PHYSICAL REVIEW

LETTERS

VOLUME 52 16 APRIL 1984 NUMBER 16

# Consistent Inference of Probabilities for Reproducible Experiments

Y. Tikochinsky,[a] N. Z. Tishby,[a] and R. D. Levine
*The Fritz Haber Molecular Dynamics Research Center, The Hebrew University, Jerusalem 91904, Israel*
(Received 3 January 1984)

The need for inducing a probability distribution from partial data and the complementary problem of the analysis of an observed distribution in terms of fewer relevant variables occur in many branches of physics. For reproducible experiments, consistency conditions which must be satisfied by any algorithm for inferring a discrete probability distribution with given averages are formulated. The only consistent algorithm is the one leading to the distribution of maximal entropy subject to the given constraints.

PACS numbers: 02.50.+s, 03.65.Bz, 05.20.−y

It is often the case that a few expectation values suffice to characterize the distribution of outcomes. Yet these values do not determine a unique distribution. An inference procedure known as the maximum entropy method has been proposed by Jaynes[1] on the basis of Shannon's axiomatic characterization of the amount of "missing information." The scope and number of effective applications of this method[2] in physics and chemistry[3,4] and engineering[5] has been constantly growing. Yet there are many scientists who are reluctant to use the procedure because of its reliance on the so-called subjective notion of missing information. Others consider that the concept of entropy function should not be used outside of its original contexts. In this Letter we offer an alternative approach to the problem of induction which does not involve Shannon's[6] entropy nor any references to subjective considerations. Rather, we start from consistency conditions which must be satisfied by any algorithm for inducing a probability distribution, for a reproducible experiment. It is assumed that the algorithm is uniform in that data of a given kind must be handled in the same way. The consistency conditions lead to strong restrictions on the algorithm. The unique ensuing distribution is the very one which results by the maximum entropy method.

The distribution of maximal entropy constrained by the average values is the only consistent induction from the data for any experiment which can be reproduced. The present derivation, we believe, should help remove the usual objections raised in connection with this method, and thus should enlarge the number of its users. Furthermore, since we do not use the concept of the entropy we can reverse the usual argument and show that the information theoretic entropy is that unique functional of the distribution which attains its maximal value for the consistent inference.[7]

The essence of the technical problem is as follows: Given a set of $n$ mutually exclusive and exhaustive alternatives, or states, and $m$ expectation values $\langle A_r \rangle$,

$$\langle A_r \rangle = \sum_{i=1}^{n} A_{ri} p_i, \qquad r = 1,\ldots,m, \quad m \leq n-1, \tag{1}$$

of the variables $A_r$ defined on these states and attaining the values $A_{ri}$ at the state $i$, a normalized probability distribution $p_1,\ldots,p_n$ is sought which fulfills (1). Here $m \leq n-1$ and usually $m \ll n$ so

 

that the expectation values do not determine via (1) a unique distribution. Required therefore is an algorithm, denoted here by (a), which selects a unique probability distribution among the set of distributions which are normalized and consistent with the data (1).

In what follows we assume that the $m$ variables $A_r = \{A_{ri}\}$, considered as $n$-dimensional vectors, and the normalization vector $A_0$ $(A_{0i} = 1)$ are linearly independent. This can always be achieved by reducing the number $m$. Say now we add other linearly independent vectors, $A_{m+1}, \ldots, A_{n-1}$ up to a complete base of the $n$-dimensional space. Then one can always expand

$$-\ln p_i = \sum_{r=0}^{n-1} \lambda_r A_{ri}. \tag{2}$$

Given a particular choice for the complementary vectors $A_{m+1}, \ldots, A_{n-1}$ and for their expansion coefficients $\lambda_{m+1}, \ldots, \lambda_{n-1}$, the coefficients $\lambda_0, \ldots, \lambda_m$ are uniquely determined by the constraints (1) and the normalization. A discrete probability distribution consistent with the constraints can thus be represented as in (2). Required is an algorithm which insures that the choice of the expansion in the complementary space is unique.

The consistency condition follows from our requirement that the experiment is reproducible. That is, it must be possible to carry out $N$ ($N$ not necessarily large), independent repetitions of the experiment. Given the induced probabilities (2), the probability of any particular sequence of outcomes, where the state $i$ might occur $N_i$ times in $N$ independent repetitions, $N = \sum N_i$, is $p_1^{N_1} \cdots p_n^{N_n}$. Collecting together all the experiments where each state $i$ occurred the number, $N_i$, of times irrespective of its order in the sequence, we have the probability of a particular distribution $\tilde{N} \equiv (N_1, \ldots, N_n)$ of outcomes

$$P_{\tilde{N}} = g_{\tilde{N}} p_1^{N_1} \ldots p_n^{N_n}. \tag{3}$$

Here $g_{\tilde{N}}$ is the number of sequences corresponding to a particular partition of $N$,

$$g_{\tilde{N}} = N!/N_1! \ldots N_n! \tag{4}$$

and plays the role of a degeneracy factor.

That the experiment is reproducible is taken here to be the requirement that the partial data (1) can be written directly in terms of $P_{\tilde{N}}$,

$$\sum_{\tilde{N}} P_{\tilde{N}} \sum_i N_i A_{ri} = N \langle A_r \rangle. \tag{5}$$

To interpret (5) we introduce the "sample average"

$$\bar{A}_r = N^{-1} \sum_i N_i A_{ri}. \tag{6}$$

Then (5) is the requirement[8] that averaging the sample average over all possible outcomes should give the expectation values over the elementary probabilities $p_i$. Defining the variables $B_{r\tilde{N}}$ by

$$B_{r\tilde{N}} = \sum_{i=1}^{n} N_i A_{ri} \tag{7}$$

we write (5) as average values over $P_{\tilde{N}}$

$$\langle B_r \rangle = \sum_{\tilde{N}} B_{r\tilde{N}} P_{\tilde{N}} = N \langle A_r \rangle. \tag{8}$$

These are $m$ constraints on the probability distribution $P_{\tilde{N}}$. They are of the same nature as the $m$ constraints (1). Given these $m$ constraints our algorithm (a) determines a unique normalized distribution $Q_{\tilde{N}}$ which is consistent with (8).

The discrete probability distribution $Q_{\tilde{N}}$ can also be expanded in a complete basis as in (2). The space is $l$ dimensional, $l = \binom{N+n-1}{n-1}$. The first $m+1$ linearly independent basis vectors are taken to be the $B_r$'s defined via (7) in terms of the $m+1$ $A_r$'s. The other $l-m-1$ linearly independent vectors are arbitrary. The algorithm (a) determines the choice of the complementary basis and the $l-m-1$ expansion coefficients $\mu_{m+1}, \ldots, \mu_{l-1}$,

$$-\ln(Q_{\tilde{N}}/g_{\tilde{N}}) = \sum_{r=0}^{l-1} \mu_r B_{r\tilde{N}}. \tag{9}$$

The constraints (8) then determine uniquely the first $m+1$ expansion coefficients $\mu_0, \ldots, \mu_m$.

We have now two alternative routes to the distribution of outcomes in $N$ independent repetitions of the experiment. Starting from the state averages $\langle A_r \rangle$ we can use (a) to derive the elementary probabilities $p_i$ and then obtain $P_{\tilde{N}}$ as the multinomial distribution (3). Alternatively, we can use the sample averages $\langle B_r \rangle$, cf. (8), and use the algorithm (a) directly to get the distribution $Q_{\tilde{N}}$.

The algorithm (a) will be called consistent if these two routes yield the same results, $Q_{\tilde{N}} = P_{\tilde{N}}$. Our consistency requirement is thus summarized by the commutative diagram

$$\left.\begin{array}{ccc} \langle A_r \rangle & \xrightarrow{\text{(a)}} & p_i \\ \downarrow & & \downarrow \\ \langle B_r \rangle & \xrightarrow[\text{(a)}]{} & P_{\tilde{N}} \end{array}\right\} \begin{array}{l} \text{sample} \\ \text{averaging} \end{array} \quad \left\{ \begin{array}{l} \text{independent} \\ \text{repetitions} \end{array} \right. . \tag{10}$$

To apply the consistency condition in a meaningful fashion it is necessary to require that the algorithm operates on all possible input in the same manner. Given the input in the form (1) the algorithm determines the corresponding probability distribu-

tion using one and the same procedure.[9] We shall take this requirement to be part of the consistency condition. The necessary and sufficient condition for consistency is our central result and can be stated as the following.

*Theorem.*—The algorithm (a) is consistent if and only if (a) is the maximum entropy procedure.

*Proof.*—We show first that the consistency condition implies the maximum entropy procedure. The statement "(a) is consistent" implies $P_{\tilde{N}} = Q_{\tilde{N}}$ or

$$\ln(Q_{\tilde{N}}/g_{\tilde{N}}) = \sum_{i=1}^{n} N_i \ln p_i. \tag{11}$$

Using (2), (3), (7), and (9) in (11),

$$\mu_0 + \sum_{r=1}^{l-1} \mu_r B_{r\tilde{N}} = \sum_{r=0}^{n-1} \mu_r \sum_{i=1}^{n} N_i A_{ri} + \sum_{r=1}^{l-1} \mu_r B_{r\tilde{N}} = \sum_{i=1}^{n} N_i \Big( \sum_{r=0}^{n-1} \lambda_r A_{ri} \Big) = \lambda_0 N + \sum_{r=1}^{n-1} \lambda_r B_{r\tilde{N}}. \tag{11'}$$

The vectors $B_{r\tilde{N}}$ are linearly independent in the $l$-dimensional vector space. Since the expansion (9) is unique it follows from (11′) that

$$\mu_0 = \lambda_0 N; \quad \mu_r = \lambda_r, \quad 1 \leq r \leq n-1;$$

$$\mu_r = 0, \quad n \leq r \leq l-1.$$

The consistency condition implies that all the expansion coefficients $\mu_r$ in (9) vanish at least for $r \geq n$. But the index $n$ has no special standing in the $l$-dimensional space. It can take up any value from $m+1$ to $l$. Therefore the consistency condition (11′) on the one hand and the requirement that (a) can handle all possible input in a uniform fashion require that $\mu_r$'s vanish for $r \geq m+1$,

$$\mu_r = 0,$$

$$\text{for all } r \text{ such that } m+1 \leq r \leq l-1. \tag{12}$$

It follows that the $\lambda_r$'s must vanish for $m+1 \leq r \leq n-1$ or

$$-\ln p_1 = \lambda_0 + \sum_{r=1}^{m} \lambda_r A_{ri} \tag{13}$$

and that

$$-\ln(Q_{\tilde{N}}/g_{\tilde{N}}) = \lambda_0 N + \sum_{r=1}^{m} \lambda_r B_{r\tilde{N}}. \tag{14}$$

The reader familiar with the maximum entropy method will recognize the $\lambda_r$'s as the Lagrange multipliers introduced therein in the process of seeking the constrained maximal value of entropy functional.

That the maximum entropy algorithm is consistent in the sense of (10) has been previously shown by Levine.[10] Indeed, the maximum entropy method applied with the constraints (8) leads to

$$Q_{\tilde{N}} = g_{\tilde{N}} \exp(-N\lambda_0 - \textstyle\sum_r \lambda_r \sum_i N_i A_{ri}), \tag{15}$$

where

$$\lambda_0 N \equiv \ln[\textstyle\sum_{\tilde{N}} g_{\tilde{N}} \exp(-\sum_r \lambda_r \sum_i N_i A_{ri})] \tag{16}$$

and the other $m$ Lagrange multipliers, $\lambda_r$, are determined from (8), by solving

$$-\frac{\partial(\lambda_0 N)}{\partial \lambda_r} = \langle B_r \rangle = N \langle A_r \rangle, \quad r = 1, \ldots, m. \tag{17}$$

The solution (15) can also be written as

$$Q_{\tilde{N}} = g_{\tilde{N}} \prod_{i=1}^{n} p_i^{N_i}, \tag{18}$$

where

$$p_i = \exp(-\lambda_0 - \sum_{r=1}^{m} \lambda_r A_{ri}). \tag{19}$$

Using the identity

$$\textstyle\sum_{\tilde{N}} g_{\tilde{N}} \prod_{i=1}^{n} [p_i \exp(\lambda_0)]^{N_i} = [p_i \exp(\lambda_0) + \ldots + p_n \exp(\lambda_0)]^N = \exp(\lambda_0 N) \tag{20}$$

together with Eqs. (16) and (17), the distribution (19) is identified as the maximum entropy solution to the problem (1). This completes the proof of the theorem.

Without reference to information theory or to a very large number of independent repetitions (the Boltzmann point of view) we have been able to derive the maximum entropy algorithm. The key elements were that the algorithm operates on all data in the same way (i.e., that it be uniform) and that it can be consistently applied for any finite number of independent repetitions. The maximum entropy method was shown to provide the one and only uniform consistent algorithm.

The consistent algorithm is already known as the maximum entropy method. But here entropy emerges as a consequence of a consistent induction and is not invoked in the specification of the algorithm. Entropy here is that unique functional of

the probabilities which is maximal for the consistent algorithm designed for a reproducible experiment.

This work was supported by the U.S. Office of Naval Research and the U.S.–Israel Binational Science Foundation. The Fritz Haber Research Center is supported by the Minerva Gesellschaft für die Forschung, mbH, München, BRD.

---

[a]Also at Racah Institute of Physics, The Hebrew University, Jerusalem, Israel.

[1]E. T. Jaynes, Phys. Rev. **106**, 620 (1975), and in *The Maximum Entropy Formalism,* edited by R. D. Levine and M. Tribus (MIT Press, Cambridge, Mass., 1979).

[2]*The Maximum Entropy Formalism,* edited by R. D. Levine and M. Tribus (MIT Press, Cambridge, Mass., 1979).

[3]For a recent review see, for example, W. T. Grandy, Jr., Phys. Rev. **62**, 175 (1980).

[4]For applications to elementary particle, nuclear, and molecular collisions see, for example, S. Dagan and Y. Dothan, Phys. Rev. D **26**, 248 (1982); Y. M. Engel and R. D. Levine, Phys. Rev. C **28**, 2321 (1983); R. D. Levine, Adv. Chem. Phys. **47,** 239 (1981).

[5]See, for example, *Modern Spectrum Analysis,* edited by D. E. Childers (IEEE Press, New York, 1978).

[6]C. E. Shannon, Bell Systems Tech. J. **27**, 379 (1948).

[7]Using the reasoning given by Jaynes (Ref. 1) this result can then be used to identify the entropy as the amount of missing information.

[8]With the frequency interpretation of probability (5) or (8) are obvious conditions. For others, these equations provide the best (in the sense of least-squares error) estimate for $\langle A_r \rangle$ given $\bar{A}_r$.

[9]Such an algorithm is called uniform. The reason for requiring uniformity is to prevent an attempt to violate the spirit of the consistency requirement by starting with the $\langle B_r \rangle$'s and then trying to construct the $\langle A_r \rangle$'s and proceeding by the upper branch of (10). Such an algorithm is not what we intend to imply by "consistent" and to prevent its use we impose the condition of uniformity.

[10]R. D. Levine, J. Phys. A **13**, 91 (1980).