*Article*

# Collision Entropy Estimation in a One-Line Formula

**Alessandro Gecchele**

**Abstract:** We address the unsolved question of how best to estimate the collision entropy, also called quadratic or second order Rényi entropy. Integer-order Rényi entropies are synthetic indices useful for the characterization of probability distributions. In recent decades, numerous studies have been conducted to arrive at their valid estimates starting from experimental data, so to derive suitable classification methods for the underlying processes, but optimal solutions have not been reached yet. Limited to the estimation of collision entropy, a one-line formula is presented here. The results of some specific Monte Carlo experiments give evidence of the validity of this estimator even for the very low densities of the data spread in high-dimensional sample spaces. The method strengths are unbiased consistency, generality and minimum computational cost.

**Keywords:** Rényi entropies; collision entropy estimation; collision entropy rate estimation

## 1. Introduction

The *information theory indices* belonging to the parametric family of *Rényi entropies* are able to express, each with a different weight, the information content of a *discrete probability distribution* (*DPD*) [1]. Typical members of this family are, for example, Shannon entropy, collision entropy and min-entropy. These indices can also be used to classify the output of *experimental processes* studied in any branch of the applied sciences, provided their reduction to pseudostationary discrete-state processes and then in the form of *DPD*s. Since usually, during the experiments, only brief *realizations* can be obtained from the process under investigation, and since the realizations give rise to *relative frequency distributions* (*RFD*s) and not to *DPD*s, then these indices, being based on probabilities, have to be *estimated* through the *elaboration of the few available data*. In this regard, the methods for the estimation of Rényi entropies are of two kinds: 1) those that first aim to estimate the probability distribution from the relative frequencies and then plug the estimated probabilities into the formulas of the entropies and 2) those that circumvent the still-open problem of the estimation of the probabilities and aim to estimate the entropy indices through the application of other elaborations to the data. Despite the numerous studies carried out in the last decades (e.g., [2], [3], [4], [5], [6], [7], [8],[9], [10], [11], [12],[13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24],[25], [26],[27], [28], [29], [30], [31], [32], [33]), optimal estimators have not been found yet. Moreover, this persistent lack of satisfactory solutions for the estimation of the indices belonging to the Rényi family (and for the estimation of their more rapidly converging derived quantities called *Rényi entropy rates*) has prompted, as a side effect, the proliferation of other similar indices conceived in many different ways (e.g. [34], [35], [36]), but all having the same purpose of classifying data with a nonparametric approach. An overview of this peculiar situation can be found in [37], where Ribeiro et al. collected and described a "galaxy" of at least thirty indices somehow functionally equivalent to those of the family initially proposed by Rényi (and to their rates). In this context, the search for the best estimators of the original indices seems most appropriate. In this regard, Skorski ([38], [39]) rightly pointed out that the estimation of those integer-order Rényi entropies that have a parameter value greater than one reduces to the estimation of the power sums of a *DPD*. Our work just starts from this latter consideration and limits its investigation only to the case of the estimation of the second power sum, which, in turn, allows the collision entropy to be estimated.

## 2. Theoretical Methods

### 2.1. Transforming a Discrete-State Stochastic Process into a DPD

Consider a discrete-state stochastic process ($DSP_q$) $x_{-\infty}, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_\infty$ whose values belong to an alphabet $A_q$ containing $q$ ordered symbols. Let $\Omega(q,d)$ be a $d$-dimensional discrete sample space resulting from the Cartesian product $d$ times of $A_q$.

$$\Omega(q,d) = \underbrace{A_q \times A_q \times ... \times A_q}_{d \ times}, \tag{1}$$

and let $n = q^d$ be the cardinality of the sample space $\Omega(q,d)$. Each elementary event $\boldsymbol{e}_k$, with $k \in \{1, 2, ..., n\}$, is uniquely identified by a vector with $d$ coordinates ($x1_k$, $x2_k$, ..., $xd_k$), with $x1_k$, $x2_k$, ..., $xd_k \in A_q$. According to the procedure indicated by Shannon in [40] at pages 5 and 6, the infinite sequence of samples constituting the $DSP_q$ can be transformed into occurrences $\#(\boldsymbol{e}_k)$ of the elementary events of $\Omega(q,d)$ by progressively considering all the $d$-grams taken from the samples as if they were the vector coordinates of the events and counting the number of times that each vector appears in the sequence. Then, according to the *frequentist definition of probability*, the final resulting $DPD$ is expressible in *set theory* notation as

$$\boldsymbol{p}(\Omega(q,d))_{DSP_q} = \left\{ p(\boldsymbol{e}_k)_{DSP_q} = \frac{\#(\boldsymbol{e}_k)_{DSP_q}}{\sum_{k=1}^{n} \#(\boldsymbol{e}_k)_{DSP_q}} \,\middle|\, \boldsymbol{e}_k \in \Omega(q,d) \right\}. \tag{2}$$

In the following, in the absence of ambiguity, $\boldsymbol{p}(\Omega(q,d))_{DSP_q}$—that is, a $DPD$ obtained by elaborating the data of a $DSP_q$— will be indicated with the bold symbol $\boldsymbol{p}$ and one of its elements with $p_k$.

### 2.2. Integer-Order Rényi α-Entropies as Synthetic Indices for the Characterization of DPDs

In general, a $DPD$ can be characterized by some indices, each of which can quantify the presence rate of a particular feature in the distribution. The parametric family of *integer-order Rényi α-entropies* is composed of synthetic indices suitable for the characterization of $DPD$s from the point of view of their informative content [1]. They are defined as

$$\begin{array}{llll} & \alpha = 1 & H_1(\boldsymbol{p}) \triangleq -\sum_{k=1}^{n} p_k \log p_k & \\ \alpha \in \mathbb{N}^+ & \alpha \neq 1 & H_\alpha(\boldsymbol{p}) \triangleq \frac{1}{1-\alpha} \log\left(\sum_{k=1}^{n} p_k^\alpha\right) & 0 \leq H_\alpha(\boldsymbol{p}) \leq \log n \\ & \alpha \to \infty & H_\infty(\boldsymbol{p}) \triangleq -\log max\{\boldsymbol{p}\}. & \end{array} \tag{3}$$

The corresponding *specific integer-order Rényi α-entropies* of the $DPD$ $\boldsymbol{p}$ are then defined as

$$\begin{array}{llll} & \alpha = 1 & \eta_1(\boldsymbol{p}) \triangleq \frac{H_1(\boldsymbol{p})}{\log n} = -\sum_{k=1}^{n} p_k \log_n p_k & \\ \alpha \in \mathbb{N}^+ & \alpha \neq 1 & \eta_\alpha(\boldsymbol{p}) \triangleq \frac{H_\alpha(\boldsymbol{p})}{\log n} = \frac{1}{1-\alpha} \log_n\left(\sum_{k=1}^{n} p_k^\alpha\right) & 0 \leq \eta_\alpha(\boldsymbol{p}) \leq 1 \\ & \alpha \to \infty & \eta_\infty(\boldsymbol{p}) \triangleq \frac{H_\infty(\boldsymbol{p})}{\log n} = -\log_n max\{\boldsymbol{p}\}. & \end{array} \tag{4}$$

Once the value of a *specific entropy* is known, it is always possible to retrieve the value of the corresponding *plain entropy*, expressed in a particular base $b$ and for a particular cardinality $n$, using the following conversion formula:

$$H_\alpha(\boldsymbol{p}, b, n) \triangleq \eta_\alpha(\boldsymbol{p}) \log_b n. \tag{5}$$

Specific entropies are preferable to plain entropies because:

1. they are the result of a *min-max normalization*, that is obtained using the minimum and the maximum possible values of plain entropies (respectively 0 and $\log n$);
2. they are formally *independent from the number of ordered symbols $q$* chosen for the quantization of the range of the output values of the process and *independent from the cardinality of the sample space $n$*; for this reason, they allow the comparison of values obtained from different distributions, even generated using different sample spaces;
3. they allow the *doubt on the choice of the base for the logarithm* present in the formula of entropies ($_2$ or $_e$ or $_{10}$) to be removed, thanks to the use of a variable base, depending on the cardinality of the considered sample space ($_n$);

### 2.3. Rényi Entropy Rates

Unlike *Rényi entropies*, whose utility is mainly related to the classification of $DPD$s, *Rényi entropy rates* are important theoretical quantities useful for the characterization of $DSP_q$s [41], [42]; they are defined as

$$H'_\alpha(DSP_q) \triangleq \lim_{d\to\infty} \frac{1}{d} H_\alpha(\boldsymbol{p}(\Omega(q,d))_{DSP_q}) \qquad 0 \le H'_\alpha(DSP_q) \le \log q. \quad (6)$$

Moreover, it is known that, for strongly stationary $DSP_q$, any *Rényi entropy rate* converges to the same limit of a sequence of Cesaro means of *conditional entropies*:

$$H'_\alpha(DSP_q) = \lim_{d\to\infty} H_\alpha(\boldsymbol{p}(A_d)|\boldsymbol{p}(A_1 \times A_2 \times \cdots \times A_{d-1})). \quad (7)$$

and, as conditional Rényi entropies preserve the chain rule [43], [44], [45], they can also be calculated as

$$H'_\alpha(DSP_q) = \lim_{d\to\infty} \Big[H_\alpha(\boldsymbol{p}(\Omega(q,d))_{DSP_q}) - H_\alpha(\boldsymbol{p}(\Omega(q,d-1))_{DSP_q})\Big]. \quad (8)$$

### 2.4. Specific Rényi Entropy Rate

Similarly to Formula (4), *specific Rényi entropy rate* is defined by the following min-max normalization:

$$\begin{aligned} \eta'_\alpha(DSP_q) &= \frac{H'_\alpha(DSP_q)}{\log q} = \\ &= \lim_{d\to\infty} \frac{\Big[H_\alpha(\boldsymbol{p}(\Omega(q,d))_{DSP_q}) - H_\alpha(\boldsymbol{p}(\Omega(q,d-1))_{DSP_q})\Big]}{\log q} = \\ &= \lim_{d\to\infty} \Big[d\,\eta_\alpha(\boldsymbol{p}(\Omega(q,d))_{DSP_q}) - (d-1)\,\eta_\alpha(\boldsymbol{p}(\Omega(q,d-1))_{DSP_q})\Big], \end{aligned} \quad (9)$$

with $0 \le \eta'_\alpha(DSP_q) \le 1$.

### 2.5. Relationship between Specific Rényi Entropy Rate and Specific Rényi Entropy

In summary, the following relationship subsists:

$$\eta'_\alpha(DSP_q) = \lim_{d\to\infty} \eta_\alpha(\boldsymbol{p}(\Omega(q,d))_{DSP_q}) \quad (10)$$

This means that, varying $d$, the *specific Rényi entropy* tends to the same value of the *specific Rényi entropy rate*, with the important difference being that the rate of convergence of the *specific Rényi entropy rate* is much faster than the rate of convergence of the *specific Rényi entropy*. For this reason, when possible, using the *specific Rényi entropy rate* is preferable to using the *specific Rényi entropy*.

## 3. Empirical Methods

### 3.1. Transforming a Realization into a Distribution of Relative Frequencies

For the practical cases, the theoretical procedure described in § 2.1 can be adapted according to the following procedure already presented with greater generality in [46] and in [47]: consider the N samples $x_1, x_2, \ldots, x_d, x_{d+1}, \ldots, x_N$ of a realization $r_q$ extracted from a $DSP_q$. Each $d$-gram composed of $d$ adjacent samples of $r_q$ is interpreted as the occurrence of the elementary event of a $d$-dimensional sample space $\Omega(q,d)$ having just those values as vector components. For example, the first two $d$-grams taken from $r_q$, $(x_1, x_2, ..., x_d)$ and $(x_2, x_3, ..., x_{d+1})$ identify the first occurrences of two elementary events. The count of the occurrences of the events is performed for all the $d$-grams progressively identified in the sequence of the samples of $r_q$. Finally, the absolute frequency of every elementary event $\#(\boldsymbol{e}_k)$ is divided by the total number of occurrences ($L = \sum_{k=1}^{n} \#(e_k)_{r_q} = N - d + 1$), yielding its relative frequency $f(\boldsymbol{e}_k)_{r_q}$. The final resulting $RFD$ is expressible in *set theory* notation as

$$\boldsymbol{f}(\Omega(q,d))_{r_q} = \Big\{ f(\boldsymbol{e}_k)_{r_q} = \frac{\#(\boldsymbol{e}_k)_{r_q}}{\sum_{k=1}^{n} \#(\boldsymbol{e}_k)_{r_q}} \;\Big|\; \boldsymbol{e}_k \in \Omega(q,d) \Big\}. \tag{11}$$

In the following, in the absence of ambiguity, an RFD $\boldsymbol{f}(\Omega(q,d))_{r_q}$ resulting from the insertion of the data of a realization in a sample space will be simply indicated with the bold symbol $\boldsymbol{f}$ and $f_k$ indicates one of its elements.

### 3.2. Estimating the Second Power Sum of a DPD

Preliminarily, the $\alpha^{th}$-power sum of a $DPD$ $\boldsymbol{p}$ and the $\alpha^{th}$-power sum of a $RFD$ $\boldsymbol{f}$ are defined as

$$S_\alpha(\boldsymbol{p}) \triangleq \sum_{k=1}^{n} p_k^\alpha, \quad S_\alpha(\boldsymbol{f}) \triangleq \sum_{k=1}^{n} f_k^\alpha \qquad\qquad \frac{1}{n^{\alpha-1}} \leq S_\alpha(\cdot) \leq 1 \tag{12}$$

Limited to the power sums of Poissonian distributions, Grassberger in 1988 [2], Formula (8), and subsequently Schürmann in 2004 [12], Formula (6), reported the theoretically demonstrable, unique unbiased estimator, repeated in Formula (13):

$$\begin{gathered} \widehat{S_\alpha(\boldsymbol{p})}_{Poisson} = \Big\langle \sum_{k=1}^{n} \widehat{p_k^\alpha} \Big\rangle_{r_q} = \sum_{k=1}^{n} \Big\langle \frac{1}{L^\alpha} \frac{\#(\boldsymbol{e}_k)_{r_q}!}{(\#(\boldsymbol{e}_k)_{r_q} - \alpha)!} \Big\rangle_{r_q} \\ \widehat{p_k^\alpha} := 0 \quad \text{for} \quad \#(\boldsymbol{e}_k)_{r_q} < \alpha, \end{gathered} \tag{13}$$

where $\langle \cdot \rangle_{r_q}$ is the mean over the infinite number of realizations that can be taken from the underlying process. For the specific case of the estimation of the second power sum, Formula (13) becomes:

$$\widehat{S_2(\boldsymbol{p})}_{Poisson} = \sum_{k=1}^{n} \Big\langle \frac{[\#(\boldsymbol{e}_k)_{r_q} - 1]\#(\boldsymbol{e}_k)_{r_q}}{L^2} \Big\rangle_{r_q} = \Big\langle \sum_{k=1}^{n} f_k^2 - \frac{1}{L} \Big\rangle_{r_q} = \langle S_2(\boldsymbol{f}) - \frac{1}{L} \rangle_{r_q}. \tag{14}$$

As far as we know, the scientific literature does not indicate whether the result of Formula (14) can also be valid for distributions different from Poissonians. So, from now on *we proceed assuming provisionally that this hypothesis is true*, and we leave the decision concerning its acceptance or rejection to the phase of the interpretation of the results of the Monte Carlo experiments described in section 5. The hypothesis can be resumed as:

$$\forall\, DSP_q \qquad \widehat{S_2(\boldsymbol{p})}_{DSP_q} = \Big\langle max\Big\{ S_2(\boldsymbol{f}) - \frac{1}{L}, \; \frac{1}{n} \Big\} \Big\rangle_{r_q} \tag{15}$$

where the lower limit $\frac{1}{n}$ is necessary because, when the cardinality of the sample space becomes high and the data density becomes too rarefied, the only possible estimate of the probability distribution results in the uniform distribution.

### 3.3. *Estimating the Specific Collision Entropy of a $DSP_q$*

*Collision entropy* is the particularization of Formula (3) for $\alpha = 2$, and it is defined as

$$H_2(\boldsymbol{p}) \triangleq -\log\Big(\sum_{k=1}^{n} p_k^2\Big) = -\log S_2(\boldsymbol{p}) \qquad 0 \le H_2(\boldsymbol{p}) \le \log n \quad (16)$$

Inserting Formula (16) into Formula (4), the *specific collision entropy* is defined as

$$\eta_2(\boldsymbol{p}) \triangleq -\frac{H_2(\boldsymbol{p})}{\log n} = -\log_n S_2(\boldsymbol{p}) \qquad 0 \le \eta_2(\boldsymbol{p}) \le 1. \quad (17)$$

In the steps of Formulas (13) and (14), the displacements of the symbol that indicates the average over different realizations $\langle\cdot\rangle_{r_q}$ from the outside to the inside of the symbol of summation $\sum$ and vice versa are mathematically indisputable. But the application of the logarithm to the second power sum for arriving at the estimation of the collision entropy does not allow these shifts anymore. In fact, although the two possible expressions for the evaluation of the mean over the realizations give similar results in the presence of *RFD*s (i.e. $-\langle \log_n S_2(\boldsymbol{f})\rangle_{r_q} \simeq -\log_n \langle S_2(\boldsymbol{f})\rangle_{r_q}$), in general they differ remarkably when the logarithm is applied to the estimate of the second power sum of probabilities:

$$\underbrace{-\left\langle \log_n max\left\{S_2(\boldsymbol{f}) - \frac{1}{L}, \frac{1}{n}\right\}\right\rangle_{r_q}}_{\text{Mean of Logs of 2-Power Sum } (MLS_2)} \neq \underbrace{-\log_n \left\langle max\left\{S_2(\boldsymbol{f}) - \frac{1}{L}, \frac{1}{n}\right\}\right\rangle_{r_q}}_{\text{Log of Mean of 2-Power Sum } (LMS_2)} . \quad (18)$$

Consequently, the estimation of the *specific collision entropy* is performed averaging the previous two possible expressions:

$$\widehat{\eta_2}(\boldsymbol{p})_{DSP_q} = -\widehat{\log_n S_2}(\boldsymbol{p}) = \frac{MLS_2 + LMS_2}{2}. \quad (19)$$

This is also **the main result of this paper**. The estimation of plain collision entropy can be obtained by inserting Formula (19) into Formula (5).

### 3.4. *Estimating the Specific Collision Entropy Rate of a DSPq*

From Formula (9) and Formula (19), it can be inferred that

$$\widehat{\eta_2'}(\boldsymbol{p}(\Omega(q,d))_{DSP_q}) = \Big[d\,\widehat{\eta_2}(\boldsymbol{p}(\Omega(q,d))_{DSP_q}) - (d-1)\,\widehat{\eta_2}(\boldsymbol{p}(\Omega(q,d-1))_{DSP_q})\Big] \quad (20)$$

and

$$\widehat{\eta_2'}(DSP_q) = min\Big\{\widehat{\eta_2'}(\boldsymbol{p}(\Omega(q,d))_{DSP_q}) \;\Big|\; 1 \le d < \infty\Big\}. \quad (21)$$

### 3.5. *Method of Validation of Entropy Estimators*

Monte Carlo simulations are the most correct experiments for observing the average effect of the application of an entropy estimator to every realization extracted from a process under examination. The protocol for the validation of the estimators of entropy and entropy rate consists of the following steps:

1. choice of a convenient $DSP_q$,
2. choice of the number of realizations $R$,
3. choice of the length $N$ of each realization,
4. transformation of the samples of any realization in a *RFD* according to § 3.1,
5. extraction of the estimated indices according to Formulas (19) and (20),
6. production of the diagrams,
7. and evaluation of the performances of the estimator.

## 4. Materials: Choice of Convenient $DSP_q$s Suitable for the Experiments

For the validation of the previous estimation formulas three completely different types of processes were used: two types, located at the opposite extreme borders of the entropy scale, are regular processes and independent, identically distributed (IID) processes exhibiting maximum entropy; the third type, located in between, is composed of simple processes with minimal memory, such as stationary, irreducible, and aperiodic Markov processes. All these types of processes have the fundamental characteristic of having known theoretical values of entropy.

1. *Regular Processes*. The first important sanity check for entropy estimators involves the use of a completely regular process, that consists of an infinitely repeating brief symbolic sequence. Once the initial sequence is known, no additional information is brought by the following samples, and the evolution of the process becomes completely determined. So, for these processes we have

$$\forall\, d \geq 2 \qquad \eta_2'(Regular) = 0. \tag{22}$$

   Then, even for short realizations of this kind of processes, any good estimator of the *specific Rényi entropy rate* has to rapidly fall to zero during the progressive increment of the dimension of the sample space.

2. *Markov Processes*. According to [41], when the $DSP_q$ is a stationary, irreducible, and aperiodic Markov process, it is possible to calculate the theoretical value of its *specific Rényi entropy rate*. In fact, given the transition matrix $\boldsymbol{p}_{qq}$ and the unique stationary distribution $\boldsymbol{\mu}^*{}_q$ obtained as the scaled (with rule $\sum \mu_i^* = 1$) right eigenvector associated to eigenvalue $\lambda = 1$ of the equation

$$\begin{vmatrix} p_{11} & p_{12} & \cdot & p_{1q} \\ p_{21} & p_{22} & \cdot & p_{2q} \\ \cdot & \cdot & \cdot & \cdot \\ p_{q1} & p_{q2} & \cdot & p_{qq} \end{vmatrix}^T \begin{vmatrix} \mu_1 \\ \cdot \\ \cdot \\ \mu_q \end{vmatrix} = \begin{vmatrix} \mu_1 \\ \cdot \\ \cdot \\ \mu_q \end{vmatrix}$$

   then

$$\eta_2'(Markov) \triangleq \lim_{d\to\infty} \frac{1}{d}\, \frac{H_2(\boldsymbol{p}(\Omega(q,d))_{Markov})}{\log q} = -\sum_{i=1}^{q} \mu_i^* \log_q \left( \sum_{j=1}^{q} p_{ij}^2 \right). \tag{23}$$

3. *Maximum Entropy IID Processes*. A third sanity check for entropy estimators involves the use of memoryless IID processes with maximum entropy, because:
   - with these processes, the distance between the entropy of the relative frequencies and the actual theoretical entropy of the process is the maximum possible (i.e., using these processes, the estimator is tested in the most severe conditions, obliging it to generate the greatest possible correction);
   - the theoretical value for the specific entropy of the processes generated is a priori known and results in being constant, regardless of the choice of the dimension of the considered sample space because the outcome of each throw is independent from the past history.
   - having an L-shaped one-dimensional distribution, with one probability bigger than the others, which remain equiprobable, the calculation of their theoretical entropy is trivial;
   - they are easily reproducible by, for example, simulating the rolls of a loaded die on which a particular preeminence of the occurrence of a side is initially imposed; the general formula is:

$$\eta_2'(MaxEnt) \triangleq \eta_2(\boldsymbol{p}(q,d))_{MaxEnt}\big|_{\forall\, d} = -\log_q \left( p_{main}^2 + \frac{(1-p_{main})^2}{q-1} \right)\Big|_{d=1}. \tag{24}$$

## 5. Results and Discussion

As part of this research, countless Monte Carlo experiments were conducted to validate the novel specific collision entropy estimator $\widehat{\eta_2}(\boldsymbol{p})$ obtained in Formula (19) and, consequently, to verify the plausibility of the hypothesis proposed for the estimation of the second power sum of any $DSP_q$ described by Formula (15). Here, only some of the most significant results are reported. Each figure presented in this section contains two diagrams that show, for an established number of realizations and for an established length of each realization, the trend of the estimated *specific collision entropy* and the trend of the estimated *specific collision entropy rate*, calculated as the dimension of the sample space varies.

### *5.1. Experiments with Realizations Coming from Completely Regular Processes*

For the experiment whose results are reported in Figure 1 the input parameters are:

- $DSP_q$ = Regular process obtained repeating the ordered numerical sequence of the values associated with the six faces of a die ($q = 6$).
- $N = 250$ and $R = 1$, because every realization is identical.

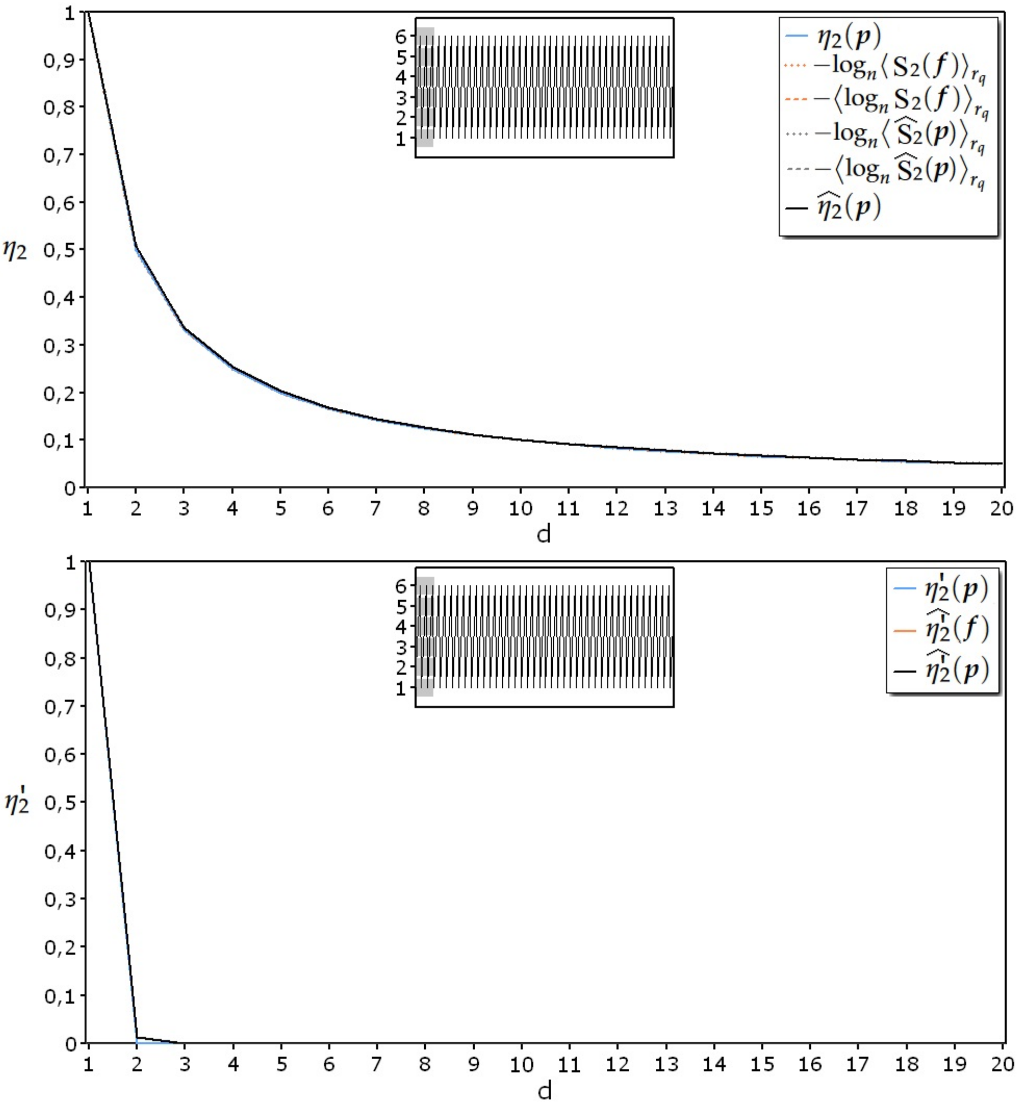

**Figure 1.** Trend of $\eta_2$ (upper diagram) and trend of $\eta_2'$ (lower diagram) for a realization composed of 250 samples taken from a regular process.

The upper diagram of Figure 1 shows that, in general, the theoretical specific collision entropy $\eta_2(\boldsymbol{p})$ decreases only asymptotically to zero and does not reach a minimum value in the dimensional range $1 \leq d \leq 20$. For this reason, this quantity is not indicated for the procedure of process classification. Instead, the lower diagram shows that the specific collision entropy rate $\eta_2'(\boldsymbol{p})$ rapidly decreases to the minimum value of zero, overlapping the theoretical trend for $d > 2$. This example shows that, as a *first necessary prerequisite*, any entropy rate estimator has to exhibit this behavior when dealing with regular processes to be able to be considered suitable for the classification of processes.

### *5.2. Experiments with Realizations Coming from Processes Presenting Some Sort of Regularity*

Consider a Markov process with six possible states (alphabet $A_q = \{1,2,3,4,5,6\}$ and $q = 6$); let the associated transition matrix $\boldsymbol{p}_{66}$ and stationary distribution $\boldsymbol{\mu}_6^*$ be

$$\boldsymbol{p}_{66} = \begin{vmatrix} 0.04 & 0.80 & 0.04 & 0.04 & 0.04 & 0.04 \\ 0.04 & 0.04 & 0.80 & 0.04 & 0.04 & 0.04 \\ 0.04 & 0.04 & 0.04 & 0.80 & 0.04 & 0.04 \\ 0.04 & 0.04 & 0.04 & 0.04 & 0.80 & 0.04 \\ 0.04 & 0.04 & 0.04 & 0.04 & 0.04 & 0.80 \\ 0.80 & 0.04 & 0.04 & 0.04 & 0.04 & 0.04 \end{vmatrix} \qquad \boldsymbol{\mu}_6^* = \begin{vmatrix} \frac{1}{6} \\ \frac{1}{6} \\ \frac{1}{6} \\ \frac{1}{6} \\ \frac{1}{6} \\ \frac{1}{6} \end{vmatrix}.$$

For this process, the theoretical value of *specific collision entropy rate* $\eta_2'(\boldsymbol{p})$ results:

$$\forall d \geq 2 \qquad \eta_2'(\boldsymbol{p}) = \frac{H_2'(\boldsymbol{p})}{\log q} = -\frac{1}{6}\, 6\, \frac{\log(0.8^2 + 5 \cdot 0.04^2)}{\log 6} \simeq 0.242\,.$$

The upper diagram of Figure 2 shows that, in general, for processes whose samples have a dependence from the past, the trend of the *estimated specific collision entropy*, calculated using Formula (19), presents, at the beginning, a decrease, which depends on the progressive reduction of the topological ambiguity encountered during the detection of recurrences hidden in the data when the dimension of the sample space is increased. The curve subsequently rises due to the reduction of the density of the occurrences in the sample space. This corresponds to a reduction in the reliability of the information supplied by the relative frequencies; as a consequence, the uncertainty contained in the probability estimates grows, and the entropy increases accordingly. This ability to ramp up the curve when the estimate is no longer reliable is the *second necessary prerequisite* for an estimator. The observation of the diagrams of Figure 2 allows also to infer that *RFDs cannot be used in place of DPDs* because they intrinsically lack this capability. In fact, the use of the *RFD*s in the estimator gives poor results because their mean specific collision entropy seamlessly decreases even when the density of the data is actually no longer sufficient for producing any kind of estimation. In the middle of the curve, the minimum value of the specific collision entropy represents the best possible compromise between the request to observe in ever greater detail the regularities contained in the data and the limitations imposed by the shortness of the data. From Figure 2 it is also possible to establish a *third necessary prerequisite* that an entropy estimator must fulfill: in fact its output has always to be greater or equal than the corresponding theoretical value, because otherwise the estimator would erroneously signal the presence of an excessive amount of regularities in the process, thus violating the fundamental precaution principle required by all those situations in which statistical fluctuations are present. In a sentence: *an estimator that expresses values of entropy higher than the correct theoretical ones is preferable to an estimator that expresses lower values.* Moreover, when the trend of the *estimated specific collision entropy* is compared with the trend of the *estimated specific collision entropy rate*, it becomes clear once again that this second index produces an impressively more rapid convergence towards the theoretical value (blue line) than the first one.

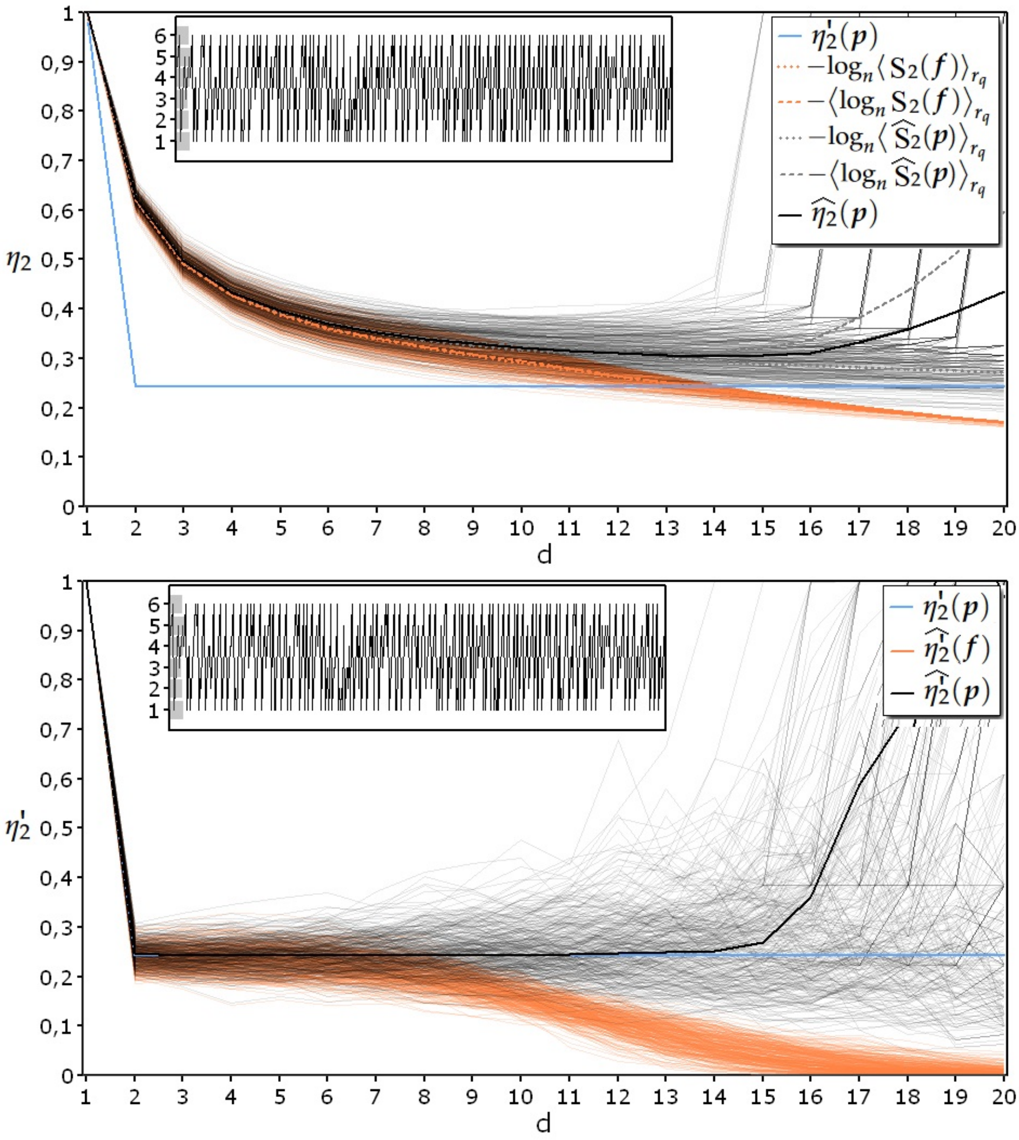

**Figure 2.** Trends of $\eta_2$ (upper diagram) and $\eta'_2$ (lower diagram) for 300 realizations, each composed of 500 samples taken from the Markovian process previously described by the transition matrix $\boldsymbol{p}_{66}$ and the stationary distribution $\boldsymbol{\mu}^*{}_6$.

In the lower diagram of Figure 2 it is possible to see that the adherence of $\widehat{\eta'_2}(\boldsymbol{p})$ to $\eta'_2(\boldsymbol{p})$ persists up to dimension $d = 11$, and in this case the data density results:

$$\delta_{min}(Markov, R = 300, N = 500) = \frac{L}{n} = \frac{N - d + 1}{q^d} = \frac{490}{6^{11}} = 1.35 \cdot 10^{-6}$$

### 5.3. *Experiments with Realizations Coming from Maximum Entropy Memoryless IID Processes*

For the experiment whose results are reported in Figure 3, the input parameters are:

- $DSP_q$ = process generated by tossing a loaded die ($q = 6$) with 50% of the outcomes equal to "1";
- Upper diagram: $R = 2000$ and $N = 250$;
- Lower diagram: $R = 500$ and $N = 1000$.

From Formula (24) it results that

$$\eta'_2(MaxEnt50\%) = \eta_2(MaxEnt50\%) = -\log_6(0.5^2 + \frac{0.5^2}{5}) = -\log_6 0.3 \simeq 0.672.$$

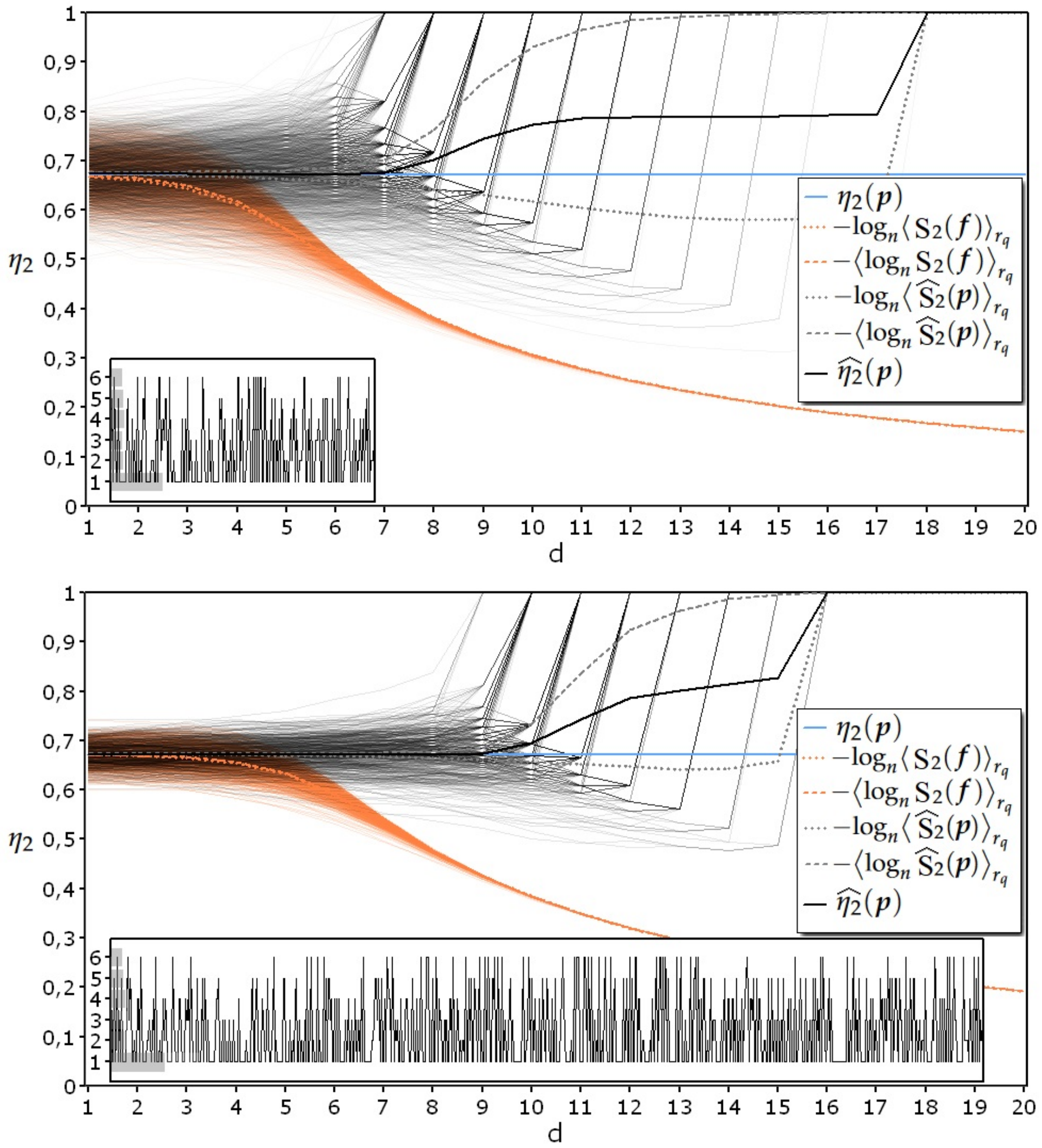

**Figure 3.** Trends of $\eta_2$ for the realizations of a process generated by tossing a loaded die with 50% of the outcomes equal to "1". Upper diagram: 2000 realizations, each 250 samples long; lower diagram: 500 realizations, each 1000 samples long.

Both diagrams of Figure 3 show that:

- the proposed estimator satisfies the aforementioned third prerequisite of never falling below the theoretical line, even in the heaviest test conditions, represented by the elaboration of data coming from a maximum entropy IID process;
- when using *RFD*s to estimate specific collision entropy, there is only a slight difference between the two possible ways of averaging the logarithm of the second power sum (dotted and dashed lines in orange); on the contrary, there is a remarkable difference between the two possible ways of averaging the estimates of the logarithm of the second power sum (dotted and dashed lines in grey) as indicated in Formula (18);
- when the data density in the sample space becomes insufficient for a reliable estimate of the entropy, its value rises toward the value corresponding to the uniform distribution.

In the upper diagram of Figure 3 it is possible to see that considering 250 samples per realization the adherence of $\widehat{\eta_2}(\boldsymbol{p})$ to $\eta_2(\boldsymbol{p})$ persists up to dimension 6; for this dimension the data density in the sample space results:

$$\delta_{min}(MaxEnt\,50\%, R = 2000, N = 250) = \frac{L}{n} = \frac{N-d+1}{q^d} = \frac{245}{6^6} = 5.25 \cdot 10^{-3}$$

and the statistical fluctuations are considerable because of the shortness of the realizations. In the lower diagram of Figure 3 it is possible to see that considering 1000 samples per realization the adherence of $\widehat{\eta_2}(\boldsymbol{p})$ to $\eta_2(\boldsymbol{p})$ persists up to dimension 9 (three dimensions more than the other situation); for this dimension the data density in the sample space results:

$$\delta_{min}(MaxEnt\,50\%, R = 500, N = 1000) = \frac{L}{n} = \frac{N-d+1}{q^d} = \frac{992}{6^9} = 9.84 \cdot 10^{-5}$$

and the statistical fluctuations are reduced because of the greater number of samples of each realization. From the comparison of the two diagrams, it can be seen that the increment in the availability of the data improves all the performance indicators of the estimator, and this fact proves its consistency even in the most severe test conditions represented by this kind of processes. In general, to obtain an adequate horizontal trend of $\widehat{\eta_2'}$ for at least two consecutive dimensions, it is necessary to rely on a sufficiently large number of samples per realization $N$ or, alternatively, on a sufficiently high number of realizations $R$. The total number of aggregated samples (i.e., $R$ x $N$) necessary for a good result of the estimation depends on the effective degree of irregularity of the signal. In fact, for completely regular processes with an alphabet composed of $q$ symbols, even only $5\,q$ samples are sufficient for a correct estimate. Vice versa, for almost random processes, at least $1,000,000$ aggregated samples seem to be necessary.

Finally, concerning the hypothesis made at the beginning about the possibility of estimating the second power sum of the $DPD$s coming from any kind of $DSP_q$ using Formula (15), the evidences that emerged from the results of the experiments made for the validation of the estimator have not provided any counterexample that may exclude its validity. For this reason, the following statistics postulate is proposed:

**Postulate**. *Given a sample space $\Omega(q,d)$ with cardinality $n = q^d$, and given a set of relative frequency distributions $\{\boldsymbol{f}(\Omega(q,d))_{r_q}\}$, each composed of L occurrences, resulting from the transformation of R short realizations $r_q$ taken from the underlying discrete stochastic process $DSP_q$, to which an unknown discrete probability distribution $\boldsymbol{p}(\Omega(q,d))$ is associated, then the unbiased and consistent estimator of the second power sum of $\boldsymbol{p}(\Omega(q,d))$ is inferred as*

$$\forall\, DSP_q \qquad \widehat{S_2}(\boldsymbol{p}(\Omega(q,d)))_{DSP_q} = \lim_{R\to\infty} \left\langle max\left\{S_2(\boldsymbol{f}(\Omega(q,d))_{r_q}) - \frac{1}{L},\ \frac{1}{n}\right\}\right\rangle_{r_q}.$$

## 6. Conclusions

Figures 2 and 3 show that the proposed *specific collision entropy rate estimator* $\widehat{\eta_2'}$ allows a very prolonged and consistent stay of its output, exactly at the values expected by the theory. This highly desirable and very rare feature, the simplicity of its formula and its complete usability with any discrete stationary process make this estimator a valid tool, suitable for measuring the degree of irregularity in experimental data from the perspective given by the collision entropy. Possible future research directions include:

- the evaluation of the admissibility of this estimator by comparing it to other similar collision entropy estimators and by using the same kind of processes for the tests;
- the characterization of the variability of the values returned by the estimator $\widehat{\eta_2'}$ as the number of aggregated samples and the irregularity of the processes vary;
- further studies on the methods of estimation in presence of the logarithm operator.

**Funding:** This research received no external funding.

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

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| $A_q$ | alphabet composed of $q$ ordered symbols |
| $\Omega(q,d)$ | Sample space resulting from the Cartesian product $d$ times of the alphabet $A_q$ |
| $n = q^d$ | cardinality of the sample space $\Omega(q,d)$ |
| $DSP_q$ | Discrete-state stochastic process whose samples belong to an alphabet $A_q$ |
| $r_q$ | Realization of a $DSP_q$ |
| $N$ | Number of samples of $r_q$ |
| $L = N - d + 1$ | Number of occurrences inserted in the events of $\Omega(q,d)$ |
| $RFD$ | Relative frequency distribution |
| $DPD$ | Discrete probability distribution |
| $\boldsymbol{f}(\Omega(q,d))_{r_q}$ | $RFD$ obtained from a realization $r_q$ of a $DSP_q$ whose $d$-grams are inserted in $\Omega(q,d)$ |
| $\boldsymbol{p}(\Omega(q,d))_{DSP_q}$ | $DPD$ obtained from a $DSP_q$ whose $d$-grams are inserted in $\Omega(q,d)$ |
| $\widehat{\boldsymbol{p}}(\Omega(q,d))_{DSP_q}$ | Estimate of the $DPD$ obtainable from a $DSP_q$ whose $d$-grams are inserted in $\Omega(q,d)$ |
| $S_2(\boldsymbol{f})$ | Second power sum of an $RFD$ |
| $S_2(\boldsymbol{p})$ | Second power sum of a $DPD$ |
| $\widehat{S_2}(\boldsymbol{p})$ | Estimate of the second power sum of a $DPD$ |
| $H_2(\boldsymbol{f})$ | Collision entropy of an $RFD$ |
| $H_2(\boldsymbol{p})$ | Collision entropy of a $DPD$ |
| $\widehat{H_2}(\boldsymbol{p})$ | Estimated collision entropy of a $DPD$ |
| $\eta_2(\boldsymbol{f})$ | Specific collision entropy of an $RFD$ |
| $\eta_2(\boldsymbol{p})$ | Specific collision entropy of a $DPD$ |
| $\widehat{\eta_2}(\boldsymbol{p})$ | Estimated specific collision entropy of a $DPD$ |
| $\eta_2'(\boldsymbol{f})$ | Specific collision entropy rate of an $RFD$ |
| $\eta_2'(\boldsymbol{p})$ | Specific collision entropy rate of a $DPD$ |
| $\widehat{\eta_2'}(\boldsymbol{p})$ | Estimated specific collision entropy rate of a $DPD$ |

## References

1. Rényi, A. On measures of entropy and information. In Proceedings of the Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics, Berkeley, CA, USA, 1961; pp. 547–561.
2. Grassberger, P. Finite sample corrections to entropy and dimension estimates. *Physics Letters A* **1988**, *128*, 369–373. https://doi.org/10.1016/0375-9601(88)90193-4.
3. Cachin, C. Smooth Entropy and Rényi Entropy. In Proceedings of the Advances in Cryptology — EUROCRYPT '97; Fumy, W., Ed. Springer-Verlag, 5 1997, Vol. 1233, *Lecture Notes in Computer Science*, pp. 193–208.
4. Schmitt, A.; Herzel, H. Estimating the Entropy of DNA Sequences. *Journal of theoretical biology* **1997**, *188*, 369–77. https://doi.org/10.1006/jtbi.1997.0493.
5. Holste, D.; Große, I.; Herzel, H. Bayes' estimators of generalized entropies. *Journal of Physics A: Mathematical and General* **1998**, *31*, 2551–2566. https://doi.org/10.1088/0305-4470/31/11/007.
6. Strong, S.P.; Koberle, R.; de Ruyter van Steveninck, R.R.; Bialek, W. Entropy and Information in Neural Spike Trains. *Phys. Rev. Lett.* **1998**, *80*, 197–200. https://doi.org/10.1103/PhysRevLett.80.197.
7. de Wit, T.D. When do finite sample effects significantly affect entropy estimates? *The European Physical Journal B - Condensed Matter and Complex Systems* **1999**, *11*, 513–516. https://doi.org/10.1007/s100510050963.
8. Antos, A.; Kontoyiannis, I. Convergence properties of functional estimates for discrete distributions. *Random Structures & Algorithms* **2001**, *19*, 163–193. https://doi.org/10.1002/rsa.10019.
9. Nemenman, I.; Shafee, F.; Bialek, W. Entropy and inference, revisited. *In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, Advances in Neural Information Processing Systems* **2002**, *14*, 471–478.
10. Paninski, L. Estimation of entropy and mutual information. *Neural Computation* **2003**, *15*, 1191—-1253. https://doi.org/10.1162/089976603321780272.
11. Chao, A.; Shen, T.J. Non parametric estimation of Shannon's index of diversity when there are unseen species. *Environ. Ecol. Stat.* **2003**, *10*, 429–443. https://doi.org/10.1023/A:1026096204727.
12. Schürmann, T. Bias analysis in entropy estimation. *Journal of Physics A: Mathematical and General* **2004**, *37*, L295. https://doi.org/10.1088/0305-4470/37/27/L02.
13. Paninski, L. Estimating entropy on m bins given fewer than m samples. *IEEE Transactions on Information Theory* **2004**, *50*, 2200–2203. https://doi.org/10.1109/TIT.2004.833360.

14. Bonachela, J.; Hinrichsen, H.; Muñoz, M. Entropy estimates of small data sets. *Journal of Physics A: Mathematical and Theoretical* **2008**, *41*, 9. https://doi.org/10.1088/1751-8113/41/20/202001.
15. Grassberger, P. Entropy Estimates from Insufficient Samplings, 2008, [arXiv:physics.data-an/physics/0307138].
16. Hausser, J.; Strimmer, K. Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. *J. Mach. Learn. Res.* **2009**, *10*, 1469–1484.
17. Lesne, A.; Blanc, J.; Pezard, L. Entropy estimation of very short symbolic sequences. *Physical Review E* **2009**, *79*, 046208. https://doi.org/10.1103/PhysRevE.79.046208.
18. Xu, D.; Erdogmuns, D., Renyi's Entropy, Divergence and Their Nonparametric Estimators. In *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*; Springer New York: New York, NY, 2010; pp. 47–102. https://doi.org/10.1007/978-1-4419-1570-2_2.
19. Vinck, M.; Battaglia, F.; Balakirsky, V.; Vinck, A.; Pennartz, C. Estimation of the entropy based on its polynomial representation. *Phys. Rev. E* **2012**, *85*, 051139. https://doi.org/10.1103/PhysRevE.85.051139.
20. Valiant, G.; Valiant, P. Estimating the Unseen: Improved Estimators for Entropy and Other Properties. *J. ACM* **2017**, *64*. https://doi.org/10.1145/3125643.
21. Zhang, Z.; Grabchak, M. Bias Adjustment for a Nonparametric Entropy Estimator. *Entropy* **2013**, *15*, 1999–2011. https://doi.org/10.3390/e15061999.
22. Archer, E.; Park, I.; Pillow, J. Bayesian entropy estimation for countable discrete distributions. *The Journal of Machine Learning Research* **2014**, *15*, 2833–2868.
23. Li, L.; Titov, I.; Sporleder, C. Improved estimation of entropy for evaluation of word sense induction. *Computational Linguistics* **2014**, *40*, 671–685. https://doi.org/10.1162/COLI_a_00196.
24. Acharya, J.; Orlitsky, A.; Suresh, A.; Tyagi, H. The Complexity of Estimating Rényi Entropy. In Proceedings of the The twenty-sixth annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015, San Diego, CA, USA, January 4-6, 2015. SIAM, SIAM, 2015, pp. 1855–1869. https://doi.org/10.1137/1.9781611973730.124.
25. Zhang, Z.; Grabchak, M. Entropic representation and estimation of diversity indices. *Journal of Nonparametric Statistics* **2016**, *28*, 563–575. https://doi.org/10.1080/10485252.2016.1190357.
26. Acharya, J.; Orlitsky, A.; Suresh, A.; Tyagi, H. Estimating Rényi entropy of discrete distributions. *IEEE Transactions on Information Theory* **2017**, *63*, 38–56. https://doi.org/10.1109/TIT.2016.2620435.
27. de Oliveira, H.; Ospina, R. A Note on the Shannon Entropy of Short Sequences **2018**. https://doi.org/10.14209/sbrt.2018.8.
28. Berrett, T.; Samworth, R.; Yuan, M. Efficient multivariate entropy estimation via $k$-nearest neighbour distances. *The Annals of Statistics* **2019**, *47*, 288–318. https://doi.org/10.1214/18-AOS1688.
29. Verdú, S. Empirical estimation of information measures: a literature guide. *Entropy* **2019**, *21*, 720. https://doi.org/10.3390/e21080720.
30. Goldfeld, Z.; Greenewald, K.; Niles-Weed, J.; Polyanskiy, Y. Convergence of smoothed empirical measures with applications to entropy estimation. *IEEE Transactions on Information Theory* **2020**, *66*, 4368–4391. https://doi.org/10.1109/TIT.2020.2975480.
31. Contreras Rodríguez, L.; Madarro-Capó, E.; Legón-Pérez, C.; Rojas, O.; Sosa-Gómez, G. Selecting an effective entropy estimator for short sequences of bits and bytes with maximum entropy. *Entropy* **2021**, *23*. https://doi.org/10.3390/e23050561.
32. Kim, Y.; Guyot, C.; Kim, Y. On the efficient estimation of Min-entropy. *IEEE Transactions on Information Forensics and Security* **2021**, *16*, 3013–3025. https://doi.org/10.1109/TIFS.2021.3070424.
33. Grassberger, P. On Generalized Schürmann Entropy Estimators. *Entropy* **2022**, *24*. https://doi.org/10.3390/e24050680.
34. Pincus, S. Approximate entropy as a measure of system complexity. *Proc Nati.Acad.Sci.USA* **1991**, *88*, 2297–2301. https://doi.org/10.1073/pnas.88.6.2297.
35. Richman, J.S.; Moorman, J.R. Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology-Heart and Circulatory Physiology* **2000**, *278*, H2039–H2049. https://doi.org/10.1152/ajpheart.2000.278.6.H2039.
36. Manis, G.; Aktaruzzaman, M.; Sassi, R. Bubble entropy: An entropy almost free of parameters. *IEEE Transactions on Biomedical Engineering* **2017**, *64*, 2711–2718. https://doi.org/10.1109/TBME.2017.2664105.
37. Ribeiro, M.; Henriques, T.; Castro, L.; Souto, A.; Antunes, L.; Costa-Santos, C.; Teixeira, A. The entropy universe. *Entropy* **2021**, *23*. https://doi.org/10.3390/e23020222.
38. Skorski, M. Improved estimation of collision entropy in high and low-entropy regimes and applications to anomaly detection. Cryptology ePrint Archive, Paper 2016/1035, 2016.
39. Skorski, M. Towards More Efficient Rényi Entropy Estimation. *Entropy* **2023**, *25*, 185. https://doi.org/10.3390/e25020185.
40. Shannon, C. A mathematical theory of communication. *The Bell System Technical Journal* **1948**, *27*, 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.
41. Kamath, S.; Verdú, S. Estimation of entropy rate and Rényi entropy rate for Markov chains. In Proceedings of the 2016 IEEE International Symposium on Information Theory (ISIT), 2016, pp. 685–689. https://doi.org/10.1109/ISIT.2016.7541386.
42. Golshani, L.; Pasha, E.; Yari, G. Some properties of Rényi entropy and Rényi entropy rate. *Information Sciences* **2009**, *179*, 2426–2433. https://doi.org/10.1016/j.ins.2009.03.002.
43. Golshani, L.; Pasha, E. Rényi entropy rate for Gaussian processes. *Information Sciences* **2010**, *180*, 1486–1491. https://doi.org/10.1016/j.ins.2009.12.012.

44. Teixeira, A.; Matos, A.; Antunes, L. Conditional Rényi Entropies. *IEEE Transactions on Information Theory* **2012**, *58*, 4273–4277. https://doi.org/10.1109/TIT.2012.2192713.
45. Fehr, S.; Berens, S. On the Conditional Rényi Entropy. *IEEE Transactions on Information Theory* **2014**, *60*, 6801–6810. https://doi.org/10.1109/TIT.2014.2357799.
46. Packard, N.H.; Crutchfield, J.P.; Farmer, J.D.; Shaw, R.S. Geometry from a Time Series. *Phys. Rev. Lett.* **1980**, *45*, 712–716. https://doi.org/10.1103/PhysRevLett.45.712.
47. Takens, F. Detecting strange attractors in turbulence. In Proceedings of the Dynamical Systems and Turbulence, Warwick 1980: proceedings of a symposium held at the University of Warwick 1979/80. Springer, 1981-2006, pp. 366–381.