International Tables for Crystallography (2006). Vol. F, Section 16.2.2, pp. 346–348.

16.2. The maximum-entropy method

By G. Bricogne

16.2.1. Introduction

The modern concept of entropy originated in the field of statistical thermodynamics, in connection with the study of large material systems in which the number of internal degrees of freedom is much greater than the number of externally controllable degrees of freedom. This concept played a central role in the process of building a quantitative picture of the multiplicity of microscopic states compatible with given macroscopic constraints, as a measure of how much remains unknown about the detailed fine structure of a system when only macroscopic quantities attached to that system are known. The collection of all such microscopic states was introduced by Gibbs under the name 'ensemble', and he deduced his entire formalism for statistical mechanics from the single premise that the equilibrium picture of a material system under given macroscopic constraints is dominated by that configuration which can be realized with the greatest combinatorial multiplicity (*i.e.* which has maximum entropy) while obeying these constraints.

The notions of ensemble and the central role of entropy remained confined to statistical mechanics for some time, then were adopted in new fields in the late 1940s. Norbert Wiener studied Brownian motion, and subsequently time series of random events, by similar methods, considering in the latter an ensemble of messages, *i.e.* 'a repertory of possible messages, and over that repertory a measure determining the probability of these messages' (Wiener, 1949). At about the same time, Shannon created information theory and formulated his fundamental theorem relating the entropy of a source of random symbols to the capacity of the channel required to transmit the ensemble of messages generated by that source with an arbitrarily small error rate (Shannon & Weaver, 1949). Finally, Jaynes (1957, 1968, 1983) realized that the scope of the principle of maximum entropy could be extended far beyond the confines of statistical mechanics or communications engineering, and could provide the basis for a general theory (and philosophy) of statistical inference and 'data processing'.

The relevance of Jaynes' ideas to probabilistic direct methods was investigated by the author (Bricogne, 1984). It was shown that there is an intimate connection between the maximum-entropy method and an enhancement of the probabilistic techniques of conventional direct methods known as the 'saddlepoint method', some aspects of which have already been dealt with in Section 1.3.4.5.2 in Chapter 1.3 of *IT* B (Bricogne, 2001).

16.2.2. The maximum-entropy principle in a general context

16.2.2.1. *Sources of random symbols and the notion of source entropy*

Statistical communication theory uses as its basic modelling device a discrete source of random symbols, which at discrete times $t = 1, 2, \ldots$, randomly emits a 'symbol' taken out of a finite alphabet $\mathcal{A} = \{s_i | i = 1, \ldots, n\}$. Sequences of such randomly produced symbols are called 'messages'.

An important numerical quantity associated with such a discrete source is its *entropy per symbol* H, which gives a measure of the amount of uncertainty involved in the choice of a symbol. Suppose that successive symbols are independent and that symbol i has probability q_i. Then the general requirements that H should be a continuous function of the q_i, should increase with increasing uncertainty, and should be additive for independent sources of uncertainty, suffice to define H uniquely as

$$H(q_1, \ldots, q_n) = -k \sum_{i=1}^{n} q_i \log q_i, \qquad (16.2.2.1)$$

where k is an arbitrary positive constant [Shannon & Weaver (1949), Appendix 2] whose value depends on the unit of entropy chosen. In the following we use a unit such that $k = 1$.

These definitions may be extended to the case where the alphabet $\mathcal{A}$ is a continuous space endowed with a uniform measure μ: in this case the entropy per symbol is defined as

$$H(q) = -\int_{\mathcal{A}} q(\mathbf{s}) \log q(\mathbf{s}) \, \mathrm{d}\mu(\mathbf{s}), \qquad (16.2.2.2)$$

where q is the probability density of the distribution of symbols with respect to measure μ.

16.2.2.2. *The meaning of entropy: Shannon's theorems*

Two important theorems [Shannon & Weaver (1949), Appendix 3] provide a more intuitive grasp of the meaning and importance of entropy:

(1) H is approximately the logarithm of the reciprocal probability of a typical long message, divided by the number of symbols in the message; and

(2) H gives the rate of growth, with increasing message length, of the logarithm of the number of reasonably probable messages, regardless of the precise meaning given to the criterion of being 'reasonably probable'.

The entropy H of a source is thus a direct measure of the strength of the restrictions placed on the permissible messages by the distribution of probabilities over the symbols, lower entropy being synonymous with greater restrictions. In the two cases above, the maximum values of the entropy $H_{\max} = \log n$ and $H_{\max} = \log \mu(\mathcal{A})$ are reached when all the symbols are equally probable, *i.e.* when q is a uniform probability distribution over the symbols. When this distribution is not uniform, the usage of the different symbols is biased away from this maximum freedom, and the entropy of the source is lower; by Shannon's theorem (2), the number of 'reasonably probable' messages of a given length emanating from the source decreases accordingly.

The quantity that measures most directly the strength of the restrictions introduced by the non-uniformity of q is the difference $H(q) - H_{\max}$, since the proportion of N-atom random structures which remain 'reasonably probable' in the ensemble of the corresponding source is $\exp\{N[H(q) - H_{\max}]\}$. This difference may be written (using continuous rather than discrete distributions)

$$H(q) - H_{\max} = -\int_{\mathcal{A}} q(\mathbf{s}) \log[q(\mathbf{s})/m(\mathbf{s})] \, \mathrm{d}\mu(\mathbf{s}), \qquad (16.2.2.3)$$

where $m(\mathbf{s})$ is the uniform distribution which is such that $H(m) = H_{\max} = \log \mu(\mathcal{A})$.

16.2.2.3. *Jaynes' maximum-entropy principle*

From the fundamental theorems just stated, which may be recognized as Gibbs' argument in a different guise, Jaynes' own maximum-entropy argument proceeds with striking lucidity and constructive simplicity, along the following lines:

(1) experimental observation of, or 'data acquisition' on, a given system enables us to progress from an initial state of uncertainty to a state of lesser uncertainty about that system;

(2) uncertainty reflects the existence of numerous possibilities of accounting for the available data, viewed as constraints, in terms of a physical model of the internal degrees of freedom of the system;

(3) new data, viewed as new constraints, reduce the range of these possibilities;

(4) conversely, any step in our treatment of the data that would further reduce that range of possibilities amounts to applying extra constraints (even if we do not know what they are) which are not warranted by the available data;

(5) hence Jaynes's rule: '*The probability assignment over the range of possibilities* [*i.e.* our picture of residual uncertainty] *shall be the one with maximum entropy consistent with the available data, so as to remain maximally non-committal with respect to the missing data*'.

The only requirement for this analysis to be applicable is that the 'ranges of possibilities' to which it refers should be representable (or well approximated) by ensembles of abstract messages emanating from a random source. The entropy to be maximized is then the entropy per symbol of that source.

The final form of the maximum-entropy criterion is thus that $q(\mathbf{s})$ should be chosen so as to maximize, under the constraints expressing the knowledge of newly acquired data, its entropy

$$\mathcal{S}_m(q) = -\int_V q(\mathbf{s}) \log[q(\mathbf{s})/m(\mathbf{s})] \, \mathrm{d}\mu(\mathbf{s}) \qquad (16.2.2.4)$$

relative to the 'prior prejudice' $m(\mathbf{s})$ which maximizes H in the absence of these data.

16.2.2.4. *Jaynes' maximum-entropy formalism*

Jaynes (1957) solved the problem of explicitly determining such maximum-entropy distributions in the case of general linear constraints, using an analytical apparatus first exploited by Gibbs in statistical mechanics.

The maximum-entropy distribution $q^{\mathrm{ME}}(\mathbf{s})$, under the prior prejudice $m(\mathbf{s})$, satisfying the linear constraint equations

$$\mathcal{C}_j(q) \equiv \int_{\mathcal{A}} q(\mathbf{s}) C_j(\mathbf{s}) \, \mathrm{d}\mu(\mathbf{s}) = c_j \quad (j = 1, 2, \ldots, M), \qquad (16.2.2.5)$$

where the $\mathcal{C}_j(q)$ are linear *constraint functionals* defined by given *constraint functions* $C_j(\mathbf{s})$, and the c_j are given *constraint values*, is obtained by maximizing with respect to q the relative entropy defined by equation (16.2.2.4). An extra constraint is the normalization condition

$$\mathcal{C}_0(q) \equiv \int_{\mathcal{A}} q(\mathbf{s}) \, 1 \, \mathrm{d}\mu(\mathbf{s}) = 1, \qquad (16.2.2.6)$$

to which it is convenient to give the label $j = 0$, so that it can be handled together with the others by putting $C_0(\mathbf{s}) = 1$, $c_0 = 1$.

By a standard variational argument, this constrained maximization is equivalent to the unconstrained maximization of the functional

$$\mathcal{S}_m(q) + \sum_{j=0}^{M} \lambda_j \mathcal{C}_j(q), \qquad (16.2.2.7)$$

where the λ_j are Lagrange multipliers whose values may be determined from the constraints. This new variational problem is readily solved: if $q(\mathbf{s})$ is varied to $q(\mathbf{s}) + \delta q(\mathbf{s})$, the resulting variations in the functionals $\mathcal{S}_m$ and $\mathcal{C}_j$ will be

$$\begin{aligned} \delta\mathcal{S}_m &= \int_{\mathcal{A}} \{-1 - \log[q(\mathbf{s})/m(\mathbf{s})]\} \, \delta q(\mathbf{s}) \, \mathrm{d}\mu(\mathbf{s}) \quad \text{and} \\ \delta\mathcal{C}_j &= \int_{\mathcal{A}} \{C_j(\mathbf{s})\} \, \delta q(\mathbf{s}) \, \mathrm{d}\mu(\mathbf{s}), \end{aligned} \qquad (16.2.2.8)$$

respectively. If the variation of the functional (16.2.2.7) is to vanish for arbitrary variations $\delta q(\mathbf{s})$, the integrand in the expression for that variation from (16.2.2.8) must vanish identically. Therefore the maximum-entropy density distribution $q^{\mathrm{ME}}(\mathbf{s})$ satisfies the relation

$$-1 - \log[q(\mathbf{s})/m(\mathbf{s})] + \sum_{j=0}^{M} \lambda_j C_j(\mathbf{s}) = 0 \qquad (16.2.2.9)$$

and hence

$$q^{\mathrm{ME}}(\mathbf{s}) = m(\mathbf{s}) \exp(\lambda_0 - 1) \exp\left[\sum_{j=1}^{M} \lambda_j C_j(\mathbf{s})\right]. \qquad (16.2.2.10)$$

It is convenient now to separate the multiplier λ_0 associated with the normalization constraint by putting

$$\lambda_0 - 1 = -\log Z, \qquad (16.2.2.11)$$

where Z is a function of the other multipliers $\lambda_1, \ldots, \lambda_M$. The final expression for $q^{\mathrm{ME}}(\mathbf{s})$ is thus

$$q^{\mathrm{ME}}(\mathbf{s}) = \frac{m(\mathbf{s})}{Z(\lambda_1, \ldots, \lambda_M)} \exp\left[\sum_{j=1}^{M} \lambda_j C_j(\mathbf{s})\right]. \qquad (\mathrm{ME1})$$

The values of Z and of $\lambda_1, \ldots, \lambda_M$ may now be determined by solving the initial constraint equations. The normalization condition demands that

$$Z(\lambda_1, \ldots, \lambda_M) = \int_{\mathcal{A}} m(\mathbf{s}) \exp\left[\sum_{j=1}^{M} \lambda_j C_j(\mathbf{s})\right] \mathrm{d}\mu(\mathbf{s}). \qquad (\mathrm{ME2})$$

The generic constraint equations (16.2.2.5) determine $\lambda_1, \ldots, \lambda_M$ by the conditions that

$$\int_{\mathcal{A}} [m(\mathbf{s})/Z] \exp\left[\sum_{k=1}^{M} \lambda_k C_k(\mathbf{s})\right] C_j(\mathbf{s}) \, \mathrm{d}\mu(\mathbf{s}) = c_j \qquad (16.2.2.12)$$

for $j = 1, 2, \ldots, M$. But, by Leibniz's rule of differentiation under the integral sign, these equations may be written in the compact form

$$\frac{\partial(\log Z)}{\partial\lambda_j} = c_j \quad (j = 1, 2, \ldots, M). \qquad (\mathrm{ME3})$$

Equations (ME1), (ME2) and (ME3) constitute the *maximum-entropy equations*.

The maximal value attained by the entropy is readily found:

$$\begin{aligned} \mathcal{S}_m(q^{\mathrm{ME}}) &= -\int_{\mathcal{A}} q^{\mathrm{ME}}(\mathbf{s}) \log\left[q^{\mathrm{ME}}(\mathbf{s})/m(\mathbf{s})\right] \mathrm{d}\mu(\mathbf{s}) \\ &= -\int_{\mathcal{A}} q^{\mathrm{ME}}(\mathbf{s}) \left[-\log Z + \sum_{j=1}^{M} \lambda_j C_j(\mathbf{s})\right] \mathrm{d}\mu(\mathbf{s}), \end{aligned}$$

i.e. using the constraint equations

$$\mathcal{S}_m(q^{\mathrm{ME}}) = \log Z - \sum_{j=1}^{M} \lambda_j c_j. \qquad (16.2.2.13)$$

The latter expression may be rewritten, by means of equations (ME3), as

$$\mathcal{S}_m(q^{\mathrm{ME}}) = \log Z - \sum_{j=1}^{M} \lambda_j \frac{\partial(\log Z)}{\partial\lambda_j}, \qquad (16.2.2.14)$$

which shows that, in their dependence on the λ's, the entropy and $\log Z$ are related by Legendre duality.

Jaynes' theory relates this maximal value of the entropy to the prior probability $\mathcal{P}(\mathbf{c})$ of the vector $\mathbf{c}$ of simultaneous constraint values, *i.e.* to the size of the sub-ensemble of messages of length N that fulfil the constraints embodied in (16.2.2.5), relative to the size of the ensemble of messages of the same length when the source operates with the symbol probability distribution given by the prior prejudice *m*. Indeed, it is a straightforward consequence of Shannon's second theorem (Section 16.2.2) as expressed in equation (16.2.2.3) that

$$\mathcal{P}^{\rm ME}(\mathbf{c}) \propto \exp(\mathcal{S}), \tag{16.2.2.15}$$

where

$$\mathcal{S} = \log Z^N - \lambda \cdot \mathbf{c} = N\mathcal{S}_m(q^{\rm ME}) \tag{16.2.2.16}$$

is the total entropy for *N* symbols.

16.2.3. Adaptation to crystallography

16.2.3.1. *The random-atom model*

The standard setting of probabilistic direct methods (Hauptman & Karle, 1953; Bertaut, 1955*a*,*b*; Klug, 1958) uses implicitly as its starting point a source of random atomic positions. This can be described in the terms introduced in Section 16.2.2.1 by using a continuous alphabet $\mathcal{A}$ whose symbols **s** are fractional coordinates **x** in the asymmetric unit of the crystal, the uniform measure μ being the ordinary Lebesgue measure $d^3\mathbf{x}$. A message of length *N* generated by that source is then a random *N*-equal-atom structure.

16.2.3.2. *Conventional direct methods and their limitations*

The traditional theory of direct methods assumes a *uniform* distribution $q(\mathbf{x})$ of random atoms and proceeds to derive joint distributions of structure factors belonging to an *N*-atom random structure, using the asymptotic expansions of Gram–Charlier and Edgeworth. These methods have been described in Section 1.3.4.5.2.2 of *IT* B (Bricogne, 2001) as examples of applications of Fourier transforms. The reader is invited to consult this section for terminology and notation. These joint distributions of complex structure factors are subsequently used to derive conditional distributions of phases when the amplitudes are assigned their observed values, or of a subset of complex structure factors when the others are assigned certain values. In both cases, the *largest* structure-factor amplitudes are used as the conditioning information.

It was pointed out by the author (Bricogne, 1984) that this procedure can be problematic, as the Gram–Charlier and Edgeworth expansions have good convergence properties only in the vicinity of the expectation values of each structure factor: as the atoms are assumed to be uniformly distributed, these series afford an adequate approximation for the joint distribution $\mathcal{P}(\mathbf{F})$ only near the origin of structure-factor space, *i.e.* for *small* values of all the structure amplitudes. It is therefore incorrect to use these local approximations to $\mathcal{P}(\mathbf{F})$ near $\mathbf{F} = \mathbf{0}$ as if they were the global functional form for that function 'in the large' when forming conditional probability distributions involving large amplitudes.

16.2.3.3. *The notion of recentring and the maximum-entropy criterion*

These limitations can be overcome by recognizing that, if the locus $\mathcal{T}$ (a high-dimensional torus) defined by the large structure-factor amplitudes to be used in the conditioning data is too extended in structure-factor space for a single asymptotic expansion of $\mathcal{P}(\mathbf{F})$ to be accurate everywhere on it, then $\mathcal{T}$ should be broken up into sub-regions, and different local approximations to $\mathcal{P}(\mathbf{F})$ should be constructed in each of them. Each of these sub-regions will consist of a 'patch' of $\mathcal{T}$ surrounding a point $\mathbf{F}^* \neq \mathbf{0}$ located on $\mathcal{T}$. Such a point $\mathbf{F}^*$ is obtained by assigning 'trial' phase values to the known moduli, but these trial values do not necessarily have to be viewed as 'serious' assumptions concerning the true values of the phases: rather, they should be thought of as pointing to a patch of $\mathcal{T}$ and to a specialized asymptotic expansion of $\mathcal{P}(\mathbf{F})$ designed to be the most accurate approximation possible to $\mathcal{P}(\mathbf{F})$ on that patch. With a sufficiently rich collection of such constructs, $\mathcal{P}(\mathbf{F})$ can be accurately calculated anywhere on $\mathcal{T}$.

These considerations lead to the notion of *recentring*. Recentring the usual Gram–Charlier or Edgeworth asymptotic expansion for $\mathcal{P}(\mathbf{F})$ away from $\mathbf{F} = \mathbf{0}$, by making trial phase assignments that define a point $\mathbf{F}^*$ on $\mathcal{T}$, is equivalent to using a non-uniform prior distribution of atoms $q(\mathbf{x})$, reproducing the individual components of $\mathbf{F}^*$ among its Fourier coefficients. The latter constraint leaves $q(\mathbf{x})$ highly indeterminate, but Jaynes' argument given in Section 16.2.2.3 shows that there is a uniquely defined 'best' choice for it: it is that distribution $q^{\rm ME}(\mathbf{x})$ having *maximum entropy* relative to a uniform prior prejudice $m(\mathbf{x})$, and having the corresponding values $\mathbf{U}^*$ of the unitary structure factors for its Fourier coefficients. This distribution has the unique property that it rules out as few random structures as possible on the basis of the limited information available in $\mathbf{F}^*$.

In terms of the statistical mechanical language used in Section 16.2.1, the trial structure-factor values $\mathbf{F}^*$ used as constraints would be the macroscopic quantities that can be controlled externally; while the 3*N* atomic coordinates would be the internal degrees of freedom of the system, whose entropy should be a maximum under these macroscopic constraints.

16.2.3.4. *The crystallographic maximum-entropy formalism*

It is possible to solve explicitly the maximum-entropy equations (ME1) to (ME3) derived in Section 16.2.2.4 for the crystallographic case that has motivated this study, *i.e.* for the purpose of constructing $q^{\rm ME}(\mathbf{x})$ from the knowledge of a set of trial structure-factor values $\mathbf{F}^*$. These derivations are given in §3.4 and §3.5 of Bricogne (1984). Extensive relations with the algebraic formalism of traditional direct methods are exhibited in §4, and connections with the theory of determinantal inequalities and with the maximum-determinant rule of Tsoucaris (1970) are studied in §6, of the same paper. The reader interested in these topics is invited to consult this paper, as space limitations preclude their discussion in the present chapter.

16.2.3.5. *Connection with the saddlepoint method*

The saddlepoint method constitutes an alternative approach to the problem of evaluating the joint probability $\mathcal{P}(\mathbf{F}^*)$ of structure factors when some of the moduli in $\mathbf{F}^*$ are large. It is shown in §5 of Bricogne (1984), and in more detail in Section 1.3.4.5.2.2 of Chapter 1.3 of *IT* B (Bricogne, 2001), that there is complete equivalence between the maximum-entropy approach to the phase problem and the classical probabilistic approach by the method of joint distributions, provided the latter is enhanced by the adoption of the saddlepoint approximation.

references