Alternative approach to maximum-entropy inference

Y. Tikochinsky,* N. Z. Tishby,* and R. D. Levine
The Fritz Haber Molecular Dynamics Research Center, The Hebrew University of Jerusalem, Jerusalem 91904, Israel
(Received 24 May 1984)

A consistent approach to the inference of a probability distribution given a limited number of expectation values of relevant variables is discussed. There are two key assumptions: that the experiment can be independently repeated a finite number (not necessarily large) of times and that the theoretical expectation values of the relevant observables are to be estimated from their measured sample averages. Three independent but complementary routes for deriving the form of the distribution from these two assumptions are reviewed. All three lead to a unique distribution which is identical with the one obtained by the maximum-entropy formalism. The present derivation thus provides an alternative approach to the inference problem which does not invoke Shannon's notion of missing information or entropy. The approach is more limited in scope than the one proposed by Jaynes, but has the advantage that it is objective and that the operational origin of the "given" expectation values is specified.

I. INTRODUCTION

Maximum entropy is a procedure,[1-3] stated in its most sweeping form by Jaynes,[1,2] for inducing an unknown probability distribution given only partial data. The procedure is finding an increasing number of incisive applications in different branches of science.[4] It is therefore worthwhile to clarify the rationale for its use. Jaynes' proposal was based on maximizing Shannon's[5] measure of missing information subject to given expectation values (or, in general, given constraints). There are therefore many who regard the procedure as subjective in that it relies on the notion of missing information. A more technical objection is that the origin of the constraining conditions is not specified. Hence they need not be the results of observations. The resulting distribution reflects therefore the "state of knowledge" of the observer rather than necessarily a statement about nature. The purpose of this paper is to discuss three alternative approaches to the problem of inference. All three lead to the same operational procedure as the usual maximum-entropy method, but none invokes Shannon's notion of missing information.

The technical problem under consideration can be stated (for the simplest case of a discrete distribution) as follows: Let $i=1,\ldots,n$ be a set of n mutually exclusive and together exhaustive alternatives. In view of the typical applications in physics we refer to these n alternatives as states. Let $\{A_r\}$ be a set of m (linearly independent, see below) variables defined on these states, A_r obtaining the value A_{ri} on the state i. Given the m expectation values $\langle A_r\rangle$,

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

a normalized probability distribution, $p_1,\ldots,p_n$, is required which fulfills (1). Inference is necessary whenever $m<n-1$ so that the m given expectation values do not determine, via (1), a unique distribution. One requires therefore a procedure [or an algorithm, denoted here by (a)] which will select a single probability distribution from the set of all normalized distributions which are consistent with (1).

In the maximum-entropy procedure the (unique) distribution is selected by maximizing the entropy, $-\sum_i p_i\ln(p_i/g_i)$ (g_i is the multiplicity of the state i), subject to normalization $\sum_i p_i=1$ and the m additional constraints given by (1). The solution is[1-3]

$$p_i=g_i\exp\left[-\sum_{r=1}^{m}\lambda_r A_{ri}\right]/Z(\lambda_1,\ldots,\lambda_m)\,,$$
$$Z(\lambda_1,\ldots,\lambda_m)=\sum_i g_i\exp\left[-\sum_{r=1}^{m}\lambda_r A_{ri}\right]\,. \tag{2}$$

The m Lagrange multipliers, $\lambda_1,\ldots,\lambda_m$ are determined by the condition that the solution (2) satisfy the m constraints (1). The resulting set of implicit equations can be written as

$$\langle A_r\rangle=-\partial\ln Z(\lambda_1,\ldots,\lambda_m)/\partial\lambda_r,\quad r=1,\ldots,m\,. \tag{3}$$

The three approaches reviewed below will lead to the very same solution. Only the rationale will be different.

A basic requirement of a scientific experiment is reproducibility (of the experiment, not the results). The present derivation is therefore limited to such experiments that can be independently repeated a finite (but not necessarily large) number of times. It is thus narrower in scope than the method proposed by Jaynes. The constraints used here are in the form of expectation values $\langle A_r\rangle$ which we assume to be measured in terms of the sample averages

$$\bar{A}_r=\sum_{i=1}^{n}A_{ri}N_i/N\,,$$
$$\sum_{i=1}^{n}N_i=N\,. \tag{4}$$

Here N_i is the number of times the state i has been realized in N independent repetitions of the experiment. The present approach has therefore the advantage that it is free of any reference to the state of knowledge of the observer and that the origin of the "given" constraints is spelled out. In what follows we shall use relations between the (measured) sample averages and the (theoretical) expectation values to induce the desired probability distribution.

In the first approach, the requirement that the given averages represent possible outcomes of a reproducible experiment, leads to a confrontation between inferences for two distributions: the probability p_i for entering the state $i=1,\ldots,n$ and the probability $P_{\underline{N}}$ for obtaining (in N independent repetitions of the elementary experiment) the set of occupation numbers $\underline{N}=(N_1,\ldots,N_n)$, where $\sum_{i=1}^{n}N_i=N$. The latter can be regarded either as a probability distribution for an elementary experiment in (the much larger) occupation number space, or as a compound (multinomial) distribution determined by the elementary distribution p_i. Comparison between the two points of view leads to a consistency condition which must be satisfied by any algorithm (a) for inducing a probability distribution from given averages. This consistency condition together with the requirement that the algorithm (a) treats all problems uniformly (regardless of the dimension of the probability space), determine the algorithm uniquely.

The second approach selects (a) as the algorithm leading to the most stable inference from the data $\overline{A}_r$. That is, among all possible normalized distributions p_i consistent with (1), the distribution which is least sensitive to statistical errors in the data is chosen.

The third approach is the oldest one and, unfortunately, perhaps the least meaningful to physicists, namely, that of sufficient statistics. Loosely speaking, a function $T(x_1,\ldots,x_N)$ is called "sufficient statistic for the parameter θ'' in the distribution $p(x\mid\theta)$, if all the information conveyed by the sample $x_1,\ldots,x_N$ about the unknown value of the parameter θ can be summarized by the single number $T(x_1,\ldots,x_N)$. Returning to our inference problem, a choice of an algorithm (a) determines $p_i=p_i(\langle A_1\rangle,\ldots,\langle A_m\rangle)$ as a known function of the (unknown) m values of the parameters $\langle A_r\rangle$. Demanding that the m sample averages

$$\overline{A}_r=\sum_{i=1}^{N}A_{ri}/N$$

serve (together) as sufficient statistics for the m parameters $\langle A_1\rangle,\ldots,\langle A_m\rangle$ is enough to identify p_i as the maximum-entropy solution. Because of our notion of a reproducible experiment we are, however, able to obtain a stronger version of the more familiar results and, specifically, our inferred distribution has a unique functional form.

Since all the approaches above lead to the maximum-entropy solution and none uses the concept of entropy, we can reverse the usual argument to identify $-\sum_i p_i\ln(p_i/g_i)$ as that unique functional of the distribution which attains its maximal value [subject to the constraints (1)] for the "correct" inference. Furthermore, accepting the reasoning given by Jaynes,[1,2] the entropy $-\sum_i p_i\ln(p_i/g_i)$ must be interpreted as the amount of missing information in a situation characterized by a probability distribution $\{p_i\}$.[6]

II. CONSISTENT UNIFORM INFERENCE

Let an experiment with n possible (mutually exclusive and collectively exhaustive) outcomes be repeated N times. We then form the sample averages $\overline{A}_r$ defined by (4). Our requirement[7] is that averaging the sample average over all possible occupation numbers N_i [cf. Eq. (4)] should give the expectation values $\langle A_r\rangle$ over the probabilities p_i.

Assume that we already have an algorithm (a) for inferring the elementary probabilities $p_1,\ldots,p_n$. Then the probability $P_{\underline{N}}$ for observing a set of occupation numbers $\underline{N}=(N_1,\ldots,N_n)$, where $\sum_{i=1}^{n}N_i=N$, is determined by the multinomial distribution[8]

$$P_{\underline{N}}=g_{\underline{N}}p_1^{N_1}p_2^{N_2}\cdots p_n^{N_n}\ . \tag{5a}$$

Here

$$g_{\underline{N}}=N!/\prod_{i=1}^{n}N_i! \tag{5b}$$

is the multiplicity (or degeneracy factor) of the compound state $\underline{N}=(N_1,\ldots,N_n)$. Introducing the variables B_r defined on the occupation number space

$$B_{r\underline{N}}=\sum_{i=1}^{n}N_iA_{ri},\quad r=1,\ldots,m \tag{6}$$

with expectation values

$$\begin{aligned}\langle B_r\rangle&=\sum_{\underline{N}}P_{\underline{N}}B_{r\underline{N}}=\sum_{\underline{N}}P_{\underline{N}}\sum_{i=1}^{n}N_iA_{ri}\\&=N\sum_{i=1}^{n}p_iA_{ri}=N\langle A_r\rangle\ ,\end{aligned} \tag{7}$$

the result (5) can be regarded as a solution to the inference problem (7) in the (much larger) occupation number space. The number of states in the space of occupation numbers is $l=\binom{N+n-1}{n-1}$. But using the same data, namely, the m expectation values $\langle B_r\rangle=N\langle A_r\rangle$ of the variables B_r, we could apply the algorithm (a) directly in the occupation number space to infer a probability distribution $Q_{\underline{N}}$ agreeing with the data. The algorithm (a) will be called consistent if the two alternative routes to the distribution of outcomes in N independent repetitions of the experiment lead to the same distribution. That is,

$$Q_{\underline{N}}=P_{\underline{N}}\ . \tag{8}$$

We now show that the consistency condition (8), together with the requirement that the algorithm (a) treats all problems uniformly, regardless of the dimension of the probability space involved, suffice to determine the algorithm.

Without loss of generality, we may assume that the m variables $A_r=(A_{ri})$ considered as n-dimensional vectors together with the normalization vector A_0 $(A_{0i}=1)$, are linearly independent. This can always be achieved by reducing m. Completing these $m+1$ vectors to a full

base of the n-dimensional space, we can always expand

$$\ln p_i = -\sum_{r=0}^{n-1} \lambda_r A_{ri} \ . \tag{9}$$

Given a complement $A_{m+1}, \ldots, A_{n-1}$ and the expansion coefficients $\lambda_{m+1}, \ldots, \lambda_{n-1}$, the values of $\lambda_0, \ldots, \lambda_m$ are uniquely[9] determined by the constraints (1). Thus, a choice of the complement $\{A_s\}$ and the expansion coefficients $\{\lambda_s\}$, $s=m+1, \ldots, n-1$ defines an algorithm (a). Similarly, the variables

$$B_r = (B_{r\underline{N}}) = \left[\sum_{i=1}^{n} N_i A_{ri} \right], \quad r=0, \ldots, n-1 \ ,$$

considered as vectors in $l=\binom{N+n-1}{N-1}$ dimensional space, can be complemented to a full base of the l-dimensional space. [It is easy to check that linear independency of $A_0, \ldots, A_{n-1}$ as vectors in n-dimensional space implies linear independency of $B_0, \ldots, B_{n-1}$ in the l-dimensional space. Indeed,

$$\sum_{r=0}^{n-1} \alpha_r B_{r\underline{N}} = \sum_{i=1}^{n} N_i \sum_{r=0}^{n-1} \alpha_r A_{ri}$$

for each $\underline{N}$, implies, in particular, for

$$\underline{N} = (0, \ldots, 0, N, 0, \ldots, 0) \ ,$$

that

$$N \sum_{r=0}^{n-1} \alpha_r A_{ri} = 0 \ .$$

Hence, by the linear independency of $A_0, \ldots, A_{n-1}$ we have $\alpha_r = 0$ for $r=0, \ldots, n-1$.] Expanding

$$\ln \frac{Q_{\underline{N}}}{g_{\underline{N}}} = -\sum_{r=0}^{l-1} \mu_r B_{r\underline{N}} \ , \tag{10}$$

and using the consistency condition (8) together with Eqs. (5) and (9), we obtain

$$\sum_{r=0}^{l-1} \mu_r B_{r\underline{N}} = \sum_{i=1}^{n} N_i \sum_{r=0}^{n-1} \lambda_r A_{ri} = \sum_{r=0}^{n-1} \lambda_r \sum_{i=1}^{n} N_i A_{ri} = \sum_{r=0}^{n-1} \lambda_r B_{r\underline{N}} \ . \tag{11}$$

But the vectors B_r, $r=0, \ldots, l-1$ were chosen to be linearly independent, hence

$$\mu_r = \lambda_r \text{ for } r=0, \ldots, n-1$$
$$\text{and} \tag{12}$$
$$\mu_r = 0 \text{ for } r=n, \ldots, l-1 \ .$$

Up to now we have used only the consistency condition (8). If, in addition, we require that the algorithm (a) treats all inference problems uniformly as problems in elementary spaces regardless of the dimensions of the spaces involved, we must have (since n is not an input and is unknown to the problem in the l-dimensional space) $\mu_r = 0$ for $r=m+1, \ldots, l-1$. Hence, by Eq. (12), $\lambda_r = 0$ for $r=m+1, \ldots, n-1$. This completes the identification of the consistent uniform algorithm with the maximum-entropy algorithm.[6]

The fact that the maximum-entropy algorithm is consistent in the sense of Eq. (8) has been demonstrated by Levine.[10] Indeed, the maximum-entropy solution to the problem (7) is

$$Q_{\underline{N}} = g_{\underline{N}} \exp\left[-\sum_{r=0}^{m} \lambda_r B_{r\underline{N}} \right] , \tag{13}$$

where

$$\lambda_0 N = \lambda_0 B_{0\underline{N}} = \ln\left[\sum_{\underline{N}} g_{\underline{N}} \exp\left[-\sum_{r=1}^{m} \lambda_r \sum_{i=1}^{n} N_i A_{ri} \right] \right] \tag{14}$$

and the other m Lagrange multipliers λ_r are determined by solving

$$-\frac{\partial(\lambda_0 N)}{\partial \lambda_r} = \langle B_r \rangle = N \langle A_r \rangle, \quad r=1, \ldots, m \ . \tag{15}$$

The solution (13) can be rewritten as

$$Q_{\underline{N}} = g_{\underline{N}} \prod_{i=1}^{n} p_i^{N_i} \ , \tag{16}$$

where

$$p_i = \exp\left[-\lambda_0 - \sum_{r=1}^{m} \lambda_r A_{ri} \right] . \tag{17}$$

Using the identity

$$\sum_{\underline{N}} g_{\underline{N}} \prod_{i=1}^{n} (p_i e^{\lambda_0})^{N_i} = (p_1 e^{\lambda_0} + \cdots + p_n e^{\lambda_0})^N = \left[\sum_{i=1}^{n} p_i \right]^N e^{\lambda_0 N} \ , \tag{18}$$

together with Eqs. (14) and (15), the distribution (17) is identified as the maximum-entropy solution to problem (1).

III. MOST STABLE INFERENCE

We begin this section by establishing an inequality satisfied by the class C of normalized distributions consistent with (1). Interpretation of the result as a stability criterion will lead us to identify the distribution least sensitive to statistical errors in the input $\langle A_r \rangle$, as the maximum-entropy distribution. Conversely, the maximum-entropy distribution using the "natural representation" for the constraints (to be defined below), will be shown to be the most stable distribution in the class C.

Given an algorithm (a) for selecting a probability distribution from the class C, the chosen distribution p_i becomes a function of the parameters $\langle A_1 \rangle, \ldots, \langle A_m \rangle$. This function satisfies

$$1 = \sum_i p_i \text{ and hence } 0 = \sum_i \frac{\partial p_i}{\partial \langle A_r \rangle} \ , \tag{19}$$

and

$$\langle A_r \rangle = \sum_i p_i A_{ri} \text{ yielding } 1 = \sum_i \frac{\partial p_i}{\partial \langle A_r \rangle} A_{ri} \ . \tag{20}$$

Rewriting (20) with the help of (19) as

$$1=\sum_i \frac{\partial p_i}{\partial\langle A_r\rangle}A_{ri}=\sum_i \frac{\partial p_i}{\partial\langle A_r\rangle}(A_{ri}-\langle A_r\rangle)$$
$$=\sum_i \sqrt{p_i}\frac{\partial \ln p_i}{\partial\langle A_r\rangle}\sqrt{p_i}(A_{ri}-\langle A_r\rangle)\ , \tag{20'}$$

we have, by the Cauchy-Schwarz inequality,

$$1\le\left\langle\left[\frac{\partial \ln p_i}{\partial\langle A_r\rangle}\right]^2\right\rangle^{1/2}\Delta A_r\ , \tag{21}$$

where

$$\Delta A_r=\langle(A_{ri}-\langle A_r\rangle)^2\rangle^{1/2}$$
$$=[\mathrm{Var}(A_r)]^{1/2},\quad r=1,\ldots,m\ . \tag{22}$$

Equality in Eq. (21) holds if and only if the vector $(\partial \ln p_i/\partial\langle A_r\rangle)$ is proportional to the vector $(A_{ri}-\langle A_r\rangle)$, that is

$$\partial \ln p_i/\partial\langle A_r\rangle=\alpha_r(\langle A_1\rangle,\ldots,\langle A_m\rangle)(A_{ri}-\langle A_r\rangle)$$
$$\text{for } i=1,\ldots,n\ . \tag{23}$$

Equation (21) is a special case of the Rao-Cramer inequality well known in statistics.[8] The case of a single constraint ($m=1$) has been discussed by Alhassid and Levine.[11] Confining our interest to the situation where equality prevails [Eq. (23)], we shall now show the following.

(a) The covariance matrix

$$C_{rs}=\langle(A_{ri}-\langle A_r\rangle)(A_{si}-\langle A_s\rangle)\rangle$$

is diagonal.

Indeed, multiplying Eq. (23) by $p_i(A_{si}-\langle A_s\rangle)$ and summing over all states i, we have by Eqs. (19) and (20)

$$\alpha_r C_{rs}=\sum_i \frac{\partial p_i}{\partial\langle A_r\rangle}(A_{si}-\langle A_s\rangle)=\frac{\partial\langle A_s\rangle}{\partial\langle A_r\rangle}=\delta_{rs}\ . \tag{24}$$

In particular,

$$\alpha_r=1/C_{rr}=1/\mathrm{Var}(A_r)\ . \tag{25}$$

(b) The proportionality constant α_r depends only on $\langle A_r\rangle$.

Taking the derivative of Eq. (23) by $\langle A_s\rangle$, we have

$$\frac{\partial^2 \ln p_i}{\partial\langle A_s\rangle\,\partial\langle A_r\rangle}=\frac{\partial\alpha_r}{\partial\langle A_s\rangle}(A_{ri}-\langle A_r\rangle)-\alpha_r\delta_{rs}$$
$$=\frac{\partial^2 \ln p_i}{\partial\langle A_r\rangle\partial\langle A_s\rangle}$$
$$=\frac{\partial\alpha_s}{\partial\langle A_r\rangle}(A_{si}-\langle A_s\rangle)-\alpha_s\delta_{rs}\ .$$

Hence, multiplying both sides by $p_i(A_{ri}-\langle A_r\rangle)$ and summing over all states i, we obtain [using (a)]

$$\frac{\partial\alpha_r}{\partial\langle A_s\rangle}C_{rr}=\frac{\partial\alpha_s}{\partial\langle A_r\rangle}C_{rs}=0 \text{ for } s\neq r\ .$$

(c) The distribution p_i is the maximum entropy distribution.

Integrating Eq. (23) once, we have

$$\ln p_i=A_{ri}\int^{\langle A_r\rangle}\alpha_r d\langle A_r\rangle$$
$$-\int^{\langle A_r\rangle}\alpha_r\langle A_r\rangle d\langle A_r\rangle+h_i\ ,$$

where h_i is independent of $\langle A_r\rangle$. Taking the derivative of the last equation by $\langle A_s\rangle\neq\langle A_r\rangle$, we secure

$$\partial h_i/\partial\langle A_s\rangle=\partial \ln p_i/\partial\langle A_s\rangle$$
$$=\alpha_s(\langle A_s\rangle)(A_{si}-\langle A_s\rangle)\ .$$

Integrating again over $\langle A_s\rangle$, we obtain

$$h_i=A_{si}\int^{\langle A_s\rangle}d\langle A_s\rangle-\int^{\langle A_s\rangle}\alpha_s\langle A_s\rangle+k_i\ ,$$

where k_i is independent of $\langle A_r\rangle$ and $\langle A_s\rangle$. Continuing in the same fashion we finally obtain

$$\ln p_i=\sum_{r=1}^m A_{ri}\int^{\langle A_r\rangle}\alpha_r d\langle A_r\rangle$$
$$-\sum_{r=1}^m\int^{\langle A_r\rangle}\alpha_r\langle A_r\rangle d\langle A_r\rangle+l_i\ ,$$

where l_i is independent of the data $\langle A_1\rangle,\ldots,\langle A_m\rangle$. Using the notation

$$\lambda_r=-\int^{\langle A_r\rangle}\alpha_r d\langle A_r\rangle\ ,$$
$$\ln Z=\sum_{r=1}^m\int^{\langle A_r\rangle}\alpha_r\langle A_r\rangle d\langle A_r\rangle\ , \tag{26}$$

and

$$\ln g_i=l_i\ ,$$

we have

$$p_i=g_i\exp\left[-\sum_{r=1}^m\lambda_r A_{ri}\right]/Z\ . \tag{27}$$

It is easy to check [using the chain rule, the normalization $\sum_i p_i=1$ and Eqs. (24) and (25)] that $\ln Z$, defined by Eq. (26), satisfies $-\partial \ln Z/\partial\lambda_r=\langle A_r\rangle$, as it should. This completes the proof of our claim that equality in Eq. (21) implies that p_i is the maximum-entropy distribution. The converse is not true unless the covariance matrix $\underline{C}=(C_{rs})$ happens to be diagonal. But $\underline{C}$ is a real symmetric matrix. Hence, there exists a real orthogonal matrix $\underline{O}$ such that $\underline{C}'=\underline{O}\underline{C}\underline{\tilde{O}}$ is diagonal. Transforming to the "natural representation"

$$A'_{ri}=\sum_k O_{rk}A_{ki} \tag{28}$$

and

$$\lambda'_r=\sum_l O_{rl}\lambda_l\ , \tag{29}$$

we have

$$\sum_r\lambda'_r A'_{ri}=\sum_r\lambda_r A_{ri}\ . \tag{30}$$

Thus the probability p_i in Eq. (27) remains unchanged, while the covariance matrix becomes diagonal,

$$C'_{rs}=\sum_i p_i(A'_{ri}-\langle A'_r\rangle)(A'_{si}-\langle A'_s\rangle)$$
$$=\sum_{k,l} O_{rk}C_{kl}O_{sl}=(\underline{O}\underline{C}\underline{\tilde{O}})_{rs}\ . \tag{31}$$

We now turn to the interpretation of the inequality (21). If $\langle A_r\rangle$ is estimated by the sample average

$$\bar{A}_r=\frac{1}{N}\sum_{i=1}^{N} A_{ri}\ ,$$

then the statistical error is proportional to the width ΔA_r, that is

$$\delta\langle A_r\rangle=\delta\bar{A}_r\propto\frac{1}{\sqrt{N}}\Delta A_r\ . \tag{32}$$

More generally, if the estimated error in $\langle A_r\rangle$ is known to be proportional to ΔA_r, that is $\delta\langle A_r\rangle=\beta_r\Delta A_r$, Eq. (21) can be rewritten as

$$\beta_r\le\left\langle\left[\frac{\delta_r p_i}{p_i}\right]^2\right\rangle^{1/2}\ , \tag{33}$$

where

$$\delta_r p_i=\frac{\partial p_i}{\partial\langle A_r\rangle}\beta_r\Delta A_r=\frac{\partial p_i}{\partial\langle A_r\rangle}\delta\langle A_r\rangle \tag{34}$$

is the (first-order) deviation of p_i due to the estimated error $\delta\langle A_r\rangle$ in $\langle A_r\rangle$. The right-hand side of Eq. (33) is naturally interpreted as a measure of the sensitivity of the distribution p_i to (statistical) errors in the data $\langle A_r\rangle$.[12] Recalling our discussion above, we conclude the following. Among all possible normalized distributions p_i consistent with (1), the least sensitive to (statistical) errors in the data is that of maximum entropy using the natural representation for the constraints. It is interesting to note that equality in Eq. (21) also assures the fulfillment of a necessary condition for p_i to be useful as a predictive tool. Indeed, if the inferred distribution p_i is to be used to predict an expectation value $\langle B\rangle$ for a variable B (not included in the data $\langle A_r\rangle$), we must have $(\delta\langle B\rangle/\Delta B)\ll 1$. Here $\delta\langle B\rangle$ and ΔB are the estimated error (due to errors $\delta\langle A_r\rangle$) and the estimated width of B. Using the first-order estimate

$$\delta_r\langle B\rangle=\frac{\partial\langle B\rangle}{\partial\langle A_r\rangle}\delta\langle A_r\rangle=\sum_{i=1}^{n}\frac{\partial p_i}{\partial\langle A_r\rangle}B_i\delta\langle A_r\rangle \tag{35}$$

and repeating the steps leading to Eq. (21), we obtain

$$\frac{\delta_r\langle B\rangle}{\Delta B}\le\frac{\delta\langle A_r\rangle}{\Delta A_r}\left[\left\langle\left[\frac{\partial\ln p_i}{\partial\langle A_r\rangle}\right]^2\right\rangle^{1/2}\Delta A_r\right]\ . \tag{36}$$

Equality in Eq. (21) now assures

$$\frac{\delta_r\langle B\rangle}{\Delta B}\le\frac{\delta\langle A_r\rangle}{\Delta A_r}=\beta_r\ . \tag{37}$$

By employing a large enough sample we can always secure $\beta_r\ll 1$.

The intuitive meaning of the results of this section is that the maximum-entropy distribution is as "spread out" as possible subject to the constraints. Hence small changes in the values of the constraints do not lead to appreciable changes in the distribution.

IV. SUFFICIENT STATISTICS INFERENCE

The concept of sufficient statistics has been introduced by Fisher[13] in the twenties for the problem of parameter estimation. In the Introduction we have loosely defined $T(x_1,\ldots,x_N)$ as a sufficient statistic for the parameter θ in the known distribution $p(x\mid\theta)$, if all the information concerning the unknown value of θ, conveyed by the sample $x_1,\ldots,x_N$, can be summarized by the single number $T(x_1,\ldots,x_N)$. We now have to be more precise. Given the sample distribution

$$P(x_1,\ldots,x_N\mid\theta)=\prod_{i=1}^{N}p(x_i\mid\theta)\ , \tag{38}$$

we can use Bayes' theorem to calculate the inverse distribution

$$F(\theta\mid x_1,\ldots,x_N)=P(x_1,\ldots,x_N\mid\theta)f_0(\theta)/\int Pf_0 d\theta \tag{39}$$

for the parameter θ. Here $f_0(\theta)$ is a "prior" distribution independent of the sample $x_1,\ldots,x_N$. Following Dynkin,[14] we call $T(x_1,\ldots,x_N)$ a sufficient statistic for the parameter θ if, for any prior $f_0(\theta)$, the inverse distribution $F(\theta\mid x_1,\ldots,x_N)$ depends on the sample $x_1,\ldots,x_N$ only through the value of $T(x_1,\ldots,x_N)$, that is

$$F(\theta\mid x_1,\ldots,x_N)=\phi(T(x_1,\ldots,x_N),\theta)f_0(\theta)\ . \tag{40}$$

With this definition it is easy to check that T is a sufficient statistic for the parameter θ if and only if the (direct) sample distribution $P(x_1,\ldots,x_N\mid\theta)$ factorizes as[15]

$$P(x_1,\ldots,x_N\mid\theta)=\prod_{i=1}^{N}p(x_i\mid\theta)$$
$$=k(x_1,\ldots,x_N)l(T(x_1,\ldots,x_N),\theta) \tag{41}$$

with k and l independent of the sample. A celebrated theorem by Pitman,[16] Koopman,[17] and Darmois[18] now states that if $p(x\mid\theta)$ is known to have a sufficient statistic T, then necessarily p is of the exponential form

$$p(x\mid\theta)=m(x)e^{-\lambda(\theta)B(x)}/Z(\lambda(\theta))\ . \tag{42}$$

[Conversely, if p is of the form (42), then, by Eqs. (38) and (41), $T=\sum_{i=1}^{N}B(x_i)$ is a sufficient statistic for θ]. The definition (40) of a sufficient statistic for the parameter θ can be naturally generalized to define a set of sufficient statistics $T_1,T_2,\ldots$ for the set of parameters $\theta_1,\theta_2,\ldots$. In the proof given below for the Pitman-Koopman-Darmois theorem, we shall follow essentially the work of Pitman.[16]

Let $T(x_1,\ldots,x_N)$ be a sufficient statistic for the parameter θ in the distribution $p(x\mid\theta)$. Then the sample distribution $P(x_1,\ldots,x_N\mid\theta)$ factorizes as per Eq. (41). Differentiating Eq. (41) with respect to the parameter θ, we have for each admissible θ,

$$\frac{\partial \ln P}{\partial \theta}=\sum_{i=1}^{N}\frac{\partial \ln p(x_i \mid \theta)}{\partial \theta}=\frac{\partial I(T,\theta)}{\partial \theta}\equiv h(T,\theta)\ . \quad (43)$$

In particular, substituting $\theta=\theta_0$ we can solve Eq. (43) for T to obtain

$$T(x_1,\ldots,x_N)=f(\bar{B}(x_1,\ldots,x_N))\ , \quad (44)$$

where

$$\bar{B}(x_1,\ldots,x_N)=\sum_{i=1}^{N}B(x_i)\equiv\sum_{i=1}^{N}\frac{\partial \ln p(x_i \mid \theta_0)}{\partial \theta}\ . \quad (45)$$

Substituting the solution (44) in Eq. (43), we have

$$\sum_i \frac{\partial \ln p(x_i \mid \theta)}{\partial \theta}=h(T,\theta)\equiv H(\bar{B},\theta)\ . \quad (46)$$

Hence, by differentiation with respect to x_i,

$$\frac{\partial^2 \ln p(x_i \mid \theta)}{\partial x_i \partial \theta}=\frac{\partial H}{\partial \bar{B}}(\bar{B},\theta)\frac{dB(x_i)}{dx_i}\ . \quad (47)$$

Now, for a given θ, the left-hand side of (47) depends only on x_i. Hence, the right-hand side of (47) can depend only on x_i and $\partial H/\partial \bar{B}$, being a symmetric function of $x_1,\ldots,x_N$, must be independent of $x_1,\ldots,x_N$. Thus

$$\frac{\partial^2 \ln p(x_i \mid \theta)}{\partial x_i \partial \theta}=\alpha(\theta)\frac{dB(x_i)}{dx_i}\ , \quad (48)$$

and by integration

$$\frac{\partial \ln p}{\partial \theta}=\alpha(\theta)B(x_i)+\beta(\theta)\ ,$$
$$\ln p(x_i \mid \theta)=\alpha_1(\theta)B(x_i)+\beta_1(\theta)+\gamma(x_i)\ . \quad (49)$$

Taking the exponential of the last equation, we recover Eq. (42). Note that the Pitman-Koopman-Darmois theorem *does not* establish the specific form of $B(x)$ in the exponent of the distribution (42). Since the form of $p(x \mid \theta)$ is not known, Eqs. (44) and (45) determine $B(x)$, via the known sufficient statistic $T(x)$, only up to an unknown (one-to-one function) f. If, however, $T=\sum A(x_i)$ is an *additive* sufficient statistic, then, by Eq. (44),

$$\frac{dA(x_i)}{dx_i}=\frac{df}{d\bar{B}}(\bar{B})\frac{dB(x_i)}{dx_i} \quad (50)$$

and, by the reasoning leading to Eq. (48),

$$A(x_i)=\gamma B(x_i)+\delta\ , \quad (51)$$

where γ and δ are constants. The unknown constants γ and δ can be absorbed in the definition of the Lagrange multiplier $\lambda(\theta)$ and the partition function $Z(\lambda(\theta))$ appearing in the distribution (42), thus leaving only the density of states (or prior) $m(x)$ unspecified. The proof given above for the Pitman-Koopman-Darmois theorem can be easily extended to the multiparameter case.

Coming back to our inference problem (1), a choice of an algorithm (a) determines $p_i=p_i(\langle A_1\rangle,\ldots,\langle A_m\rangle)$ as a function of the parameters $\langle A_1\rangle,\ldots,\langle A_m\rangle$. Demanding that the sample averages $\bar{A}_r=\sum_{i=1}^{N}A_{ri}/N$ serve together as sufficient statistics for the parameters $\langle A_1\rangle,\ldots,\langle A_m\rangle$ is enough to identify p_i as the maximum-entropy distribution (2). Note that the degeneracy factors g_i [or the density of states $m(x)$ in the continuous case] are considered as known, being part of the specification of states for the inference problem.

V. CONCLUDING REMARKS

Three complementary points of view, each of which invokes the concept of a reproducible experiment, lead to the same, unique, procedure for inducing a probability distribution. The procedure is the one known as the maximum-entropy procedure. None of the three characterizations presented here invokes the concept of entropy nor that of information. Rather, the requirement that the experiment is reproducible is translated into relations between the measured sample averages and the expectation values for the (unknown) probability distribution. In the first approach we require that averaging (in a Gedanken experiment) the sample average over all possible samples yields the expectation value. In the second approach we use a particular (measured) value of the sample average for the (unknown) expectation value but recognize that in so doing there may be a statistical error (due to the finite size of the sample). The third approach stems from the assumption that the sample average is all that can be extracted from the observations regarding the expectation value. These derivations should help remove the "subjective" character that is sometimes associated with the procedure of maximum entropy. They also clearly show that physics comes in not by the choice of the procedure (which is universal) but by the choice of the variables whose average is the input to the method.

ACKNOWLEDGMENTS

We are indebted to D. Shalitin for a critical discussion of the ideas presented in this work. This work was partly supported by the U.S. Office of Naval Research. The Fritz Haber Research Center is supported by the Minerva Gesellschaft für die Forschung, mbH, München, Federal Republic of Germany.

---

*Also at Racah Institute of Physics, The Hebrew University of Jerusalem, Jerusalem 91904, Israel.

[1]E. T. Jaynes, Phys. Rev. **106**, 620 (1957); **108**, 171 (1957).

[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. Rep. **62**, 175 (1980). For earlier work see W. M. Elsasser, Phys. Rev. **52**, 987 (1937); R. L. Stratonovich, Zh. Eksp. Teor. Fiz. **28**, 547 (1955) [Sov. Phys.—JETP **1**, 426 (1955)]; U. Fano, Rev. Mod. Phys. **29**, 74 (1957). See also E. H. Wichmann, J. Math. Phys. **4**, 884 (1963).

[4]For applications to elementary particle, nuclear, atomic, 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); T. Aberg, A. Blomberg, J. Tulkki, and O. Goscinski, Phys. Rev. Lett. **52**, 1207 (1984); R. D. Levine, Adv. Chem. Phys. **47**, 239 (1981). See also R. K. Nesbet in *Theoretical Chemistry: Theory of Scattering,* edited by D. Henderson (Academic, New York, 1981). For applications to image processing, see, for example, R. Kikuchi and B. H. Soffer, J. Opt. Soc. Am. **67**, 1656 (1978); S. F. Gull and G. J. Daniell, Nature (London) **272**, 686 (1978); B. R. Frieden, *Probability, Statistical Optics and Data Testing* (Springer, Berlin, 1983). For line-shape problems, see, for example, J. G. Powles, Mol. Phys. **48**, 1083 (1983), B. J. Berne and G. D. Harp, Phys. Rev. A **2**, 2514 (1970).

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

[6]An axiomatic derivation of the procedure of maximum entropy has been provided before [J. Shore and R. Johnson, IEEE Trans. Inf. Theory **IT-26**, 26 (1980) ; **IT-29**, 942 (1983)]. It should be noted however that the present approach is inherently different in that we do not assume that the solution is obtained by maximizing a functional of the distribution. If we do use this assumption then the proof in Sec. II can be considerably simplified and, in particular, we do not need to impose uniformity. Another difference is the explicit use of sample averages and expectation values.

[7]Y. Tikochinsky, N. Z. Tishby, and R. D. Levine, Phys. Rev. Lett. **52**, 1357 (1984).

[8]See, for example, R. V. Hogg and A. T. Craig, *Introduction to Mathematical Statistics* (Macmillan, New York, 1978).

[9]Y. Alhassid, N. Agmon, and R. D. Levine, Chem. Phys. Lett. **53**, 22 (1978).

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

[11]Y. Alhassid and R. D. Levine, Chem. Phys. Lett. **73**, 16 (1980).

[12]Note that $\langle(\delta p/p)^2\rangle$ is $-\delta^2 S$, the second variation of the entropy and is closely related to the concept of fluctuations.

[13]R. A. Fisher, Philos. Trans. R. Soc. London, Ser. A **222**, 309 (1922).

[14]E. B. Dynkin, Usp. Mat. Nauk. N. S. **6**, 68 (1951), English translation: Selected Transl. Math. Stat. Prob. (I.M.S.) **1**, 17 (1961).

[15]This result is known in statistics as the Neyman factorization theorem; see, e.g., Ref. 8.

[16]E.J.G. Pitman, Proc. Cambridge Philos. Soc. **32**, 567 (1936).

[17]B. O. Koopman, Trans. Am. Math. Soc. **39**, 399 (1936).

[18]G. Darmois, Rev. Inst. Int. Stat. **13**, 9 (1945).