliver $S(g_a) = S(g_{1-a})S(g_a) \vee \Phi(1 - a)$, so that $\Phi(a)$ is finite for $a \in (0, 1)$. Since $S(g_{1-a}) \in (0, 1)$ by Condition 2(e), it follows that

$$S(g_a) = \Phi(1 - a) \in (0, 1), \qquad a \in (0, 1). \tag{48}$$

Thus, the recursion Formula (47) determines H uniquely as soon as Φ is determined. Since S is continuous, the function Φ is continuous on $(0, 1)$. Lemma 3 yields

$$n^{-c} = S(U_n) = \Phi(1 - 1/n)(n - 1)^{-c} \vee \Phi(1/n), \qquad n \in \mathbb{N},$$

for some $c \geq 0$. In particular, $\Phi(1/n) \leq n^{-c}$ and $\Phi(1/2) = 2^{-c}$. Assuming $c = 0$ leads to $1 = \Phi(1 - 1/n) \vee \Phi(1/n)$, contradicting Equation (48). Equation (10) implies that $0 \leq \Phi(0) \leq \inf_{n \in \mathbb{N}} S(U_n) = 0$. Equations (47), (22), and (48) yield

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

On the other hand, Equations (23) and (48) and iterating Equation (47) yield

$$S(G_{n,a}) = \Phi(a)S(G_{n-1,a}) \vee \Phi(1 - a) = \Phi^n(a) \vee \Phi(1 - a).$$

We define $U_{n,n} = \{0\} \cup U_n$ and $\Phi(1) = 1$. Then, Equations (24) and (25) yield

$$S(U_{n-k,n}) = \Phi(1 - k/n)(n - k)^{-c} \vee \Phi(k/n)$$

and

$$\begin{aligned}
S(U_{n-k,n}) &= \Phi(1 - 1/n)S(U_{n-k-1,n-1}) \vee \Phi(1/n) \\
&= \prod_{m=k+1}^{n} \Phi(1 - 1/m) \vee \bigvee_{\ell=k+1}^{n} \Phi(1/\ell) \prod_{m=\ell+1}^{n} \Phi(1 - 1/m)
\end{aligned}$$

for $0 \leq k < n$. Lemma 2 finalizes the proof.

3.6. Proof of Theorem 4

Assume that H is the Hartley entropy and let $f(a) = \mathbf{1}_{(0,1]}(a)$. Then, the second condition in the theorem holds true. Reversely, assume that the second condition applies. Equations (12) and (19) yield

$$S(g_{2^n,p_1})S(\hat{\partial}_1^0 p) = S(\hat{\partial}_1^n p) = S(\hat{\partial}_1^0 p)F_n(p_1), \quad p_1 < 1, p \in P,$$

so that

$$S(g_{2^n,a}) = F_n(a), \qquad a \in (0,1), \tag{49}$$

as $S(p) \in [1, \infty)$ for $p \in P_0$. Since $F_n(a)$ is monotone in n, the limit $F(a) = \lim_{n \to \infty} F_n(a)$ always exists, i.e., $F(a) \in [1, \infty]$. Assume that $F(a_*) < \infty$ for some $a_* \in (0,1)$. Condition 2(f) and Equations (29), (17), (49), and (11) yield for $p \in P_0$ that

$$\begin{aligned}
\infty &> (S(\{a_*\} \cup (1 - a_*)p) - 1)F(a_*) + 1 \\
&= \lim_{n \to \infty} S(\partial_1^n(\{a_*\} \cup (1 - a_*)p)) \\
&= \lim_{n \to \infty} \frac{S(a_*^{2^n}\{1\} \cup (1 - a_*^{2^n})g_{2^n,a_*} \times p)}{S(g_{2^n,a_*} \times p)} \lim_{n \to \infty} F_n(a_*)S(p) \\
&= F(a_*)S(p). \tag{50}
\end{aligned}$$

Assume that $F(a_*) = 1$. Equations (50) and (11) deliver for $p \in P_0$

$$\begin{aligned}
\infty > S(p) &= S(\{a_*\} \cup (1 - a_*)p) = \ldots = S((1 - (1 - a_*)^n)g_{n,1-a_*} \cup (1 - a_*)^n p) \\
&= \lim_{n \to \infty} S(g_{n,1-a_*}). \tag{51}
\end{aligned}$$

Lemma 3 yields $S(U_n) = n^c$ for $c \geq 0$. Now, Equation (51) implies that $c = 0$, while the assumption in the theorem, $H(U_2) > 0$, implies that $c > 0$. Hence, $F(a_*)$ cannot be equal to 1. Plugging $p = G_{k,1-a_*}$ into Equation (50) gives

$$S(G_{k-1,1-a_*})F(a_*) = \lim_{n \to \infty} S(\partial_2^n G_{k,1-a_*}) = (S(G_{k,1-a_*}) - 1)F(a_*) + 1,$$

by Equation (28). As $F \geq 1$, iterating yields $S(G_{k,1-a_*}) = 1 + k(1 - 1/F(a_*))$, so that

$$1 + 2k(1 - 1/F(a_*)) = (1 + f(a_*^k))(1 + k(1 - 1/F(a_*))) - f(a_*^k), \qquad k \in \mathbb{N},$$

by Equation (30). Since $F(a_*) \neq 1$, we obtain $f(a_*^k) = 1$, $k \in \mathbb{N}$, contradicting the finiteness of $F(a_*)$. Hence, $F(a) = \infty$ for all $a \in (0,1)$. Equations (29), (11), and (49) yield

$$S(p) - 1 = \frac{S(\partial_1^n p) - 1}{F_n(p_1) S(\hat{\partial}_1^0 p)} S(\hat{\partial}_1^0 p) \to S(\hat{\partial}_1^0 p), \qquad p_1 \in (0,1), S(p) < \infty,$$

i.e.,

$$S(p) = S(\hat{\partial}_1^0 p) + 1, \qquad p_1 \in (0,1), S(p) < \infty. \tag{52}$$

The assertion of the theorem follows for $p \in P_0$ by Conditions 2(b)–2(d) and the fact that $S(\{1\}) = 1$. In particular, $f(a) = \mathbf{1}_{(0,1]}(a)$. Assume now that $S(p) < \infty$ for some $p \in P \setminus P_0$. Without loss of generality, $p_i > 0$ for all $i \in \mathbb{N}$. Then, by induction, Equation (52) delivers $S((\hat{\partial}_1^0)^n p) < \infty$ and $S(p) = S((\hat{\partial}_1^0)^n p) + n$, $n \in \mathbb{N}$. This contradicts $S \geq 1$ and $S(p) < \infty$. Hence, $S(p) = \infty$ for $p \in P \setminus P_0$.

4. Discussion

Although all conditions in the characterizations of Shannon's and Rényi's entropy have intrinsic motivations and interpretations, their relevance might differ. Continuity and finiteness offer advantages also from a technical point of view and will usually be retained. Condition 2(b) is intrinsic for probability distributions. The additivity might be considered to be axiomatic. Therefore, we would like to discuss here the importance of the following three points:

1. The discrete nature of the objects;
2. The hierarchical, self-similar structure of the objects, cf. Condition 2(g);
3. The parity among the objects at each hierarchical level, cf. Condition 2(b).

4.1. Practical Implications

The three conditions above are obviously tailormade for tree-like graphs and other hierarchical structures, without the necessity to appeal information-theorical arguments. Examples that may fit our approach are, for instance,

- Classification systems: forms of art (e.g., music genres), languages and their dialects.
- Biology: taxonomie, diversity, genome.
- Administration units: districts, corporate structure, military.
- Universe: clustering.
- Material sciences [18,19].
- Networks [20].
- Statistics: graphs and cluster analysis.

For all these examples, Shannon's or Rényi's entropy may at least be a default choice. The min-entropy may suit graphical models for extremes [21].

4.2. Modifications

The three properties above might be the most questionable from a nonideal, practical point of view. While we believe that nondiscrete objects lead to different theories, modifications and variants of the self-similarity property and the parity property might be accessible to some extent by theoretical advancements of this framework. Also modifications of the operator ∂ could be of practical interest. Since the tree built by ∂^∞ can be seen as a deterministic two-type Kesten tree [22], randomizations of ∂^∞ bridge to the Galton–Watson processes.

4.3. The Geometric Distribution

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. The important role of the geometric distribution is already known in information theory: it maximizes the Shannon entropy among discrete probability distributions given the mean ([23], Theorem 5.8) and minimizes the min-entropy under fixed variance, among all discrete log-concave random variables in $\mathbb{Z}$ ([12], Theorem 1). The geometrical distribution also appears in branching processes [24] and in physics to describe the canonical ensemble [25]. By our research, we add that the geometric distribution is central to describe the self-similarity law of the Shannon entropy. Obviously, the operator ∂ can be inverted, so that the distribution is not refined at a leaf of the graph, but the structure is extended at the root, i.e., $\partial^{-1}(q, a, p) := (1 - a)p \cup aq$, so that $\partial^{-1}(\partial_1^\infty p, p_1, \hat{\partial}_1^0 p) = \partial_1^\infty p$. Hence, $\partial_1^\infty p$ can be interpreted as a kind of limit distribution that is invariant to ∂^{-1}. Whether this leads to any formulation of a limit law including a domain of attraction is an open question. Note that the operator ∂^{-1} has already been addressed in [8].

4.4. On the Proportionality Relation Characterizing the Rényi Entropy

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 (15)) does not play a role in the case of the Shannon entropy. Also, the functional characterization in Equation (8) of the Shannon entropy is not the limit of the characterization of the Rényi entropy (see Equation (9)). However, Equation (8) might be understood as the derivative of Equation (9) in the following sense, cf. [26]. Assume that f in Equation (9) 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 $S(p) = \lim_{\alpha \to 1} \|p\|_\alpha^\alpha$. Then, we obtain

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

Another interpretation of our results is that the order-α information energy S is not a natural transformation of the Rényi entropy, in contrast to

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

which equals the Harvda–Charvát entropy [27] or Tsallis entropy [28], up to a multiplicative constant. Therefore, T bridges to the generalized information function ([5], Section 6.2), but also to the definite functions ([29], Chapter 3, Corollary 3.3 and proof of Theorem 2.2, and Chapter 6, Example 5.16). Indeed, Equation (9) can be rewritten in full analogy to Equation (8), 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.

4.5. Tsallis and Other Entropies

It is an open question whether further practically relevant chain-rule-like conditions such as Condition 2(g) can be found, or, equivalently, recursive formulae such as Equations (34), (36), and (47), and, if so, whether they will be based on the geometric distribution. Reversely, it is also unclear which other existing definitions of an entropy can

be described by conditions similar to Conditions 2(a)–2(g). Reference [15] considers additivity, i.e., Condition 2(f), as the only axiomatic property that an entropy must have. The prevalent interpretation of the Tsallis entropy [28] is that it is not additive; hence, the Tsallis entropy and many others are, per se, excluded from this framework, or the framework would have to be largely extended. Reference [15] remarks that the additivity law has two ingredients—the independence of the components before being joined and the requirement that the components do not interact when being joined. In other words, one may add an additivity law to the Tsallis entropy so that the Tsallis entropy indeed fits this framework. The practical use of such an approach may show the future.

4.6. Shannon's Notion of Entropy

Shannon [30] took a very general perspective on which properties an entropy should have, namely, that "any monotonic function of this number [the finite number of messages] can be regarded as a measure of the information produced when one message is chosen from the set, all choices being equally likely". Further, Shannon justified his and Hartley's choice by the statement that "the logarithmic measure is more convenient". Shannon added three arguments, namely, "it is practically more useful", "it is nearer to our intuitive feeling as to the proper measure", and "it is mathematically more suitable". Shannon is known for his repeated warnings against the blind use of his approach outside the framework of telecommunication [31]. The approach of [15] is, in this spirit, trying to make Shannon's three arguments mathematically more concrete and generally applicable. To this end, the entropy is characterized in algebraic terms only, which reflects the desired properties. Ideally, these properties are rich enough such that the choice of entropy becomes unique. Preferably, the desired properties stem from common sense, invariance considerations, and limit laws. This paper and the presented characterizations of the entropies follow these requirements, giving a deep justification for the use of the Shannon entropy in a large spectrum of applications. But it also indicates its limits of application, unless further deep characterizations are found.

5. Conclusions

The characterizations given here are not only novel with respect to splitting the chain rule into its natural components, they are also novel in the sense that only a rather weak relation is required for the chain rule, i.e., the proportionality factor itself is not given, but arises naturally. Furthermore, the recursion formulae have the advantage that they allow an interpretation as an intrinsic property, since only a single distribution is involved. Further, the recursion formulae may allow additional practical interpretations such as self-similarity or scale invariance. Lastly, 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 broad spectrum of applications outside telecommunication.

References

1. Khinchin, A. The concept of entropy in the theory of probability. *Uspekhi Mat. Nauk* **1953**, *8*, 3–20.
2. Faddeev, D. On the concept of entropy of a finite probabilistic scheme. *Uspekhi Mat. Nauk* **1956**, *11*, 227–231.
3. Rényi, A. On measures of entropy and information. In *Fourth Berkeley Symposium on Mathematical Statistics and Probability: Contributions to the Theory of Statistics*; University of California Press: Berkeley, CA, USA, 1961; pp. 547–562.
4. Leinster, T. *Entropy and Diversity—The Axiomatic Approach*; Cambridge University Press: Cambrdige, UK, 2021.
5. Aczél, J.; Daróczy, Z. *On Measures of Information and Their Characterizations*; Academic Press: New York, NY, USA, 1975.
6. Ebanks, B.; Sahoo, P.K.; Sander, W. *Characterization of Information Measures*; World Scientific: Singapore, 1998.
7. Carcassi, G.; Aidala, C.; Barbour, J. Variability as a better characterization of Shannon entropy. *Eur. J. Phys.* **2021**, *42*, 045102. [CrossRef]
8. Baez, J.; Fritz, T.; Leinster, T. A characterization of entropy in terms of information loss. *Entropy* **2011**, *13*, 1945–1957. [CrossRef]
9. Onicescu, O. Theorie de l'information energie informationelle. *Comptes Rendus De L'Acad. Des Sci. Ser. A-B* **1966**, *263*, 841–842.
10. Pardo, L. Order-α weighted information energy. *Inf. Sci.* **1986**, *40*, 155–164. [CrossRef]
11. Hartley, R. Transmission of information 1. *Bell Syst. Tech. J.* **1928**, *7*, 535–563. [CrossRef]
12. Jakimiuk, J.; Murawski, D.; Nayar, P.; Słobodianiuk, S. Log-concavity and discrete degrees of freedom. *Discret. Math.* **2024**, *347*, 114020. [CrossRef]
13. Aczél, J.; Forte, B.; Ng, C. Why the Shannon and Hartley entropies are 'natural'. *Adv. Appl. Probab.* **1974**, *6*, 131–146. [CrossRef]
14. Arunachalam, S.; Chakraborty, S.; Koucký, M.; Saurabh, N.; De Wolf, R. Improved bounds on Fourier entropy and min-entropy. *ACM Trans. Comput. Theory* **2021**, *13*, 22. [CrossRef]
15. Schlather, M. An algebraic generalization of the entropy and its application to statistics. *arXiv* **2024**, arXiv:2404.05854.
16. Gradshteyn, I.; Ryzhik, I. *Table of Integrals, Series, and Products*, 6th ed.; Academic Press: London, UK, 2000.
17. Erdös, P. On the distribution function of additive functions. *Ann. Math.* **1946**, *47*, 1–20. [CrossRef]
18. Rickman, J.; Barmak, K.; Chen, B.; Patrick, M. Evolving information complexity of coarsening materials microstructures. *Sci. Rep.* **2023**, *13*, 22390. [CrossRef] [PubMed]
19. Du, C.; Li, X.; Liu, C.; Song, C.; Yuan, J.; Xin, Y. Combining ultrasonic guided wave and low-frequency electromagnetic technology for defect detection in high-temperature Cr–Ni alloy furnace tubes. *Sci. Rep.* **2023**, *13*, 18592. [CrossRef] [PubMed]
20. Song, C.; Havlin, S.; Makse, H. Self-similarity of complex networks. *Nature* **2005**, *433*, 392–395. [CrossRef]
21. Engelke, S.; Hitz, A. Graphical models for extremes. *J. R. Stat. Soc. Ser. B* **2020**, *82*, 871–932. [CrossRef]
22. Abraham, R.; Delmas, J.F. An introduction to Galton–Watson trees and their local limits, 2015. *arXiv* **2014**, arXiv:1506.05571.
23. Conrad, K. Probability Distributions and Maximum Entropy. Available online: https://kconrad.math.uconn.edu/blurbs/analysis/entropypost.pdf (accessed on 14 November 2012).
24. Janson, S. Simply generated trees, conditioned Galton-Watson trees, random allocations and condensation. *Discret. Math. Theor. Comput. Sci.* **2012**, *9*, 103–252. [CrossRef]
25. Wikipedia. Canonical Ensemble. Available online: https://en.wikipedia.org/wiki/Canonical_ensemble (accessed on 28 November 2012).
26. Życzkowski, K. Rényi extrapolation of Shannon entropy. *Open Syst. Inf. Dyn.* **2003**, *10*, 297–310. [CrossRef]
27. Havrda, J.; Charvát, F. Quantification method of classification processes. Concept of structural a-entropy. *Kybernetika* **1967**, *3*, 30–35.
28. Tsallis, C. Entropy. *Encyclopedia* **2022**, *2*, 264–300. [CrossRef]
29. Berg, C.; Christensen, J.P.R.; Ressel, P. *Harmonic Analysis on Semigroups. Theory of Positive Definite and Related Functions*; Springer: New York, NY, USA, 1984.
30. Shannon, C. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
31. Shannon, C. The bandwagon. *IRE Trans. Inf. Theory* **1956**, *2*, 3. [CrossRef]