THE MAXIMUM ENTROPY FORMALISM

A Conference Held at the Massachusetts Institute of Technology on May 2-4, 1978

Edited by Raphael D. Levine and Myron Tribus

The MIT Press
Cambridge, Massachusetts, and London, England
AN ALGORITHM FOR DETERMINING THE LAGRANGE PARAMETERS IN THE MAXIMAL ENTROPY FORMALISM

N. Agmon, Y. Alhassid and R. D. Levine

This conference discussed a formalism where one needs to determine the maximum of the 'missing information' or entropy function

\[ S = \sum_{i=1}^{n} x_i \ln x_i \]  

subject to \( m \) constraints

\[ \sum_{i=1}^{n} x_i A_{ri} = b_r \quad r = 1, \ldots, m. \]  

Here we consider the practical problem of explicitly determining the distribution \( p_1, \ldots, p_n \) which gives the maximum of \( S \) subject to the stated constraints. A flow chart, user's guide and a listing of the actual computer program is available from us upon request.

In principle, the problem is solved by the conventional Lagrange multipliers method. The distribution \( p \) of maximal entropy is then given by

\[ p_i = z^{-1} \exp\left(- \sum_{r=1}^{m} \lambda_r A_{ri} \right) \]  

where \( z = \sum_{i=1}^{n} \exp\left(- \sum_{r=1}^{m} \lambda_r A_{ri} \right) \) is the partition function. To calculate the probability function \( p \) one needs yet to solve the set of \( m \) coupled, implicit nonlinear equations

\[ \sum_{i=1}^{m} p_i A_{ri} = b_r \quad r = 1, \ldots, m \]  

for the Lagrange parameters \( \lambda = (\lambda_1, \ldots, \lambda_m) \). So, although in principle we know how to solve the maximal entropy problem, in practice it is not so, unless one has an efficient algorithm for solving the set of nonlinear equations 4 for the Lagrange parameters.

Consider equations 3 with a set of 'trial' Lagrange parameters, \( \lambda^t = (\lambda_1^t, \ldots, \lambda_m^t) \). Then the left-hand side of equation 4 can be written as \( f(\lambda^t) \) where \( f = (f_1, \ldots, f_m) \). The desired set of Lagrange parameters \( \lambda \) is the solution of the implicit, nonlinear equation

\[ f(\lambda^t) = 0 \]  
Without invoking any special properties of \( f \), one can try some of the standard methods for solving nonlinear equations. For example, the function

\[
\psi(\lambda^t) \equiv \sum_{i=1}^{m} f_i^2(\lambda^t)
\]

is non-negative and vanishes if and only if \( f \) vanishes. Therefore, one can try to find the minimum of \( \psi(\lambda^t) \), using the Newton-Raphson iteration procedure, which converges in principle (i.e. overlooking problems of computer's round-off errors) for any initial guess of \( \lambda^t \) if \( \psi \) is a strictly convex function. In the case of the maximal entropy problem, \( \psi \) is not convex even for an example as simple as a single constraint given by the first moment (i.e. \( m=1, A_{11}=1 \)).

Thus, blindly applying a conventional procedure cannot guarantee a solution. The reason is that we have ignored the special properties of the function \( f \) of the present case.

More explicitly, the Jacobian matrix \( M \) of \( f \) which is given by

\[
M_{rs} = \frac{\partial^2 f}{\partial \lambda_r \partial \lambda_s} = -\frac{\partial <A_r - b>_r}{\partial \lambda_r} = \langle (A_r - A_r^t)(A_s - A_s^t)^t \rangle^t
\]

(6)

is a symmetric, positive definite matrix. The first property is immediately transparent from equation 6, and the second property (which is not much harder to show) is proved in (1).

The first property entails the existence of a scalar 'potential function' \( P(\lambda^t) \) such that \( f = \nabla P \). The second property means that \( P \) is a strictly convex function (it has a positive definite Hessian \( M \)). Therefore, \( P \) has a unique global minimum for the \( \lambda^t \) which solves \( \nabla P(\lambda^t) = 0 \). In this manner the problem of solving equation 4 is transformed to finding the minimum of the strictly convex function \( P \). To do that we use the modified Newton-Raphson procedure which is now guaranteed to converge for any initial guess. The algorithm is simple: if for \( k \)th iteration the Lagrange multipliers have the value \( \lambda(k) \), \( \lambda(k+1) \) is determined as follows: first one chooses a vector \( u \) which is directed as closely as possible towards the minimum \( \lambda \). \( u = -M^{-1}\nabla P = -M^{-1}f \) is a better choice than \( u = -\nabla P \) because it is the minimum of the second order Taylor expansion of \( P \). Next a number \( \alpha \) is chosen such that \( \lambda^t(k+1) \), defined by

\[
\lambda^t(k+1) = \lambda^t(k) - \alpha u,
\]

(7)

is the minimum of \( P \) along the direction \( u \). It is precisely this determination of \( \alpha \) which cannot be accomplished in a
An Algorithm for Determining the Lagrange Parameters

conventional iteration method where \( F \) is unknown.

Finally, let us mention the explicit form of \( F \) (discussed in (2)), namely:

\[
F(\lambda^t) = \ln z(\lambda^t) + \sum_{r=1}^{m} \lambda^t_r b_r
\]

(8)

(It is easily seen that indeed \(- (\nabla F)_r = <A_r>^t - b_r\). It should be emphasized that in equation 8 the \( b_r \)'s are the given 'experimental' averages which are held constant while the \( \lambda^t_r \)'s are varied independently. Only at the minimum of \( F \) one has that \( \lambda^t = \lambda \), the desired Lagrange parameters conjugate to the \( b \)'s. Then \( F(\lambda) = S \) (so that \( F \) is an upper bound to the entropy) and the Hessian \( M \) is the correlation matrix.

A computer program has been written utilizing the above algorithm. It includes also necessary consistency checks, of which a fuller account is given in (3). The program's source, listing, flow chart and user's guide are available from the authors.

*Work supported by the Office of Scientific Research, USAF, under grant AFOSR 77-3135.

**Permanent address: The Hebrew University of Jerusalem, Jerusalem, Israel.

