BULLETIN DE L'ACADÉMIE
POLONAISE DES SCIENCES
Série des sciences math., astr.
et phys. — Vol. XI, No. 8, 1963

THEORETICAL PHYSICS

Information Theory and Variational Principles in Statistical Theories

by

R. S. INGARDEN

Presented by W. RUBINOWICZ on June 3, 1963

1. Extremal methods have been used in phenomenological and statistical thermodynamics from the time of Clausius and Boltzmann, and in later times variational methods have found important application in quantum statistics (cf. e.g., [1]). A similar idea was recently used in statistical optics (cf. [2] in connection with [3]). In general statistics and probability theory the question of an objective principle of finding probability distributions is one of central problems, especially if practical applications are in mind. It seems that the first who found the right way of solving this problem was R. A. Fisher [4] (cf. also [5]). There is nothing surprising in the fact that his method is variational and that the most important of its features is in common with the Boltzmann idea of using the maximum of the logarithm of probability as a criterion for the desired distribution. Roughly speaking Fisher's and Boltzmann's expressions appeared to be special cases (or modifications) of the general concept of information formulated by Shannon [6]. Shannon's information theory gave the most general and simplest tool for solving problems of this kind, in general statistics, (cf. [7]), as well as in statistical physics (cf. [8], [9]).

The aim of the present paper is to show how the known method of information theory may be yet generalized to a simple principle which unifies physical and general-statistical concepts and extends them on a new range of possible applications.

2. In [10]—[12] it was shown that the concept of information can be defined independently of that of probability, but, that, on the other hand, complete knowledge of the first determines the other, and reversely.

Both concepts belong to different logical levels. In the present paper it will be shown how their interrelation may be used for obtaining a principle of statistics. The usefulness of this principle, in physics and in problems of statistical estimation and inference, may be considered as a heuristical proof of the righthness of our choice of axioms of information.

If we fix our attention on some Boolean ring B with n atoms (n can be finite or denumerably infinite), information can be expressed by Boltzmann's formula

(1) $$H[p] = -\sum_{k=1}^{n} p_k \log p_k,$$

where p_k is the probability determined on the k-th atom, $k = 1, 2, ..., n$. We see that information is a functional of the probability distribution, which we express by the sign $H[p]$, $p = \{p_k\} = p(k)$. We can formulate. therefore, the following variational principle

(2) $$\frac{\delta H}{\delta p} = 0, \quad \frac{\delta^2 H}{\delta p^2} < 0^{*)}.$$

Conditions (2) express the *principle of maximum uncertainty* which according to Jaynes [9] should replace the rather ambiguous Laplacean "principle of insufficient reason". It is, of course, not only a natural generalization of the thermodynamical principle of the maximum of entropy in adiabatically isolated systems (2nd law of thermodynamics), but also an interpretation and logical explanation of the latter. In view of that correspondence we will call information H the *adiabatic information* (in contradistinction to the other which will be defined later on). To (2) we have to add, of course, the additional conditions

(3) $$\sum_{k=1}^{n} p_k = 1, \quad p_k > 0.$$

Further, if we consider some systems of real numbers defined on atoms of B (spectra of random variables, in particular, of physical quantities)

(4) $$u = \{u_k\} = u(k), \quad w = \{w_k\} = w(k), \quad \text{etc.},$$

we may add the following additional conditions to (3):

(5) $$\sum_{k=1}^{n} u_k p_k = U_1, \quad \sum_{k=1}^{n} u_k^2 p_k = U_2, ..., \sum_{k=1}^{n} u_k^r p_k = U_r,$$

(6) $$\sum_{k=1}^{h} w_k p_k = W_1, \quad \sum_{k=1}^{n} w_k^2 p_k = W_2, ..., \sum_{k=1}^{n} w_k^s p_k = W_s, \text{ etc.},$$

where $r, s = 0, 1, 2, ...$. More generally, we may add additional conditions of the form

(7) $$f_1(p_1, ..., p_n) = A_1, ..., f(p_1, ..., p_n) = A_m \quad (m = 0, 1, 2, ...),$$

where $f_1, ..., f_m$ are some given functions of $p_1, ..., p_n$, and $A_1, ..., A_m$ are some given constants. For the sake of simplicity and concreteness we shall consider here only additional conditions linear in p_k's of the form (5), (6). We see that the problem (2) with (3), (5) is a generalization of the known momentum problem in the probability theory.

*) This formulation may be weakened to the form

(2') $$\frac{\delta H}{\delta p} = \frac{\delta^2 H}{\delta p^2} = ... = \frac{\delta^{2l-1} H}{\delta p^{2l-1}} = 0,$$

$$\frac{\delta^{2l} H}{\delta p^{2l}} < 0 \quad (l = 1, 2, ..),$$

but (2) suffices for our purposes.

3. Let us take as a special case of (4) $u_k = k$ $(k = 0, 1, 2, \ldots)$, i.e. $n = \infty$, $s =$ etc. $= 0$, all U_i's $(i = 1, \ldots, r)$ are positive. Introducing instead of H

(8) $$L_H = -\sum_{k=1}^{\infty} p_k (\log p_k + \lambda + \mu_1 k + \mu_2 k^2 + \ldots + \mu_r k^r),$$

where $\lambda, \mu_1, \mu_2, \ldots, \mu_r$ are Lagrange multipliers, we get from (2), (3), (5)

(9) $$\frac{\partial L_H}{\partial p_k} = -(\log p_k + \lambda + \mu_1 k + \mu_2 k^2 + \ldots + \mu_r k^r) - 1 = 0,$$

and hence

(10) $$p_k = \exp(-\mu_0 - \mu_1 k - \mu_2 k^2 - \ldots - \mu_r k^r),$$

where $\mu_0 = \lambda + 1$. We see that the condition $p_k > 0$ $(k = 0, 1, \ldots)$ is satisfied and further that

(11) $$\frac{\partial^2 L_H}{\partial p_k^2} = -\frac{1}{p_k} < 0 \quad (k = 0, 1, \ldots),$$

i.e. that the second condition in (2) is satisfied, too. Denoting

(12) $$Z = \sum_{k=0}^{\infty} \exp(-\mu_1 k - \mu_2 k^2 - \ldots - \mu_r k^r) = Z(\mu_1, \mu_2, \ldots, \mu_r)$$

(we call Z the *statistical sum* or the *sum over states*) we get from (3)

(13) $$\mu_0 = \log Z$$

and from (5)

(14) $$U_1 = -\frac{\partial}{\partial \mu_1} \log Z,\ U_2 = -\frac{\partial}{\partial \mu_2} \log Z, \ldots, U_r = -\frac{\partial}{\partial \mu_r} \log Z.$$

Solving (14) with respect to $\mu_1, \mu_2, \ldots, \mu_r$ we get $\mu_1 = \mu_1(U_1, \ldots, U_r), \ldots, \mu_r = \mu_r(U_1, \ldots, U_r)$, then from (12) $Z = Z(U_1, \ldots, U_r)$, from (13) $\mu_0 = \mu_0(U_1, \ldots, U_r)$, from (10) $p_k = p_k(U_1, \ldots, U_r)$, and finally from (1) $H_{\max} = H(U_1, \ldots, U_r)$. Now we may define quantities

(15) $$\frac{1}{T_i} = \frac{\partial H(U_1, \ldots, U_r)}{\partial U_i} \quad (i = 1, \ldots, r).$$

We call T_i the *i-th* (*generalized*) *temperature* and μ_i the *i-th statistical potential.* From (1), (3), (5) we get easily the following important connection between them

(16) $$\frac{1}{T_i} = \mu_i \quad (i = 1, \ldots, r).$$

Substituting (10) into (1) we obtain further

(17) $$H_{\max} = \mu_0 + \mu_1 U_1 + \mu_2 U_2 + \ldots + \mu_r U_r.$$

From (12) and (16) we can express Z as a function of temperatures T_i's:

$$Z = \sum_{k=0}^{\infty} \exp\left(-\frac{k}{T_1} - \frac{k^2}{T_2} - \dots - \frac{k^r}{T_r}\right) = Z(T_1, \dots, T_r) \tag{18}$$

and then define the quantity

$$I = -\mu_0 = -\log Z(T_1, \dots, T_r) = I(T_1, \dots, T_r) \tag{19}$$

which we shall call the *free information* or *isothermic information*. Because of (17) we may write

$$I = \mu_1 U_1 + \dots + \mu_r U_r - H \tag{20}$$

what together with $\mu_i = (\partial H/\partial U_i)$ resembles the Legendre transformation of analytical mechanics. For $r = 0$ I corresponds exactly to Brillouin's "negentropy" [13], $I = -H$, and it seems that in such a way the controversy between Shannon's and Brillouin's definitions of information can be settled. From (2), (1) and (5) I can be expressed as a functional of p by given $T_1, \dots, T_r$:

$$I = \sum_{k=0}^{\infty} p_k\left(\frac{k}{T_1} + \frac{k^2}{T_2} + \dots + \frac{k^r}{T_r} + \log p_k\right). \tag{21}$$

Now we can formulate a variational principle for I considering temperatures $T_1, \dots, T_r$ as determined by the external conditions of the statistical system (*isothermic problem*). In such a way we have a generalization of a physical system in a thermostat and we may speak in our case about the *thermostat of the k-th order*. We put

$$\frac{\delta I}{\delta p} = 0, \quad \frac{\delta^2 I}{\delta p^2} > 0 \tag{22}$$

(we require the minimum of I) with the single additional condition (3). Denoting

$$L_I = \sum_{k=0}^{\infty} p_k\left(\frac{k}{T_1} + \frac{k^2}{T_2} + \dots + \frac{k^r}{T_r} + \log p_k + \lambda\right), \tag{23}$$

where λ is the Lagrange multiplier connected with (3), we get finally

$$p_k = \exp\left(-\Lambda - \frac{k}{T_1} - \dots - \frac{k^r}{T_r}\right) = \frac{\exp\left(-\sum_{i=1}^{r} \frac{k^i}{T_i}\right)}{Z(T_1, \dots, T_r)} \tag{24}$$

($\Lambda = \lambda + 1$). We see that also now p_k's are positive and that

$$\frac{\partial^2 L_I}{\partial p_k^2} = \frac{1}{p_k} > 0 \tag{25}$$

as it should be.

4. Now we shall discuss in greater detail the first three simplest cases: $r = 0, 1, 2$.

a) $r = 0$ (for an obvious reason we take, however, a finite n). We get from (10)

$$p_k = e^{-\mu_0} = \frac{1}{Z} = \frac{1}{n}, \tag{26}$$

i.e. the *uniform distribution* of probability, as might be expected. Further we have

$$H_{\max} = -I_{\min} = \log n. \tag{27}$$

Of course, concepts of generalized temperatures and statistical potentials do not occur in this case;

b) $r = 1$ ($n = \infty$, $U_1 = U$, $T_1 = T$, $\mu_1 = \mu$). We finally get

$$p_k = \frac{U^k}{(1+U)^{k+1}} \tag{28}$$

in the adiabatic case and

$$p_k = \frac{\exp\left(-\frac{k}{T}\right)}{1 - \exp\left(-\frac{1}{T}\right)} \tag{29}$$

in the isothermic case. Eqs. (28) correspond to the so-called *geometric distribution* well-known in statistics, (cf., e.g., [14] Ch. III Eq. (5.2)), when (29) give a special case of the *Gibbs distribution* of statistical mechanics. Then,

$$\mu = \frac{1}{T} = \log\frac{U+1}{U}. \tag{30}$$

Since $U > 0$, T (and μ) is always positive (the asymptotic case $T = 0$ is excluded by our assumption of strictly positive probabilities, cf. (3), according to the recent point of view in probability theory). The extremal values of H and I are:

$$H_{\max} = (U+1)\log(U+1) - U\log U, \; I_{\min} = -\log\left(1 - e^{-\frac{1}{T}}\right); \tag{31}$$

c) $r = 2$ ($n = \infty$, $U_1 = U$, $U_2 - U_1^2 = \sigma^2$). In this case no exact summation of the statistical sum (12) is possible. However, for the case

$$U \gg \sqrt{2}\,\sigma \tag{32}$$

we can approximatively replace summation by integration and finally get for the adiabatic distribution

$$p_k = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(k-U)^2}{2\sigma^2}\right), \tag{33}$$

i.e. the *Gauss distribution*. From (14) we obtain

$$\mu_1 = \frac{1}{T_1} = -\frac{U}{\sigma^2}, \quad \mu_2 = \frac{1}{T_2} = \frac{1}{2\sigma^2} \tag{34}$$

and, therefore, the isothermic distribution is

$$p_k = \frac{\exp\left(-\dfrac{T_2}{4T_1^2}\right)}{\sqrt{\pi T_2}} \exp\left(-\frac{k}{T_1} - \frac{k^2}{T_2}\right). \tag{35}$$

We see from (34) that the second temperature T_2 is proportional to the square of the fluctuation (standard deviation) σ or to the variance of the distribution, as statististicians call σ^2. At the same time the first temperature T_1 is negative ($U > 0$). The latter fact is intuitively clear, since probabilities for larger k are greater than those for smaller k, when $k < U$, i.e. we have then the so-called *inverted population* from the point of view of the usual thermodynamical Gibbs distribution (29). Such a situation occurs, for instance, due to the process of excitation (e.g. "optical pumping") of masers and lasers. From this example we see that the concept of equilibrium with keeping two first moments constant has a quite different meaning (of an *enforced equilibrium*) than of that with merely the first moment constant. In usual situations only the latter case occurs in physics (*natural equilibrium*) and this explains why T_2 was hitherto not in use, (cf. [10]).

5. Coming back to the general form (4) of the random variable u, we see that the above formalism can be carried out only under assumption that (for $n = \infty$)

$$Z = \sum_{k=1}^{\infty} \exp\left(-\sum_{i=1}^{1} \mu_i u_k^i\right) = \sum_{k=1}^{\infty} \exp\left(-\sum_{i=1}^{r} \frac{u_k^i}{T_i}\right) < \infty, \tag{36}$$

i.e. that the statistical sum is convergent for given $U_1, \ldots, U_r$ or $T_1, \ldots, T_r$. (This condition was fulfilled in points **3.** and **4.**). Accordingly, we call a random variable u *statistically regular* if there exist such real numbers $\mu_1, \ldots, \mu_r$ (or $T_1, \ldots, T_r$) that (36) is convergent, (cf. [11]).

If we consider not one, but, e.g., two compatible (comeasurable) random variables u, w with conditions (5) and (6), we get

$$p_k = \frac{1}{Z} \exp\left(-\sum_{k=1}^{r} \frac{1}{T_i} u_k^i - \sum_{i=1}^{s} \nu_j w_k^j\right) \tag{37}$$

under the assumption of convergence of

$$Z = \sum_{k=1}^{\infty} \exp\left(-\sum_{i=1}^{r} \frac{u_k^i}{T_i} - \sum_{j=1}^{s} \nu_j w_k^j\right) < \infty. \tag{38}$$

We have used above, in analogy to physics, temperatures for u (e.g., energy) and potentials for w (e.g. number of particles in the system; statistical potentials are then generalizations of the concept of the chemical potential). Correspondingly, we may have then the mixed problem: isothermic with respect to u, and adiabatic with respect to w, etc. It is easy to get the generalized thermodynamic identity

$$dH = \sum_{i=1}^{r} \frac{1}{T_i} dU_i + \sum_{j=1}^{s} \nu_j dW_j. \tag{39}$$

If spectral values u_k's and w_k's depend on some external parameter, V say (as volume in physics), we get an additional term (or terms if there are more of such parameters) in (39):

$$dH = \sum_{i=1}^{r} \frac{1}{T_i} dU_i + \sum_{j=1}^{s} v_j\, dW_j + PdV, \tag{40}$$

where P represents a (*generalized*) *force* connected with parameter V (pressure in this case).

The author wishes to thank the Ford Foundation, the Courant Institute of Mathematical Sciences of the New York University and the Department of Physics and Astronomy of the University of Rochester for grants and visiting professorship, respectively, which helped him appreciably to start and develop the present work. The author is also indebted to Dr. S. Watanabe and Dr. B. Mandelbrot of the I.B.M. Research Center, Yorktown Heights, New York, for a very stimulating discussion.

INSTITUTE OF PHYSICS, POLISH ACADEMY OF SCIENCES
(INSTYTUT FIZYKI, PAN)

INSTITUTE OF THEORETICAL PHYSICS, UNIVERSITY, WROCŁAW
(INSTYTUT FIZYKI TEORETYCZNEJ, UNIWERSYTET, WROCŁAW)

REFERENCES

[1] H. Koppe, *Werner Heisenberg und die Physik unserer Zeit*, Braunschweig, 1961, p. 182.
[2] R. S. Ingarden, Optica Acta, 7 (1960), 179.
[3] — , Bull. Acad. Polon. Sci., Sér. sci. math., astr. et phys., **7** (1959), 687; **9** (1961), 549.
[4] R. A. Fisher, Mess. of Math., **41** (1912), 155.
[5] — , *Statistical methods and scientific inference*, 2nd ed., Edinburgh, 1959.
[6] C. E. Shannon, Bell System Techn. J., **27** (1948), 379, 623.
[7] S. Kullback, *Information theory and statistics*, New York, 1959.
[8] E. T. Jaynes, Phys. Rev., **106** (1957), 620.
[9] R. S. Ingarden and K. Urbanik, Acta Phys. Polon., **21** (1962), 281.
[10] — — , Bull. Acad. Polon. Sci., Sér. sci. math., astr. et phys., **9** (1961), 313.
[11] — — , Colloquium Math., **9** (1962), 132.
[12] R. S. Ingarden, Bull. Acad. Polon. Sci., Sér. sci. math., astr. et phys., **11** (1963), 209.
[13] L. Brillouin, *Science and information theory*, New York, 1956.
[14] W. Feller, *An introduction to probability theory and its applications*, New York, 1950.