The Maximum-Entropy Method without the Positivity Constraint – Applications to the Determination of the Distance-Distribution Function in Small-Angle Scattering

BY STIG STEENSTRUP
Niels Bohr Institute, University of Copenhagen, Universitetsparken 5, 2100 KBH Ø, Denmark
AND STEEN HANSEN
Department of Mathematics and Physics, Royal Veterinary and Agricultural University, Thorvaldsensvej 40, 1871 FRB C, Denmark

(Received 5 March 1993; accepted 25 January 1994)

Abstract
The maximum-entropy principle (MaxEnt) is well suited to the solution of underdetermined problems by inclusion of prior knowledge in a logically consistent way. In most applications of MaxEnt, a set of numbers – pixel densities in an image or counts in a spectrum – are determined. In these cases, the set of numbers can be interpreted as a probability distribution and is, as such, all positive. It is the purpose here to show that MaxEnt is able to provide estimates of a set of quantities that cannot necessarily be interpreted as probabilities and that may become negative, as is the case for the correlation function in small-angle scattering. The method is illustrated both by analysis of simulated data and by measurements on sodium dodecylsulfate.

1. Introduction
The maximum-entropy method (MaxEnt) has proved to be powerful for handling underdetermined problems. Typical uses are encountered when the 'state' of a system, described e.g. by a state vector \( \mathbf{f} \), is not uniquely determined by measured data \( \mathbf{d} \). The maximum-entropy principle may be applied in such cases to provide a probability distribution for the state of the system. From such a probability, a 'best' estimate of the state is provided by the mean value of the state over the probability distribution.

For a large class of problems, the state vector \( \mathbf{f} \) lends itself to being interpreted as a probability distribution (see e.g. Jaynes, 1986). This would be the case, for example, for the number of counts measured in a channel of a one-dimensional spectrum or the number of photons detected in a pixel of a two-dimensional picture. In these cases, the state vector has often been taken as being proportional to the probability distribution calculated by MaxEnt.

It is the aim here to show that MaxEnt does not require such an identification and hence is not restricted to cases where the estimated state vector has to be positive. It is only the probability distribution of the state vector that necessarily must be non-negative. Furthermore, this distribution allows evaluation of a goodness of fit as a variance can be calculated.

Previously, the maximum-entropy method has been applied to nonpositive distributions (see e.g. Laue, Skilling & Staunton, 1985; David, 1990; Sivia, Hamilton, Smith, Rieker & Pynn, 1991) using a splitting of the estimated distribution in a positive and a negative channel. The approach suggested in the present paper is different: by rewriting the fundamental equations for MaxEnt, it is demonstrated how nonpositive states can be included in the original MaxEnt framework, leading to a well known quadratic constraint.

The usefulness of the method is shown by the inclusion of examples from small-angle scattering using simulated and experimental data. Intensity in a small-angle scattering experiment is related to the distribution function (correlation function) for the scattering medium. By a suitable normalization, the scattering intensity is the Fourier transform of this distribution function, which can take positive and negative values.

2. Basic equations
The basis of MaxEnt has been described extensively, e.g. by Jaynes (1983), and only the main features are given here.

It is assumed that data \( d_1, \ldots, d_m \) are measured and are related to the state vector \( f_1, \ldots, f_n \) by

\[
d_i = \sum_{j=1}^{n} A_{ij} f_j + e_i, \quad i = 1, \ldots, m,
\]

where \( A \) is a given matrix (not necessarily square or invertible) and the data have errors \( e_1, \ldots, e_n \) so,
even in the cases where \( m = n \), (1) does not provide a unique determination of \( \mathbf{f} \).

MaxEnt consists of settling for less, namely in evaluating a probability distribution \( P(\mathbf{f}) \) by requiring that its entropy is maximum. It is appropriate to include a prior probability distribution \( P^0(\mathbf{f}) \); this is in fact mandatory when the components of \( \mathbf{f} \) are continuous variables, in which case one should use the Kullback directed divergence

\[
I[\mathbf{P}; P^0] = \sum P(\mathbf{f}) \ln \frac{P(\mathbf{f})}{P^0(\mathbf{f})},
\]

where the sum is over all possible values of \( \mathbf{f} \) (being replaced by integrals for the continuous case). Note that, for \( P^0(\mathbf{f}) \) uniform, \( I[\mathbf{P}; P^0] \) is equal to a constant minus the normal entropy \( S = -\sum P(\mathbf{f}) \ln P(\mathbf{f}) \). Maximizing the entropy or minimizing \( I[\mathbf{P}; P^0] \) is equivalent. The MaxEnt distribution is now obtained by minimizing (2) with the constraints (1), where \( f_j \) now has to be replaced by \( \langle f_j \rangle \), the mean value. Standard Lagrange-multiplier techniques with Lagrange multipliers \( \lambda_1, \ldots, \lambda_m \) yield for the MaxEnt distribution

\[
P(\mathbf{f}) = Z^{-1} P^0(\mathbf{f}) \exp \left( -\sum_i \lambda_i \sum_j A_{ij} f_j \right),
\]

where \( Z \) is a normalization constant given by

\[
Z = \sum \mathbf{P}(\mathbf{f}) \exp \left( -\sum_i \sum_j \lambda_i A_{ij} f_j \right).
\]

Determination of the Lagrange multipliers from the constraints now yields the complete probability distribution and a set of best values of \( \mathbf{f} \) is obtained as a set of the mean values \( \langle \mathbf{f} \rangle = \sum \mathbf{f} P(\mathbf{f}) \).

In practice, however, it is not very convenient numerically to evaluate a function in \( n \) variables. Fortunately, for a number of cases of interest it is possible to solve part of the problem and recast the optimization into one involving only mean values. The ease with which this can be done depends on the space in which \( \mathbf{f} \) takes values and on the choice of the prior.

For instance, it has been shown (Steenstrup & Wilkins, 1984; Steenstrup, 1985) that, if each \( f_j \) can only take on integer values and the prior is a product of Poisson distributions

\[
P^0(\mathbf{f}) = \prod_{j=1}^n \exp \left( -m_j \right) \frac{m_j^{f_j}}{f_j!},
\]

then the \( I[\mathbf{P}; P^0] \) for the MaxEnt distribution \( \mathbf{P} \) has the value

\[
I[\mathbf{P}; P^0] = \sum_{j=1}^n \left[ \langle f_j \rangle \ln \left( \frac{\langle f_j \rangle}{m_j} \right) - \langle f_j \rangle + m_j \right].
\]

The assumption of Poisson distributions applies well to the situations of counting statistics. Expression (6) (without the mean value signs) has also been obtained by Skilling (1988, 1989) on the basis of four axioms and interpreting \( \mathbf{f} \) as being proportional to a probability distribution.

In many cases, however, the assumptions that \( f_j \in \mathbb{N} \) and that the prior is a product of Poisson distributions are hardly appropriate. Among other reasons, the fact that the physics of the problem may allow negative values of \( \mathbf{f} \) is the most important for the example considered in the following section. Furthermore, in this case the measurements (and counting statistics) take place in reciprocal space, while the distribution to be estimated lies in direct space.

We instead assume that each \( f_j \) is a real variable with range from \( -\infty \) to \( \infty \) and that the prior is a product of Gaussians:

\[
P^0(\mathbf{f}) = \prod_{j=1}^n \frac{1}{\sqrt{2\pi \sigma_j^2}} e^{-\frac{(f_j - m_j)^2}{2\sigma_j^2}},
\]

or

\[
P^0(\mathbf{f}) = (2\pi)^{-n/2} (\sigma_1 \ldots \sigma_n)^{-1} \exp \left\{ -\sum_{j=1}^n \frac{(f_j - m_j)^2}{2\sigma_j^2} \right\}.
\]

With the assumption of the Gaussian prior, we have an equal probability of a deviation \( +\delta \) and a deviation \( -\delta \) from the expected mean value \( m \). The MaxEnt distribution (3) and the state sum \( Z \) [[4]] can now be evaluated in closed form:

\[
Z = \exp \left\{ \sum_{j=1}^n \left( \mu_j^2 \sigma_j^2 / 2 \right) - \mu_j m_j \right\},
\]

and

\[
P(\mathbf{f}) = (2\pi)^{-n/2} (\sigma_1 \ldots \sigma_n)^{-1} \exp \left\{ -\sum_{j=1}^n \frac{(f_j - m_j - \mu_j \sigma_j^2)^2}{2\sigma_j^2} \right\},
\]

where the abbreviation \( \mu_j = \sum_{j=1}^m \lambda_j A_{ij} \) has been introduced.

From (10), it is seen that \( \langle f_j \rangle = m_j - \mu_j \sigma_j^2 \). Solving for \( \mu_j \) and inserting it into the expression

\[
I[\mathbf{P}; P^0] = -\ln Z - \sum_{i=1}^m \lambda_i d_i
\]

then

\[
I[\mathbf{P}; P^0] = -\sum_{j=1}^n \left( \mu_j^2 \sigma_j^2 / 2 \right) - \mu_j m_j + \mu_j \langle f_j \rangle,
\]

one obtains the following simple expression for the directed divergence:

\[
I[\mathbf{P}; P^0] = \sum_{j=1}^n \left( \langle f_j \rangle - m_j \right)^2 / 2\sigma_j^2,
\]

Expression (6) (without the mean value signs) has also been obtained by Skilling (1988, 1989) on the basis of four axioms and interpreting \( \mathbf{f} \) as being proportional to a probability distribution.
i.e. a form of the well known quadratic constraint [for an alternative derivation of (6) and (12) see Kullback (1959)].

Instead of evaluating a probability distribution in \( n \) variables, it is now possible to work directly with mean values. It suffices to minimize (12) subject to constraints (1) or if errors are included by constraining the value of \( \chi^2 \):

\[
\chi^2 = \sum_{i=1}^{n} \left[ \left( d_i - \sum_j A_{ij} \langle f_j \rangle \right)^2 / 2s_i^2 \right],
\]

where \( s_i \) is the standard deviation of the error of data point \( d_i \). This is obviously done by solving

\[
\nabla \left\{ \sum_{j=1}^{n} \left[ (\langle f_j \rangle - m_j)^2 / 2\sigma_j^2 \right] + \lambda \chi^2 \right\} = 0,
\]

where \( \lambda \) is a Lagrange multiplier allowing the \( \chi^2 \) to reach a predetermined value. This predetermined value should be provided \textit{a priori} for a simple application of MaxEnt (and will usually be provided from \textit{a priori} information about the nature of the experiment and experience). Alternatively (and numerically slightly more complicated), the Lagrange multiplier can be found by Bayesian estimation, as shown by Gull (1989). However, as demonstrated by Bryan (1990), the Bayesian estimation of the Lagrange multiplier in MaxEnt has to be performed with caution to avoid over- or underfitting the data.

Equation (14) can be solved numerically by a successive over/under-relaxation algorithm [as described by Steenstrup (1985)]. For estimation of one-dimensional spectra (having just a few hundred points), this is done in seconds on a PC.

It is noted that a second-order Taylor approximation of (6) gives

\[
I[P; P^0] \simeq \sum_{j=1}^{n} \left[ (\langle f_j \rangle - m_j)^2 / 2m_j \right],
\]

for \( \langle f \rangle \simeq m \). In consequence, (6) and (12) will give identical results when \( c_j^2 = m_j \) and \( \langle f \rangle \simeq m \), expressing the fact that the Poisson distribution may be approximated by a Gaussian (provided of course that \( m_j \) is large enough so that integrals over \( f_j \) can be extended to \(-\infty\) to a good approximation).

We stress that (12) in itself does \textit{not} give an entropy. It just happens that, for the given space and the given prior, the values given by (12) and for the entropy by (2) are the same when the MaxEnt distribution is inserted in (2) and when the values of \( \langle f_j \rangle \) minimizing (12) subject to the constraints are inserted in (12). Whatever the nature of the problem, \textit{i.e.} the space, the prior and the explicit form of the constraints, MaxEnt can be invoked using (2). In particular cases, simplifications are possible, and as we have shown the simplification can be extended to cases with nonpositive state vectors.

When using MaxEnt, it is important first to consider how its basic assumptions apply to the problem addressed. It must be considered whether it is sensible to assume that each \( f_j \) is Poisson distributed – or if the four axioms of Skilling are fulfilled. If this is not the case, a new prior probability distribution for \( f_j \) should be found – and, subsequently, the entropy with respect to \textit{this} probability distribution should be maximized. Second, it is important to express the ‘default’ values – the prior – \( m \) so as to incorporate the maximum possible prior information in \( P(f) \). In many of the cases where MaxEnt gives disappointing results, these two points might not have been given sufficient consideration.

It is noted that, in the absence of data points, the estimated function will be identical to the \textit{a priori} estimated mean value, \textit{e.g.} identically zero if \( m = 0 \) has been chosen as the prior estimate.

### 3. Results for small-angle scattering

Among its many possible applications, small-angle scattering (SAS) is used for obtaining structural information about molecules in solution. The interest of performing solution experiments appears, for example, in biophysics, where it may be crucial to preserve the exact and functionally active structure of the biomolecule. The loss of information owing to the random orientation of the molecules in the solution makes it important to extract the maximum information from the measured scattering profile. For interpretation of the experimental results, it is often relevant to represent the scattering data in direct space, which requires a Fourier transformation of the data preserving the full information content of the experimental data. A direct Fourier transform is of limited use, owing to noise, smearing and truncation.

Attempts to take these effects into account by indirect Fourier transformation (IFT) have been suggested in the literature \textit{e.g.} Glatter, 1977; Moore, 1980; Svergun, Semenyuk & Feigin, 1988; Hansen & Pedersen, 1991).

In small-angle scattering, the intensity \( I \) is measured as a function of the length of the scattering vector \( q = 4\pi \sin (\theta) / \lambda \), where \( \lambda \) is the wavelength of the radiation and \( \theta \) is half the scattering angle. For scattering from a dilute solution of monodisperse molecules of maximum dimension \( D \), the intensity can be written in terms of the distance distribution function \( p(r) \):

\[
I(q) = 4\pi \int_0 p(r) \left[ \sin \left( qr / qr \right) \right] dr.
\]

\[16\]
The distance distribution function is related to the density-density correlation \( \gamma(r) \) of the scattering-length density \( \rho(r) \) by

\[
\rho(r) = r^2 \gamma(r) = r^2 \left( \int_V \rho(r') \rho(r + r') \, dr' \right),
\]

(17)

where \( \rho(r) \) is the scattering contrast, given by the difference in scattering density between the scatterer \( \rho_{\text{sc}}(r) \) and the solvent \( \rho_{\text{so}} \), i.e. \( \rho(r) = \rho_{\text{sc}}(r) - \rho_{\text{so}} \) means averaging over all orientations of the molecule and \( V \) is the volume of the molecule. The possibility of negative contrast may give negative regions in the distance-distribution function, which will also become negative in the region around \( D \) for high concentrations of the scatterer (see e.g. Glatter, 1982). For uniform scattering density of the molecule, the distance distribution function is proportional to the probability distribution for the distance between two arbitrary scattering points within the molecule.

For the case of polydispersity, the size distribution for the molecules can be calculated by a similar indirect Fourier transformation if the shape of the molecules is known (usually spheres are assumed). The mathematics of this problem is completely analogous to the problem of estimation of the distance-distribution function for monodisperse systems. The actual calculations may cause additional difficulties, as the system of equations to be solved for calculations of a size distribution is often more ill conditioned than for the determination of the distance-distribution function (as will appear by a singular value decomposition of the transformation matrix). However, the numerical problems encountered are similar.

The maximum-entropy method has been used for estimation of size distributions (Daniell, Potton & Rainford, 1988; Morrison, Corcoran & Lewis, 1992) and distance-distribution functions (Hansen & Pedersen, 1991) in small-angle scattering. These previous applications used (6) as the constraint on the estimated distributions. However, there is nothing in the nature of these problems that requires the estimated distributions to obey Poisson statistics as expressed by maximization of (6). Furthermore, for the estimation of distance-distribution functions, it is frequently possible for the distribution to take negative values, as noted above.

Minimizing the square of the distance from a priori estimation as a constraint on IFT in small-angle scattering was previously suggested by Glatter (1977); however, this approach was discarded as the influence of the Lagrange multiplier (balancing the importance of the \( \chi^2 \) and the additional constraint) not negligible. But, according to the derivation above, it should not be negligible as it governs the extent to which the prior estimate is allowed to influence the final solution. The ability to include prior information in an easy, transparent and logically consistent way may be considered the major advantage of maximum entropy compared to other methods of data analysis. This includes the possibility of constraining the solution to be positive if it is known a priori that the scattering contrast has the same sign for all parts of the molecule and that no concentration effects are present. Certainly, for estimation of size distributions, the optional positivity constraint must be considered an advantage.

The methods of Tikhonov & Arsenin (1977) allows the inclusion of prior knowledge by the general form of the stabilizer:

\[
\Omega(f, m, \rho) = \| f - m \|^2 + \rho \| f' \|^2
\]

(18)

The first term minimizes the deviation from the prior estimate with respect to a given norm and the second term imposes a smoothness constraint similar to that of Glatter (1977) on the distribution to be estimated. However, usually only the second term is used when applying Tikhonov regularization to underdetermined problems [see e.g. Svergun, Semenyuk & Feigin (1988) for the application of the method to small-angle scattering]. The above derivation by maximum entropy underlines the importance of the first term of the Tikhonov regularization and furthermore provides guidelines on how to choose the norm by the identification of the \( \sigma_i \)'s as the individual widths of the Gaussian priors. However, the addition of a smoothness constraint to the maximum-entropy method is of course also quite legitimate and has been tested (Charter & Gull, 1991).

In many situations, the usually applied smoothness constraint of Glatter (1977) (minimizing the curvature of the distribution function) is a very sensible way of expressing the prior information that a small-angle scattering experiment often has low resolution. What is offered by the maximum-entropy method is simply an additional way of expressing prior information. For example, given some prior information about the geometrical shape of the scatterer, MaxEnt is likely to give better results than a mere smoothness constraint and should consequently be considered as an additional analytical tool in small-angle scattering.

In MaxEnt, the estimation of the errors can be done in the conventional way using the usual curvature matrix for the \( \chi^2 \) but including the additional entropy-regularization term as described by Gull (1989).

Examples are given below for both simulated and experimental data.

**Simulated data**

Fig. 1(a) shows a simulated scattering profile obtained by adding Gaussian noise to a theoretical scattering profile calculated from a spherical model. The model consists of a sphere of radius 16 Å and
scattering density $-1.5$ inside a shell of internal radius $16\,\text{Å}$, external radius $23\,\text{Å}$ and scattering density $1$. The added noise is shown in the figure by the error bars on simulated data points. The MaxEnt fit is also shown. The corresponding MaxEnt estimate of the distance-distribution function is shown in Fig. 1(b) with the original distance-distribution function for the model. The positions of the peaks and the overall shape of the curves show good agreement between the two distance-distribution functions.

Another simulated scattering profile constructed as for Fig. 1(a) but including an extra shell in the molecule is shown in Fig. 2(a). The radii were now 20, 35 and $50\,\text{Å}$ and the scattering lengths $1$, $-2$ and $1$, respectively. The MaxEnt estimate of the distances-distribution function is presented in Fig. 2(b) and still shows good agreement with the original distribution.

The above MaxEnt calculations were done assuming that the maximum length $D_{\text{max}}$ used for the IFT was equal to the maximum dimension of the scatterer. Furthermore, the noise level $\chi^2$ was assumed to be known. If these two parameters are not known in advance all is not lost but the various methods of IFT mentioned above provide suggestions as to how $D_{\text{max}}$ and $\chi^2$ can be estimated during the calculations (see also Svergun, 1992). Exactly how sensitive the present method is to errors in the estimation of $\chi^2$ and $D_{\text{max}}$ depends of course on the quality of the experimental data.

$m = 0$ was chosen for the prior in both cases. Furthermore, the width of the Gaussian $\sigma_j$ in (7) was varied over $[0:D_{\text{max}}]$. Somewhat arbitrarily, the variation itself was chosen as a Gaussian but the main effect of this variation was to ‘tie down’ the estimate of the distance-distribution function at the end points as it is known \textit{a priori} that $p(0) = p(D_{\text{max}}) = 0$. A similar approach has been suggested by Wilkins, Steenstrup & Varghese (1985).

Fig. 1. (a) Simulated scattering profile for a two-phase sphere. Dimensions are as described in the text. Simulated data: error bars connected by full line. MaxEnt fit: dashed line. (b) Distance distribution for the two-phase sphere. Original distribution: full line. MaxEnt estimate: dashed line [corresponding to the fit shown in (a)].

Fig. 2. (a) Simulated scattering profile for a three-phase sphere. Dimensions are as described in the text. Simulated data: error bars connected by full line. MaxEnt fit: dashed line. (b) Distance distribution for the three-phase sphere. Original distribution: full line. MaxEnt fit: dashed line [corresponding to the fit shown in (a)].
In Fig. 3, the result of a previously published MaxEnt estimate obtained using (6) (Hansen & Pedersen, 1991) is compared to one obtained using (12). In this case, the scatterer was an object consisting of eight spheres and the original distance-distribution function was calculated by May & Nowotny (1989). For both of the present calculations, a spherical prior was assumed and the maximum diameter of the scatterer was found by the minimum value of (6) and (12), respectively. For the estimate using (12), it was assumed that $\sigma_r^2 = m$. It is apparent that the two methods give approximately the same result. This is expected from the comparison of the quadratic approximation for the entropy (15) and (6).

Experimental data

Fig. 4(a) shows data from small-angle X-ray scattering measurements on sodium dodecylsulfate (SDS) at a concentration of 10 mg ml$^{-1}$ (Samsø, Daban & Jones, 1994). The corresponding estimate of the distance-distribution function is shown in Fig. 4(b). For this calculation, a $D_{\text{max}}$ of 65 Å was used and the prior was chosen as for the simulated data. SDS consists of amphiphiles that have the ability to form micelles in solution. The micelles are expected to have an inner spherical region of low scattering density corresponding to the hydrophobic CH$_2$ tails of the amphiphiles and two outer shells consisting of the hydrophilic polar heads and some ordered structure of solvent around the micelle. The dimensions derived from the distance-distribution function are 16 Å for the radius of the inner sphere, 23 Å for the external radius of the shell of polar heads and an overall maximum correlation length of 65 Å owing to the ordering around the micelle. The inner part of the distance-distribution function is seen to correspond well with the simulated example in Fig. 1 [the offset at $p(0)$ is probably due to incorrect background subtraction]. Finally, as our present aim is to demonstrate the connection between MaxEnt and the well known quadratic constraint, as well as to investigate the applicability of the quadratic constraint to the analysis of small-angle scattering data, we desist from any attempt to analyze the structural information in Fig. 4(b) in further detail but simply note that the depicted distance-distribution function is in good agreement with the SDS structure obtained from previous experiments (e.g. Cabane, Duplessis & Zemb, 1985; Zemb & Charpin, 1985; Bezzobotnov et al., 1988).

4. Concluding remarks

From the expression for the Kullback directed divergence, it has been demonstrated that MaxEnt includes estimation of nonpositive distributions leading to a well known quadratic constraint. The test examples given for small-angle scattering show good
agreement between original and estimated distributions and illustrate the applicability of MaxEnt within this particular area.

The main part of this work was done at CSIRO Division of Materials Science and Technology, Australia, of which the hospitality, inspiration and financial support are greatly appreciated. We thank Stephen Wilkins (CSIRO) for inspiring discussions and Gareth Jones (BSL Daresbury Laboratory) for making SDS data available to us. Financial support from the Danish Natural Science Research Council is also acknowledged.

References


