On the Mathematical Definition of Entropy

Bo-Sture Skagerstam

Institute of Theoretical Physics, Fack, Göteborg, Sweden

Z. Naturforsch. **29 a**, 1239–1243 [1974]; received March 12, 1974)

We discuss the Baron-Jauch definition of entropy and show that it is the unique answer to an entropy which has the properties of extensivity, positivity and continuity in a weak sense. As an application we also show how one easily can derive the canonical distribution from this definition of entropy using information theoretical arguments.

Introduction

Recently Baron and Jauch [1] discussed the connection between Boltzmann's statistical definition of entropy and the similar concept used in information theory. One of their conclusions was stated in the following form: "The similarity of the formal expressions in the two cases has misled many authors to identify entropy of information (as measured by the formula of Shannon) with negative physical entropy. The origin of the confusion is traced to a seemingly paradoxical thought experiment of Szilard, which we analyze herein. The result is that this experiment cannot be considered as a justification for such identification and that there is no paradox". [1] Thus we see that the Baron's and Jauch's thesis is in sharp contradiction to Brillouins view according to which the two entropy concepts should be identified [2]. We shall however not dwell on this important question but instead investigate the problem of uniqueness. This is a question which was put forward and answered in the beginning of information theory in the case of discrete information sources [3]. Baron and Jauch were able to find a sufficiently precise mathematical definition of entropy which is general enough to include a lot of interesting applications [3–5]. The question of uniqueness was, however, not raised by them. We prove that extensivity, positivity and a weak form of continuity gives a unique entropy which is precisely of the form that Baron and Jauch proposed.

Definitions

The basic idea of Baron and Jauch is that we have associated two different probability measures with some physical system under consideration. One measure ν is given to characterize an apriori probability distribution on the possible states of the system, describing the situation before any observation is made. The other measure μ is associated with the information which is gained by the observer. To be more specific we have two probability spaces (X, S, ν) and $X, S\,\mu)$ where we for simplicity assume that the underlying spaces (X) and σ-algebras (S) are the same. This is not necessary but it simplifies the notation to some extent. By definition we then have:

$$\mu, \nu \geqq 0\,;\; 1 = \mu(X) = \int_X \mathrm{d}\mu \;= \int_X \mathrm{d}\nu = \nu(X)\,. \qquad (1)$$

The μ- and ν-measures are not assumed to be independent. In fact it seems reasonable to assume that $\mu < \nu$ i. e. μ is absolutely continuous with respect to ν [1]. In the appendix we give a very simple example of this. The precise meaning of this is the following:

$$\mu < \nu \iff 0 = \nu(\Omega) \Rightarrow \Omega\ \mu \text{ measurable and } \mu(\Omega) = 0\ \forall\, \Omega \in S\,.$$

We then have the following important theorem from integration theory:

The radon-nikodym theorem: Suppose that μ and ν are positive and bounded measures such that $\mu < \nu$. Then there exists a unique function $f \in L^1(\nu)$ such that:

$$\mu(\Omega) = \int_\Omega f\,\mathrm{d}\nu\ \forall\ \Omega \in S\,.$$

Remark: f is the so called radon-nikodym derivative and is sometimes denoted by the symbol $\mathrm{d}\mu/\mathrm{d}\nu$ (see Reference [6]).

We notice the following wellknown fact:

Lemma I: Suppose that (X, S, μ) and (X, S, ν) are probability spaces and $\mu < \nu$. Then the subset Ω defined by

$$\Omega = \left\{ x \in X \;\middle|\; \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\,(x) \leqq 0 \right\}$$

has μ- and ν-measure zero.

Reprint requests to Dr. Bo-Sture Skagerstam, Institute of Theoretical Physics, Fack, *S-40220 Göteborg 5*, Sweden.

Proof: We can realize this in the following way:

$$1 = \int_X \mathrm{d}\mu = \int_X \mathrm{d}\nu = \int_{\Omega'} \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\,\mathrm{d}\nu + \int_{\Omega} \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\,\mathrm{d}\nu \leqq \mu(\Omega') \leqq 1 .$$

Moreover since

$$\mu(\Omega) = \int_{\Omega} (\mathrm{d}\mu/\mathrm{d}\nu)\,\mathrm{d}\nu$$

we see at once that $\nu(\Omega) = 0$ (e. g. use theorem 1.39 in Reference [6]).

If we now have a sequence of probability spaces:

$$\{(X, S, \mu_k), (X, S, \nu_k)\}_{k=1}^{n}$$

we know [7] first of all that there exists a unique measure $\bigotimes_{k=1}^{n} \mu_k$ (or $\bigotimes_{k=1}^{n} \nu_k$) such that:

$$\bigotimes_{k=1}^{n} \mu_k (A_1 \otimes \ldots \otimes A_n) = \mu_1(A_1) \ldots \mu_n(A_n) \qquad (2)$$

and if $\mu_k < \nu_k$ where $k \in \{1 \ldots n\}$ the Radon-Nikodym theorem can be extented in such a way that

$$\exists\,!\, f_{1\ldots n} \in \bigotimes_{k=1}^{n} L_1(\nu_k) \quad \bigotimes_{k=1}^{n} \mu_k \left(\bigotimes_{l=1}^{n} A_l \right) = \int_{\bigotimes_{l=1}^{n} A_l} f_{1\ldots n}\, \mathrm{d}\left(\bigotimes_{k=1}^{n} \nu_k \right) \qquad (3)$$

where

$$f_{1\ldots n} = \frac{\mathrm{d}\left(\bigotimes_{k=1}^{n} \mu_k \right)}{\mathrm{d}\left(\bigotimes_{l=1}^{n} \nu_k \right)} = \prod_{k=1}^{n} f_k \qquad (4)$$

and

$$\forall\, k \in \{1 \ldots n\}\ ;\ \mu_k(\Omega) = \int_{\Omega} f_k\, \mathrm{d}\nu_k \ \forall\, \Omega \in S . \qquad (5)$$

Here we have used the fact we have σ-finite measure spaces in the sense of [6].

To simplify the notation we introduce the following subsets of $L^1(\nu_k)$ where $k \in \{1 \ldots n\}$

$$\mathcal{N}(\nu_k) = \{ f \in L^1(\nu_k) \,|\, (X, S, \mu_k),\ (X, S, \nu_k) \text{ prob. spaces};\ \mu_k < \nu_k \text{ and } \mu_k(\Omega) = \int_{\Omega} f\, \mathrm{d}\nu_k \} .$$

We need also a definition of "disjoint" systems:

Definition I: Consider n physical systems (or information sources) such that we can associate to every system a pair of measures exactly in the same way as described above. That is we have a sequence of measures

$$\{(\mu_k, \nu_k)\}_{k=1}^{n}$$

We say that the n systems are disjoint if the associated measure for the whole system can be written in the form

$$\left(\bigotimes_{k=1}^{n} \mu_k , \bigotimes_{l=1}^{n} \nu_l \right).$$

(μ_1, ν_1)	(μ_K, ν_K)	(μ_l, ν_l)	(μ_n, ν_n)

Fig 1. No "correlation" between disjoint sytems.

Definition II: The Baron-Jauch entropy function $\mathcal{H}(\ldots)$ is a real-valued function on the measures (μ_k, ν_k) where $k \in \{1 \ldots n\}$ such that

(i) $\mathcal{H}(\mu, \mu) = 0$;
(ii) $\mathcal{H}(\mu, \nu) \geqq 0$; $\mu < \nu$
(iii) $\mathcal{H}(\mu_1 \otimes \mu_2, \nu_1 \otimes \nu_2) = \mathcal{H}(\mu_1, \nu_1) + \mathcal{H}(\mu_2, \nu_2)$

where we as above assume the same underlying (X, S)-structure.

(i) is a normalization such that "complete knowledge" corresponds to zero entropy (iii) is the most crucial property stating that entropy is additive for disjoint physical systems (i. e. the entropy is extensive).

That such a definition is nonempty was shown in [1] by considering

$$\mathcal{H}(\mu, \nu) = \int_X f \ln f\, \mathrm{d}\nu \qquad (6)$$

where $f \in \mathcal{N}(\nu)$. Since the measures $\{\nu_k\}_{k=1}^{n}$ are assumed to be apriori known we see with help of the generalization of the Radon-Nikodym theorem that $\mathcal{H}$ is essentially a function defined on the classes $\mathcal{N}(\nu_k) k \in \{1 \ldots n\}$. Hence we can write

$$\mathcal{H}(\cdot) : \bigotimes_{k=1}^{n} \mathcal{N}(\nu_k) \to \mathcal{H}\left(\bigotimes_{k=1}^{n} \mathcal{N}(\nu_k) \right) \subset \mathcal{R}$$

such that

(i) $\mathcal{H}(f) = 0$ if $f = 1$;
(ii) $\mathcal{H}(f) \geqq 0$ when $f \in \mathcal{N}(\nu_k) k \in \{1 \ldots n\}$;
(iii) $\mathcal{H}(f_k f_l) = \mathcal{H}(f_k) + \mathcal{H}(f_l)$ (7)
where $k, l \in \{1 \ldots n\}$; $f_k \in \mathcal{N}(\nu_k)$.

We now also assume that $\mathcal{H}(\cdot)$ as a function

$$\mathcal{H}(\cdot) : L^1(\nu) \to \mathcal{R}$$

is continuous along rays in $L^1(\nu)$. Under certain conditions on the elements in $\mathcal{N}(\nu)$ we shall now show how to give a precise meaning to the $\ln(\cdot)$-function in (6) and also show that (6) in fact is the unique answer to the definition II.

Uniqueness

First of all we note the following fact:

Theorem I: Suppose that (X, S, ν) is a probability space and that $f \equiv 1+\varphi$ is a measurable function such that

$$0<\varepsilon<f(x)<\mathcal{M}=2 \quad \text{a. e.}$$

Then the series

$$\sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n} \varphi^n$$

converges almost everywhere. Moreover we have that:

$$\ln f = \ln(1+\varphi) = \int_0^{\varphi} \frac{\mathrm{d}\xi}{1+\xi} = \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n} \varphi^n$$

is an element in $L^1(\nu)$.

Proof: The idea is to use Lebesgue's dominated convergence theorem (theorem 1.34 in Reference [6]). Consider the following expression

$$f_k(x) = \sum_{n=1}^{K} \frac{(-1)^{n-1}}{n} \varphi^n(x) \,.$$

But since $-1<\varphi<1$ we conclude that

$$\ln f(x) = \lim_{k\to\infty} f_k(x)$$

exists a. e. in X and moreover we have that

$$|f_k(x)| \leqq \max(|\ln \varepsilon|, \ln M)$$

Lebesgue's dominated convergence theorem then gives us that

$$\lim_{k\to\infty} f_k \in L^1(\nu)$$

and the fact that

$$\lim_{k\to\infty} \int_X f_k \,\mathrm{d}\nu = \int_X \lim_{k\to\infty} f_k \, d\nu$$

which in our case means that

$$\sum_{k=1}^{\infty} \int_X \frac{(-1)^{k-1}}{k} \varphi^k \,\mathrm{d}\nu = \int_X \sum_{k=1}^{n} \frac{(-1)^{k-1}}{k} \varphi^k \,\mathrm{d}\nu \,.$$

With essentially the same arguments we also see that

$$\ln(1+\varphi) = \ln f = \int_0^{\varphi} \frac{\mathrm{d}\xi}{1+\xi} = \int_0^{\varphi} \sum_{\eta=0}^{\infty} (-\xi)^{\eta} \,\mathrm{d}\xi$$

$$\Rightarrow \ln f = \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n} \varphi^n \,.$$

By an essentially trivial extension of the notation and use of the Lebesgue-Fubinis theorem on product measures [9] we conclude that the following theorem holds:

Theorem II: Let $\{(X, S, \nu_k)\}_{k=1}^{n}$ be a sequence of probability spaces. Suppose

$$f_k = 1+\varphi_k$$

is defined on $L^1(\nu_k)$ and that

$$0<\varepsilon_k<f_k<1$$

almost everywhere where $k \in \{1 \ldots n\}$. We then have that

$$\ln[f_1(x_1)\ldots f_n(x_n)] = \sum_{m=1}^{\infty} \frac{(-1)^{m-1}}{m} \cdot \left\{ \prod_{k=1}^{n} (1+\varphi_k) - 1 \right\}^m \tag{8}$$

where $(x_1, \ldots, x_n) \in \bigotimes_{k=1}^{n} X$ is defined a. e. on $\bigotimes_{k=1}^{n} X$ and belongs to

$$\bigotimes_{k=1}^{n} L^1(\nu_k) \,.$$

By the continuity of the $\ln(\cdot)$-function we also conclude that we can define another continuous function

$$\Phi(\cdot) : L^1(\nu) \to \mathcal{R}$$

such that

$$\mathcal{H}(f) = \Phi(\ln f) \,;\; \ln f \in L^1(\nu) \,. \tag{9}$$

Lemma II: The Φ-function defined above has the following properties

$$\text{(i)} \quad \Phi(0) = 0$$

$$\text{(ii)} \quad \Phi(\ln f) \geqq 0; \quad f \in \mathcal{N}(\nu); \quad \ln f \in L^1(\nu) \tag{10}$$

$$\text{(iii)} \quad \Phi(\lambda \ln f) = \lambda \, \Phi(\ln f) \;\; \forall \lambda \in R_+; \quad f \in \mathcal{N}(\nu); \quad \ln f \in L^1(\nu) \,.$$

Proof: The only thing which needs a proof is (iii). From the properties of the Φ-function we see that

$$\Phi(\underbrace{\ln f \ldots f}_{n\text{-times}}) = \underbrace{\Phi(\ln f) + \ldots + \Phi(\ln f)}_{n\text{-times}}$$

i. e. $\forall\, n \in N$ we have that $\Phi(n \ln f) = n\, \Phi(\ln f)$. Now we take two conjugate integers p, q i. e. such that $p \cdot q = 1$

$$\Phi(\ln f) = \Phi(p \cdot q \ln f) = p\, \Phi(q \ln f)$$

$$\Rightarrow \Phi[(1/p) \ln f] = (1/p)\, \Phi(\ln f) \,.$$

Hence we conclude that $\forall\ p/q$ where $p, q \in N$ but $q \neq 0$ we have

$$\Phi[(p/q)\ln f] = (p/q)\,\Phi(\ln f)\ .$$

But since the rationals are dense in the real numbers and the Φ-function continuous we have that

$$\Phi(\lambda \ln f) = \lambda\,\Phi(\ln f)\ \forall\ \lambda \in R_+\ .$$

Hence we have constructed a continuous linear functional on the subset

$$\Omega = \{\lambda \ln f \mid f \in \mathcal{N}(\nu),\ \ln f \in L^1(X, S, \nu),\ \lambda \in R_+\} \cdot \subset L^1(X, S, \nu)\ .$$

But then we can use for example the Hahn-Banach theorem [10] to extend the functional Φ to the positive cone in $L^1(X, S, \nu)$ i. e. to the subset

$$\Omega_+ = \{\lambda f \mid f \in \mathcal{N}(\nu);\ \lambda \in R_+\} \subset L^1(X, S, \nu)\ .$$

A duality theorem in the case of L^p-spaces (theorem 6.16 in [6]) then gives the unique existence of a function g in $L^\infty(X, S, \nu)$ such that

$$\Phi(\ln f) = \int_X g \ln f\, \mathrm{d}\nu \tag{11}$$

where we assume that $\ln f \in L^1(X, S, \nu)$. But since g is unique and since by the Baron-Jauch-construction $g = f$ on the subspace Ω we have that

$$\Phi(\ln f) = \int_X f \ln f\, \mathrm{d}\nu \tag{12}$$

as a functional on the subset Ω.

Conclusion

Consider the Baron-Jauch definition of entropy i. e. areal-valued continuous (in the sense explained above) $\mathcal{H}$-function on a pair of measures satisfying absolute continuity $\mu < \nu$ such that

(i) $\mathcal{H}(\mu, \mu) = 0$,

(ii) $\mathcal{H}(\mu, \nu) \geqq 0$; if $\mu < \nu$,

(iii) $\mathcal{H}(\mu_1 \otimes \mu_2, \nu_1 \otimes \nu_2) = \mathcal{H}(\mu_1, \nu_1) + \mathcal{H}(\mu_2, \nu_2)$.

Now suppose that $\ln f \in L^1(X, S, \nu)$ then $\mathcal{H}$ is unique in the sense that $\mathcal{H}$ has the following functional structure

$$\mathcal{H}(\mu, \nu) = k \int_X f \ln f\, \mathrm{d}\nu$$

where $f \in L^\infty(X, S, \nu)$ is the Radon-Nikodym derivative $\mathrm{d}\mu/\mathrm{d}\nu$ and where k is an arbitrary positive constant. Moreover theorems I and II shows that this is self-consistent.

Remarks: It is trivial to see that (13) also has a meaning when $f = 0$ or $f = 1$ a. e. Moreover we notice that we can give a meaning to the duality theorem on Ω_+ if we investigate the proof in [6] (theorem 6.16). Finally we see that we can extend theorems I and II to the case of an arbitrary upper bound i. e.

$$0 < \varepsilon_k < f_k < M_k$$

where $k \in \{1 \ldots n\}$ if we consider the scaled set of functions $\{\tilde{f}_k\}$ such that

$$\tilde{f}_k = f_k/\mathcal{M} \qquad \text{where} \qquad \mathcal{M} = \sup M_k\ .$$

Acknowledgements

The author wishes to thank A. Din, S. Andersson, and M. Månsson for clarifying discussions and especially K. E. Eriksson for critical reading of the manuscript and helpful comments.

Appendix

The canonical distribution

We consider a phase space coarse-grained into a collection of subsets $\{A_k\}$ such that

$$\bigcup_{k \in I} A_k = \Gamma$$

where I is an index set and Γ is the phase space. Moreover we assume that we have an apriori given measure ν such that (see e. g. page 224 in [1])

$$\nu(A_k) = g_k\ \forall\ k \in I\ ;\ \sum_{k \in I} g_k = 1\ .$$

Now suppose that $\{p_k\}$ is a probability distribution such that it conforms to the given data which in our case is the mean energy i. e.

$$\langle E \rangle = \sum_{i \in I} E_i \cdot p_i$$

where E_i is a characteristic energy for the subset A_i. In the mathematical definition of entropy (13) we then put

$$\mu(A_i) = p_i\ \forall\ i \in I\ .$$

Moreover we assume that if $g_k = 0$ then $p_k = 0$, and by this assumption $\mu < \nu$. It is now straightforward to show that

$$\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(A_k) = p_k/g_k\ \forall\ g_k \neq 0\ .$$

Hence we conclude that

$$\mathcal{H}(\mu, \nu) = \int_\Gamma f \ln f\, \mathrm{d}\nu = \sum_{i \in I} p_i \ln p_i - \sum_{i \in I} p_i \ln g_i\ ,$$

where we exclude all terms with $g_i = 0$. The first principle of statistical mechanics can now be stated in the following form:

Every physically distinct microscopic distribution of a set of given particles among the various energy levels which satisfies both the condition that the total energy is $E \pm \Delta E$ (small but finite uncertainty ΔE) and the requirements of the exclusion principle, if it applies, is equally likely to occur. This means that we shall look for the extreme values for the $\mathcal{H}(\cdot)$-function under the conditions that

$$\sum_{i \in I} p_i \cdot E_i = \langle E \rangle ,$$
$$\sum_{i \in I} p_i = 1 .$$

Using a simple variational principle we then find that

$$p_k = \frac{g_k \exp\{-\beta E_k\}}{\sum_{i \in I} g_i \exp\{-\beta E_i\}}$$

where β is a Lagrange multiplier with an obvious physical interpretation. But this is just the ordinary canonical distribution form.

[1] J. M. Jauch and J. G. Baron, Helv. Phys. Acta **45**, 220–232 [1972].

[2] L. Brillouin, Science and Information Theory, American Institute of Physics, New York 1956.

[3] A. Katz, Principles of Statistical Mechanics, Freeman, San Francisco 1967.
The following uniqueness theorem is wellknown from probability theory:
Theorem: Suppose $H(p_1 \ldots, p_n)$ is a function defined for any integer n and for all values of $p_1 \ldots, p_n$ such that $p_k \geqslant 0$ $k \in \{1 \ldots n\}$ and the sum $\sum p_k = 1$.
If for any n this function is continuous with respect to all its arguments and if the following three properties hold:
(i) for any given n the function $H(p_1, \ldots, p_n)$ takes its largest value for $p_k = 1/n$ $k \in \{1 \ldots n\}$;
(ii) $H(A \circ B) = H(A) + H(B)$ where A and B are two, different or not, finite schemes.
(iii) $H(p_1, \ldots, p_n, 0) = H(p_1, \ldots, p_n)$
then this funktion H has the following functional form:
$$H(p_1, \ldots, p_n) = -a \sum_{k=1}^{n} p_k \ln p_k$$
where a is a positiv constant. For a proof see A. Khinchin, Mathematical Foundations of Informations Theory, Dover, New York 1957.

[4] E. T. Jaynes, Information Theory and Statistical Mechanics, Brandeis Lectures 1962, Benjamin, New York 1962.

[5] H. Grad, Comm. Pure. App. Math. **14**, 323 [1961].

[6] W. Rudin, Real and Complex Analysis, McGraw-Hill, London 1970.

[7] H. Bauer, Wahrscheinlichkeitstheorie und Grundzüge der Maßtheorie, § 23, Walter de Gruyter & Co., Berlin 1968.

[8] S. K. Beriberian, Measure and Integration, MacMillan & Co., New York 1962.

[9] J. Dieudonné, Treatise on Analysis, Theorem 13.21.19., American Institute of Physics, New York 1970.

[10] G. F. Simmons, Introduction to Topology and Modern Analysis, Theorem A, page 228, McGraw-Hill, London 1963.