An information theoretical approach to inversion problems

J. Phys. A: Math. Gen., **13** (1980) 91–108. Printed in Great Britain

An information theoretical approach to inversion problems†

R D Levine

Department of Physical Chemistry and Institute for Advanced Studies, The Hebrew University, Jerusalem, Israel

Received 18 December 1978, in final form 20 April 1979

Abstract. An inversion procedure which provides the most conservative inference for an unknown function in terms of partial data is discussed on the basis of information theoretic considerations. The method is based on the procedure of maximal entropy, but is not limited to the estimation of unknown probabilities. Rather, inductive inferences can be drawn regarding the values of general (if necessary, dimension-bearing) variables. The solution of an inversion problem using data linear in the unknown function is discussed in detail and explicit results are obtained. For every class of problems with common symmetry properties, the general algorithm can be reduced to a more direct procedure. When the data consist of average values for an unknown distribution, the general approach is in the spirit of the Darwin–Fowler method, while the reduced route is the procedure of maximal entropy ('method of most probable distribution') as usually employed in statistical mechanics. Other classes of problems discussed include the representation of an unknown function in a complete orthonormal basis using as input a partial set of expansion coefficients, and the inference of line shapes and power spectra.

1. Introduction

An inversion procedure is required whenever the available measurements depend on the value of a function of interest at more than one point. A well-known example is the determination of the interparticle potential from scattering (or transport) measurements. Another familiar example is in statistical mechanics, where the probabilities of occupation of the different quantum states are determined, say, from the given average energy of the system. There is, of course, an essential difference between our usual approaches to these two (and similar) problems. In the former we require extensive data before we proceed, while in the latter a seemingly insufficient characterisation of the unknown probabilities is deemed to be an appropriate starting point. The purpose of this paper is to advocate the adoption of the guiding principles of the methods made familiar by statistical mechanics to other inversion problems where a similar need arises. The present approach is similar in spirit to the Darwin–Fowler (1922) method, but for reasons which will become obvious it uses Lagrange multipliers rather than the steepest-descent technique. On the other hand, it is very much motivated by the information theoretic approach (Jaynes 1957, 1963, Tribus 1961) to statistical mechanics.

† Work supported by the Office of Naval Research.

One can argue that the methods of statistical mechanics are particularly geared to the physics of the problem and so cannot be applied indiscriminately elsewhere. Maybe so, but the prospect of an inversion procedure using incomplete data is perhaps of sufficient interest to warrant a closer examination. One can also argue that the essential difference between the two problems above is not one of physics but of a methodological point of view; whereas one wants to deduce the potential, one is quite accustomed to inducing a probability distribution. In other words, one can regard statistical reasoning as an application of the methods of inductive inference (Jaynes 1957, 1963, Tribus 1961, 1969, Levine and Tribus 1978). There is then no inherent objection to employing similar methods for other problems involving only a partial characterisation of the variable of interest.

Section 2 defines the problem in a technical sense, and shows that a solution of the inversion problem is possible. It does so by introducing a 'reciprocal' basis set, whose dimension equals the number of independent pieces of data (the 'constraints') that are available. Using this basis it is always possible to invert the implicit equations which define the variable of interest. The procedure of maximal entropy as applied to the inversion problem is discussed in § 3. Section 4 provides explicit discussion of four classes of problems, whose common denominator is that the available data are linear in the variable of interest. The discussion is also used to offer complementary interpretations of the formalism. In particular it is shown that (as in statistical mechanics) it is not necessary to go back to the full formalism every time a new problem is considered. Instead, one can use a simpler form of the entropy (or of the partition function) which is suitable for the class of situations of which the particular problem is a member and proceed from there. Work is in progress on various practical aspects of the formalism (as is discussed in § 4.3) and on concrete applications of the inversion procedure, with special reference to the determination of potentials from scattering data, in collaboration with Professor R B Gerber.

2. General considerations

Consider a variable u which assumes the value u_i at the point† i, $i = 1, 2, \ldots, N$. The purpose of the inversion procedure is to determine the set of values u_i. The data available are the values of different functions‡ of u,

$$A_r = f_r(\boldsymbol{u}), \qquad r = 1, 2, \ldots, M, \quad M < N. \tag{1}$$

The vector notation for the argument of f_r serves as a reminder that $f_r(\boldsymbol{u})$ will in general depend on the value of u at more than one point. The simplest example is that of a linear condition

$$A_r = \sum_{i=1}^{N} a_{ri} u_i, \tag{2}$$

where the matrix $\boldsymbol{a}$ is given.

† To keep the mathematics simple we shall (if necessary) proceed to the limit where i is continuous only at the final stage of the procedure. Otherwise one needs to interpret various derivatives as functional derivatives.

‡ To ensure the uniqueness of the procedure these conditions need to limit the range of possible solutions to a convex set, so that a linear combination of possible solutions is also a solution. If they do not, then the procedure may be applied by regarding $f_r(\boldsymbol{u})$ itself as the unknown function.

The problem is that the range of r is smaller (or even very much smaller) than that of i. Hence the M conditions, (1), do not in themselves suffice to determine the u_is uniquely. The conditions do however constrain the range (which we assume to be convex) of possible solutions, and the problem is to further narrow down the choice to some particular vector $\boldsymbol{y}$ which will be the 'best' inference for the unknown $\boldsymbol{u}$.

2.1. The reciprocal basis

That a solution to the inversion problem is not out of the question is suggested by the following argument. Consider for simplicity the linear case

$$\boldsymbol{ay} = \boldsymbol{A}. \tag{3}$$

If $\boldsymbol{a}$ is a square matrix and has an inverse, one can solve for $\boldsymbol{y}$ directly:

$$\boldsymbol{y} = \boldsymbol{bA} \qquad \text{or} \qquad y_i = \sum_r b_{ir} A_r, \tag{4}$$

where

$$\boldsymbol{ab} = \boldsymbol{I}. \tag{5}$$

Now the whole point is that $\boldsymbol{a}$ is not square. One can still however satisfy (5) as follows. The components of $\boldsymbol{y}$ are to be determined using the given values of the A_rs. Different values may well yield a different inference for $\boldsymbol{y}$. Hence $\boldsymbol{y}$ is a function of $\boldsymbol{A}$, and one can define a rectangular matrix $\boldsymbol{b}$ by

$$b_{ir} = \partial y_i / \partial A_r. \tag{6}$$

Then $\boldsymbol{a}$ and $\boldsymbol{b}$ are conformable, and

$$(\boldsymbol{ab})_{sr} = \partial A_s / \partial A_r. \tag{7}$$

The condition

$$\partial A_r / \partial A_s = \delta_{r,s}, \tag{8}$$

where $\delta_{r,s}$ is the Kronecker delta, is simply the condition that the constraints are linearly independent, i.e. that there is no non-trivial set of numbers α_r such that

$$\alpha_0 = \sum_{r=1}^{M} \alpha_r f_r(\boldsymbol{y}). \tag{9}$$

If in practice this is not the case, then the number of constraints can be reduced until only a linearly independent set is left. It can be shown† that the constraints that have been so eliminated are non-informative in the sense that their inclusion would not change the inference regarding $\boldsymbol{y}$.

In the Appendix it is shown that the inversion procedure introduced in § 3 defines the reciprocal set even for non-linear constraints, and a simple example is worked out.

The observation that $\boldsymbol{y}$ is a homogeneous function of degree one in the A_rs is also of importance in showing that if $\boldsymbol{y}$ bears dimensions then it will scale properly whenever the units used to express the value of A_r are changed.

† The proof is based on showing that the distribution $P(\boldsymbol{x})$ which is introduced below and which is used to define $\boldsymbol{y}$ (cf (10)) is unchanged by the elimination of the linearly dependent constraints (see e.g. Alhassid *et al* 1978). There is no harm in including linearly dependent constraints; these only lengthen the algebra, but may offer other advantages, as is discussed in § 4.

2.2. *The distribution $P(\boldsymbol{x})$*

To determine the solution $\boldsymbol{y}$ explicitly, consider expanding it in the set $\{\boldsymbol{x}\}$ of all potential solutions of the inversion problem,

$$y_i = \sum_{\boldsymbol{x}} x_i P(\boldsymbol{x}) \tag{10a}$$

if y_i assumes only discrete values, or

$$y_i = \int \mathrm{d}\boldsymbol{x}\, x_i P(\boldsymbol{x}) \tag{10b}$$

if y_i is a continuous variable. The weight function $P(\boldsymbol{x})$ is unknown except that y_i must satisfy the M conditions (1), i.e.

$$A_r = \sum_i a_{ri} y_i = \sum_{\boldsymbol{x}} \Big(\sum_i a_{ri} x_i\Big) P(\boldsymbol{x}), \tag{11a}$$

or if y_i is continuous,

$$A_r = \int \mathrm{d}\boldsymbol{x} \Big(\sum_i a_{ri} x_i\Big) P(\boldsymbol{x}). \tag{11b}$$

For non-linear constraints, we take the conditions on $P(\boldsymbol{x})$ to be

$$A_r = \int \mathrm{d}\boldsymbol{x}\, f_r(\boldsymbol{x}) P(\boldsymbol{x}), \tag{12}$$

with an equivalent form for the discrete case.

The distribution $P(\boldsymbol{x})$ can be regarded as the weight of the vector $\boldsymbol{x}$, and definition (10) implies that $\boldsymbol{y}$ can be interpreted as the average over the distribution $P(\boldsymbol{x})$. The reason for defining $\boldsymbol{y}$ as the average rather than any other statistic of $P(\boldsymbol{x})$ is the familiar one that the average is the 'best' estimate in the sense that the expected square deviance of $\boldsymbol{y}$ from the exact solution $\boldsymbol{u}$ will be minimal. The present approach is thus similar to the point of view adopted in the Darwin–Fowler (1922, see also Schrödinger 1952) formulation of statistical mechanics. It does differ however in one essential detail, namely that we do not necessarily centre attention on the limit where the variance of $P(\boldsymbol{x})$ is very small.

2.3. *Information theory*

The vector $\boldsymbol{x}$ is a sequence of numbers x_i and can thus be regarded as a message in the sense of information theory. In his fundamental study, Shannon (1948, § 7, theorem 4 in particular) introduced entropy as a measure of choice between messages. Since we are looking for a distribution which represents the widest choice (and hence does not unduly favour any particular $\boldsymbol{x}$), we should take $P(\boldsymbol{x})$ as the distribution of maximal entropy which is consistent with the constraints. This is the method of inference advocated by Jaynes (1957, 1963) except that, unlike the situation in statistical mechanics but as in Shannon's work, the probability distribution so computed has no direct observational relevance. It is simply a measure of the range of vectors x and so determines the sharpness of the inference. The narrower the distribution, the more likely is $\boldsymbol{y}$ to be an accurate estimate. The spread in $P(\boldsymbol{x})$ is simply the price one pays for inverting with only a partial characterisation of the variable. As a consistency check, one should note however that if the constraints do suffice to specify a unique solution, then the same solution will be given by the maximum entropy formulation (since $P(\boldsymbol{x})$ is then a Kronecker (or Dirac) delta). Another consistency check is that the spread is typically reduced (and definitely does not increase, cf § 3.4) upon adding constraints.

3. The procedure of maximum entropy

The general considerations of § 2 call for the determination of the distribution $P(\boldsymbol{x})$ as the (unique, cf § 3.4) distribution of maximal entropy among all those distributions which are normalised,

$$\sum_{\boldsymbol{x}} P(\boldsymbol{x}) = 1, \tag{13}$$

and consistent with the constraints

$$A_r = \sum_{\boldsymbol{x}} f_r(\boldsymbol{x}) P(\boldsymbol{x}), \qquad r = 0, 1, \ldots, M. \tag{14}$$

In what follows we define $f_0(\boldsymbol{x}) = 1$ so that (14) represents $M+1$ constraints, with A_0 being the value of the normalisation sum.

While the solution of this constrained maximum problem is standard and has been explicitly considered (Tribus 1969), it does require some comments, particularly regarding the concept of the prior distribution.

3.1. Entropy and entropy deficiency

Given a set of messages, say $\{\boldsymbol{e}\}$, such that in the absence of any data they are all equally likely, the entropy of the distribution of messages is given by (Shannon 1948, Khinchin 1957)

$$S[\boldsymbol{e}] = -\sum_{\boldsymbol{e}} P(\boldsymbol{e}) \ln P(\boldsymbol{e}). \tag{15}$$

It follows from the convexity of the logarithmic function that $S[\boldsymbol{e}]$ is indeed maximal for a uniform distribution.

It is not unusual however to have to deal with messages where even in the absence of any specific constraints our prior knowledge dictates that not all messages are equally probable. The example discussed in detail in § 4.1 is that of repeated independent experiments (so-called 'Bernoulli trials'). Say the message $\boldsymbol{n}$ is the set of numbers $\{n_i\}$, where n_i is the number of times the ith outcome was observed to occur. Then even if all outcomes (in a single experiment) are equally probable, some sets of numbers $\{n_i\}$ are more likely than others. The reason, as is well known, is that many sets $\{n_i\}$ can be realised in more than one way. Specifically, if all outcomes are equally probable, the message $\boldsymbol{n}$ can be realised in $g(\boldsymbol{n})$,

$$g(\boldsymbol{n}) = n! \Big/ \prod_{i=1}^{N} n_i!, \tag{16}$$

ways, where $n = \Sigma_i n_i$ is the total number of trials and $\Sigma_{\boldsymbol{n}}\, g(\boldsymbol{n}) = 2^n$ is the total number of distinct sequences of outcomes. In terms of the probability $P(\boldsymbol{n})$ of a particular set $\{n_i\}$,

$P(\boldsymbol{e})$ is thus given by

$$P(\boldsymbol{e}) = P(\boldsymbol{n})/g(\boldsymbol{n}), \tag{17}$$

and so, using (15),

$$S[\boldsymbol{e}] = S^0[\boldsymbol{e}] - \sum_{\boldsymbol{n}} P(\boldsymbol{n}) \ln(P(\boldsymbol{n})/P^0(\boldsymbol{n})). \tag{18}$$

Here $S^0[\boldsymbol{e}]$,

$$S^0[\boldsymbol{e}] = \ln\Big(\sum_{\boldsymbol{n}} g(\boldsymbol{n})\Big), \tag{19}$$

is the maximal value of $S[\boldsymbol{e}]$, and the second term in (18) is strictly non-negative (Levine and Bernstein 1976) and is termed 'the entropy deficiency'. It is the difference between the global maximum and the actual value of the entropy. To compute the entropy deficiency (or the entropy) one thus requires both the actual, $P(\boldsymbol{n})$, and the prior, $P^0(\boldsymbol{n})$, distributions

$$P^0(\boldsymbol{n}) = g(\boldsymbol{n})\Big/\sum_{\boldsymbol{n}} g(\boldsymbol{n}), \qquad P^0(\boldsymbol{e}) = 1\Big/\sum_{\boldsymbol{n}} g(\boldsymbol{n}). \tag{20}$$

If (and only if) $P(\boldsymbol{n}) = P^0(\boldsymbol{n})$, then $S[\boldsymbol{e}] = S^0[\boldsymbol{e}]$. The distribution of maximal entropy is thus unique, and as in other applications (Levine and Bernstein 1976, Levine 1978) we reserve the term 'prior' for the distribution determined to be of maximal entropy subject only to those constraints that are always present. As in the example above, such constraints typically reflect some symmetry that is inherent in the problem. In other words, $g(\boldsymbol{n})$ is typically a degeneracy factor, i.e. the number of distinct messages that are grouped together (that are not resolved) by the index $\boldsymbol{n}$.

The constraints imposed in determining the prior distribution are inherent in the problem and are unchanged when additional data are available (excluding of course the obvious exception where the additional data imply that the degeneracy has been broken by some means). They are taken into account by writing the entropy as in (18) and are therefore automatically included when the entropy is maximised. If additional constraints are present, the maximal value of the entropy (subject to these constraints) will then be lower than S^0 and the entropy deficiency will be positive (cf § 3.4).

3.2. *The distribution of maximal entropy*

The unique† distribution $P(\boldsymbol{x})$, which is of maximal entropy subject to the constraints of normalisation and to M data constraints (14), is readily determined, using the Lagrange undetermined multipliers procedure, to be of the form

$$P(\boldsymbol{x}) = g(\boldsymbol{x}) \exp\Big(-\sum_{r=0}^{M} \lambda_r f_r(\boldsymbol{x})\Big). \tag{21}$$

The $M+1$ (Lagrange) parameters are determined by the $M+1$ conditions (14). The resulting set of equations

$$\sum_{\boldsymbol{x}} \hat{f}_r(\boldsymbol{x}) P(\boldsymbol{x}) = 0, \tag{22}$$

† The distribution is unique whether the constraints are linearly independent or not, but uniqueness does require that the set of functions which are consistent with the value of constraints be convex.

where $\hat{f}_r(\boldsymbol{x}) = f_r(\boldsymbol{x}) - A_r$, is coupled and highly non-linear. Except for special circumstances, an analytical solution for the λ_rs is not possible. An efficient numerical procedure (Alhassid *et al* 1978) has however been described, and the program, including user's instructions, is available from the author upon request.

In principle, the Lagrange parameters can be computed from the entropy of the distribution $P(\boldsymbol{x})$, which, using (18), is

$$S[\boldsymbol{x}] = \sum_{r=0}^{M} \lambda_r A_r, \tag{23}$$

and hence (using (A1))

$$\lambda_r = \partial S[\boldsymbol{x}]/\partial A_r. \tag{24}$$

Either (23) or (24) shows that if A_r does bear dimensions then so does λ_r, and it does so in such a manner that if the units used to express the value of A_r are changed, then the value of λ_r will change by a corresponding amount, leaving $P(\boldsymbol{x})$ invariant.

The functional form (21) demonstrates explicitly that, while the distribution $P(\boldsymbol{x})$ of maximal entropy is unique, the constraints and their conjugate Lagrange parameters are not. One can always define equivalent constraints via the linear combinations

$$h_s(\boldsymbol{x}) = \sum_r d_{sr} f_r(\boldsymbol{x}), \tag{25}$$

which will yield an identical distribution $P(\boldsymbol{x})$ provided only that the Lagrange parameters conjugate to the new constraints denoted by μ_s,

$$\lambda_r = \sum_r \mu_s d_{sr}, \tag{26}$$

transform in a contragradient manner. Indeed, the matrix $\mathbf{d}$ need not even be square, and in this fashion one can eliminate (or introduce) linearly dependent constraints. This freedom allows us to offer an alternative interpretation of $P(\boldsymbol{x})$. Consider linear constraints

$$f_r(\boldsymbol{x}) = \sum_{i=1}^{N} a_{ri} x_i. \tag{27}$$

Then one can rewrite (21) as

$$P(\boldsymbol{x}) = g(\boldsymbol{x}) \exp\left(-\lambda_0 - \sum_{i=1}^{N} \mu_i x_i\right), \tag{28}$$

where

$$\mu_i = \sum_r \lambda_r a_{ri}. \tag{29}$$

The functional form (28) can be derived directly as the distribution of maximal entropy subject to the (possibly not linearly independent) $y_i = \langle x_i \rangle$, $i = 1, \ldots, N$, as constraints. This conclusion will be shown in § 4 to offer a reduced level of description where no reference to $P(\boldsymbol{x})$ need be explicitly made.

3.3. *The reciprocal basis*

The $M+1$ Lagrange parameters and hence $P(\boldsymbol{x})$ can be regarded as functions of the $M+1$ constraints A_r. Indeed the distribution of maximal entropy is a homogeneous function of degree one in the constraints (Robertson 1967):

$$P(\boldsymbol{x})=\sum_{r=0}^{M} A_r\partial P(\boldsymbol{x})/\partial A_r. \tag{30}$$

Intuitively the result is obvious. If all the constraints (including A_0) are scaled by the same factor, then clearly $P(\boldsymbol{x})$ should be unchanged. A short algebraic proof is given in the Appendix, where a simple example is also worked out. An important point to be noted is that the $M+1$ constraints (and hence the $M+1$ Lagrange parameters) must be regarded as independent variables in (30). This is in contrast to the common procedure in statistical mechanics which assigns A_0 the value unity from the very start, and thereby makes λ_0 a function of the other M Lagrange parameters.

It follows from (30) that any expectation value computed using $P(\boldsymbol{x})$, e.g. A_q, $q>M$,

$$A_q=\sum_{\boldsymbol{x}} f_q(\boldsymbol{x})P(\boldsymbol{x})=\sum_{r=0}^{M} A_r\partial A_q/\partial A_r, \tag{31}$$

and in particular $\boldsymbol{y}=\langle\boldsymbol{x}\rangle$, is a homogeneous function (of degree one) of the constraints. Note that the proof of (30) (and hence of (4)) in the Appendix is valid also for non-linear constraints.

The expansion of $\boldsymbol{y}$ in a reciprocal basis (cf (4)) can be explicitly written in terms of correlation matrices. For simplicity we limit the derivation to linear constraints. Let $\mathbf{W}$, $\mathbf{W}=\mathbf{aCa}^{\mathrm{T}}$, be the correlation matrix for the constraints

$$W_{rs}=\langle f_r(\boldsymbol{x})f_s(\boldsymbol{x})\rangle, \qquad C_{ij}=\langle x_ix_j\rangle. \tag{32}$$

Note also (say from (A1)) that $\partial\boldsymbol{A}/\partial\boldsymbol{\lambda}=\mathbf{W}$, so that using the chain rule, $\partial\boldsymbol{y}/\partial\boldsymbol{A}=(\partial\boldsymbol{y}/\partial\boldsymbol{\lambda})(\partial\boldsymbol{\lambda}/\partial\boldsymbol{A})=(\partial\boldsymbol{y}/\partial\boldsymbol{\lambda})\mathbf{W}^{-1}$. By explicit differentiation of (10), $\boldsymbol{a}(\partial\boldsymbol{y}/\partial\boldsymbol{\lambda})=\mathbf{W}$ and hence $\mathbf{a}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{a}(\partial\boldsymbol{y}/\partial\boldsymbol{\lambda})=\mathbf{a}^{\mathrm{T}}$ or

$$\boldsymbol{y}=(\partial\boldsymbol{y}/\partial\boldsymbol{A})\boldsymbol{A}=(\mathbf{a}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{a})^{-1}\mathbf{a}^{\mathrm{T}}\mathbf{W}^{-1}\boldsymbol{A}=\mathbf{Ca}^{\mathrm{T}}\mathbf{W}^{-1}\boldsymbol{A}. \tag{33}$$

3.4. Sequential inference

As a consistency check on the formalism, consider the change in the entropy upon the inclusion of additional constraints. Let $Q(\boldsymbol{x})$ be the distribution of maximal entropy subject to M' constraints $(M'>M)$

$$\begin{aligned}0&\leq\sum_{\boldsymbol{x}} Q(\boldsymbol{x})\ln(Q(\boldsymbol{x})/P(\boldsymbol{x}))\\&=\sum_{\boldsymbol{x}} Q(\boldsymbol{x})\ln(Q(\boldsymbol{x})/g(\boldsymbol{x}))-\sum_{\boldsymbol{x}} Q(\boldsymbol{x})\ln(P(\boldsymbol{x})/g(\boldsymbol{x}))\\&=-S_Q[\boldsymbol{x}]-\sum_{\boldsymbol{x}} P(\boldsymbol{x})\ln(P(\boldsymbol{x})/g(\boldsymbol{x}))\\&=S_P[\boldsymbol{x}]-S_Q[\boldsymbol{x}].\end{aligned} \tag{34}$$

The first expression is non-negative by Gibbs' inequality, and the replacement of the second by the third expression in (34) is possible since, by definition, $Q(\boldsymbol{x})$ is consistent with all the M constraints used to characterise $P(\boldsymbol{x})$, hence

$$\int \mathrm{d}\boldsymbol{x}\, f_r(\boldsymbol{x})P(\boldsymbol{x})=\int \mathrm{d}\boldsymbol{x}\, f_r(\boldsymbol{x})Q(\boldsymbol{x}), \qquad r=1,\ldots,M, \tag{35}$$

and the result follows upon use of (21). Equality in (34) holds if (and only if) $Q(\boldsymbol{x}) = P(\boldsymbol{x})$. Hence, upon the addition of further constraints, either the distribution of maximal entropy remains unchanged (i.e. the new constraints are not informative) or the distribution is changed, in which case the value of its entropy goes down.

4. Applications

Four general classes of inversion problems using linear constraints are examined in some detail. Their most important common characteristic was already implicitly derived in § 3.2. Since the entropy is a function of the constraints, and since the constraints can be recast as the values $y_i = \langle x_i \rangle$ (cf (28)), it follows that one can regard the entropy not as a function of $P(\boldsymbol{x})$ but directly as a function of the y_is. As we shall show by the examples, it is indeed possible to compute the entropy as an explicit function of the y_is and thereby obtain a simplified procedure.

4.1. The inversion of average values and of frequency data

A single experiment can result in any one of N outcomes. It is required to infer the number of times the ith outcome, $i = 1, 2, \ldots, N$, has been realised in n independent repetitions of the experiment. The data available are the M average values

$$A_r = \sum_{i=1}^{N} a_{ri} n_i, \qquad r = 1, \ldots, M, \tag{36}$$

where a_{ri} is the value of the rth observable for the ith outcome. In the more usual application of the maximum entropy formalism we are asked to infer the probability p_i of the ith outcome subject to given average values and, as is well known (Jaynes 1957, Tribus 1961), obtain

$$p_i = \exp\left(-\sum_{r=0}^{M} \lambda_r a_{ri}\right). \tag{37}$$

This problem thus serves as a check on the formalism, which should recover result (37). As a bonus for working harder we shall however obtain additional insight.

The distribution of outcome vectors $\boldsymbol{n}$ ($\sum n_i = n$) which is of maximal entropy subject to the M constraints (36) is

$$P(\boldsymbol{n}) = g(\boldsymbol{n}) \exp\left[-n\lambda_0 - \sum_{r=1}^{M} \lambda_r \left(\sum_{i=1}^{N} a_{ri} n_i\right)\right]. \tag{38}$$

Here $g(\boldsymbol{n})$ is the degeneracy factor given by (16), and for future convenience we have written the Lagrange parameter conjugate to the normalisation constraint as $n\lambda_0$. Using the definition (29) of μ_i and putting $p_i = \exp(-\lambda_0 - \mu_i)$,

$$P(\boldsymbol{n}) = g(\boldsymbol{n}) \prod_{i=1}^{N} p_i^{n_i} = n! \prod_{i=1}^{N} p_i^{n_i}/n_i! \tag{39}$$

Using (29), p_i is seen to be given by (37). We recognise (39) as the multinomial distribution, which is the standard probability theory result (e.g. Feller 1968) for the distribution $P(\boldsymbol{n})$, with p_i being the probability of the ith outcome.

The inference for n_i is

$$\langle n_i \rangle = \sum_{\boldsymbol{n}} n_i P(\boldsymbol{n}) = np_i, \tag{40}$$

and the variance of the inference is

$$\langle n_i^2 \rangle - \langle n_i \rangle^2 = n^2[p_i(1-p_i)/n]. \tag{41}$$

Both these results are as expected, with the fractional variance decreasing as n^{-1}.

For this problem the 'partition function' $\exp(n\lambda_0)$ can be explicitly computed:

$$\exp(n\lambda_0) = \sum_{\boldsymbol{n}} \frac{n!}{\Pi_i n_i!} \prod_{i=1}^{N} \exp(-\mu_i n_i) = \left(\sum_{i=1}^{N} \exp(-\mu_i)\right)^n$$

or

$$\lambda_0 = \ln\left(\sum_{i=1}^{N} \exp(-\mu_i)\right). \tag{42}$$

The result can be verified by noting that $P(\boldsymbol{n})$ can be regarded as the distribution of maximal entropy subject to the N $(N > M)$ not necessarily linearly independent constraints $\langle n_i \rangle$. The Lagrange parameter conjugate to $\langle n_i \rangle$ is μ_i. Using the general identity

$$A_r = -\partial \ln(\text{partition function})/\partial\lambda_r, \tag{43}$$

we verify (40):

$$\langle n_i \rangle = -\partial(n\lambda_0)/\partial\mu_i = \exp(-n\lambda_0)\partial \exp(n\lambda_0)/\partial\mu_i = n \exp(-\lambda_0 - \mu_i) = np_i. \tag{44}$$

The entropy of the distribution $P(\boldsymbol{n})$ is defined by (18). Using the explicit result (38) for $P(\boldsymbol{n})$,

$$S[\boldsymbol{n}] = n\lambda_0 + \sum_{r=1}^{M} \lambda_r A_r = n\lambda_0 + \sum_{i=1}^{N} \mu_i \langle n_i \rangle, \tag{45}$$

but since $\mu_i = -\ln p_i - \lambda_0$,

$$S[\boldsymbol{n}] = -n \sum_{i=1}^{N} p_i \ln p_i \tag{46}$$

The maximal value of $S[\boldsymbol{n}]$ determined via the present procedure coincides with the value determined by the more familiar approach which works with the probabilities p_i from the very start. In other problems where $\langle x_i \rangle$ cannot be interpreted as a probability, we could still express $S[\boldsymbol{x}]$ in terms of the $\langle x_i \rangle$s (e.g. equation (59) below), but the functional dependence will not be that in (46). As will become obvious, the particular dependence in (46) reflects the structure of the degeneracy factor $g(\boldsymbol{n})$. Different factors give rise to different functions $S(\langle \boldsymbol{x} \rangle)$. If we could learn to recognise these functions, then a short cut could be achieved. In the same way that one determines the probabilities directly by maximising (46) subject to average value constraints (Jaynes 1957, 1963, Tribus 1961), one could, in general, determine the optimal $\boldsymbol{y}$ by maximising $S[\boldsymbol{y}]$ subject to any available linear constraints.

4.2. Uniform prior distribution

In the absence of any prior constraints†, and when the data correspond to linear constraints, the different components of $\boldsymbol{x}$ are independent of one another:

$$P(\boldsymbol{x})=\exp\left[-\lambda_0-\sum_{r=1}^{M}\lambda_r\left(\sum_{i=1}^{N}a_{ri}x_i\right)\right]=\prod_{i=1}^{N}P(x_i), \tag{47}$$

where†, using (29),

$$P(x_i)=\exp(-\lambda_{0i}-\mu_i x_i), \tag{48}$$

$$\exp(\lambda_{0i})=\sum_{x_i}\exp(-\mu_i x_i). \tag{49}$$

As is obvious from (47) and (29), the partition function factorises,

$$\exp(\lambda_0)=\prod_{i=1}^{N}\exp(\lambda_{0i}), \tag{50}$$

while the entropy can be expressed as the sum of the entropies of the distributions of the different components:

$$S[\boldsymbol{x}]=-\sum_{\boldsymbol{x}}P(\boldsymbol{x})\ln P(\boldsymbol{x})=-\sum_{i}\sum_{x_i}P(x_i)\ln P(x_i)=\sum_{i}S[x_i]. \tag{51}$$

Problems where the partition function can be readily evaluated include: (i) x_i can assume only the values 0 or 1; (ii) x_i assumes non-negative integer values; and (iii) x_i is continuous and positive.

$$\begin{aligned}\exp(\lambda_{0i})&=\sum_{x_i}\exp(-\mu_i x_i)\\&=1+\exp(-\mu_i), && x_i=0,1\\&=[1-\exp(-\mu_i)]^{-1}, && x_i=0,1,2,\dots\\&=1/\mu_i, && x_i\geqslant 0.\end{aligned} \tag{52}$$

The results of the inversion determined using $\langle x_i\rangle=-\partial\lambda_{0i}/\partial\mu_i$ are

$$\begin{aligned}y_i=\langle x_i\rangle&=[\exp(\mu_i)+1]^{-1}, && x_i=0,1\\&=[\exp(\mu_i)-1]^{-1}, && x_i=0,1,2,\dots\\&=1/\mu_i, && x_i\geqslant 0.\end{aligned} \tag{53}$$

It is worthwhile to recall that in all these results

$$\mu_i=\sum_{r=1}^{M}\lambda_r a_{ri}, \tag{54}$$

and so they are valid for any set of linear constraints.

The entropy

$$S[\boldsymbol{x}]=\lambda_0+\boldsymbol{\mu}\,.\,\langle\boldsymbol{x}\rangle \tag{55}$$

† The results of this section remain valid as long as $g(\boldsymbol{x})$ can be expressed as a product of degeneracy factors, one for each component of $\boldsymbol{x}$. In this case (48) must have the RHS multiplied by g_i, and similar simple modifications are required in other results.

can now be explicitly computed as a function of $\langle \boldsymbol{x} \rangle$. For the first two cases,

$$\mu_i = \ln[(1 \mp \langle x_i \rangle)/\langle x_i \rangle], \tag{56}$$

while the third case corresponds to the $\langle x_i \rangle \ll 1$ limit,

$$\mu_i = 1/\langle x_i \rangle, \qquad x_i \geqslant 0. \tag{57}$$

Hence, for the first two cases[†]

$$S[\boldsymbol{x}] = \sum_{i=1}^{N} \{\langle x_i \rangle \ln[(1 \mp \langle x_i \rangle)/\langle x_i \rangle] \mp 1 \ln(1 \mp \langle x_i \rangle\}, \tag{58}$$

while for $x_i \geqslant 0$

$$S[\boldsymbol{x}] = \sum_{i=1}^{N} \ln\langle x_i \rangle + N. \tag{59}$$

One can now seek to infer $\langle x_i \rangle$ by maximising $S[\boldsymbol{x}]$ subject to the constraints $\boldsymbol{A} = \mathbf{a}\langle \boldsymbol{x} \rangle$. One readily verifies that this does lead back to (53). Clearly it is far simpler to solve the maximal entropy problem using (58) or (59). In either case, the results (53) require only about two lines of algebra. For example, varying (59), the variational problem is

$$\sum_i \delta\langle x_i \rangle \Big(\langle x_i \rangle^{-1} - \sum_r \lambda_r a_{ri} \Big) = 0, \tag{60}$$

while for the first two cases the term $\langle x_i \rangle^{-1}$ in (60) needs to be replaced by $\ln[(1 \mp \langle x_i \rangle)/\langle x_i \rangle]$.

As a check of the results one can compute $\partial S[\boldsymbol{x}]/\partial A_r$. For all three cases one verifies that it equals λ_r, e.g. for case (iii)

$$\begin{aligned} \partial S/\partial A_q &= \sum_{i=1}^{N} (\partial S/\partial y_i)(\partial y_i/\partial A_q) = \sum_{i=1}^{N} \Big(\sum_{s=1}^{M} \lambda_s a_{si} \Big) a_{qi} \\ &= \sum_{s=1}^{M} \lambda_s \delta_{s,q} = \begin{cases} \lambda_q, & q \leqslant M \\ 0\ , & q > M. \end{cases} \end{aligned} \tag{61}$$

4.3. Expansion in a basis set

A familiar procedure for obtaining an approximation for an unknown vector $\boldsymbol{y}$ is to expand it in a basis set:

$$y_i = \sum_{r=1}^{N} a_{ri} A_r \qquad \text{or} \qquad \boldsymbol{y} = \mathbf{a}^{\mathrm{T}} \boldsymbol{A}. \tag{62}$$

Here a_{ri} is the ith component of the rth vector in the set. To obtain an exact representation it will be necessary, in general, to include all N basis vectors in the expansion (where N is the dimension of the space, i.e. the range of the index i). With no loss of generality we take the basis vectors to be orthonormal:

$$\sum_{i=1}^{N} a_{ri} a_{si} = \delta_{r,s} \qquad \text{or} \qquad \mathbf{a}\mathbf{a}^{\mathrm{T}} = \mathbf{I}. \tag{63}$$

[†] In the presence of degeneracy, but provided $g(x) = \Pi\, g_i$, a similar discussion does go through except that all three ones inside the braces in (58) need to be replaced by g_i. In (59), $\ln(\langle x_i \rangle)$ is replaced by $\ln(\langle x_i \rangle / g_i)$. With these changes, (58) is the familiar result for the Fermi–Dirac and Bose–Einstein statistics (see e.g. Landsberg 1959).

The expansion coefficients are then determined, as usual, from (62) and (63) as

$$A_r = \sum_{i=1}^{N} a_{ri} y_i, \tag{64}$$

Say now that only M ($M<N$) coefficients A_r are available. The conventional approach is to replace the exact expansion (62) by the (possibly) approximate one

$$y_i = \sum_{r=1}^{M} a_{ri} A_r. \tag{65}$$

This is often described as an optimal (in a least-squares sense) choice. Indeed, among all possible linear expansions of the form

$$y_i = \sum_{r=1}^{M} a_{ri} X_r, \tag{66}$$

the choice $X_r = A_r$ does minimise the square deviances from the exact result. Yet from another point of view, the choice (65) is quite disturbing. In principle one requires N coefficients A_r to represent y. In practice only M coefficients are known. The approximation (65) then implies that all unknown $N-M$ coefficients are identically zero. But why? What possible grounds does one have for assuming that just because these numbers are unknown they must be zero? Surely a point of view that argues that if these coefficients are unknown their value should be inferred from the values of the known coefficients deserves at least a hearing.

There are, of course, practical grounds for preferring the linear expansion (65). It is therefore important to note that the maximal entropy solution can also be written as a linear sum of M terms (cf (4)),

$$y_i = \sum_{r=1}^{M} b_{ir} A_r. \tag{67}$$

The essential difference here is that the vectors of the reciprocal basis $\mathbf{b}$ are not the vectors of the original basis, i.e. $\mathbf{b} \neq \mathbf{a}^{\mathrm{T}}$. The reason is, of course, that while $\mathbf{aa}^{\mathrm{T}} = \mathbf{I}$, $\mathbf{a}^{\mathrm{T}}\mathbf{a}$ is the identity matrix only when the index r rangers over 1 to N, which is not the case in (67).

The linear expansion (67) is one way of writing the maximum entropy solution. Another is

$$y_i = \sum_{r=1}^{M} a_{ri} A_r + \sum_{q=M+1}^{N} a_{qi} A_q, \tag{68}$$

i.e. as a linear combination of all N basis vectors. To determine the maximal entropy inference for the coefficients A_q, $M<q\leqslant N$, consider the case where y_i is a continuous variable so that the entropy $S[y]$ is given by (59). In seeking the maximum, the variation is over all vectors y of the form (68), where the coefficients A_r, $r \leqslant M$, are fixed and only the coefficients A_q, $M<q\leqslant N$, may be varied. Inserting the expansion (68) in (59), the extremum is implicitly given by

$$\partial S[y]/\partial A_q = 0, \qquad M<q\leqslant N$$

or (69)

$$\sum_{i=1}^{N} (y_i)^{-1} \partial y_i/\partial A_q = 0 \qquad = \sum_{i=1}^{N} (y_i)^{-1} a_{qi}, \qquad M<q\leqslant N.$$

In other words, the *inverse* of the optimal inference must be orthogonal to all the basis vectors whose expansion coefficients are unknown. This conclusion is to be contrasted with the usual assumption that the unknown solution itself is orthogonal to those vectors whose expansion coefficients are not available.

The result (69) can be written as

$$y_i^{-1} = \sum_{r=1}^{M} \lambda_r a_{ri}, \tag{70}$$

where the M unknown multipliers λ_r are just our previous Lagrange parameters.

Lest one concludes that the only difference between the present, (70), and the conventional, (65), results is that they are inverse to one another, we hasten to reiterate that the form (70) applies only when y_i is a continuous non-negative variable. For the other two cases of § 4.2 the entropy is given by (58), and it is $\ln[(1 \mp y_i)/y_i]$ which is found to be orthogonal to the vectors a_s. Hence

$$\ln[(1 \mp y_i)/y_i] = \sum_{r=1}^{M} \lambda_r a_{ri}$$

or

$$y_i = \left[\exp\left(\sum_{r=1}^{M} \lambda_r a_{ri}\right) \pm 1\right]^{-1}, \tag{71}$$

which is just our previous result (53).

The real practical difference between the linear (equation (65)) and the maximum entropy inferences is that the latter requires the evaluation of the M Lagrange parameters (using the M values A_r). If M is comparable with N, then one may well be tempted to avoid the extra work and use the linear expansion. After all, the linear expansion is the best approximation of the form

$$y_i = \sum_{r=1}^{M} a_{ri} X_r, \qquad M < N, \tag{72}$$

while the maximum entropy result is the best approximation of the form

$$y_i = \sum_{r=1}^{N} a_{ri} X_r. \tag{73}$$

If M is near N, the additional freedom in (73) (which is an expansion in a complete set of basis vectors) is possibly not important. However if $M < N$ (and particularly if $M \ll N$), the difference can be dramatic. An objection that can be raised is that it may be that the correct solution is exactly given by (67) with only M terms. If we know this to be the case, then we know that $A_q \equiv 0$ for $q > M$, the two procedures will give identical results, and there is no paradox. If we do not know the exact solution, then the probable deviation is smaller using the maximal entropy procedure.

We have centred attention on the case where the basis vectors a_{ri} are orthonormal. There is however a celebrated problem where they are not, i.e. the problem of moments (Shohat and Tamarkin 1943), where $a_{ri} = i^r$. The problem of inversion using a finite set of moments has indeed been discussed, but its thrust was in determining upper and lower bounds (Gordon 1968, Corcoran and Langhoff 1977). We thus defer comparison with these results to a sequel paper which considers the bounds on y_i which are provided by the present formalism. There we shall also argue that it is possible to determine the expansion coefficients A_q, $q > M$, directly and in a recursive fashion (i.e. in the sequence $A_{M+1}, A_{M+2}, \ldots$) in terms of the known set of M coefficients A_r, and also pay special attention to the case where the a_{ri}s are eigenvectors of a matrix so that (62) is an eigenfunction expansion (Hall 1963).

4.4. *The determination of line shapes*

The concept of the line shape (as a function of frequency) arises in many diverse applications. There have been previous discussions of line shapes which did invoke the procedure of maximising the entropy. Mostly, they normalise the line shape to unity, and regard it therefore as a probability density function whose entropy is to be maximised subject to constraints (Powles and Carazza 1970, Berne and Harp 1970, Czajkowski 1973, Carazza 1976). In another point of view (Burg 1972, as quoted in Smylie *et al* 1973, Carazza 1976), the maximum entropy inference is used in an attempt to eliminate the effects of noise with specified statistical properties. Here we consider the problem as an example of the determination of a function (of a continuous argument) which is not fully specified by the data. The maximum entropy formalism is used to provide an optimal inference in the face of uncertainty. No special statistical properties of the uncertainty need be invoked. On the contrary, we seek the maximally non-committed functional form which is consistent with the data. We shall recover as special cases the results of the previous authors without having to include any assumptions about the distribution of noise (or of errors (Carazza 1976)).

The index i is here allowed to vary continuously (as it stands for the frequency), and y_i will be written as $\alpha(\omega)$. (The vector space of solutions is now a general normed function space.) We consider the situation where the line shape need not be regarded as a probability density, so that the results are the analogues of those discussed for the discrete case in §§ 4.2 and 4.3.

The magnitude $\alpha(\omega)$ of the line shape at a given frequency is a continuous non-negative variable. If the constraints are linear,

$$A_r = \int d\omega\, a_r(\omega)\langle\alpha(\omega)\rangle, \qquad r = 0, 1, \ldots, M, \tag{74}$$

then the discussion of case (iii) of § 4.2 applies except that a_{ri} is replaced by $a_r(\omega)$ and integration over ω replaces the summation over i. The expression for the entropy is thus (cf (59))

$$S[\alpha] = \int d\omega\, \ln\langle\alpha(\omega)\rangle, \tag{75}$$

or

$$S[\alpha] = \int d\omega\, \rho(\omega) \ln(\langle\alpha(\omega)\rangle/\rho(\omega)) \tag{76}$$

when degeneracy is present (but does not correlate different frequency components so that $\rho(\omega)$ is the continuous analogue of g_i in the discrete case). To prove that (76) is the proper generalisation of (75), note that it is invariant to change of variable, and that a measure of uncertainty should indeed be invariant under such a transformation (Jaynes 1963). Those who find the argument less than fully convincing are asked to note that a density of states in information theory is essentially a Jacobian of the transformation from f to ω, where $\alpha(f)$ is uniform in the absence of constraints (Dinur and Levine 1975). Hence

$$S[\alpha]=\int \mathrm{d}f \ln(\langle\alpha(f)\rangle). \qquad (77)$$

Changing variables, and using $\rho(\omega)=\partial f/\partial\omega$, we recover (76).

The optimal inference for $\alpha(\omega)$ using the $M+1$ constraints A_r is obtained as previously by maximising the entropy $S[\alpha]$ subject to the constraints:

$$\langle\alpha(\omega)\rangle=\left(\sum_{r=0}^{M}\lambda_r a_r(\omega)\right)^{-1}. \qquad (78)$$

If the functions $a_r(\omega)$ are orthonormal (which can always be arranged since they are linearly independent), we have as an alternative representation

$$\langle\alpha(\omega)\rangle=\sum_{r=0}^{\infty}A_r a_r(\omega). \qquad (79)$$

Here the first $M+1$ coefficients A_r have the values assigned by the data, and the rest are the inferred values, determined in principle by equating (78) and (79). The two equations can be combined in one, namely

$$A_q=\int \mathrm{d}\omega\, a_q(\omega)\bigg/\left(\sum_{r=0}^{M}\lambda_r a_r(\omega)\right). \qquad (80)$$

For $q\leqslant M$, this equation defines the values of the $M+1$ Lagrange parameters λ_r in terms of the $M+1$ given values A_r, $r=0,1,\ldots,M$. For $q>M$, this equation defines the values of A_q. For practical purposes it is clearly preferable to determine the value of A_q, $q>M$, directly in terms of the A_r's, $r\leqslant M$, as mentioned already in § 4.3.

The result (78) generalises the expression of Burg (1972, see e.g. Smylie *et al* 1973), who examined the problem of a Fourier series representation ($a_r(\omega)=\exp(\mathrm{i}r\omega)$). The essential point in the generalisation is however in delineating the limitation on the result. Burg was particularly concerned with inference in the presence of noise, for a time series which is a Gaussian process. As is clear from the present discussion, the result is independent of the origin of the uncertainty. Thus while the previous derivation (Smylie *et al* 1973) of expression (75) for the entropy appears to be strictly limited to Gaussian processes, the present derivation demonstrates that it is not. The assumptions that do go into (75) are $\alpha(\omega)\geqslant 0$, the linear nature of the constraints, and a uniform $g(\omega)$, as otherwise (76) is to be used.

The Lorentzian line shape

$$\alpha(\omega)=(\Gamma/\pi)[\omega^2-2\omega_0\omega+(\omega_0^2+\Gamma^2)]^{-1}=(\Gamma/\pi)[(\omega-\omega_0)^2+\Gamma^2]^{-1} \qquad (81)$$

is a special case of (78), where the Lagrange parameters can be explicitly evaluated. The three constraints are

$$1=A_0=\int \mathrm{d}\omega\,\alpha(\omega), \qquad \omega_0=A_1=\int \mathrm{d}\omega\,\omega\alpha(\omega), \qquad \Gamma^2+\omega_0^2=\int \mathrm{d}\omega\,\omega^2\alpha(\omega), \qquad (82)$$

where the last two integrals need to be defined with care (e.g. between finite limits). The present derivation differs from the recent one by Carazza (1976) in that experimental errors need not be invoked to prove the result.

5. Summary

The procedure of maximal entropy is usually employed to induce a probability distribution given only partial data. This paper argued that the procedure need not be limited to inferences regarding probabilities, but can be used as a general inversion procedure and that inductive inferences can be drawn regarding the most reasonable values of general (if necessary, dimension-bearing) variables. Particular attention was given to the solution of an incompletely specified inversion problem using data linear in the unknown variable of interest (so called 'linear constraints'). Three equivalent forms of the solution were presented:

(i) As an average over the distribution of maximal entropy, where the explicit form (e.g. (70)) is dependent on the Lagrange parameters conjugate to the constraints.

(ii) As an expansion in functions of the reciprocal basis set, e.g. (4). The number of terms in such an expansion is finite and equals the number of constraints, but the reciprocal basis needs to be determined as it is tailor-made for the constraints.

(iii) As an expansion in a fixed basis set, where the number of terms equals the number of points which the function need be determined.

Several types of applications were examined in some detail, and the use of a 'reduced' formalism was demonstrated for linear constraints. The reduction enables one to solve the inversion problem directly without having first to consider the distribution of maximal entropy as an intermediate construct.

Acknowledgments

I thank several colleagues, and in particular Professor S Baer, for their comments. This work was supported by the Office of Naval Research.

Appendix

To prove (30), consider the condition of linear independence of the constraints

$$\partial A_0/\partial A_r = \delta_{0,r} = \sum_{\boldsymbol{x}} \partial P(\boldsymbol{x})/\partial A_r = \sum_{\boldsymbol{x}}\sum_{s} (\partial P(\boldsymbol{x})/\partial \lambda_s)(\partial \lambda_s/\partial A_r)$$

$$= \sum_{\boldsymbol{x}}\sum_{s} -f_s(\boldsymbol{x})P(\boldsymbol{x})(\partial \lambda_s/\partial A_r) = -\sum_{s} A_s \partial \lambda_s/\partial A_r = -\sum_{s} A_s \partial \lambda_r/\partial A_s. \tag{A1}$$

Hence, recalling that $f_0(\boldsymbol{x}) = 1$,

$$P(\boldsymbol{x}) = \sum_{r} f_r(\boldsymbol{x})P(\boldsymbol{x})\delta_{r,0} = \sum_{s,r} -f_r(x)P(\boldsymbol{x})A_s(\partial \lambda_r/\partial A_s) = \sum_{s} A_s(\partial P(\boldsymbol{x})/\partial A_s). \tag{A2}$$

As a simple example, take a one-dimensional problem where x is a non-negative integer and the constraints are $A_0 = \langle 1\rangle$ and $A_1 = \langle x\rangle$:

$$\langle 1\rangle = \sum_{x=0}^{\infty} \exp(-\lambda_0 - \mu x) = \exp(-\lambda_0)/[1-\exp(-\mu)], \tag{A3}$$

$$\langle x\rangle = \sum_{x=0}^{\infty} x \exp(-\lambda_0 - \mu x) = \langle 1\rangle/[\exp(\mu)-1] \tag{A4}$$

One now eliminates λ_0 and μ in favour of $\langle 1\rangle$ and $\langle x\rangle$, so that

$$P(x)=\exp(-\lambda_0-\mu x)=\langle 1\rangle[\langle 1\rangle/(\langle 1\rangle+\langle x\rangle)][\langle x\rangle/(\langle 1\rangle+\langle x\rangle)]^x. \qquad \text{(A5)}$$

The distribution is clearly homogeneous of degree one in the constraints.

References

Alhassid Y, Agmon N and Levine R D 1978 *Chem. Phys. Lett.* **53** 22
Berne B J and Harp G D 1970 *Phys. Rev.* A **2** 2514
Burg J P 1972 *Geophys. J.* **37** 375
Carazza B 1976 *J. Phys. A: Math. Gen.* **9** 1069
Corcoran C T and Langhoff P W 1977 *J. Math. Phys.* **18** 651
Czajkowski G Z 1973 *J. Phys. A: Math., Nucl. Gen.* **6** 906
Darwin C G and Fowler R H 1922 *Phil. Mag.* **44** 450
Dinur U and Levine R D 1975 *Chem. Phys.* **9** 17
Edward J A and Fitelson M M 1973 *IEEE Trans. Inf. Theory* IT-19 232
Feller W 1968 *Introduction to Probability Theory and its Applications* (New York: Wiley)
Gordon R G 1968 *J. Math. Phys.* **9** 655
Hall G G 1963 *J. Chem. Phys.* **38** 1104
Jaynes E T 1957 *Phys. Rev.* **106** 620
—— 1963 *Statistical Physics* (New York: Benjamin) p 181
Khinchin A I 1957 *Mathematical Foundations of Information Theory* (New York: Dover)
Landsberg P T 1959 *Proc. Phys. Soc.* **74** 486
Levine R D 1978 *Ann. Rev. Phys. Chem.* **29** 59
Levine R D and Bernstein R B 1976 *Dynamics of Molecular Collisions* (New York: Plenum) p 323
Levine R D and Tribus M 1978 *The Maximum Entropy Formalism* (Cambridge: MIT)
Powles J G and Carazza B 1970 *J. Phys. A: Gen. Phys.* **3** 335
Robertson B 1967 *Phys. Rev.* **160** 175
Schrödinger E 1952 *Statistical Thermodynamics* (Cambridge: UP)
Shannon C E 1948 *Bell Syst. Tech. J.* **27** 379
Shohat J A and Tamarkin J D 1943 *The Problem of Moments* (New York: AMS)
Smylie D E, Clarke G K C and Ulrych T J 1973 *Meth. Comput. Phys.* **13** 391
Tribus M 1961 *Thermodynamics and Thermostatics* (Princeton: Von Nostrand)
—— 1969 *Rational Descriptions Decisions and Design* (New York: Pergamon)