Journal of Universal Computer Science, vol. 16, no. 5 (2010), 749-794
submitted: 20/9/09, accepted: 12/2/10, appeared: 1/3/10 © J.UCS

Entropy and Higher Moments of Information[1]

Helmut Jürgensen
(Department of Computer Science, The University of Western Ontario
London, Ontario, Canada N6A 5B7)

David E. Matthews
(Department of Statistics and Actuarial Science, University of Waterloo
Waterloo, Ontario, Canada N2L 3G1)

Abstract: The entropy of a finite probability space or, equivalently, a memoryless source is the average information content of an event. The fact that entropy is an expectation suggests that it could be quite important in certain applications to take into account higher moments of information and parameters derived from these like the variance or skewness. In this paper we initiate a study of the higher moments of information for sources without memory and sources with memory. We derive properties of these moments for information defined in the sense of Shannon and indicate how these considerations can be extended to include the concepts of information in the sense of Aczél or Rényi. For memoryless sources, these concepts are immediately supported by the usual definitions of moments; for general stationary sources, let alone general sources, no such applicable framework seems to exist; on the other hand, the special properties of stationary Markov sources suggest such definitions which are both, well-motivated and mathematically meaningful.

Key Words: information, entropy, moments of information, variance of information
Category: F.1.0, F.2.0

1 Introduction

Consider a finite probability space $\mathcal{X} = (X, p)$ with set X of elementary events and probabilities $p(x)$ for $x \in X$. A small set of intuitively convincing axioms defines the *average information content of an event in* X uniquely as

$$H(\mathcal{X}) = \sum_{x \in X} p(x) \cdot \log \frac{1}{p(x)}$$

up to a constant factor $c > 0$. Here the constant c can be related to the base of the logarithm and, thus, to the unit of measurement for information[2]. To define the constant one may require – as is often done – that, for $|\, X \,| = 2$ and $p(x) = \frac{1}{2}$ for $x \in X$, $H(\mathcal{X}) = a$ for some constant $a \in \mathbb{R}_+$ as a *normalization condition;*

[1] This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada.

[2] In information theory one usually assumes that the base is 2; for the purpose of mathematical manipulations it is more convenient to use natural logarithms. For base-2 logarithms the unit of information is *bits,* for natural logarithms it is often called *nats.*

as $H(\mathcal{X}) = \log 2$, this determines the base of the logarithm. The quantity $H(\mathcal{X})$ measures information in the sense of Shannon [Sha48].

Formally, $H(\mathcal{X})$ looks like the *entropy* in thermodynamics. There are some crucial differences between the information theoretic and physical notions of entropy, however, despite this formal similarity. Moreover, there are also formal similarities to expressions for and properties of, descriptional complexity in the sense of Kolmogorov or Chaitin [Cal02] and to properties of energy; see [R8́2, Jür08] for a discussion of these and related issues. Some of these problems become apparent when one considers continuous rather than discrete probability spaces.

Entropy is often used as a statistic in decision making. An early important example is that of cryptographic security due to Shannon [Sha49]. We outline this application as it reveals several important problems.

A cryptographic system is said to achieve *perfect (key or message) secrecy* if the (key or message) equivocation conditional on the received cryptogram is large, essentially equal to the *a priori* entropy of the keys or messages, respectively. On the basis of this idea, one computes a quantity, called the *unicity distance,* for a cryptographic system; in a simplified interpretation often found in the cryptographic literature, the unicity distance is the length of encrypted messages up to which unique decryption is *impossible.*

This interpretation is not supported by the mathematics without some careful clarifications. More importantly, however, referring to something as being *impossible* in the context of statistics, even when it has a probability of 0, is rather stretching the terminology.

To clarify this point, here is an example taken from [Jür08], but originally stated in [Rob98, JR96]:

> *Consider the cryptanalysis of a message E encrypted by a perfectly secure cryptographic system (in the sense of Shannon). In such a system, the encrypted message can be the result of any possible message with the same probability. Suppose that there are* $2^{1000000}$ *messages altogether. Hence the probability of the encrypted message being E is* $2^{-1000000}$*. With some work, however, we have found that E might say that there will be a devastating terrorist attack in Berlin tomorrow at 9 a.m. The probability that our reading of the encrypted message is correct is* $2^{-1000000}$*. The entropy of the information space of correct versus incorrect reading is nearly zero. What should we do?*

As shown in [Rob98, JR96], the dilemma is not restricted to information theory, but recurs in other settings, like complexity theory or risk analysis, in which taking global measurements seems to be insufficient for an adequate description of reality as it is perceived. Some quite different recent attempts to deal with this problem include [Flü95, Lyr02, Lyr04]; whether these solve the problem is unclear. Information as defined in information theory may not describe what we intend it to mean. In particular, information may be subjective.

In the context of cryptography, this issue was raised already by us in [JM84]. We proposed to base decisions not on entropy alone, but to involve at least its variance in the evaluation process.

Formally, the entropy $H(\mathcal{X})$ is an expectation of the random variable

$$I(x) = \log \frac{1}{p(x)},$$

called the *information content* of the event $x \in X$ (see [Rén61, Rén65]). This interpretation suggests alternative definitions of information which differ from $H(\mathcal{X})$ mainly in the way the information content of an event is expressed [AD75, San87, Csi08, ESS98]. We restrict our attention to information defined as $H(\mathcal{X})$ in this paper, and only briefly outline the extension of our results to other notions of information in Section 7 below.

As an expectation, $H(\mathcal{X})$ is the first moment of the probability distribution of the random variable $I(x)$; it is natural to consider higher moments of this random variable to address the usual problems associated with the expectation. For the application of entropy in cryptography, using the variance was proposed and briefly discussed by us in [JM84]. It is surprising to note that, despite it being a natural idea in statistics, there are no studies of the higher moments of information except a very recent paper by da Fontoura Costa [FC].

In the present paper we initiate a systematic exploration of the concept of higher moments of information in the sense of Shannon. We also briefly discuss how this work can be extended to other concepts of information.

Our paper is structured as follows: In Section 2 we establish the notation and review some elementary facts from probability and information theory. We then turn to the case of finite probability spaces or memoryless sources in Section 3. We derive formulæ for the moments of information and consider their behaviours as the parameters vary. We briefly discuss the longer-term behaviour of memoryless sources in Section 4. We investigate the possibility of extending the ideas to arbitrary stationary sources in Section 5. It turns that there are serious mathematical and intuitive obstacles to this attempt. In Section 6 we show that the theory can be extended to stationary Markov sources in a meaningful way. In Section 7 we outline how our ideas generalize to other notions of information, that is, changes to the formula defining the random variable I. We conclude with general comments in Section 8.

2 Notation and Basic Notions

In this section we introduce the notation to be used and we review some basic concepts and some required background. For general results regarding information theory we refer to the books by Csiszár and Körner [CK81], Guiaşu [Gui77], MacKay [Mac07] or by Yaglom and Yaglom [YY73]. Comprehensive treatments of the axiomatic for information measures are available in the books by Aczél and Daróczy [AD75] and by Ebanks, Sahoo and Sander [ESS98]. For background about functional equations, we refer to a book by Aczél [Acz61].

We employ the usual notation for sets. Consider sets S and T: $|\,S\,|$ is the cardinality of S; 2^S is the set of all subsets of S; for S being a subset of T, a proper subset of T or not a subset of T, we write $S \subseteq T$, $S \subsetneq T$ and $S \not\subseteq T$, respectively; by $T \backslash S$ we denote the difference set, that is the set $\{t \mid t \in T, t \notin S\}$. When S is a singleton set, we often omit the set brackets, that is, we write x instead of $\{x\}$ when there is no risk of confusion.

By $\mathbb{N}$ we denote the set $\{1, 2, \ldots\}$ of positive integers; let $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. As usual, $\mathbb{Z}$ and $\mathbb{R}$ denote the sets of integers and real numbers. Let $\mathbb{R}_+ = \{r \mid r \in \mathbb{R}, r > 0\}$.

All logarithms in this paper are natural logarithms unless stated otherwise explicitly. The change of the base of a logarithm is afforded by the equation

$$\log_a x = \frac{\log_b x}{\log_b a}.$$

Thus

$$\log_2 x = \frac{\log x}{\log 2}.$$

In the context of information theory, changing the base of the logarithms amounts to changing the unit of measurement: *nats* for natural logarithms versus *bits* for logarithms with base 2.

An *alphabet* is a finite non-empty set. Throughout this paper, let Σ be an alphabet with $|\,\Sigma\,| > 1$. Elements of an alphabet are called *letters* or *symbols.* We assume that a and b are distinct symbols in Σ. A word over Σ is a finite sequence of symbols from Σ. By Σ^* one denotes the set of all words over Σ including the empty word ε; let $\Sigma^+ = \Sigma^* \setminus \varepsilon$. With the concatenation of words as multiplication, Σ^* is a monoid and Σ^+ is a semigroup, freely generated by Σ.

A finite probability space $\mathcal{X}$ is a construct $\mathcal{X} = (X, p^X)$ where X is a finite non-empty set, the set of elementary events, and p^X is a probability measure on 2^X. For $x \in X$, the value of $I^X(x) = \log \frac{1}{p^X(x)}$ is the *information content* of x and

$$H(\mathcal{X}) = \mathsf{E}\, I^X(x) = \sum_{x \in X} I^X(x) \cdot p^X(x)$$

is the *entropy* or the *information content* of $\mathcal{X}$. We omit the superscript X referring to the probability space in question when there is no risk of confusion.

Note that $I(x) \cdot p(x) \to 0$ as $p(x) \to 0$. One has $0 \le H(\mathcal{X}) \le \log|\,X\,|$. In particular, $H(\mathcal{X}) = 0$ if and only if $p(x) = 1$ for exactly one $x \in X$ and, hence,

$p(x') = 0$ for all $x' \in X \setminus x$; on the other hand, $H(\mathcal{X}) = \log |\, X \,|$ if and only if $p(x) = |\, X \,|^{-1}$ for all $x \in X$.

This definition of information was proposed by Shannon [Sha48]. Occasionally, we need to distinguish this notion of information from other notions of information; in such cases we write I_{Shannon} and H_{Shannon} instead of I and H.

Let X and Y be finite non-empty sets, let $Z = X \times Y$ and let $\mathcal{Z} = (Z, p^Z)$ be a probability space. From $\mathcal{Z}$ one obtains probabilities p^X and p^Y on X and Y and conditional probabilities as usual: for $x \in X$ and $y \in Y$,

$$\begin{aligned} p^X(x) &= p^Z(x \times Y), & p^{X|y}(x \mid y) &= \frac{p^Z(x,y)}{p^Y(y)}, \\ p^Y(y) &= p^Z(X \times y), \text{ and } & p^{Y|x}(y \mid x) &= \frac{p^Z(x,y)}{p^X(x)} \end{aligned}$$

with the appropriate exceptions to avoid division by 0. One obtains the probability spaces $\mathcal{X} = (X, p^X)$, $\mathcal{Y} = (Y, p^Y)$, $\mathcal{X}|y = (X, p^{X|y})$ and $\mathcal{Y}|x = (Y, p^{Y|x})$ and their entropies. In particular,

$$H(\mathcal{X} \mid y) = \sum_{x \in X} I^{X|y}(x) \cdot p^{X|y}(x \mid y) = \sum_{x \in X} \log \frac{1}{p^{X|y}(x \mid y)} \cdot p^{X|y}(x \mid y)$$

and similarly

$$H(\mathcal{Y} \mid x) = \sum_{y \in Y} I^{Y|x}(y) \cdot p^{Y|x}(y \mid x) = \sum_{y \in Y} \log \frac{1}{p^{Y|x}(y \mid x)} \cdot p^{Y|x}(y \mid x).$$

The *conditional entropy* (of $\mathcal{X}$ given $\mathcal{Y}$) is defined as

$$H(\mathcal{X} \mid \mathcal{Y}) = \sum_{y \in Y} H(\mathcal{X} \mid y) \cdot p^Y(y),$$

that is, as the expectation $\mathsf{E}\, H(\mathcal{X} \mid y)$. With $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ one has

$$\begin{aligned} H(\mathcal{Z}) = H(\mathcal{X} \times \mathcal{Y}) &= H(\mathcal{Y}) + H(\mathcal{X} \mid \mathcal{Y}) \\ &= H(\mathcal{X}) + H(\mathcal{Y} \mid \mathcal{X}) \leq H(\mathcal{X}) + H(\mathcal{Y}) \end{aligned}$$

with equality if and only if $\mathcal{X}$ and $\mathcal{Y}$ are independent. We refer to this property as the *additivity property* of H.

The quantity $I(\mathcal{X} : \mathcal{Y}) = H(\mathcal{Y}) - H(\mathcal{Y} \mid \mathcal{X})$ is the *information* of $\mathcal{X}$ about $\mathcal{Y}$. As

$$\begin{aligned} I(\mathcal{X} : \mathcal{Y}) &= H(\mathcal{Y}) - H(\mathcal{Y} \mid \mathcal{X}) = H(\mathcal{Y}) - H(\mathcal{X} \times \mathcal{Y}) + H(\mathcal{X}) \\ &= H(\mathcal{X}) - H(\mathcal{X} \mid \mathcal{Y}) = I(\mathcal{Y} : \mathcal{X}) \end{aligned}$$

this quantity is also called the *mutual information.*

For a probability space $\mathcal{X} = (X, p)$ and $k \in \mathbb{N}$, we consider the kth *power* $\mathcal{X}^k = (X^k, p^k)$ of $\mathcal{X}$ where X^k is the set of k-tuples

$$\overrightarrow{(x)}_k = (x_1, x_2, \ldots, x_k)$$

of elements $x_1, x_2, \ldots, x_k$ in X and

$$p^k\left(\overrightarrow{(x)}_k\right) = \prod_{i=1}^{k} p(x_i).$$

One has $H(\mathcal{X}^k) = k \cdot H(\mathcal{X})$.

Let Σ be an alphabet as above. A *message* is a mapping $\mu : \mathbb{Z} \to \Sigma$. We consider $i \in \mathbb{Z}$ as a moment in time; then $\mu(i)$ is the symbol in μ generated at time i. We often write a message μ as a sequence or a part of a sequence in the form

$$\mu = \ldots \sigma_{-2}\sigma_{-1}\sigma_0\sigma_1\sigma_2 \ldots$$

where $\sigma_i = \mu(i)$ and where we omit separating commas. In the theories of codes and languages, such sequences are called *bi-infinite words* or *ζ-words* (see [JK97] for additional background information).

For $t \in \mathbb{N}_0$, let $[\,t\,] = \{1, 2, \ldots, t\}$. In particular, $[\,0\,] = \emptyset$. A finite strictly increasing sequence $T = (i_1, i_2, \ldots, i_t)$ of integers with $t \in \mathbb{N}_0$ is called a *sequence of time instances.* Thus $i_1, i_2, \ldots, i_t \in \mathbb{Z}$ and $i_1 < i_2 < \cdots < i_t$; note that T is the empty sequence ( ) when $t = 0$. The sequence T can be considered as a strictly increasing mapping of $[\,t\,]$ into $\mathbb{Z}$, that is, $T(j) = i_j$ for $j \in [\,t\,]$. For $j \in [\,t\,]$, let $T \ominus j = (i_1, \ldots, i_{j-1}, i_{j+1}, \ldots, i_t)$, that is

$$T \ominus j(k) = \begin{cases} T(k), & \text{if } k < j, \\ T(k+1), & \text{if } k \geq j, \end{cases}$$

for $k \in \{1, 2, \ldots, t-1\}$. For $h \in \mathbb{Z}$, $T + h$ is the sequence T *shifted* by h steps, that is, the sequence $(i_1 + h, i_2 + h, \ldots, i_t + h)$. Let $\mathfrak{T}$ be the set of sequences of time instances.

For $T \in \mathfrak{T}$ and $\overrightarrow{(\sigma)}_t \in \Sigma^t$, the set

$$C_T\left(\overrightarrow{(\sigma)}_t\right) = \{\mu \mid \mu : \mathbb{Z} \to \Sigma \text{ with } \mu(T(j)) = \sigma_j \text{ for } j \in [\,t\,]\}$$

is called a *cylindre set.* Let

$$\mathcal{C}_T = \left\{ C_T\left(\overrightarrow{(\sigma)}_t\right) \mid \overrightarrow{(\sigma)}_t \in \Sigma^t \right\}.$$

Each set $\mathcal{C}_T$ is finite.

A *source* is a construct $\mathfrak{S} = (\{\mathcal{S}_T \mid T \in \mathfrak{T}\}, \Sigma)$ with the following two properties:

1. For all $T \in \mathfrak{T}$, $\mathcal{S}_T = (\mathcal{C}_T, p_T)$ is a probability space.
2. For all $T \in \mathfrak{T}$, for all $j \in [\,t\,]$ and for all $\overrightarrow{(\sigma)}_{t-1} \in \Sigma^{t-1}$,

$$p_{T \ominus j}\left(\overrightarrow{(\sigma)}_{t-1}\right) = \sum_{\sigma \in \Sigma} p_T\left(\sigma_1, \ldots, \sigma_{j-1}, \sigma, \sigma_j, \ldots, \sigma_{t-1}\right).$$

We refer to property (2) as the *consistency condition.*

A source $\mathfrak{S} = (\{\mathcal{S}_T \mid T \in \mathfrak{T}\}, \Sigma)$ is said to be *stationary* if and only if, for all $T \in \mathfrak{T}$, $p_{T+1} = p_T$.

When $\mathfrak{S}$ is stationary, one has $p_{T+h} = p_T$ for any shift $h \in \mathbb{Z}$. A stationary source is, therefore, defined completely by the probability spaces $\mathcal{S}_{[\,t\,]}$ with $t \in \mathbb{N}$. For a stationary source, one can give a convincing definition of entropy because of the following well-known result.

Theorem 1 *Let $\mathfrak{S} = (\{\mathcal{S}_T \mid T \in \mathfrak{T}\}, \Sigma)$ be a stationary source. The limits*

$$\lim_{t\to\infty} \frac{H\left(\mathcal{S}_{[\,t\,]}\right)}{t}$$

and

$$\lim_{t\to\infty} H\left(\mathcal{S}_{t+1} \mid \mathcal{S}_{[\,t\,]}\right)$$

exist and are equal.

For a proof see, for instance, [CK81, Gui77]. Note that the theorem only holds for the information measure in the sense of Shannon and related ones.

Because of Theorem 1 one defines the entropy of a stationary source $\mathfrak{S}$ as $H(\mathfrak{S}) = \lim_{t\to\infty} H(\mathcal{S}_{t+1} \mid \mathcal{S}_{[\,t\,]})$. Intuitively, $H(\mathfrak{S})$ is the average information content of an output symbol of the source after the source has been observed for a long time or, alternatively, as the average information content of an output symbol in a very long message.

The assumption of stationarity allows one to simplify the notation for cylindre sets. For $T = (1, 2, \ldots, t)$, a message in $C_T\left(\overrightarrow{(\sigma)}_t\right)$ can be described by a finite word $\overrightarrow{\sigma} = \sigma_1\sigma_2\cdots\sigma_t$ instead of the t-tuple $\overrightarrow{(\sigma)}_t$. Thus, instead of $p_T(\overrightarrow{(\sigma)}_t)$ we simply write $p(\overrightarrow{\sigma})$. The notation $\overrightarrow{\sigma}$ implies that σ_i is the ith symbol in this word.

A stationary source $\mathfrak{S}$ is called a *Markov source* if it satisfies the following conditions[3]:

[3] We do not consider non-stationary Markov sources nor Markov sources with a memory (or order) greater than 1 in this paper; hence, by a ‘Markov source’ we mean a stationary Markov source of order 1. Moreover, unless stated otherwise, we assume that the initial distribution of such a source is steady-state.

1. For all $t \in \mathbb{N}$ with $t > 1$ and all $\sigma_1, \sigma_2, \ldots, \sigma_t \in \Sigma$,

$$p_{[\,t\,]}\left(\sigma_t \mid \sigma_1\sigma_2\cdots\sigma_{t-1}\right) = p_{(t-1,t)}\left(\sigma_t \mid \sigma_{t-1}\right).$$

2. For all $i \in \mathbb{N}_0$ and all $\sigma \in \Sigma$, $p_i(\sigma) = p_0(\sigma)$.

A stationary Markov source can be represented conveniently by a matrix and a vector as follows. Let P be the square matrix indexed by $\Sigma \times \Sigma$ such that the entry p_{σ_1,σ_2} of P is the probability $p(\sigma_2 \mid \sigma_1)$, that is, the probability of σ_2 being the next symbol when σ_1 has just been output by the source. For $n \in \mathbb{N}_0$, let $\pi(n)$ be the vector describing the distribution of the output symbol at time n, that is, $\pi(n)_\sigma = p_n(\sigma)$. Then $\pi(n) = \pi(0)P^n$. The second condition, called *steady-state condition,* is equivalent to $\pi(0)P = \pi(0)$. If $\mathfrak{S}$ is a stationary Markov source, one has

$$H(\mathfrak{S}) = \sum_{\sigma_0,\sigma_1 \in \Sigma} p_0(\sigma_0) \cdot p(\sigma_1 \mid \sigma_0) \cdot \log \frac{1}{p(\sigma_1 \mid \sigma_0)}.$$

Note that the steady-state condition is crucial for the proof of this equality.

A source $\mathfrak{S}$ is said to be memoryless if it is a Markov source with a defining matrix in which all rows are equal. Thus, a memoryless source is completely defined by a probability space $\mathcal{S} = (\Sigma, p)$. The entropy $H(\mathfrak{S})$ of a memoryless source $\mathfrak{S}$ is equal to $H(\mathcal{S})$.

3 Information Moments of Memoryless Sources

In this section we consider only sources without memory. Such a source is a finite probability space $\mathcal{S} = (\Sigma, p)$ in which Σ is the output alphabet and, for $\sigma \in \Sigma$, $p(\sigma) = \mathsf{Prob}\,(\sigma)$ is the probability of output σ. The information content

$$I(\sigma) = \log \frac{1}{p(\sigma)} = -\log p(\sigma)$$

of σ is a random variable with $H(\mathcal{S}) = \mathsf{E}\,I(\sigma)$ as its expectation. We assume that $|\,\Sigma\,| > 1$.

For the definition of the moments of information and for their analysis one needs the following generalization of a well-known fact about the information function.

Lemma 2 *For all* $i \in \mathbb{N}$,

$$\lim_{x \to +0} x\,(\log x)^i = 0.$$

PROOF. Applying L'Hôpital's rule j times with $1 \leq j \leq i$ yields

$$\lim_{x\to+0} x\,(\log x)^i = \lim_{x\to+0} (-1)^j\, i(i-1)\cdots(i-j+1)\,x\,(\log x)^{i-j}\,.$$

For $j = i$ this results in

$$\lim_{x\to+0} x\,(\log x)^i = \lim_{x\to+0} (-1)^i\; i\,!\; x = 0.$$

□

Remark 3 *Using Lemma 2, let* $0\,(\log 0)^i = 0$ *for all* $i \in \mathbb{N}$*.*

Let $i \in \mathbb{N}$. The i-th moment of the information is computed as

$$M^{(i)}(I) = \mathsf{E}\, I(\sigma)^i = \sum_{\sigma\in\Sigma} p(\sigma)\cdot(-\log p(\sigma))^i\,.$$

In particular, $M^{(1)}(I) = H(\mathcal{S})$, the entropy of $\mathcal{S}$; moreover, $M^{(2)}(I) - H(\mathcal{S})^2 = \mathsf{Var}\, I$ is the variance of I.

Graphs of the first six moments of I are shown in Figure 1 for the case of $|\,\Sigma\,| = 2$. In Figure 2 we show graphs of the first six moments of I for the case of $|\,\Sigma\,| = 3$. In Figure 3 we show graphs of the variance of I for the cases of $|\,\Sigma\,| = 2$ and $|\,\Sigma\,| = 3$ (see also [JM84]).

Let $n = |\,\Sigma\,| > 1$ and $\Sigma = \{\sigma_1, \sigma_2, \ldots, \sigma_n\}$. For $i = 1, 2, \ldots, n$, let $p_i = p(\sigma_i)$. The n-tuple $(p_1, p_2, \ldots, p_n)$ is an element of the compact space

$$\Delta_n = \left\{ (q_1, q_2, \ldots, q_n) \mid 0 \leq q_i \leq 1 \text{ for } i = 1, 2, \ldots, n,\ \sum_{i=1}^{n} q_i = 1 \right\}.$$

The moments of information are independent of the names of the symbols in Σ and may thus be considered as functions mapping Δ_n into $\mathbb{R}$. When it is convenient we emphasize this interpretation by writing $H_n(\pi)$ for the entropy, $I(p_i)$ instead of $I(\sigma_i)$ and $M_n^{(i)}(\pi)$ for the i-th moment, where $\pi = (p_1, p_2, \ldots, p_n) \in \Delta_n$. As Δ_n is compact, $M_n^{(i)}(\pi)$ has maximal and minimal values on Δ_n.

By Lemma 2 and $-x\log x > 0$ for $0 < x \leq 1$, one computes that $M^{(i)}(\pi) > 0$ when $\pi = (p_1, p_2, \ldots, p_n)$ is such that $p_i \neq 0$ and $p_j \neq 0$ for some i, j with $1 \leq i < j \leq n$; otherwise $M_n^{(i)}(\pi) = 0$. The latter are the minima. When $p_1 = p_2 = \cdots = p_n$, that is, $p_i = 1/n$ for all i, then $M_n^{(i)}(\pi) = (\log n)^i$.

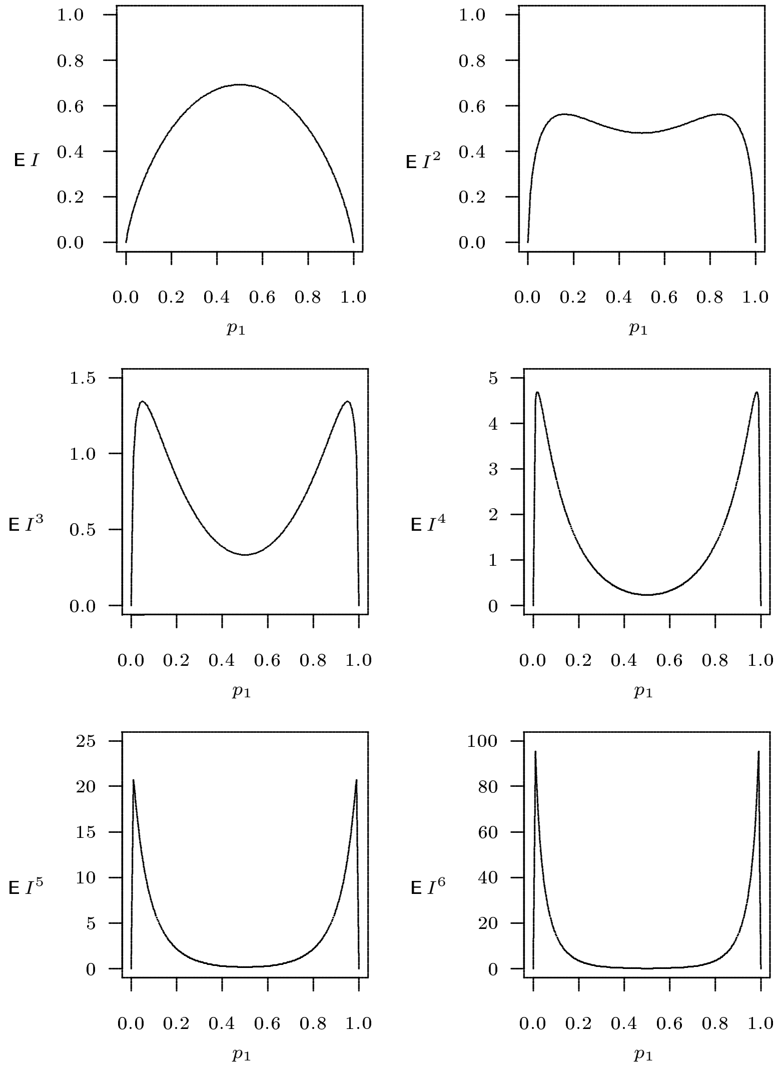

Figure 1: The first six moments of I about the origin, when $|\,\Sigma\,| = 2$. The probability distribution is $(p_1, 1 - p_1)$. Note the scale changes on the vertical axes. The logarithm base is e.

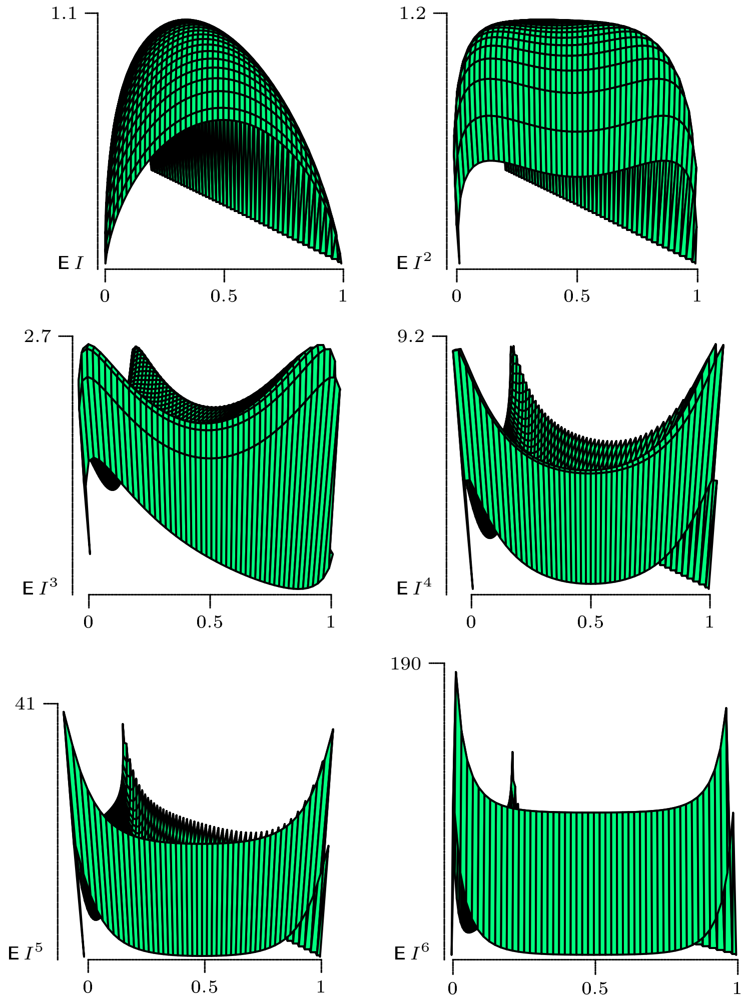

Figure 2: The first six moments of I about the origin, when $|\Sigma| = 3$. The logarithm base is e.

As $\sum_{j=1}^{n} p_j = 1$, $p_n = 1 - \sum_{j=1}^{n-1} p_j$. Therefore, $M_n^{(i)}$ can be considered as a function of the $n-1$ variables $p_1, p_2, \ldots, p_{n-1}$ in the closed interval $[\,0:1\,]$ in the form

$$\begin{aligned} M_n^{(i)}(\pi) &= \sum_{j=1}^{n-1} p_j \cdot \bigl(I(p_j)\bigr)^i + p_n \cdot \bigl(I(p_n)\bigr)^i \\ &= \sum_{j=1}^{n-1} p_j \cdot \bigl(I(p_j)\bigr)^i + \bigl(1 - \sum_{k=1}^{n-1} p_k\bigr) \cdot \bigl(I\bigl(1 - \sum_{k=1}^{n-1} p_k\bigr)\bigr)^i. \end{aligned}$$

We now determine the first derivatives of $M_n^{(i)}$ with respect to p_j where $j = 1, 2, \ldots, n-1$.

Lemma 4 *For $j = 1, 2, \ldots, n-1$, one has*

$$\begin{aligned} \frac{\partial}{\partial p_j} M_n^{(i)}(\pi) = \bigl(I(p_j)\bigr)^i + i \cdot p_j \cdot \bigl(I(p_j)\bigr)^{i-1} \cdot \frac{\partial}{\partial p_j} I(p_j) \\ -\bigl(I(p_n)\bigr)^i + i \cdot p_n \cdot \bigl(I(p_n)\bigr)^{i-1} \cdot \frac{\partial}{\partial p_j} I(p_n). \end{aligned}$$

For $I(p) = -\log p$, this specializes to

$$\begin{aligned} \frac{\partial}{\partial p_j} M_n^{(i)}(\pi) = (-\log p_j)^i - i \cdot (-\log p_j)^{i-1} \\ - (-\log p_n)^i + i \cdot (-\log p_n)^{i-1}. \end{aligned}$$

PROOF. For $j = 1, 2, \ldots, n-1$ one has

$$\frac{\partial}{\partial p_j} M_n^{(i)}(\pi) = \frac{\partial}{\partial p_j} \Bigl(p_j \cdot \bigl(I(p_j)\bigr)^i\Bigr) + \frac{\partial}{\partial p_j} \Bigl(p_n \cdot \bigl(I(p_n)\bigr)^i\Bigr).$$

One computes

$$\begin{aligned} \frac{\partial}{\partial p_j} \Bigl(p_j \cdot \bigl(I(p_j)\bigr)^i\Bigr) &= \bigl(I(p_j)\bigr)^i + i \cdot p_j \cdot \bigl(I(p_j)\bigr)^{i-1} \cdot \frac{\partial}{\partial p_j} I(p_j) \\ &= (-\log p_j)^i - i \cdot (-\log p_j)^{i-1}. \end{aligned}$$

With p_n as above,

$$\frac{\partial}{\partial p_j} p_n = -1$$

and, using $I(p_n) = -\log p_n$,

$$\frac{\partial}{\partial p_j} I(p_n) = \frac{1}{p_n}.$$

One has

$$\begin{aligned} \frac{\partial}{\partial p_j} \Bigl(p_n \cdot \bigl(I(p_n)\bigr)^i\Bigr) &= -\bigl(I(p_n)\bigr)^i + i \cdot p_n \cdot \bigl(I(p_n)\bigr)^{i-1} \cdot \frac{\partial}{\partial p_j} I(p_n) \\ &= -(-\log p_n)^i + i \cdot (-\log p_n)^{i-1}. \end{aligned}$$

This completes the proof. □

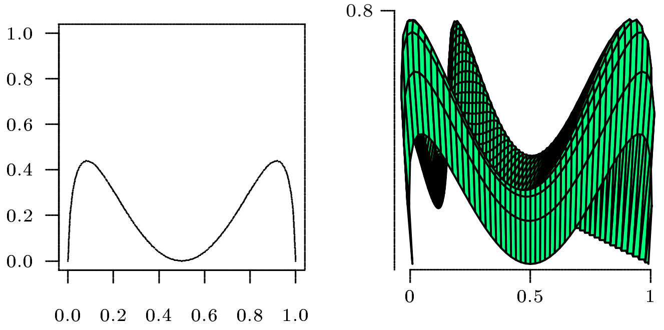

Figure 3: The variance of I when $|\,\Sigma\,|$ is 2 or 3. The logarithm base is e.

Remark 5 *When logarithms are taken with respect to base B instead of base e the derivatives in Lemma 4 have the form*

$$\begin{aligned}\frac{\partial}{\partial p_j} M_n^{(i)}(\pi) &= \Big(\left(-\log_B p_j\right)^i - \left(-\log_B p_n\right)^i \Big) \\ &\qquad - \frac{i}{\ln B} \cdot \Big(\left(-\log_B p_j\right)^{i-1} - \left(-\log_B p_n\right)^{i-1} \Big) \\ &= \left(-\log_B p_j + \log_B p_n\right) \\ &\qquad \cdot \left(\sum_{k=0}^{i-1} \left(-\log_B p_j\right)^k \left(-\log_B p_n\right)^{i-1-k} \right. \\ &\qquad\qquad \left. - \frac{i}{\ln B} \cdot \sum_{k=0}^{i-2} \left(-\log_B p_j\right)^k \left(-\log_B p_n\right)^{i-2-k} \right).\end{aligned}$$

PROOF. In the proof of the lemma one computes

$$\frac{\partial}{\partial p_j} I(p_j) = \frac{-1}{p_j \ln B} \quad \text{and} \quad \frac{\partial}{\partial p_j} I(p_n) = \frac{1}{p_n \ln B}.$$

One verifies the factorization by multiplication. □

The factorization of the partial derivatives of $M_n^{(i)}$ provided above is used on several occasions in the sequel.

We list a few important properties of $M_n^{(i)}$.

Proposition 6 *Let $n, i \in \mathbb{N}$ with $n > 1$. The function $M_n^{(i)}(\pi) : \Delta_n \to \mathbb{R}$ has the following properties:*

1. $M_n^{(i)}(\pi) = \sum_{j=1}^{n} p_j \left(-\log p_j\right)^i$ *where* $\pi = (p_1, p_2, \ldots, p_n) \in \Delta_n$.

2. $M_n^{(i)}$ *is continuous.*

3. $0 \leq M_n^{(i)}(\pi)$.

4. *$M_n^{(i)}(\pi) = 0$ if and only if, for some j with $1 \le k \le n$, $p_j = 1$ and, hence, $p_k = 0$ for all k with $1 \le k \le n$ and $k \ne j$.*

5. *For $j = 1, 2, \dots, n-1$,*
$$\frac{\partial}{\partial p_j} M_n^{(i)}(\pi) = 0$$
if and only if
$$(-\log p_j)^i - i \cdot (-\log p_j)^{i-1} = (-\log p_n)^i - i \cdot (-\log p_n)^{i-1}.$$
In particular, this is true when $p_1 = p_2 = \dots = p_n = \frac{1}{n}$, and in this case $M_n^{(i)}(\pi) = (\log n)^i$.

6. *The entropy $M_n^{(1)}(\pi)$ has a unique maximum when $p_1 = p_2 = \dots = p_n = \frac{1}{n}$, and in this case $M_n^{(1)}(\pi) = \log n$. It has no other extrema.*

PROOF. The first two statements are obviously true. When $0 < p_j < 1$ then $p_j\,(-\log p_j)^i > 0$. This proves the third and fourth statements. By Lemma 4, the partial derivatives are 0 if and only if

$$\begin{aligned}\big(I(p_j)\big)^i + i \cdot p_j \cdot \big(I(p_j)\big)^{i-1} \cdot \frac{\partial}{\partial p_j} I(p_j) \\ = \big(I(p_n)\big)^i - i \cdot p_n \cdot \big(I(p_n)\big)^{i-1} \cdot \frac{\partial}{\partial p_j} I(p_n).\end{aligned}$$

With $I(p_j) = -\log p_j$ and $I(p_n) = -\log p_n$, one obtains the formula stated. The last statement is well-known. □

Remark 7 *By Remark 5, the condition in Proposition 6(5) is*

$$(-\log_B p_j)^i - \frac{i}{\ln B} \cdot (-\log_B p_j)^{i-1} = (-\log_B p_n)^i - \frac{i}{\ln B} \cdot (-\log_B p_n)^{i-1}.$$

when one uses logarithms with base B.

Some of the statements in Lemma 4 and Proposition 6 do not depend on the specific choice of the information function I. These are applicable also to other information measures as discussed briefly below in Section 7.

We now determine the extrema of the second moment $M_n^{(2)}$ as n varies.

Proposition 8 *The following statements hold true for $M_n^{(2)}$ with base-B logarithms, where $B > 1$ and $c = \frac{2}{\ln B}$:*

1. *For all n, the derivatives $\frac{\partial}{\partial p_j} M_n^{(2)}(\pi)$ for $j = 1, 2, \dots, n-1$ are equal to 0 for $\pi = (\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n})$.*

2. *For $n = 2$, $M_n^{(2)}(\pi)$ has two maxima and one minimum as follows:*

   (a) *The maxima are obtained for $\pi = (p_1, 1 - p_1)$ with*

   $$p_1 = \frac{1}{2} \pm \frac{1}{2e}\sqrt{e^2 - 4}$$

   for every base B.

   (b) *The minimum is obtained for $\pi = (\frac{1}{2}, \frac{1}{2})$.*

3. *For $n \geq 3$, $M_n^{(2)}(\pi)$ has a single extremum which is the unique maximum at $\pi = (\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n})$.*

PROOF. All calculations in this proof are carried out using a base B for the logarithms where $B > 1$.

For $i = 2$, the formulæ in Proposition 6(5) have the form

$$a_j^2 - 2a_j - b^2 + 2b = 0$$

where $j = 1, 2, \ldots, n-1$, $a_j = -\log p_j$, $b = -\log p_n$ and $p_n = 1 - \sum_{k=1}^{n-1} p_k$. By Remark 5, when we use base B instead of base e, the formulæ turn into

$$a_j^2 - \frac{2}{\ln B} a_j - b^2 + \frac{2}{\ln B} b = 0$$

with $a_j = -\log_B p_j$ and $b = -\log_B p_n$.

To allow for a change in the base of the logarithm, we write c instead of $2/\ln B$ in this proof. Thus, the equations have the general form

$$a_j^2 - ca_j - b^2 + cb = 0.$$

Note that $B^c = e^2$. This can be factorized into

$$(a_j - b)(a_j + b - c) = 0.$$

For a solution, one of the two factors is equal to 0.

Note that

$$a_j - b = -\log p_j + \log p_n = \log \frac{p_n}{p_j} = 0$$

if and only if $p_j = p_n$. In particular, $a_j - b = 0$ for all j if and only if $p_1 = p_2 = \cdots = p_n = 1/n$. This solution is called *trivial* in the remainder of the proof.

All other solutions of the system of equations are obtained from various combinations of which factors are equal to 0. Because of the symmetry, it suffices to distinguish the cases according to how many factors of the form $(a_j - b)$ are equal to 0. Permutations of the co-ordinates will yield all solutions from these.

First, consider $n = 2$. The only non-trivial solutions can be the solutions of the equation $a_1 + b - c = 0$. As $p_2 = 1 - p_1$, one has

$$a_1 + b - c = -\log p_1 - \log(1 - p_1) - c = -\log p_1(1 - p_1) - c,$$

hence

$$p_1^2 - p_1 = -B^{-c}$$

using $\log = \log_B$ with $B = e$ not excluded. Thus

$$p_1 = \frac{1}{2} \pm \sqrt{\frac{1}{4} - \frac{1}{B^c}} = \frac{1}{2} \pm \frac{1}{2B^{c/2}}\sqrt{B^c - 4}.$$

Both roots are real as $e^2 = B^c \geq 4$.

Next, consider $n = 3$. Then $p_3 = 1 - p_1 - p_2$. To obtain the non-trivial solutions we need to consider the following three cases:

1. $a_1 - b = 0$ and $a_2 + b - c = 0$.
2. $a_1 + b - c = 0$ and $a_2 - b = 0$.
3. $a_1 + b - c = 0$ and $a_2 + b - c = 0$.

Case 1: Then $p_1 = 1 - p_1 - p_2$, hence $p_1 = (1 - p_2)/2$. The second equation yields

$$\begin{aligned} a_2 + b - c &= -\log p_2 - \log\left(\frac{1}{2} - \frac{p_2}{2}\right) - c \\ &= -\log\frac{p_2(1 - p_2)}{2} - c = 0. \end{aligned}$$

Hence, equivalently,

$$p_2^2 - p_2 + \frac{2}{B^c} = 0$$

with

$$p_2 = \frac{1}{2} \pm \sqrt{\frac{1}{4} - \frac{2}{B^c}} = \frac{1}{2} \pm \frac{1}{2B^{c/2}}\sqrt{B^c - 8}.$$

As $e^2 - 8 = B^c - 8 < 0$, these solutions are not real-valued. Case 2 is similar to Case 1.

Case 3: We consider the equations

$$a_1 + b - c = -\log p_1 - \log(1 - p_1 - p_2) - c = 0$$

and

$$a_2 + b - c = -\log p_2 - \log(1 - p_1 - p_2) - c = 0.$$

Solving the first one for p_1 yields

$$p_1 = \frac{1 - p_2}{2} \pm \sqrt{\frac{(1 - p_2)^2}{4} - \frac{1}{B^c}}.$$

Substituting this into the second equation results in

$$p_2^2 - \left(1 - \frac{1-p_2}{2} \mp \sqrt{\frac{(1-p_2)^2}{4} - \frac{1}{B^c}}\right) \cdot p_2 + \frac{1}{B^c} = 0.$$

One solves this to find

$$p_2 = \frac{1}{4} \pm \frac{1}{4B^{c/2}}\sqrt{B^c - 8}.$$

As above this solution is not real-valued.

Finally, we consider $n > 3$. As in the case of $n = 3$, we consider the two equations

$$(a_1 - b)(a_1 + b - c) = 0 \text{ and } (a_2 - b)(a_2 + b - c) = 0,$$

where, however,

$$b = -\log(1 - p_1 - p_2 - r)$$

for some r with $0 \leq r \leq 1$. The idea is that $r = \sum_{k=3}^{n-1} p_k$. We distinguish the three cases as above.

Case 1: One computes that

$$p_1 = \frac{1 - p_2 - r}{2}$$

and

$$p_2 = \frac{1-r}{2} \pm \frac{1}{2B^{c/2}}\sqrt{(1-r)^2 B^c - 8}.$$

As $0 \leq r \leq 1$, by

$$(1-r)^2 B^c - 8 = ((1-r)e)^2 - 8 \leq e^2 - 8 < 0$$

p_2 is not real-valued.

Case 2 is analogous to Case 1.

Case 3: One computes

$$p_1 = \frac{1 - p_2 - r}{2} \pm \frac{1}{2B^{c/2}}\sqrt{(1 - p_2 - r)^2 B^c - 4}$$

and

$$p_2 = \frac{1-r}{4} \pm \frac{1}{4B^{c/2}}\sqrt{(1-r)^2 B^c - 8}.$$

As above, p_2 is not real-valued.

This completes the proof. □

Statements like the ones in Proposition 8 are independent of the choice of the base of the logarithms. This needed to be proved, but was not completely unexpected. Indeed, features, like extrema, of a function intended to describe a physical phenomenon should not depend on the unit of measurement.

The significant difference in the shapes of the second moment between the cases $n = 2$ and $n > 2$, as stated in Proposition 8 was a surprise to us. The issue is clarified, at least in part, by the following results concerning the variance $\mathsf{Var}_n\, I$. One observes a similar phenomenon for the third moment. The details for that case are stated in Proposition 11.

Proposition 9 *The variance* $\mathsf{Var}_n\, I$ *has the following properties:*

1. $\mathsf{Var}_n\, I(\pi)$ *is a continuous function of* $\pi \in \Delta_n$ *into* $\mathbb{R}_+ \cup 0$.

2. $\mathsf{Var}_n\, I(\pi) = 0$ *if and only if* π *satisfies the following condition: Let* $k \in \mathbb{N}$ *be the number of components* p_j *of* π *with* $0 < p_j$*; then, for* $j = 1, 2, \ldots, n$*, if* $p_j \neq 0$ *then* $p_j = 1/k$.

3. $\mathsf{Var}_2\, I(\pi)$ *has exactly two maxima assuming the base* B *of the logarithms satisfies* $B > 1$.

4. *Assume that the base* B *of the logarithms satisfies* $B > 1$*. There are* $2^n - 1$ *distinct values of* $\pi = (p_1, p_2, \ldots, p_n)$ *for which the gradient of* $\mathsf{Var}_n\, I(\pi)$ *is equal to the* 0*-vector; these values are as follows:*

   – $p_1 = p_2 = \cdots = p_n = \frac{1}{n}$*. This corresponds to the unique global minimum of* $\mathsf{Var}_n\, I(\pi)$.

   – *There is an integer* k*, such that* $1 \leq k \leq n-1$*, a set* $K \subseteq \{1, 2, \ldots, n\}$ *with* $|\,K\,| = k$ *and a unique real number* x *with* $0 < x < \frac{1}{n}$ *and* $x < \frac{1}{2k}$ *such that*

$$p_j = \begin{cases} x, & \textit{if } j \in K, \\ \frac{1-kx}{n-k}, & \textit{if } j \notin K. \end{cases}$$

PROOF. The first statement follows from $\mathsf{Var}_n\, I = \mathsf{E}\,(I - \mathsf{E}\,I)^2 = \mathsf{E}\,I^2 - (\mathsf{E}\,I)^2$ and the continuity of the moments.

When π satisfies the condition of the second statement, then $M_n^{(1)}(\pi) = \log k$ and $M_n^{(2)}(\pi) = (\log k)^2$, hence $\mathsf{Var}_n\, I(\pi) = 0$. To prove the converse, suppose, without loss of generality, that the k strictly positive components of π are $p_1, \ldots, p_k$. Since

$$\mathsf{Var}_n\, I(\pi) = \sum_{j=1}^{k} p_j \left(-\log p_j - M_n^{(1)}(\pi)\right)^2 \geq 0$$

with equality occurring only when $-\log p_j = M_n^{(1)}(\pi)$ for $j = 1, 2, \ldots, k$, the result follows immediately.

We turn to the proof of (3). Assume that $n = 2$. One proves that

$$\mathsf{Var}_2\, I(\pi) = p_1(1-p_1)\left(\log_B\left(\frac{p_1}{1-p_1}\right)\right)^2 .$$

The first derivative of $\mathsf{Var}_2\, I(\pi)$ with respect to p_1 is

$$\frac{\partial}{\partial p_1}\mathsf{Var}_2\, I(\pi) = \log_B\left(\frac{p_1}{1-p_1}\right)\left((1-2p_1)\log_B\left(\frac{p_1}{1-p_1}\right)+c\right)$$

where, as before, $c = \frac{2}{\ln B}$. Thus, local extrema occur when $p_1 = 1 - p_1 = \frac{1}{2}$, which corresponds to minimum variance, or when

$$(1-2p_1)\log_B\left(\frac{p_1}{1-p_1}\right)+c = 0.$$

Let

$$f(p_1) = (1-2p_1)\log_B\left(\frac{p_1}{1-p_1}\right)+c.$$

The following facts concerning $f(p_1)$ establish that $f(p_1)$ has exactly two real roots, $\bar{p}_1$ and $\bar{p}_2$, such that $0 < \bar{p}_1 < \frac{1}{2} < \bar{p}_2 < 1$, which are symmetric with respect to $\frac{1}{2}$. The function $f(p_1)$ is continuous on the open interval $(0, 1)$; for $p_1 > 0$, $\lim_{p_1\to 0} f(p_1) < 0$; for $p_1 < 1$, $\lim_{p_1\to 1} f(p_1) < 0$; moreover, when $B > 1$, $\lim_{p_1\to 1/2} f(p_1) = c > 0$. Thus, by the intermediate-value theorem, $f(p_1)$ has at least two real roots $\bar{p}_1$ and $\bar{p}_2$ with $0 < \bar{p}_1 < \frac{1}{2} < \bar{p}_2 < 1$. As the first derivative

$$\frac{\partial}{\partial p_1} f(p_1) = -2\log_B\left(\frac{p_1}{1-p_1}\right) + \frac{(1-2p_1)}{\ln B\; p_1(1-p_1)}$$

is positive for $0 < p_1 < \frac{1}{2}$ and negative for $\frac{1}{2} < p_1 < 1$, there are no more roots in that interval.

As $\mathsf{Var}_2\, I(\pi)$ is symmetric with respect to $\frac{1}{2}$, one has $\frac{1}{2} - \bar{p}_1 = \bar{p}_2 - \frac{1}{2}$.

To compute $\bar{p}_1$ and $\bar{p}_2$, let

$$B^x = \frac{p_1}{1-p_1}.$$

Thus

$$p_1 = \frac{B^x}{1+B^x}$$

and

$$1 - 2p_1 = \frac{1-B^x}{1+B^x}.$$

Then

$$f(p_1) = \frac{1-B^x}{1+B^x}\cdot x + c.$$

With x_1 and x_2 the two real roots of

$$x(1-B^x)+c(1+B^x)=0,$$

one has

$$\bar{p}_1=\frac{B^{x_1}}{1+B^{x_1}} \text{ and } \bar{p}_2=\frac{B^{x_2}}{1+B^{x_2}}$$

or vice versa. This completes the proof of (3).

For (4) we consider the partial derivatives

$$\begin{aligned}\frac{\partial}{\partial p_j}\mathsf{Var}_n\, I(\pi) &= \frac{\partial}{\partial p_j}M_n^{(2)}(\pi)-\frac{\partial}{\partial p_j}\left(M_n^{(1)}(\pi)\right)^2\\ &= (I(p_j))^2+2\cdot p_j\cdot I(p_j)\cdot\frac{\partial}{\partial p_j}I(p_j)\\ &\qquad -(I(p_n))^2+2\cdot p_n\cdot I(p_n)\cdot\frac{\partial}{\partial p_j}I(p_n)\\ &\qquad -2\cdot M_n^{(1)}(\pi)\cdot\\ &\qquad\quad \Bigl(I(p_j)+p_j\cdot\frac{\partial}{\partial p_j}I(p_j)\\ &\qquad\quad -I(p_n)+p_n\cdot\frac{\partial}{\partial p_j}I(p_n)\Bigr)\end{aligned}$$

for $j=1,2,\ldots,n-1$. For the rest of this proof, let $\log=\log_B$ where $B>1$. With $I(x)=-\log x$, $c=\frac{2}{\ln B}$ and $H=H_n(\pi)=M_n^{(1)}(\pi)$, the entropy, one finds

$$\begin{aligned}\frac{\partial}{\partial p_j}\mathsf{Var}_n\, I(\pi) &= (-\log p_j)^2-c(-\log p_j)-(-\log p_n)^2+c(-\log p_n)\\ &\qquad -2H\cdot(-\log p_j)-2H\cdot\log p_n\\ &= (-\log p_j+\log p_n)\cdot\\ &\qquad (-\log p_j-\log p_n-c-2H)\,.\end{aligned}$$

Without loss of generality we assume that $p_j>0$ for all j. It follows that

$$\frac{\partial}{\partial p_j}\mathsf{Var}_n\, I(\pi)=0$$

if and only if

$$p_j=p_n \text{ or } \log p_j+\log p_n+c+2H=0.$$

Thus, if all partial derivatives are equal to 0 then there is a k with $1\le k\le n$ such that, for k values of j one has $p_j=p_n$ while for the other $n-k$ values of j one has $p_j\neq p_n$, but $\log p_j+\log p_n+c+2H=0$. There are 2^n-1 possible choices of the k values of j. Without loss of generality, we assume that $p_{n-k+1}=p_{n-k+2}=\cdots=p_n$ and that $p_j\neq p_n$ for $j=1,2,\ldots,n-k$.

Let $p_n = x$ where $0 < x < 1$; then

$$\sum_{j>n-k} p_j = kx \text{ and } \sum_{j\leq n-k} p_j = 1 - kx,$$

hence $x \leq \frac{1}{k}$. For $j \leq n-k$ one has $p_j \neq x$; define

$$\begin{aligned} f_{k,j}(x) &= \log p_j + \log p_n + c + 2H \\ &= \log p_j + \log x + c \\ &\qquad + 2 \sum_{i\leq n-k} p_i \log \frac{1}{p_i} + 2kx \log \frac{1}{x} = 0. \end{aligned}$$

When $k = n$ then $p_j = x = \frac{1}{n}$, and the variance is zero.

For the rest of the proof we assume that $1 \leq k \leq n-1$. When $k = n-1$ then $p_1 = 1 - (n-1)x$, hence

$$\begin{aligned} f_{n-1,1}(x) &= \log p_1 + \log p_n + c + 2H \\ &= \log(1-(n-1)x) + \log x + c \\ &\qquad + 2\left((1-(n-1)x)\log\frac{1}{1-(n-1)x} + (n-1)x\log\frac{1}{x}\right) \\ &= (1-2(n-1)x)\log\frac{x}{1-(n-1)x} + c. \end{aligned}$$

Intuition and the shape of the graphs in Figure 3 suggest that for all other values of k one has

$$p_j = \frac{1-kx}{n-k}, \text{ but } p_j \neq x,$$

for $j = 1, 2, \ldots, n-k$. In that case, $p_j \neq x$ implies $x \neq \frac{1}{n}$; moreover,

$$x < \frac{1}{n} \text{ if and only if } \frac{1-kx}{n-k} > \frac{1}{n}.$$

For these assumptions, define

$$\begin{aligned} f_k(x) &= \log\frac{1-kx}{n-k} + \log x + c \\ &\qquad + 2\left((1-kx)\log\frac{n-k}{1-kx} + kx\log\frac{1}{x}\right) \\ &= (1-2kx)\log\frac{(n-k)x}{1-kx} + c. \end{aligned}$$

The function f_k has the following properties:

1. $\lim_{x\to 0} f_k(x) = -\infty$ for $x > 0$;
2. $f_k(x)$ is continuous for $0 < x < \frac{1}{n}$;

3. $f_k(x) = c > 0$ for $x = \frac{1}{2k}$ or $x = \frac{1}{n}$ when $B > 1$;
4. The derivative of f_k is given by

$$\begin{aligned} f_k'(x) &= -2k \log \frac{(n-k)x}{1-kx} \\ &\quad + \frac{(1-kx)(1-2kx)}{(n-k)x} \cdot \left(\frac{n-k}{1-kx} + \frac{k(n-k)x}{(1-kx)^2} \right) \\ &= -2k \log \frac{(n-k)x}{1-kx} \\ &\quad + \frac{(1-2kx)}{(n-k)x} \cdot \left(n-k + \frac{k(n-k)x}{1-kx} \right) \\ &= -2k \log \frac{(n-k)x}{1-kx} + \frac{1-2kx}{x(1-kx)}. \end{aligned}$$

For $0 < x < \frac{1}{n}$ one has

$$0 \leq \frac{(n-k)x}{1-kx} < 1,$$

hence

$$-2k \log \frac{(n-k)x}{1-kx} > 0.$$

For $0 < x < \frac{1}{2k}$ one has

$$\frac{1-2kx}{x(1-kx)} > 0.$$

Thus the derivative of f_k is strictly positive for

$$0 < x < \min\left(\frac{1}{2k}, \frac{1}{n}\right).$$

By the intermediate-value theorem, f_k has exactly one root x_0 satisfying

$$0 < x_0 < \min\left(\frac{1}{2k}, \frac{1}{n}\right).$$

There are $2^n - 1$ non-empty sets $K \subseteq \{1, 2, \ldots, n\}$. Let $k = |K|$. For $k = n$ one gets the global minimum. With $1 \leq k < n$, our assumptions imply

$$p_j = \begin{cases} x_0, & \text{if } j \in K, \\ \frac{1-kx_0}{n-k}, & \text{if } j \notin K. \end{cases}$$

As $x_0 < \frac{1}{n}$, one has

$$\frac{1-kx_0}{n-k} > \frac{1}{n}.$$

Therefore, the k-element subsets of $\{1, 2, \ldots, n\}$ are in one-to-one correspondence with the vectors π defined by the conditions above. □

For Proposition 9(3) and $n = 2$, by numerical calculations we can determine that $x_1 \approx -2.399356$ and $x_2 \approx 2.399356$, so that the corresponding values of p_1 at which the maximal variance of approximately 0.4392288 occurs are approximately 0.08322182 and 0.91677818, respectively. Here the symbols x_1, x_2 and p_1 refer to the symbols used in the proof above.

The statement of Proposition 9(4) does not exclude the possibility that there could be further values of π for which the gradient of the variance is the 0-vector. The graphs in Figure 3 suggest that the n values of π resulting from $k = n - 1$ correspond to local maxima; moreover, we believe that, in general, these are the only local maxima of $\mathsf{Var}_n\, I$. For $n = 3$, Proposition 9(4) establishes at least $2^3 - 1 = 7$ distinct values of π for which the gradient of $\mathsf{Var}_3\, I(\pi)$ is the 0-vector as visible in Figure 3: (1) the global minimum at the point with all co-ordinates equal to $\frac{1}{3}$ and the variance equal to 0; (2) the three local maxima at the points with two co-ordinates approximately equal to 0.06165176, the remaining co-ordinate approximately equal to 0.8766965 and the variance approximately equal to 0.7618022; (3) the three saddle points with two co-ordinates approximately equal to 0.4744505, the remaining co-ordinate approximately equal to 0.051099 and the variance approximately equal to 0.2407779. For $n = 4$, by Proposition 9(4) we find 15 points as follows:

1. all co-ordinates are equal to $\frac{1}{4}$ and the variance is equal to 0;
2. four points for $k = 1$, three co-ordinates of which are approximately equal to 0.3209829 with the remaining co-ordinate approximately equal to 0.03705117 and the variance being approximately equal to 0.16663207.
3. six points for $k = 2$, two co-ordinates of which are approximately equal to 0.458394 with the remaining two co-ordinates approximately equal to 0.04160604 and the variance approximately equal to 0.4392288.
4. four points for $k = 3$, three co-ordinates of which are approximately equal to 0.04950273 with the remaining co-ordinate approximately equal to 0.8514918 and the variance approximately equal to 1.023491.

Numerical calculations suggest: the points according to (2) and (3) correspond to local minima; the points according to (4) correspond to local maxima. Clearly, the point of (1) corresponds to a global minimum.

At this moment we do not see a pattern emerging from the calculations which might support a plausible conjecture beyond the following claims based on the graphs shown in Figures 1–3:

Conjecture 10 *Let $i, n \in \mathbb{N}$ with $n \geq 2$ and $\pi \in \Delta_n$.*

1. *Let $n = 2$. For $i > 1$, $M_n^{(i)}(\pi)$ has two maxima at*
$$\pi_{1,2}^{(i)} = \left(\frac{1}{2} \pm x^{(i)}, \frac{1}{2} \mp x^{(i)}\right)$$
with $0 < x^{(i)} < \frac{1}{2}$. Moreover,
$$x^{(2)} = \frac{1}{2e}\sqrt{e^2 - 4},$$
and $x^{(i)}$ is strictly increasing as $i \to \infty$ with
$$\lim_{i \to \infty} x^{(i)} = \frac{1}{2}.$$

2. *Let $n = 2$ and let $\pi_{1,2}^{(i)}$ and $x^{(i)}$ be as above. Then, for $i \geq 2$*
$$M_n^{(i)}(\pi_1^{(i)}) = M_n^{(i)}(\pi_2^{(i)}) > (\log 2)^i$$
is strictly increasing and unbounded as $i \to \infty$.

3. *Let $n > 2$. For $i > 2$, $M_n^{(i)}(\pi)$ has n local maxima.*

4. *For $n \geq 2$, as $i \to \infty$, the maximal values of $M_n^{(i)}$ are strictly increasing and unbounded.*

The value of $x^{(2)}$ has been obtained in Proposition 8. For $n = 2$ the graphs suggest not only that $M_2^{(i)}(\pi)$ has the shape of a trough but also that its inner walls become extremely steep as $i \to \infty$. In the limit the shape seems to be that of the symbol $\sqcup$ with walls of infinite height. For the claim (3) of Conjecture 10, we have a partial answer as follows.

Proposition 11 *Let $n \in \mathbb{N}$, $n \geq 2$ and let $B > 1$ be the base of the logarithms. The points $\pi = (p_1, p_2, \ldots, p_n) \in \Delta_n$ at which the gradient of $M_n^{(3)}(\pi)$ is the 0-vector have the following properties:*

1. *For all n, the gradient of $M_n^{(3)}(\pi)$ is the 0-vector when $p_1 = p_2 = \cdots = p_n = \frac{1}{n}$.*

2. *For $n = 2$, $M_n^{(3)}(\pi)$ has a global minimum at $p_1 = p_2 = \frac{1}{2}$ and two local maxima at*
$$\pi_{1,2}^{(3)} = \left(\frac{1}{2} \pm x^{(3)}, \frac{1}{2} \mp x^{(3)}\right)$$
with $0 < x^{(3)} < \frac{1}{2}$. Moreover,
$$\frac{9}{20} < x^{(3)} < \frac{19}{42}.$$
These are the only points at which the gradient of $M_2^{(3)}(\pi)$ is the 0-vector.

3. *For $n = 3$, $M_n^{(3)}(\pi)$ has a global minimum at $p_1 = p_2 = p_3 = \frac{1}{3}$, three local maxima and three saddle points. They are at points π the components of which satisfy the following conditions:*

   (a) *two of the components are equal to some value x and the remaining component is equal to $1 - 2x$;*

   (b) *for the local maxima,*
   $$\frac{1}{21} < x < \frac{1}{20}.$$

   (c) *for the saddle points, $0.4708 < x < 0.4709$.*

   These are the only points at which the gradient of $M_3^{(3)}(\pi)$ is the 0-vector.

4. *For all $n \in \mathbb{N}$ with $n \geq 2$, if the gradient of $M_n^{(3)}(\pi)$ is the 0-vector then exactly one of the following two conditions is met:*

   (a) *π is such that $p_1 = p_2 = \cdots = p_n = \frac{1}{n}$.*

   (b) *There is an $x \in \mathbb{R}$ with $0 < x < 1$ and a set $K \subseteq \{1, 2, \ldots, n\}$ with $k = |K| = n - 1$ such that the following statements hold true:*

      i. *$0 < x < \frac{1}{k}$, $p_j = x$ for all $j \in K$ and $p_j \neq x$ for $j \notin K$;*

      ii. *for $j \notin K$,*
      $$\ln p_j = -\frac{3 + \ln x}{2} - \frac{\sqrt{3}}{2} \cdot \sqrt{4 - (1 + \ln x)^2};$$

      iii. *for at least one $j \in \{1, 2, \ldots, n\}$ one has $p_j < \frac{1}{n}$;*

      iv. *for all $j \in \{1, 2, \ldots, n\}$ one has $p_j > e^{-3}$.*

5. *For $n > 20$ the gradient of $M_n^{(3)}(\pi)$ is the 0-vector if and only if $p_1 = p_2 = \cdots = p_n = \frac{1}{n}$. At this point $M_n^{(3)}(\pi)$ is a global maximum.*

PROOF. Using Remark 5 with $\log = \log_B$ and $d = 1/\ln B$, one has

$$\begin{aligned}
\frac{\partial}{\partial p_j} M_n^{(3)}(\pi) &= ((-\log p_j) - (-\log p_n)) \\
&\quad \cdot \big((-\log p_j)^2 + (-\log p_j)(-\log p_n) + (-\log p_n)^2 \\
&\qquad - 3d \cdot ((-\log p_j) + (-\log p_n))\big) \\
&= (-\log p_j + \log p_n) \\
&\quad \cdot \big((\log p_j p_n)^2 + 3d \log p_j p_n - \log p_j \log p_n\big).
\end{aligned}$$

for $j = 1, 2, \ldots, n-1$. As in the proof of Proposition 9(4), for the gradient to be the 0-vector, there is a k with $1 \leq k \leq n$ and a set $K \subseteq \{1, 2, \ldots, n\}$ with $|\,K\,| = k$ and $n \in K$ such that $p_j = p_n$ for $j \in K$ and $p_j \neq p_n$, but

$$(\log p_j p_n)^2 + 3d \log p_j p_n - \log p_j \log p_n = 0$$

for $j \notin K$.

When $k = n$, then $p_1 = p_2 = \cdots = p_n = \frac{1}{n}$ and $M_n^{(3)} = (\log n)^3$. This proves (1).

For the rest of the proof let $k \leq n-1$ and, without loss of generality, let $K = \{n-k+1, n-k+2, \ldots, n\}$. As in the proof of Proposition 9(4), let $p_n = x$ with $0 < x < 1$. As $k < n$, $x < \frac{1}{k}$. For $j \leq n-k$ one has $p_j \neq x$ and $\sum_{j=1}^{n-k} p_j = 1 - kx$. It follows that there is a $j \in \{1, 2, \ldots, n\}$ with $p_j < \frac{1}{n}$.

Define the function

$$h(x,y) = (\ln(xy))^2 + 3\ln(xy) - \ln x \cdot \ln y.$$

With x as above, $p_j = y$ for $j \notin K$ only if $h(x,y) = 0$. Note that $h(x,y)$ is given by the condition above when $B = e$. One computes

$$h(x,y) = (\ln x)^2 + (\ln y)^2 + \ln x \cdot \ln y + 3\ln x + 3\ln y.$$

Solving this for $\ln x$ and $\ln y$ yields

$$\ln x = -\frac{3 + \ln y}{2} \pm \frac{\sqrt{3}}{2} \cdot \sqrt{4 - (1 + \ln y)^2}$$

and

$$\ln y = -\frac{3 + \ln x}{2} \pm \frac{\sqrt{3}}{2} \cdot \sqrt{4 - (1 + \ln x)^2}.$$

For $\ln y$ to be real-valued, it is necessary that

$$4 - (1 + \ln x)^2 \geq 0.$$

This holds if and only if

$$-2 \leq 1 + \ln x \leq 2.$$

As $x < 1$, $1 + \ln x \leq 2$ is always true. Thus, x has to satisfy $e^{-3} \leq x$. Similarly, also $e^{-3} \leq y$. As $p_j < \frac{1}{n}$ for some j, one has $e^{-3} \leq \frac{1}{n}$, that is, $n \leq e^3$. This implies $n \leq 20$. Hence, for $n > 20$, there is no solution of $h(x,y)$ satisfying the conditions. This proves (5).

When $e^{-3} = y$ then $x = 1$; similarly, $e^{-3} = x$ implies $y = 1$. Therefore, for all values $p_j > e^{-3}$.

The fact that $\ln x < 0$ implies $1 + \ln x < 1$. Hence

$$2 - (1 + \ln x) - (1 + \ln x)^2 > 0$$

which implies

$$\frac{\sqrt{3}}{2}\cdot\sqrt{4-(1+\ln x)^2} > \frac{3+\ln x}{2},$$

where both sides are positive because of $x > e^{-3}$. Therefore, the case of

$$\ln y = -\frac{3+\ln x}{2} + \frac{\sqrt{3}}{2}\cdot\sqrt{4-(1+\ln x)^2}$$

is impossible as it implies that $\ln y \geq 0$. By symmetry this also applies to $\ln x$. As a consequence,

$$\ln y = -\frac{3+\ln x}{2} - \frac{\sqrt{3}}{2}\cdot\sqrt{4-(1+\ln x)^2}.$$

This implies that, for all $j \notin K$, there is only a single possible value y for p_j and

$$y = \frac{1-kx}{n-k}.$$

As this result is independent of n it follows that $k = n-1$. This proves (4).

We now turn to the cases of $n = 2$ and $n = 3$. For $n = 2$ we are looking for x and y with $0 < x < \frac{1}{2} < y < 1$ and $x + y = 1$ such that $h(x, y) = 0$. Define $g(x) = h(x, 1-x)$. The function g has the following properties:

- $\lim_{x\to 0} g(x) = +\infty$ for $x > 0$;
- $\lim_{x\to 1} g(x) = +\infty$ for $x < 1$;
- $g(\frac{1}{20}) = g(\frac{19}{20}) < 0$;
- $g(x)$ is continuous for $0 < x < 1$;
- the derivative of g is negative for $0 < x < \frac{1}{2}$ and positive for $\frac{1}{2} < x < 1$.

By the intermediate-value theorem the function $g(x)$ has exactly two zeroes $x_{1,2}$ which satisfy $0 < x_1 < \frac{1}{20}$ and $\frac{19}{20} < x_2 = 1 - x_1 < 1$. Numerical calculations show that $x_1 > \frac{1}{21}$ and $x_2 < \frac{20}{21}$. One computes

$$\begin{aligned} M_2^{(3)}\left(\frac{1}{20},\frac{19}{20}\right) &\approx 1.34437 > 0.33302 \\ &\approx (\ln 2)^3 = M_2^{(3)}\left(\frac{1}{2},\frac{1}{2}\right). \end{aligned}$$

Therefore, at (x_1, x_2) and (x_2, x_1) the moment $M_2^{(3)}(\pi)$ is maximal and at $(\frac{1}{2},\frac{1}{2})$ it is minimal. This proves Statement (2) with $x^{(3)} = \frac{1}{2} - x_1$.

For $n = 3$ we need to consider the situation when exactly two of the probabilities are the same, say equal to x, and the remaining probability y is equal to $1 - 2x$. Define $g(x) = h(x, 1-2x)$. The function g has the following properties:

- $\lim_{x\to 0} g(x) = +\infty$ for $x > 0$;
- $\lim_{x\to\frac{1}{2}} g(x) = +\infty$ for $x < \frac{1}{2}$;
- $g(\frac{1}{20}) = g(\frac{9}{20}) < 0$;
- $g(x)$ is continuous for $0 < x < \frac{1}{2}$;
- there is an x' with $\frac{1}{4} \leq x' < \frac{1}{2}$ such that the derivative of g is negative for $0 < x < x'$ and positive for $x' < x < \frac{1}{2}$.

To apply the intermediate value theorem and to find good approximations of the roots (if they exist) we need to find x_1 and x_2 such that

$$0 < x_1 < x' < x_2 < \frac{1}{2},$$

$g(x_1) < 0$ and $g(x_2) < 0$.

Observe that, if

$$(\ln x(1-2x))^2 + 3\ln x(1-2x) < 0,$$

then also $g(x) < 0$ as $\ln x \cdot \ln(1-2x) > 0$. In more general terms, when

$$(\ln x(1-mx))^2 + 3\ln x(1-mx) < 0,$$

then also $h(x, 1-mx) < 0$ where $m \in \mathbb{N}$ and $0 < mx < 1$. One has

$$(\ln x(1-mx))^2 + 3\ln x(1-mx) < 0,$$

if and only if

$$x(1-mx) > e^{-3}.$$

As solutions of $x(1-mx) \approx e^{-3}$ one computes

$$x \approx \frac{1}{2m} \cdot \left(1 \pm \sqrt{1-4me^{-3}}\right).$$

This is real-valued for $4m \leq 20$ and complex for $4m \geq 21$.

For the case of $n = 3$ we need to consider $m = 2$. One computes

$$x \approx \frac{1}{4} \cdot \left(1 \pm \sqrt{1-8e^{-3}}\right) \approx \begin{cases} 0.443924, \\ 0.056076. \end{cases}$$

The smaller of these values is approximately equal to $\frac{1}{18}$. Numerical calculation shows that

$$h\left(\frac{1}{20}, \frac{18}{20}\right) < 0$$

and

$$h\left(\frac{1}{21}, \frac{19}{21}\right) > 0,$$

which shows there is a solution x_1 with

$$\frac{1}{21} < x_1 < \frac{1}{20}.$$

Approximating x_1 by $\frac{1}{20}$ one computes

$$M_3^{(3)}\left(\frac{1}{20}, \frac{1}{20}, \frac{18}{20}\right) \approx 2.689546154$$

while

$$(\ln 3)^3 \approx 1.32596896.$$

Using $\frac{9}{20}$, which is what one would try from $\frac{1}{2} - \frac{1}{20}$ by symmetry, as an approximation of x_2 is insufficient. Numerical calculation shows that $0.4708 < x_2 < 0.4709$. Using $x_2 \approx 0.47085$ as an approximation, one finds that the third moment is approximately equal to 1.740889037.

Thus, in summary: $M_3^{(3)}(\pi)$ has a global minimum when all components of π are equal to $\frac{1}{3}$; it has three local maxima when two of the components of π are equal to x_1 and the remaining one is equal to $1 - 2x_1$; it has three saddle points when two of the components of π are equal to x_2 and the remaining one is equal to $1 - 2x_2$. □

Proposition 11 asserts the existence of 3 distinct points at which the gradient of $M_n^{(3)}$ is the 0-vector for $n = 2$; of 7 such points for $n = 3$; of just 1 such point for $n > 20$. For $4 \leq n \leq 20$ we only have the necessary condition (4). Numerical calculations lead us to believe that there is only a single point with gradient equal to the 0-vector even for values of n which are much smaller than 20.

We extract some observations from the proof of Proposition 11.

Remark 12 *Let $m \in \mathbb{N}$ and $x \in \mathbb{R}$ with $0 < mx < 1$.*

1. $(\ln x(1-mx))^2 + 3\ln x(1-mx) = 0$ *has a real-valued solution for the product $x(1-mx)$ if and only if $m \leq 5$. In that case there are the two solutions*

$$\frac{1}{2m} \cdot \left(1 \pm \sqrt{1 - 4me^{-3}}\right).$$

2. *If they exist, solutions to $h(x, 1-mx) = 0$ for $m \geq 2$ are nearly symmetrical around $\frac{1}{2m}$. The value of the smaller one is approximately equal to*

$$\frac{1}{2m} \cdot \left(1 - \sqrt{1 - 4me^{-3}}\right).$$

PROOF. The first statement was proved above. The second statement follows from the fact that, as x being small implies that $1 - mx$ is nearly equal to 1, hence also $|\ln x \cdot \ln(1 - mx)|$ is small. □

4 Memoryless Sources with Words as Outputs

We now turn to a different set of problems. We consider the case of $\mathcal{S}^k = (\Sigma^k, p^k)$ with $k > 1$; this probability space models the behaviour of a memoryless source, when one considers output words of length k rather than single output symbols. It is well known that $H(\mathcal{S}^k) = kH(\mathcal{S})$. This is a consequence of the additivity of H for products of independent probability spaces. Note that additivity holds true when information is measured in the sense of Shannon, but can be violated when other information measures are used.

To derive the higher moments of information, we consider the following more general situation: Consider k independent finite probability spaces $\mathcal{X}_j = (X^{(j)}, p^{(j)})$ for $j = 1, 2, \ldots, k$ with $k \geq 1$, and let

$$\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_k = (X, p)$$

where

$$X = X_1 \times X_2 \times \cdots \times X_k$$

and

$$p(x_1, x_2, \ldots, x_k) = \prod_{j=1}^{k} p^{(j)}(x_j)$$

where $(x_1, x_2, \ldots, x_k) \in X$. For information in the sense of Shannon,

$$I(x_1, x_2, \ldots, x_k) = \sum_{j=1}^{k} I^{(j)}(x_j) = \sum_{j=1}^{k} \left(-\log p^{(j)}(x_j)\right).$$

Hence, for $i \in \mathbb{N}$,

$$\begin{aligned} I(x_1, x_2, \ldots, x_k)^i &= \left(\sum_{j=1}^{k} I^{(j)}(x_j)\right)^i = \left(\sum_{j=1}^{k} -\log p^{(j)}(x_j)\right)^i \\ &= \sum_{\substack{i_1, i_2, \ldots, i_k \\ i_1+i_2+\cdots+i_k=i}} \binom{k}{i_1, i_2, \ldots, i_k} \prod_{j=1}^{k} \left(I^{(j)}(x_j)\right)^{i_j} \\ &= \sum_{\substack{i_1, i_2, \ldots, i_k \\ i_1+i_2+\cdots+i_k=i}} \binom{k}{i_1, i_2, \ldots, i_k} \prod_{j=1}^{k} \left(-\log p^{(j)}(x_j)\right)^{i_j}. \end{aligned}$$

This proves:

Proposition 13 *Let $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_k$ be a product of k independent probability spaces as above. The i-th moment of the information is*

$$M^{(i)}(I) = \sum_{\substack{i_1, i_2, \ldots, i_k \\ i_1+i_2+\cdots+i_k=i}} \binom{k}{i_1, i_2, \ldots, i_k} \prod_{j=1}^{k} M^{(i_j)}(I^{(j)}).$$

For the specific case of $i = 1$ one finds the well-known additivity of H. For $i = 2$ and $k = 2$, one gets

$$M^{(2)}(I) = M^{(2)}(I^{(1)}) + 2H(\mathcal{X}_1)H(\mathcal{X}_2) + M^{(2)}(I^{(2)}).$$

When $\mathcal{X}_1 = \mathcal{X}_2$ this simplifies to

$$M^{(2)}(I) = 2M^{(2)}(I^{(1)}) + 2H(\mathcal{X}_1)^2$$

and

$$\mathsf{Var}\, I = 2M^{(2)}(I^{(1)}) - 2H(\mathcal{X}_1)^2 = 2\mathsf{Var}\, I^{(1)}.$$

Applying these results to the k-symbol output behaviour of a memoryless source $\mathcal{S}$ yields the following formulæ for the moments and variance of information in the sense of Shannon.

Corollary 14 *Let $\mathcal{S} = (\Sigma, p)$ be a memoryless source, let $k, i \in \mathbb{N}$. Let I^k denote the information measure in the sense of Shannon for $\mathcal{S}^k$. Then*

$$M^{(i)}(I^k) = \sum_{\substack{i_1, i_2, \ldots, i_k \\ i_1 + i_2 + \cdots + i_k = i}} \binom{k}{i_1, i_2, \ldots, i_k} \prod_{j=1}^{k} M^{(i_j)}(I^1)$$

and

$$\mathsf{Var}\, I^k = M^{(2)}(I^k) - \left(M^{(1)}(I^k)\right)^2 = k\mathsf{Var}\, I^1.$$

The variance of I^k is unbounded as $k \to \infty$ when $\mathsf{Var}\, I^1 > 0$.

5 Information Moments of General Sources

Recall that a source

$$\mathfrak{S} = (\{\mathcal{S}_T \mid T \in \mathfrak{T}\}, \Sigma)$$

with alphabet Σ is a family of finite probability spaces connected by a consistency condition. Just using the formal definitions, one can define the moments of information and derived parameters like variance and skewness for each of the spaces $\mathcal{S}_T$ with $T \in \mathfrak{T}$. Having all these values is not helpful in general. Intuitively, the following quantities promise better insights:

1. average per-symbol parameter values;

2. conditional parameter values.

We explain what we mean by this only in words as the technicalities are not relevant for the rest of this paper. By parameters, we mean moments of information, variance and similar quantities derived from the moments. In particular, entropy, the first moment, is one of the parameters under consideration. Let P be any such parameter.

1. By an *average per-symbol value* of P, we mean the value of
$$\frac{P(\mathcal{S}_T)}{|\,T\,|}$$
for $T \in \mathfrak{T}$.

2. By a *conditional value* of P, we mean the value of
$$P(\mathcal{S}_{T'} \mid \mathcal{S}_T)$$
for $T', T \in \mathfrak{T}$.

In particular, one would consider long-term values, that is, these values when $|\,T\,|$ is very large and $|\,T'\,|$ is small.

For $P = H$, the entropy, it is helpful to assume that $\mathfrak{S}$ is stationary. A stationary source is completely specified by the probability spaces $\mathcal{S}_{[\,t\,]}$ with $t \in \mathbb{N}$. By the additivity property of H and the consistency condition one obtains expressions for $H(\mathcal{S}_T)$ for any $T \in \mathfrak{T}$ and the connection

$$H\left(\mathcal{S}_{[\,t+1\,]}\right) = H\left(\mathcal{S}_{[\,t\,]}\right) + H\left(\mathcal{S}_{t+1} \mid \mathcal{S}_{[\,t\,]}\right) \leq H\left(\mathcal{S}_{[\,t\,]}\right) + H\left(\mathcal{S}_{t+1}\right)$$

between the conditional and unconditional entropies. Note that this connection is valid for the information measure H in the sense of Shannon, but not for information measures in general (see [AD75, ESS98]). It is crucial in the proof of Theorem 1 to show that the limits

$$\lim_{t\to\infty} \frac{H\left(\mathcal{S}_{[\,t\,]}\right)}{t} \quad \text{and} \quad \lim_{t\to\infty} H\left(\mathcal{S}_{t+1} \mid \mathcal{S}_{[\,t\,]}\right)$$

exist and are equal. For the information measure $H = H_{\text{Shannon}}$, this limiting behaviour of both the average per-symbol value and the conditional value justifies defining the entropy of $\mathfrak{S}$ as

$$H(\mathfrak{S}) = \lim_{t\to\infty} H(\mathcal{S}_{t+1} \mid \mathcal{S}_{[\,t\,]}).$$

This idea does not generalize to other information measures.

We continue assuming that $\mathfrak{S}$ is stationary. For $T \in \mathfrak{T}$, let I^T denote the random variable representing the information content of events of $\mathcal{S}_T$ with respect to some information measure I. With $T' \in \mathfrak{T}$, let $I^{(T'|T)}$ be the random variable representing the information content of events of $(\mathcal{S}_{T'} \mid \mathcal{S}_T)$ with respect to I.

We consider the moments $M^{(i)}(I^T)$, for $i \in \mathbb{N}$, and the variance $\mathsf{Var}\, I^T$. In particular, when $I = I_{\text{Shannon}}$, then $M^{(1)}(I^T)$ is the entropy of the space $\mathcal{S}_T$. In view of Theorem 1, it seems natural to consider the limiting behaviour of

$$\frac{M^{(i)}(I^{[\,t\,]})}{t} \quad \text{and} \quad \frac{\mathsf{Var}\, I^{[\,t\,]}}{t}$$

and of

$$M^{(i)}(I^{(t+1|[t])}) \text{ and } \mathsf{Var}\, I^{(t+1|[t])}$$

as $t \to \infty$. Our attempts to prove the existence and equality of the corresponding limits as suggested by Theorem 1 have been unsuccessful so far. In the example below we show that the proof method of Theorem 1 is not likely to apply as, when $I = I_{\text{Shannon}}$, not even $\mathsf{Var}\, I^{[2]}$ nor $M^{(2)}(I^{[2]})$ satisfy the additivity condition.

We consider the following stationary Markov source $\mathfrak{S}$ with alphabet $\Sigma = \{1,2\}$. As noted before, such a source can be represented by a matrix P and a vector π such that $\pi P = \pi$. Let

$$\pi = \left(\tfrac{1}{2}\ \tfrac{1}{2}\right) \text{ and } P = \begin{pmatrix} p_{1,1} & p_{1,2} \\ \\ p_{2,1} & p_{2,2} \end{pmatrix}$$

such that

$$\frac{1}{2} > p_{1,1} = p_{2,2} > 0 \text{ and } p_{1,2} = 1 - p_{1,1} = p_{2,1}$$

and

$$H_2(p_{1,1}, 1 - p_{1,1}) = -p_{1,1} \log p_{1,1} - p_{1,2} \log p_{1,2} > \left(\sqrt{2} - 1\right) \cdot \log 2.$$

Here H_2 is the entropy function on Δ_2. Hence, $H_2(p_{1,1}, 1 - p_{1,1})$ is the entropy of the first row of P and

$$H_2(p_{1,1}, 1 - p_{1,1}) = H_2(p_{2,1}, 1 - p_{2,1}).$$

This choice of P is possible as H_2 maps Δ_2 continuously onto the real numbers between 0 and $\log 2$. Let $c = H_2(p_{1,1}, 1 - p_{1,1})$.

Thus π_i is the probability of i being the first symbol; $p_{i,j}$ is the probability of j being the next symbol given that i was the previous symbol. The probability of the two-symbol sequence (i,j) is $\frac{p_{i,j}}{2}$.

One computes

$$M^{(2)}(I^1) = (\log 2)^2.$$

As the source is stationary, one has

$$M^{(2)}(I^2) = M^{(2)}(I^1)$$

and, therefore,

$$M^{(2)}(I^1) + M^{(2)}(I^2) = 2(\log 2)^2.$$

One also computes

$$\begin{aligned} M^{(2)}(I^{[2]}) &= \frac{1}{2} \cdot \sum_{i=1,2} \sum_{j=1,2} p_{i,j} \left(\log \frac{1}{2p_{i,j}}\right)^2 \\ &= (c + \log 2)^2 + \mathsf{Var}\, I^{[2]} \\ &> 2(\log 2)^2 = M^{(2)}(I^1) + M^{(2)}(I^2). \end{aligned}$$

Moreover,

$$\mathsf{Var}\, I^{[2]} > 0 = \mathsf{Var}\, I^{1} + \mathsf{Var}\, I^{2}.$$

This proves that additivity does not hold in general for $M^{(i)}$ and Var, not even when $\mathfrak{S}$ is a stationary Markov source. If an analogue of Theorem 1 for higher information moments holds true at all, its proof cannot rely on the additivity property; it could, however, rely on the convergence of the sequences of the underlying probability spaces $(\mathcal{S}_{t+1} \mid \mathcal{S}_{[t]})$ and $\mathcal{S}_{[t]}$ or the limit of the Cesaro averages of the latter.

6 Information Moments of Markov Sources

To define the moments of information for a memoryless source with a finite alphabet one simply uses the standard definition of moments, but specialized to the case when the random variable is an information measure. For an arbitrary stationary source, a convincing definition of information can be derived from Theorem 1; however, this is not necessarily the first moment of a probability distribution and, moreover, the techniques employed do not lend themselves to plausible definitions of analoga of higher moments of information. In this section we focus on stationary Markov sources. For such sources, one can provide a convincing definition of the higher moments of information.

We assume the reader to be familiar with elementary facts concerning finite Markov chains. A Markov source (with finite alphabet) is just a finite Markov chain with states called symbols. Keeping this in mind, the terminology of Markov chains applies to Markov sources. Let $\mathfrak{S} = (\{\mathcal{S}_T \mid T \in \mathfrak{T}\}, \Sigma)$ be a Markov source[4]. Such a source is conveniently described by a row vector π and a matrix P as follows:

1. π is indexed by Σ, such that
$$0 \leq \pi_\sigma \leq 1 \text{ and } \sum_{\sigma \in \Sigma} \pi_\sigma = 1.$$
2. P is indexed by $\Sigma \times \Sigma$, such that, for each entry,
$$0 \leq p_{\sigma_0,\sigma_1} \leq 1 \text{ and } \sum_{\sigma_1 \in \Sigma} p_{\sigma_0,\sigma_1} = 1.$$

The value of π_σ is interpreted as the probability of output of $\mathfrak{S}$ at time 0 being σ. The matrix entry p_{σ_0,σ_1} is the probability $p(\sigma_1 \mid \sigma_0)$ of the next output symbol being σ_1 given the present one has been σ_0.

[4] We only consider Markov sources with memory size 1. Allowing for a larger memory size does not make an essential difference apart from leading to a more complicated notation.

For $n \in \mathbb{N}_0$, let $\pi(n)$ be the row vector, indexed by Σ, in which $\pi(n)_\sigma$ is the probability of σ being the output at time n. Then $\pi(0) = \pi$ and $\pi(n) = \pi P^n$. We assume that $\mathfrak{S}$ is stationary; this implies that $\pi = \pi(n)$ for all n and, hence, $\pi P = \pi$. A vector π with this property need not exist for a given matrix P; we refer the reader to the literature on Markov chains for detailed analyses of this issue; however, when it exists uniquely it represents $\lim_{n\to\infty} \pi(n)$.

For the remainder of this section we assume that $\mathfrak{S}$ is a stationary Markov source given by π and P. The Markov property implies that

$$(\mathcal{S}_{t+1} \mid \mathcal{S}_{[\,t\,]}) = (\mathcal{S}_{t+1} \mid \mathcal{S}_t)$$

for all t. By stationarity one has

$$(\mathcal{S}_{t+1} \mid \mathcal{S}_t) = (\mathcal{S}_1 \mid \mathcal{S}_0).$$

Therefore, slightly abusing notation,

$$\lim_{t\to\infty} (\mathcal{S}_{t+1} \mid \mathcal{S}_{[\,t\,]}) = (\mathcal{S}_1 \mid \mathcal{S}_0).$$

Using the definition of entropy for stationary sources as suggested by Theorem 1, one finds

$$H(\mathfrak{S}) = \sum_{\sigma_0,\sigma_1\in\Sigma} p_0(\sigma_0) \cdot p(\sigma_1 \mid \sigma_0) \cdot \log \frac{1}{p(\sigma_1 \mid \sigma_0)}$$

which is the first central moment of Shannon information for the source $(\mathcal{S}_1 \mid \mathcal{S}_0)$.

As an alternative, one could try to base the definition of $H(\mathfrak{S})$ on the other limit in Theorem 1. Of course, this works; however, in that case, $H(\mathfrak{S})$ cannot be interpreted as an expectation of a given probability space. The obvious limit to consider is $\lim_{t\to\infty} \mathcal{S}_{[\,t\,]}$ or a variant of it modelled as a Cesaro average. Using these we get $H(\pi)$ as the entropy, not a satisfactory solution.

The important observation is: for a stationary Markov source one has a limiting finite probability space on which to base definitions.

Definition 1 *Let $\mathfrak{S}$ be a stationary Markov source with finite alphabet Σ. Let I be an information measure. For $i \in I$, the i-th moment of I of $\mathfrak{S}$ is the i-th moment of I on the space $(\mathcal{S}_1 \mid \mathcal{S}_0)$.*

Given this definition of moments, it is clear how to derive formulæ for parameters like variance or skewness from these. The results of Section 3 apply. A more detailed analysis would reveal specific properties of the information moments of stationary Markov sources. We postpone this study to a later time.

7 Moments for Other Kinds of Entropy

There have been many conceptually different definitions of the notion of information. For a brief survey see [Csi08]. While most definitions are based on probability spaces, there have also been attempts to use logical relationships, expressed in algebraic terms, as a foundation (for example [IU62, Ing65]; see also [Gui77]). These definitions result in a general class of information measures which includes Shannon's H as a highly significant special case.

In that class one also finds the information measure H_{Hartley} proposed by Hartley in 1928 [Har28]. For a finite probability space $\mathcal{S} = (\Sigma, p)$, let

$$c_S = |\,\{\sigma \mid \sigma \in \Sigma, p(\sigma) > 0\}\,|.$$

Then $H_{\mathrm{Hartley}}(\mathcal{S}) = \log c_S$. Note that, like H_{Shannon}, H_{Hartley} can be interpreted as an expectation, albeit in a trivial way, as $\mathsf{E}\, I_{\mathrm{Hartley}}(\sigma)$ where $I_{\mathrm{Hartley}}(\sigma) = \log c_S$ for all $\sigma \in \Sigma$.

Variants, intuitively well motivated, of the sets of axioms defining the information measure H_{Shannon} lead to the following two parameterized classes of information measures:

- Aczél's measure (entropy of degree α)

$$H^{\alpha}_{\mathrm{Aczél}}(\mathcal{S}) = \frac{\sum_{\sigma\in\Sigma} p(\sigma)^{\alpha} - 1}{\left(\left(\frac{1}{2}\right)^{\alpha-1} - 1\right)\cdot(\log 2)^{-1}}$$

  where $\alpha \in \mathbb{R}$, $\alpha \neq 1$ and, usually, $\alpha \geq 0$.

- Rényi's measure (entropy of order β)

$$H^{\beta}_{\mathrm{Rényi}}(\mathcal{S}) = \frac{\log\left(\sum_{\sigma\in\Sigma} p(\sigma)^{\beta}\right)}{1-\beta}$$

  where $\beta \in \mathbb{R}_{+}$ and $\beta \neq 1$.

For details see [Acz84, AFN74, AD75, Rén65, Rén61, ESS98, San87]. Both in $H_{\mathrm{Aczél}}$ and $H_{\mathrm{Rényi}}$, the denominator achieves the normalization. One proves that

$$\lim_{\alpha\to 1} H^{\alpha}_{\mathrm{Aczél}}(\mathcal{S}) = \lim_{\beta\to 1} H^{\beta}_{\mathrm{Rényi}}(\mathcal{S}) = H_{\mathrm{Shannon}}(\mathcal{S})$$

and

$$\lim_{\beta\to 0} H^{\beta}_{\mathrm{Rényi}}(\mathcal{S}) = H_{\mathrm{Hartley}}(\mathcal{S}).$$

It is occasionally necessary to deal with the case of $c_S < |\,\Sigma\,|$ separately.

The measure $H_{\mathrm{Aczél}}$ can be interpreted as an expectation as follows: Let

$$I^{\alpha}_{\mathrm{Aczél}}(\sigma) = \frac{p(\sigma)^{\alpha-1} - 1}{\left(\left(\frac{1}{2}\right)^{\alpha-1} - 1\right)\cdot(\log 2)^{-1}}.$$

We need to assume that $\alpha \neq 1$. Then

$$\mathsf{E}\, I^{\alpha}_{\mathrm{Aczél}}(\sigma) = \sum_{\sigma\in\Sigma} p(\sigma)\cdot\frac{p(\sigma)^{\alpha-1} - 1}{\left(\left(\frac{1}{2}\right)^{\alpha-1} - 1\right)\cdot(\log 2)^{-1}} = H^{\alpha}_{\mathrm{Aczél}}(\mathcal{S}).$$

The measure $H_{\text{Rényi}}$ cannot be understood as an expectation in such a simple fashion; it can, however, be obtained in a form which is similar to that of an arithmetic mean, called a quasiarithmetic mean. Let ψ be a continuous and strictly increasing mapping of $\mathbb{R}_+ \cup 0$ into $\mathbb{R}$. Under some well motivated conditions, the measures H_{Shannon} and $H^{\beta}_{\text{Rényi}}$ are the only functions H satisfying

$$H(\mathcal{S}) = \psi^{-1}\left(\mathsf{E}\,\psi(I_{\text{Shannon}}(\sigma))\right)$$

where, as before,

$$I_{\text{Shannon}}(\sigma) = \log \frac{1}{p(\sigma)}.$$

With H_{Shannon} being a special case of both $H_{\text{Aczél}}$ and $H_{\text{Rényi}}$, the fact that the latter have the form of an expectation or a related form suggests to extend our investigation of the higher moments of information to these more general cases.

We indicate this for the case of information in the sense of Aczél. For $i \in \mathbb{N}$, the i-th moment of $I^{\alpha}_{\text{Aczél}}$ is given by the formula

$$\mathsf{E}\ \left(I^{\alpha}_{\text{Aczél}}(\sigma)\right)^i = \sum_{\sigma \in \Sigma} p(\sigma) \cdot \left(\frac{p(\sigma)^{\alpha-1} - 1}{\left(\left(\frac{1}{2}\right)^{\alpha-1} - 1\right) \cdot (\log 2)^{-1}} \right)^i .$$

Graphs of the first two moments of $I^{\alpha}_{\text{Aczél}}$ about the origin and its variance for several values of α and $|\,\Sigma\,| = 2$ are shown in Figures 4, 5 and 6. The corresponding graphs for $|\,\Sigma\,| = 3$ are shown in Figures 7, 8 and 9.

8 Final Observations

Taken on its own, information in the statistical sense, being an average, need not be a sufficient basis for decisions. This statement is independent of the specific measure of information – be it in the sense of Shannon or Aczél or Rényi or any other proposed notion of information. In 1983, in the context of cryptography, we proposed to consider higher moments of information and, in particular, also the variance of information for a better foundation of assessments of cryptographic security. In this paper we provide a first systematic investigation of the moments of information measures.

For information theory, by Theorem 1, stationarity of sources constitutes a natural boundary for the applicability of methods. We looked for meaningful concepts corresponding to moments in probability theory within this realm, leaving the notion of information open in general, but focussing on information in the sense of Shannon, when required.

The basic concepts can be defined as usual as most sources are based on finite probability spaces. We presented these ideas in the section on memoryless sources; obviously, they can can be re-phrased for finite sources of more complicated types. The intriguing outcome of this investigation is that the behaviour of the moments $M^{(}i)_n$ of information – and consequently of derived quantities like variance and skewness – is highly dependent on the parameters i and n; we could prove a partial characterization of these dependencies.

For stationary sources in general, the mechanism, which leads to the definition of their entropy, seems not to be applicable to higher moments. At this point, we do not have a mathematically well-motivated definition of the higher moments of information for such sources.

The intermediate case of stationary Markov sources is special. There is a limit for one of the two sequences of probability spaces considered in the proof of Theorem 1, though not the other one. Hence, the definition of moments of information can be based on this limiting probability space. These definitions are consistent with our intuition.

We have laid some groundwork for the investigation of moments of information. To some of the more important problems left open in this paper we have partial solutions which we hope to complete in the near future.

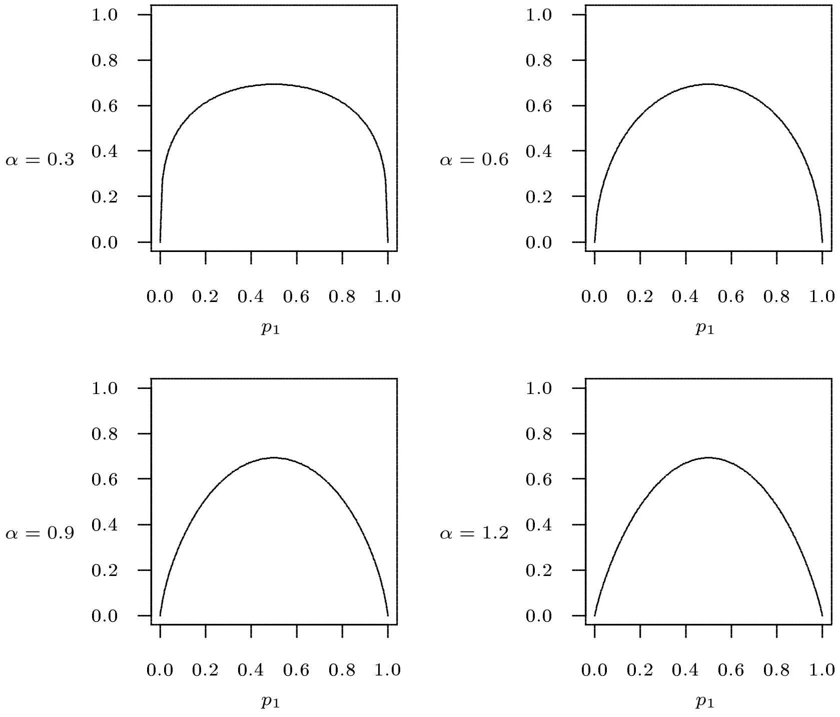

Figure 4: The first moment (entropy) of $I^{\alpha}_{\text{Aczél}}$ about the origin for different values of α, when $|\,\Sigma\,| = 2$. The proability distribution is $(p_1, 1 - p_1)$.

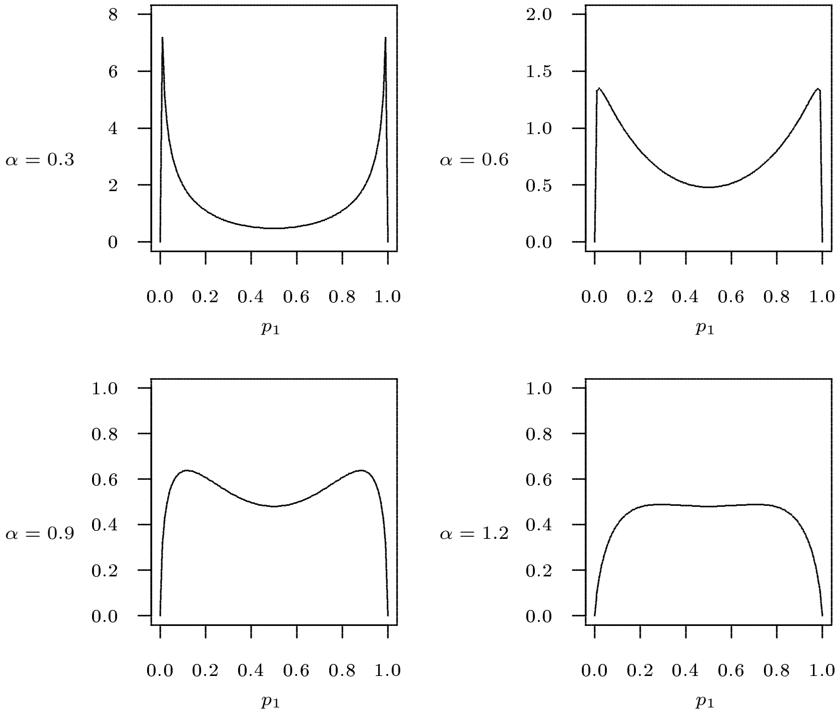

Figure 5: The second moment of $I^{\alpha}_{\text{Aczél}}$ about the origin for different values of α, when $|\,\Sigma\,| = 2$. The proability distribution is $(p_1, 1-p_1)$. Note the scale changes on the vertical axes.

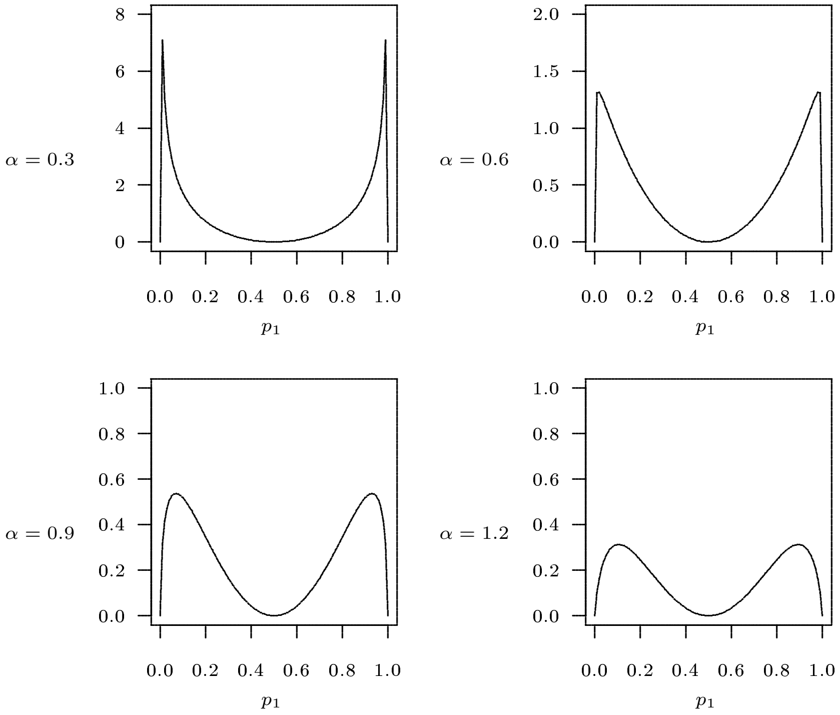

Figure 6: The variance of $I^{\alpha}_{\text{Aczél}}$ for different values of α, when $|\,\Sigma\,| = 2$. The probability distribution is $(p_1, 1 - p_1)$. Note the scale changes on the vertical axes.

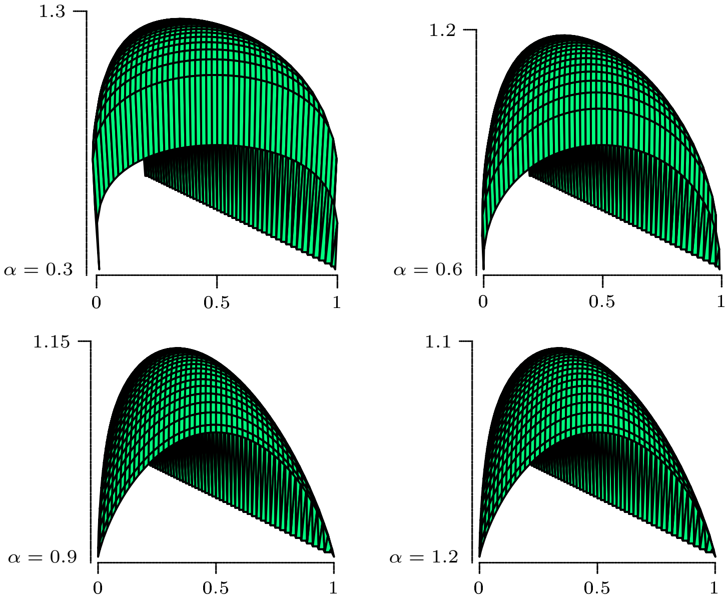

Figure 7: The first moment (entropy) of $I^{\alpha}_{\text{Aczél}}$ about the origin for different values of α, when $|\,\Sigma\,| = 3$.

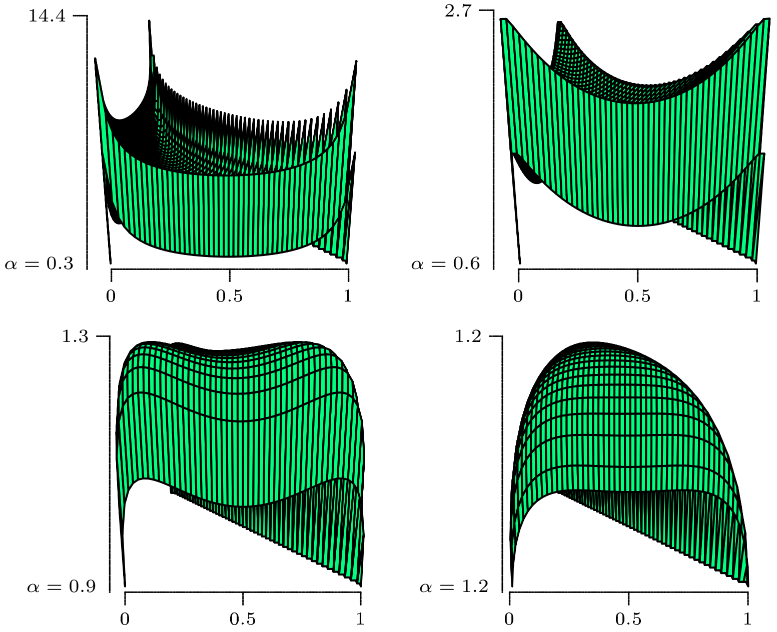

Figure 8: The second moment of $I^{\alpha}_{\text{Aczél}}$ about the origin for different values of α, when $|\,\Sigma\,| = 3$.

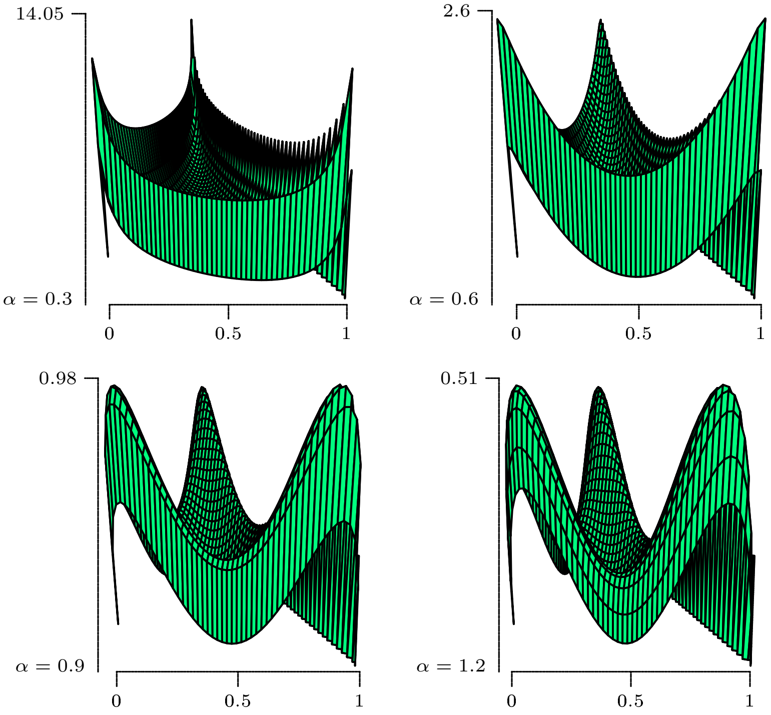

Figure 9: The variance of $I^{\alpha}_{\text{Aczél}}$ for different values of α, when $|\,\Sigma\,| = 3$.

References

[Acz61] Aczél, J.: *Vorlesungen über Funktionalgleichungen und ihre Anwendungen.* Lehrbücher und Monographien aus dem Gebiete der exakten Wissenschaften, Mathematische Reihe 25. Birkhäuser Verlag, Basel, 1961.

[Acz84] Aczél, J.: Measuring information beyond communication theory. Some probably useful and some almost certainly useless generalizations; *Information Processing & Management*, 20:383–395, 1984.

[AD75] Aczél, J. and Daróczy, Z.: *On Measures of Information and Their Characterizations.* Mathematics in Science and Engineering 115. Academic Press, New York, 1975.

[AFN74] Aczél, J., Forte, B., and Ng, C. T.: Why the Shannnon and Hartley entropies are 'natural'; *Adv. Appl. Prob.*, 6:131–146, 1974.

[Cal02] Calude, C. S.: *Information and Randomness — An Algorithmic Perspective.* Springer-Verlag, Berlin, second edition, 2002.

[CK81] Csiszár, I. and Körner, J.: *Information Theory: Coding Theorems for Discrete Memoryless Systems.* Disquisitiones mathematicæ Hungaricæ 12. Akadémiai Kiadó, Budapest, 1981.

[Csi08] Csiszár, I.: Axiomatic characterizations of information measures; *Entropy*, 10:261–273, 2008.

[ESS98] Ebanks, B., Sahoo, P., and Sander, W.: *Characterizations of Information Measures.* World Scientific, Singapore, 1998.

[Flü95] Flückiger, D. F.: *Contribution Towards a Unified Concept of Information* Dissertation, Universität Bern, 1995.

[FC] da Fontoura Costa, L.: Entropy moments characterization of statistical distributions; hal-00266512, version 1, 6 Apr 2008; available at `http://hal.archives-ouvertes.fr/hal-00266512/en/`.

[Gui77] Guiaşu, S.: *Information Theory with Applications.* McGraw-Hill, New York, 1977.

[Har28] Hartley, R. V.: Transmission of information; *Bell System Tech. J.*, 7:535–563, 1928.

[Ing65] Ingarden, R. S.: Simplified axioms for information without probability; *Prace Matematyczne*, 9:273–282, 1965.

[IU62] Ingarden, R. S. and Urbanik, K.: Information without probability; *Colloquium Mathematicum*, 9:131–151, 1962.

[JK97] Jürgensen, H. and Konstantinidis, S.: Codes; In Rozenberg, G. and Salomaa, A., editors, *Handbook of Formal Languages*, volume 1, pages 511–607. Springer-Verlag, Berlin, 1997.

[JM84] Jürgensen, H. and Matthews, D.: Some results on the information theoretic analysis of cryptosystems; In Chaum, D., editor, *Advances in Cryptology, Proceedings of CRYPTO 83, Santa Barbara, 1983*, pages 303–356. Plenum Press, New York, 1984.

[JR96] Jürgensen, H. and Robbins, L.: Towards foundations of cryptography: Investigation of perfect secrecy; *J. UCS*, 2:347–379, 1996 Special issue: C. Calude (ed.), *The Finite, the Unbounded and the Infinite, Proceedings of the Summer School "Chaitin Complexity and Applications,"* Mangalia, Romania, 27 June – 6 July, 1995.

[Jür08] Jürgensen, H.: Complexity, information, energy; *Internat. J. Foundations Comput. Sci.*, 19:781–793, 2008.

[Lyr02] Lyre, H.: *Informationstheorie, eine philosophisch-naturwissenschaftliche Einführung.* Wilhelm-Fink-Verlag, München, 2002.

[Lyr04] Lyre, H.: *Quantentheorie der Information. Zur Naturphilosophie der Theorie der Ur-Alternativen und einer abstrakten Theorie der Information. Mit einem Geleitwort von Carl Friedrich von Weizsäcker.* mentis, Paderborn, 2004.

[Mac07] MacKay, D. J. C.: *Information Theory, Inference, and Learning Algorithms.* Cambridge University Press, Cambridge, 6th edition, 2007.

[R8́2] Rényi, A.: *Tagebuch über die Informationstheorie.* VEB Deutscher Verlag der Wissenschaften, Berlin, 1982.

[Rén61] Rényi, A.: On measures of entropy and information; In Neyman, J., editor, *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, June 20 – July 30, 1960, Statistical Laboratory of the University of California, Berkeley*, Volume 1, Contributions to the Theory of Statistics, pages 547–561. University of California Press, Berkeley, California, 1961. Reprinted in [Tur76a], number 180, 565–580, with a comment by I. Csiszár.

[Rén65] Rényi, A.: On the foundations of information theory; *Rev. Inst. Internat. Statist.*, 33:1–14, 1965 Reprinted in [Tur76b], number 242, 304–318, with a comment by I. Csiszár.

[Rob98] Robbins, L. E.: *Modelling Cryptographic Systems.* PhD thesis, The University of Western Ontario, London, Canada, 1998.

[San87] Sander, W.: The fundamental equation of information and its generalizations; *Aequationes Math.*, 33:150–182, 1987.

[Sha48] Shannon, C. E.: A mathematical theory of communication; *Bell System Tech. J.*, 27:379–423, 623–656, 1948.

[Sha49] Shannon, C. E.: Communication theory of secrecy systems; *Bell System Tech. J.*, 28:656–715, 1949.

[Tur76a] Turán, P., editor *Selected Papers of Alfréd Rényi, 2, 1956–1961* Akadémiai Kiadó, Budapest, 1976.

[Tur76b] Turán, P., editor *Selected Papers of Alfréd Rényi, 3, 1962–1970* Akadémiai Kiadó, Budapest, 1976.

[YY73] Yaglom, A. M. and Yaglom, I. M.: Вероятность и информация Издательство "Наука", Moskow, 1973 Издание третье, переработанное и допольненное; English translation: *Probability and Information,* Kluwer, Boston, 1983; German translation: A. M. Jaglom, I. M. Jaglom, *Wahrscheinlichkeit und Information,* Verlag Harri Deutsch, Thun, 1984.