Journal of Statistical Physics, Vol. 26, No. 4, 1981

The Maximum Entropy Principle as a Consequence of the Principle of Laplace

N. Hadjisavvas[1]

Received February 27, 1981

The maximum entropy principle states that the probability distribution which best represents our information is the one which maximizes the entropy with the given evidence as constraints. We prove that this principle is implied from the Laplace principle of equiprobabilities applied to the set S of all N-term sequences of results which are compatible with the given evidence. We generalize to the "information gain" method of Kullback.

KEY WORDS: Maximum entropy; "Jaynes;" statistical inference.

1. INTRODUCTION

Since its proposal in 1957 by E. J. Jaynes[1,2] the "maximum entropy principle" (MEP) has been a subject of controversy. An extensive account of the "debate" between the advocates and the adversaries of the MEP has been given by J. Cyranski.[3] The present paper is written by an advocate, and its scope is to develop a simple, *quantitative* argument in favor of the MEP.

As is well known, this principle serves to determine a "subjective" probability distribution when little is known. It states that, for a given amount of information, the probability distribution which best describes our knowledge is the one which maximizes the Shannon informational entropy subject to the given evidence as constraints. The usual objection of the adversaries consists in applying the MEP to tricky situations and in thus finding results which seem paradoxical.[4] Although there exist answers to these objections, we shall not be concerned by such problems since we intend to focus our attention on the simple case to which the MEP is

[1] Laboratoire de Mécanique Quantique, BP 347, Reims, France.

usually applied, i.e. when the available information consists in the giving of the mean values of some random variables.

Explicitly, the problem is presented as follows: let $\Omega = \{\epsilon_1, \epsilon_2, \ldots, \epsilon_k\}$ be a finite set of mutually exclusive elementary events. The ϵ_i can be seen, for instance, as the possible outcomes of an experiment. Suppose now that the appearance of the ϵ_i is governed by a probability law μ which is unknown, and that we only know the mean values $\hat{f}_l$ of certain random variables $f_l(\epsilon_i)$, $l = 1, 2, \ldots, m$ defined on Ω. In general the knowledge of $\hat{f}_l$ does not fix the probability $\mu(\epsilon_i) = \mu_i$ for there may exist an infinite number of different probability assignments $p(\epsilon_i) = p_i$ giving the same mean values to $f_l(\epsilon_i)$, so that we have

$$\sum_{i=1}^{k} p_i f_l(\epsilon_i) = \hat{f}_l, \qquad l = 1, 2, \ldots, m \tag{1}$$

Since the true probability μ_i cannot be found on the basis of our knowledge of $\hat{f}_l$, Jaynes put the problem of choosing between all p_i satisfying Eq. (1) one which best represents the available information by avoiding bias.

There is a special case to this problem which is well known from the everyday life: it is possible to have absolutely no information on the ϵ_i except that they belong to Ω, i.e., that they are possible outcomes. One then usually assigns to all ϵ_i equal probabilities $p_i = 1/k$. This attitude is strongly suggested by intuition, since any other choice evidently implies bias, and was raised by Laplace to the status of a principle which bears his name.

For the more general case in which our available information is given by Eq. (1), Jaynes proposed the MEP as a generalization of the principle of Laplace: the most unbiased probability assignment is, by the MEP, the one which maximizes the entropy subject to the constraints (1). He advanced several arguments to justify the MEP.[1,5,6] Unfortunately, these arguments have a qualitative character and leave many physicists unconvinced. In what follows we shall propose a new, quantitative argument in favor of the MEP. We shall show that if one accepts Laplace's principle, then the MEP follows automatically as a special case rather than as a generalization. We think it advisable to begin our discussion on an heuristic level which shall make transparent the physical argument.

2. HEURISTIC INVESTIGATION

Suppose that we repeat the initial experiment N times. The result will be a sequence $\epsilon_{j_1}, \epsilon_{j_2}, \ldots, \epsilon_{j_N} \in \Omega^N$, $(j_i = 1, 2, \ldots, k)$.

Since our only knowledge is that the mean values of $f_l(\epsilon_i)$ are $\hat{f}_l$, we ignore which sequence will be actually realized or with what relative frequencies n_i/N for the elementary events ϵ_i. We only know that, if N is sufficiently great, we shall have almost surely

$$\forall l = 1,2, \ldots, m : \sum_{i=1}^{k} \frac{n_i}{N} f_l(\epsilon_i) \simeq \hat{f}_l \tag{2}$$

where the approximation $\simeq$ can be made as good as we wish by increasing N.

Let S be the set of all sequences in Ω^N which satisfy (2). Since our only information is that the sequence which will be actually realized will belong almost surely to S, it is now evident how to choose the probability law p_i: by Laplace's principle we should choose, if possible, a p_i which assigns to all members of S the same probability. One can now show that such a probability exists, is unique, and maximizes the entropy subject to the constraints (1). This is our justification of the MEP.

Indeed, it is easy to see that if p_i maximizes the entropy subject to (1), then the probabilities of all the members of S are the same. In fact, as is well known, there must exist constants X, Y_l (Lagrange multipliers) such that

$$\frac{\partial}{\partial p_i}\left(-\sum_i p_i \log p_i\right) + X\frac{\partial}{\partial p_i}\left(\sum_i p_i\right) + \sum_l Y_l \frac{\partial}{\partial p_i}\left[\sum_i p_i f_l(\epsilon_i)\right] = 0$$

which implies, if $p_i \neq 0$,

$$\log p_i = -1 + X + \sum_l Y_l f_l(\epsilon_i) \tag{3}$$

The constants X, Y_l can be determined from $\sum p_i = 1$ and the constraints (1). Now any element $C = \epsilon_{j_1}\epsilon_{j_2} \cdots \epsilon_{j_N}$ of Ω^N has probability $p(C) = \prod_i p_i^{n_i}$, where n_i is the number of times that ϵ_i appears in the sequence C. If we take $C \in S$, so that the relative frequencies n_i/N obey (2), we shall have by virtue of Eq. (3):

$$\frac{\log p(C)}{N} = \sum_i \frac{n_i}{N} \log p_i \simeq -1 + X + \sum_l Y_l \hat{f}_l$$

We thus see that indeed, the probabilities $p(C)$ of the elements of S are independent of n_i/N (they depend only on $\hat{f}_l$ and N) so that they are all equal. It is also possible to prove the converse: if the $p(C)$ {or $\log[p(C)]/N$} is independent of the n_i/N for all $C \in S$ and N sufficiently great, then the probability distribution p_i on Ω maximizes the entropy. We shall present the proof of this assertion in the next section, where this discussion will be made on a rigorous level.

3. RIGOROUS DEVELOPMENT

We now translate all that has been said in the preceding section in the usual epsilon–delta language. For example, relation (2) should be a statement of the type

$$\left|\sum_i \frac{n_i}{N} f_l(\epsilon_i) - \hat{f}_l\right| < \delta$$

for any small positive δ. Likewise, the set S introduced above should depend on N, δ. So let us define, for any $N \in \mathbb{N}$ and $\delta > 0$ the set

$$S_{N,\delta} = \left\{ C \in \Omega^N : \left|\sum_i \frac{n_i}{N} f_l(\epsilon_i) - \hat{f}_l\right| < \delta, \forall l \right\} \tag{4}$$

where, as usual, n_i counts the number of appearances of ϵ_i in C. As is well known, a probability distribution p on Ω satisfies (1) if and only if

$$\forall \delta > 0 : p(S_{N,\delta}) \underset{N\to\infty}{\to} 1 \tag{5}$$

We want to choose the most unbiased probability law p_i among all those satisfying Eq. (1) [or equivalently, Eq. (5)]. We now repeat the argument of the preceding section: since almost all sequences C which can actually appear shall belong to $S_{N,\delta}$ [cf. Eq. (5)] and since furthermore this is *all* we know, we should choose, according to Laplace's principle, a probability law p_i which assigns approximately equal probabilities to all members of $S_{N,\delta}$. In addition, this approximation should be as good as we wish, if we take δ sufficiently small and N sufficiently great. The exact translation of this statement into mathematical language is the following:

$$\forall \epsilon > 0, \qquad \exists \delta > 0 \qquad \text{and } \exists N_0 \in \mathbb{N} \text{ such that}$$

$$\forall N > N_0, \qquad \forall C, C' \in S_{N,\delta} : \qquad \left| \frac{\log p(C)}{N} - \frac{\log p(C')}{N} \right| < \epsilon \tag{6}$$

We use $\log[p(C)]/N$ instead of $p(C)$ for two distinct reasons: The first is that we are guided by the preceding heuristic investigation. The second is that comparing $p(C)$ would be useless. Indeed, it is obvious that $p(C)$ tends uniformly to zero as N increases, for any $C \in \Omega^N$ and any probability law p_i. Thus $|p(C) - p(C')| < \epsilon$ would be trivially satisfied independently of the choice of p_i.

Condition (6) is a rather involved mathematical statement which has a simple physical interpretation. On the other hand, the MEP is a simple mathematical statement whose exact physical interpretation is unknown. If we succeed in establishing the equivalence of condition (6) to the MEP, we shall have shown, that "maximization of the entropy" is a simple mathematical way of saying "assignment of (almost) equal probabilities to all sequences which are likely to appear."

The equivalence of (6) to the MEP will be established by the following theorem. Let us first make a remark: if the evidence (1) implies that some of the probabilities of ϵ_i are necessarily zero, it is advisable to exclude those ϵ_i from Ω. This is physically natural, since we know in advance that the excluded ϵ_i are impossible to appear, and in addition it preserves us from unnecessary mathematical complications since, for example, the nice relation (3) holds only for nonzero p_i. That is why we shall make the assumption that for each ϵ_j there exists a probability measure $p^j(\epsilon_i) \equiv p_i^j$ on Ω satisfying (1) and such that $p_j^j \neq 0$. It follows immediately that there exists a probability measure p satisfying (1) and such that $p(\epsilon_i) \neq 0$ for all i (take, for instance, $p = \sum_j p^j / k$).

Theorem 1. Let $\Omega = \{\epsilon_1, \epsilon_2, \ldots, \epsilon_k\}$ be a set and $f_l(\epsilon_i), l = 1, 2, \ldots, m$ be random variables on Ω. Let furthermore $\hat{f}_l$ be real numbers and assume that there exists at least one probability measure p_i' satisfying

$$\sum_i p_i' f_l(\epsilon_i) = \hat{f}_l, \qquad l = 1, 2, \ldots, m \tag{1'}$$

and such that $p_i' \neq 0, \forall i$. Let finally p_i be any probability measure satisfying (1). Then p_i fulfills condition (6) if and only if it maximizes the entropy subject to the constraints (1).

Proof. (a) Suppose that p_i maximizes the entropy. Then $p_i \neq 0$ for all i. Indeed, we have by hypothesis $p_i' \neq 0$, $\forall i$. Consider now the function

$$g(a) = -\sum_i \left[a p_i + (1-a) p_i' \right] \log \left[a p_i + (1-a) p_i' \right], \qquad a \in [0,1]$$

If we had $p_i = 0$ for some i, then it is easy to see that $g'(a) \to -\infty$ when $a \to 1$, so that $g(1)$ could not be a maximum, as it should. Thus $p_i \neq 0$ for all i.

Since p_i maximizes the entropy, we already know that there exist constants X, Y_l such that (3) holds. Now for any $\delta > 0$ and $N \in \mathbb{N}$ let $C \in S_{N,\delta}$. Then

$$\frac{\log p(C)}{N} = \sum_i \frac{n_i}{N} \log p_i = -1 + X + \sum_l Y_l \left[\sum_i \frac{n_i}{N} f_l(\epsilon_i) \right]$$

so that if $C, C' \in S_{N,\delta}$ we find

$$\begin{aligned} \left| \frac{\log p(C)}{N} - \frac{\log p(C')}{N} \right| &= \left| \sum_l Y_l \left[\sum_i \frac{n_i}{N} f_l(\epsilon_i) - \sum_i \frac{n_i'}{N} f_l(\epsilon_i) \right] \right| \\ &\leqslant \sum_l |Y_l| \left[\left| \sum_i \frac{n_i}{N} f(\epsilon_i) - \hat{f}_l \right| + \left| \sum_i \frac{n_i'}{N} f_l(\epsilon_i) - \hat{f}_l \right| \right] \\ &< 2\delta \sum_l |Y_l| \end{aligned}$$

Thus if for any $\epsilon > 0$ we chose $\delta = \epsilon/(2\sum_l |Y_l|)$, we deduce (6).

(b) Suppose that p_i satisfies (6) together with relation (1). We set $\hat{f}_0 = 1$ and define the following vectors inside $\mathbb{R}^k$:

$$\mathbf{f}_0 = \underbrace{(1, \ldots, 1)}_{k \text{ times}}, \qquad \mathbf{f}_l = (f_l(\epsilon_1), \ldots, f_l(\epsilon_k)), \qquad \mathbf{q} = (\log p_1, \ldots, \log p_k) \tag{7}$$

Let $\mathbf{r}$, $\mathbf{r}'$ be any two vectors in $\mathbb{R}^k$ satisfying

$$\mathbf{r} \cdot \mathbf{f}_l = \mathbf{r}' \cdot \mathbf{f}_l = \hat{f}_l, \qquad l = 0, 1, \ldots, m \tag{8}$$

and $r_i > 0$, $r_i' > 0$.

We intend to show that $\mathbf{r} \cdot \mathbf{q} = \mathbf{r}' \cdot \mathbf{q}$. To do this, we proceed as follows. For any $\epsilon > 0$, let $\delta > 0$ and N_0 be defined so that (6) is true. We can take $\delta < \epsilon$. Now set $\delta' = \delta/[\max_l \sum_i |f_l(\epsilon_i)|]$. Since the rationals are dense in R, we can find nonnegative integers n_i, n_i' and an integer $N > N_0$ such that

$$\forall i = 1, \ldots, k : \left|\frac{n_i}{N} - r_i\right| < \delta', \qquad \left|\frac{n_i'}{N} - r_i'\right| < \delta', \qquad \sum_i n_i = \sum_i n_i' = N \tag{9}$$

Consider now inside Ω^N any two sequences C, C' having, respectively, n_i/N and n_i'/N as relative frequencies for the ϵ_i. Then

$$\begin{aligned} \left|\sum_i \frac{n_i}{N} f_l(\epsilon_i) - \hat{f}_l\right| &= \left|\sum_i \frac{n_i}{N} f_l(\epsilon_i) - \sum_i r_i f_l(\epsilon_i)\right| \\ &\leqslant \sum_i \left|r_i - \frac{n_i}{N}\right| \cdot |f_l(\epsilon_i)| \leqslant \delta \end{aligned}$$

and thus $C, C' \in S_{N,\delta}$. We deduce from the choice of δ that

$$\left|\sum_i \frac{n_i}{N} \log p_i - \sum_i \frac{n_i'}{N} \log p_i\right| < \epsilon \tag{10}$$

One now has

$$\begin{aligned} |\mathbf{r} \cdot \mathbf{q} - \mathbf{r}' \cdot \mathbf{q}| &= |\sum_i r_i \log p_i - \sum_i r_i' \log p_i| \\ &\leqslant \left|\sum_i r_i \log p_i - \sum_i \frac{n_i}{N} \log p_i\right| + \left|\sum_i \frac{n_i}{N} \log p_i - \sum_i \frac{n_i'}{N} \log p_i\right| \\ &\quad + \left|\sum_i \frac{n_i'}{N} \log p_i - \sum r_i' \log p_i\right| \leqslant \sum_i \left|\frac{n_i}{N} - r_i\right| |\log p_i| \\ &\quad + \left|\sum \frac{n_i}{N} \log p_i - \sum \frac{n_i'}{N} \log p_i\right| + \sum_i \left|\frac{n_i'}{N} - r_i'\right| |\log p_i| \\ &< 2\delta' \sum_i |\log p_i| + \epsilon \leqslant \epsilon\left[1 + \frac{2\sum_i |\log p_i|}{\max_l \sum_i |f_l(\epsilon_i)|}\right] \end{aligned}$$

We conclude that for any $\epsilon' > 0$ one has $|\mathbf{r} \cdot \mathbf{q} - \mathbf{r}' \cdot \mathbf{q}| < \epsilon'$, which implies $\mathbf{r} \cdot \mathbf{q} = \mathbf{r}' \cdot \mathbf{q}$. We have thus proved

$$(\mathbf{r} \cdot \mathbf{f}_l = \mathbf{r}' \cdot \mathbf{f}_l = \hat{f}_l \; \forall l \quad \text{and} \quad r_i > 0, \quad r_i' > 0) \Rightarrow \mathbf{q} \cdot \mathbf{r} = \mathbf{q} \cdot \mathbf{r}' \tag{11}$$

Let us define the sets

$$A = \{\mathbf{r} \in \mathbb{R}^k : \mathbf{r} \cdot \mathbf{f}_l = 0, \quad l = 0, 1, \ldots, m\}$$

$$B = [\mathbf{f}_0, \mathbf{f}_1, \ldots, \mathbf{f}_m] \qquad \text{(subspace spanned by } \mathbf{f}_l\text{)}$$

It is obvious that A is the orthogonal complement of B; $A = B^{\perp}$ and thus $B = A^{\perp}$. We shall now show that for all $\mathbf{r} \in A$ one has $\mathbf{r} \cdot \mathbf{q} = 0$. If $\mathbf{r} = 0$, this is obvious. If $\mathbf{r} \neq 0$, consider the vector $\mathbf{p}' = (p_1', \ldots, p_k')$ from the assertion of the theorem and set for any $b \in \mathbb{R}$: $\mathbf{r}(b) = \mathbf{p}' + b\mathbf{r}$. Since $r_i(0) = p_i' > 0$, we can find $b \neq 0$ such that $r_i(b) > 0$. Then $\mathbf{r} = b^{-1}(\mathbf{r}(b) - \mathbf{p}')$.

Since we have obviously $\mathbf{r}(b) \cdot \mathbf{f}_l = \mathbf{p}' \cdot \mathbf{f}_l = \hat{f}_l$, we conclude from (11) that $\mathbf{r} \cdot \mathbf{q} = 0$. This is true for all $\mathbf{r} \in A$, so that $\mathbf{q} \in A^{\perp} = B$, and thus there exist constants a_l, $l = 0, 1, \ldots, m$, such that $\mathbf{q} = \sum_{l=0}^{m} a_l \mathbf{f}_l$, or equivalently

$$\forall i : \log p_i = a_0 + \sum_{l=1}^{m} a_l f_l(\epsilon_i)$$

This is equivalent to relation (3) characterizing the maximum entropy distribution, via an obvious identification of terms. Thus, p_i maximizes the entropy subject to the constraints (1). ■

4. COMMENTS

We anticipate a possible question of the reader: How can the preceding theorem be compatible with the law of large numbers? For suppose we take as probability measure p_i on Ω the one which maximizes the entropy subject to the constraints (1). Then the law of large numbers assures us that the sequence C which will appear after N repetitions of the experiment will have, almost surely, relative frequencies $n_i/N \simeq p_i$. On the other hand, the theorem says that for the same p_i, *all* the sequences with relative frequencies such that $\sum_i (n_i/N) f_l(\epsilon_i) \simeq \hat{f}_l$ have approximately the same probability, even those for which n_i/N is quite different from p_i. But then why is a sequence with $n_i/N \simeq p_i$ more likely to be realized? The answer to this question is simple: All the sequences satisfying $\sum_i (n_i/N) f_l(\epsilon_i) \simeq \hat{f}_l$ have, indeed, equal probabilities, but the *number* of the sequences for which in addition $n_i/N \simeq p_i$ is overwhelmingly greater than the number of all other sequences for which $\sum_i (n_i/N) f_l(\epsilon_i) \simeq \hat{f}_l$ holds. The proof of this fact is very simple and is contained in essence in many papers on the MEP (see, e.g., Ref. 5, p. 231).

5. THE "INFORMATION GAIN" OF KULLBACK

A few years after the fundamental article of Jaynes, Kullback[7] proposed a generalization of the MEP. Consider again the set $\Omega = \{\epsilon_1, \epsilon_2, \ldots, \epsilon_k\}$ of all possible outcomes of an experiment of which we ignore the exact probability law. Suppose that at a given moment our knowledge is correctly represented by a "subjective" probability measure $q(\epsilon_i) \equiv q_i$ on Ω. If we acquire new evidence of the form "the mean values of some random variables $f_l(\epsilon_i)$ is $\hat{f}_l$," what will be the "subjective" probability measure p_i after the acquisition of this evidence? Kullback proposed to choose the p_i which minimizes the "information gain" or "discrimination information" $\sum_i p_i \log(p_i/q_i)$.

Again, the arguments put forward in favor of this choice are somewhat hazy. In the light of the discussion on the MEP, we can propose a new argument which seems stronger because of its quantitative character. Consider again the set S of all sequences $C \in \Omega^N$ such that $\sum_i (n_i/N) f_l(\epsilon_i) \simeq \hat{f}_l$. The new probability p satisfies (1), thus $p(S) \simeq 1$, so that all the possible N-term outcomes shall belong almost surely to S. On the other hand, p should avoid bias. Since the prior probability is q, this is accomplished if and only if the *relative* probabilities of the elements of S do not change, i.e., if

$$\frac{p(C)}{p(C')} \simeq \frac{q(C)}{q(C')}, \qquad \forall C, C' \in S \tag{12}$$

As before, it can be again proved that p satisfies (12) + (1) if and only if it minimizes the information gain subject to the constraints (1). More precisely, we have the following theorem:

Theorem 2. Let $\Omega = \{\epsilon_1, \epsilon_2, \ldots, \epsilon_k\}$ be a set, $f_l(\epsilon_i)$ $(l = 1, 2, \ldots, m)$ be functions of Ω, and $\hat{f}_l$ real numbers. Let q be a probability measure on Ω such that $q(\epsilon_i) \equiv q_i > 0$ $\forall i$, and p another probability measure such that

$$\sum_i p_i f_l(\epsilon_i) = \hat{f}_l, \qquad \forall l = 1, 2, \ldots, m \tag{1}$$

and $p_i > 0$ $\forall i$. Then p_i minimizes the Kullback "information gain" $\sum_i p_i \log(p_i/q_i)$ if and only if it satisfies the condition $\forall \epsilon > 0$ $\exists \delta > 0$ and $\exists N_0 \in \mathbb{N}$ such that $\forall N \geqslant N_0$

$$\forall C, C' \in S_{N,\delta} : \left| \frac{\log[p(C)/q(C)]}{N} - \frac{\log[p(C')/q(C')]}{N} \right| < \epsilon \tag{13}$$

where $p(C)$, $q(C)$ are the probabilities of the sequence C for the measures p and q on Ω, and $S_{N,\delta}$ the set defined by relation (4).

The proof of the theorem is exactly the same as the proof of Theorem 1, so we omit it.

6. CONCLUSION

Our goal was to prove that the MEP is a consequence of the principle of Laplace. In order to achieve it, we passed from the space Ω to the space Ω^N and then to the subset $S \subset \Omega^N$ defined in Section 2. This passage from the space Ω to the metaspaces Ω and S is indicated by Russel's methodology, which is based on the hierarchy of the logical types, and was also used in a related article of M. Mugur-Schächter.[8] The reason for doing so is here obvious. We cannot apply the principle of Laplace on Ω since our information contains more than the mere knowledge of the set of all possible outcomes. On the other hand, all the available information can be expressed by the fact that almost surely $C \in S$, so that the principle of Laplace is applicable on S. We reached our goal by showing that this implies the MEP.

The theorem of Section 4 gives also a precise meaning to the frequent assertion that the probability distribution which maximizes the entropy is more "spread" than all others.[6] In fact this distribution is uniform on S.

It is of course possible to desire an even deeper consideration of the conceptual roots of the MEP. In any case, our result shows that it is sufficient to reduce one's attention to the conceptually and mathematically much simpler principle of Laplace.

ACKNOWLEDGMENT

The author wishes to thank Professor M. Mugur-Schächter for many fruitful discussions on the subject.

REFERENCES

1. E. T. Jaynes, *Phys. Rev.* **106**:620 (1957); **108**:171 (1957).
2. E. T. Jaynes, in *Colloquium Lectures in Pure and Applied Science*, No. 4 (February 1958).
3. J. Cyranski, *Found. Phys.* **8**(5/6):493 (1978).
4. K. Friedman and A. Shimony, *J. Stat. Phys.* **3**:381 (1971).
5. E. T. Jaynes, *I.E.E.E. Trans. Syst. Scien. Cyb.* **SSC-4**:227 (1968).
6. E. T. Jaynes, in *Statistical Physics*, Vol. 3, K. W. Ford, ed. (W. A. Benjamin, New York, 1963).
7. S. Kullback, *Information Theory and Statistics* (Wiley, New York, 1959).
8. M. Mugur-Schächter, *Ann. Inst. Henri Poincaré A* **32**(1):33 (1980).