Acta Cryst. (1983). A39, 622–631

Constrained Density Modifications by Variational Techniques

BY JORGE NAVAZA, E. E. CASTELLANO* AND G. TSOUCARIS

Laboratoire de Physique, Tour B, Centre Universitaire Pharmaceutique, 92290 Chatenay Malabry, France

(*Received* 1 *November* 1982; *accepted* 11 *March* 1983)

Abstract

A general method for phase extension and refinement is described. It consists of minimizing a conveniently defined functional M by the steepest-descent or conjugate-gradient techniques. This functional consists of a term $R = \frac{1}{2}\sum(|F| - |F^{obs}|)^2$ plus any given number of constraint equations selected according to the requirements of the particular problem considered. Several examples of the introduction of constraints are described in detail, both in direct and reciprocal space, and a few one-dimensional tests provide insight in the behavior of the different alternatives. Also, some general features that the calculated electron density function should fulfil, such as positivity and boundness, are directly introduced without the need of terms other than R in the functional M. The connection between this approach and the ones which use the principle of maximum entropy is discussed. A critical analysis shows that those methods are just different ways of minimizing R with the constraints of positivity.

Introduction

We wish to discuss here a general variational procedure for phase determination and refinement which is operationally related to the methods known under the broad denomination of 'electron density modifications' [see, for example, the review article by Sayre (1980)].

In general, density modifications are iterative procedures that can be schematically described, borrowing from the notation of Zwick, Bantz & Hughes (1976), in the form:

$$\rho^j \xrightarrow{M} \rho_m^j \xrightarrow{T} [F_c,\varphi_c]^j \xrightarrow{S} [F_s,\varphi_s] \xrightarrow{T^{-1}} \rho^{j+1}.$$

That is, electron density at step j is modified to ρ_m^j and its Fourier transform calculated; then a synthesis is performed (often consisting in simply restoring the moduli of the calculated F's to their observed values) and finally the electron density for step $j + 1$ is obtained by an inverse Fourier transform. The nature of the modification operator M and the synthesis operator S will in general depend upon the type of information that is available or the 'model' of the electron density function one would like to adopt. Although this procedure may eventually converge in favourable circumstances, there is no guarantee that it will do so in general and, moreover, there is no way of determining the 'size' of the modifications M and S along the refining trajectory in order to optimize convergence.

The method we shall describe is a simple variational formalism which aims to overcome these deficiencies. The calculus of variations is concerned with the problem of determining maxima or minima, or in general stationary values of *functionals* of specific *argument functions* defined in some given domain and, as we shall show, virtually all problems related to the lack of knowledge of the structure-factor phases in crystallography, such as electron density refinement, or phase extension, or even *ab initio* phase calculations, can be formulated in its framework.

A few related methods have been proposed. Khachaturyan, Semenovskaya & Vainshtein (1981) have described a procedure for *ab initio* phase determination based on equations derived from statistical thermodynamic analogies which are solved by variational techniques. In general terms, our present formalism is closely related to that of these authors. We believe, however, that the eventual power of their technique lies more on its variational foundations than on the physical analogs from which the basic equations are derived. In particular, we shall show that their method is equivalent to some form of adjusting the observed and calculated moduli of the structure factors with some specific constraints, which impose positivity on the electron density function and bound it between the values zero and one.

Also related to this method are the ones, recently introduced in connection with the phase problem, which involves the *principle of maximum entropy* as their fundamental basis. Because of the considerable expectation that this last approach has raised due to its potential usefulness in protein phase extension and

* Present address: Instituto de Física e Química de São Carlos, Departamento de Física e Ciência dos Materiais, *CP* 369, 13.560 São Carlos, SP, Brasil.

0108-7673/83/050622-10$01.50

refinement (Gull & Daniel, 1978; Collins, 1982) and even as an alternative procedure for *ab initio* phase determination,* we carried out some three-dimensional test calculations in order to compare their performance with similar numerical methods based on different physical hypotheses.

In practical numerical calculations the number of variables has to be kept necessarily finite (typical examples are the values of the electron density at the points of a grid in the unit cell) and essentially variational problems take the form of ordinary extremum problems. The Sayre's well known least-squares phase refinement (Sayre, 1972, 1974) is an example of a variational calculation which, owing to the finite discretization introduced by resolution, has been naturally formulated as an ordinary extremum problem.

The procedure that we shall describe aims at a new general way of attacking the phase problem. We shall show that once a few operational rules are stated, a specific procedure can be generated for almost any set of conditions that define a particular problem, either in direct or reciprocal space. Several previously proposed methods can be combined into a general unified frame thus gaining in understanding and physical insight.

A fuller account of practical applications will be the subject of a forthcoming publication.

The variational formalism

We start be recalling the main concepts of variational calculus, which has its origin in the search of extrema of *functionals* rather than extrema of functions of a finite number of variables.

The *domain* of a functional is a set of admissible functions rather than a region of a coordinate space. A functional associates a number to the entire course of a function which is called the *argument function*. A simple example in crystallography is the value of a particular structure factor which can be considered as a functional of the electron density ρ.

We state now the following variational problem: In a given domain of admissible electron density functions ρ find the function for which a given functional is a minimum with respect to all argument functions of the domain.

As a specific example let us consider the quadratic crystallographic unnormalized R factor

$$R = \frac{1}{2} \sum_h [|F(h)| - |F(h)|^{\text{obs}}]^2, \tag{1}$$

where $|F(h)|$ is the observed structure-factor magnitude measured in units of $|F(0)|^{\text{obs}}$. The $F(h)$ are then non-dimensional quantities which we shall relate to a non-dimensional electron density function through:

$$F(h) = \frac{1}{V} \int \rho(r) \exp(-2\pi i h r)\, \mathrm{d}^3 r \tag{2}$$

(where $\mathrm{d}^3 r$ denotes the differential volume element), and its inverse relationship

$$\rho(r) = \sum_h F(h) \exp(2\pi i h r). \tag{3}$$

This normalization proves very convenient for subsequent numerical and formal analysis and we shall adopt it throughout the paper. As a consequence of (2) the true electron density function satisfies the normalization condition

$$\frac{1}{V} \int \rho(r)\, \mathrm{d}^3 r = 1. \tag{4}$$

Owing to (2), R can be considered as a functional of ρ and our problem is to find the particular electron density function which makes R a minimum. We start by calculating the variation of R when ρ is changed to the new function $\rho + \delta\rho$. $\delta\rho$ is called the *variation* of ρ. According to (1), R changes because each $F(h)$ changes with $\delta\rho$ so that what we need are the variations $\delta F(h)$:

$$\delta F(h) = \frac{1}{V} \int \exp(-2\pi i h r)\, \delta\rho(r)\, \mathrm{d}^3 r. \tag{5}$$

We then have

$$\begin{aligned} \delta R &= \sum_h [|F(h)| - |F^{\text{obs}}(h)|]\, \delta |F(h)| \\ &= \frac{1}{V} \int \sum_h [|F(h)| - |F^{\text{obs}}(h)|] \\ &\quad \times \frac{1}{2|F(h)|} [F(h) \exp(2\pi i h r) \\ &\quad + F^*(h) \exp(-2\pi i h r)]\, \delta\rho(r)\, \mathrm{d}^3 r. \end{aligned} \tag{6}$$

Recalling the inversion formula (3) and denoting

$$\tilde{\rho}(r) = \sum_h |F(h)|^{\text{obs}} \frac{F(h)}{|F(h)|} \exp(2\pi i h r), \tag{7}$$

we obtain

$$\delta R = \frac{1}{V} \int [\rho(r) - \tilde{\rho}(r)]\, \delta\rho(r)\, \mathrm{d}^3 r, \tag{8}$$

which is called the first variation of R, from which the *functional derivative* is defined as

$$\frac{\delta R}{\delta \rho(r)} = \rho(r) - \tilde{\rho}(r). \tag{9}$$

* For the connection between the principle of maximum entropy and the maximum determinant rule see Britten & Collins (1982), Nayaran & Nityananda (1982), and Piro (1983).

If no restrictions are imposed on the admissible ρ's, then $\delta\rho$ is any arbitrary variation and a necessary condition for a minimum is

$$\rho(r) - \tilde{\rho}(r) = 0, \tag{10}$$

which amounts to stating that the final difference map should be flat. A less trivial result is obtained by imposing some restrictions on the set of admissible density functions. For instance, the very general condition that ρ should be everywhere positive can be forced by choosing $\rho = g^2/4$, g being now an arbitrary function. As another example, the choice

$$\rho = \frac{1}{1 + e^g} \tag{11}$$

forces ρ to take values in the range between zero and one. Considering now the R factor as a functional of the arbitrary function g we get

$$\frac{\delta R}{\delta g(r)} = \frac{\delta R}{\delta \rho(r)} \frac{\mathrm{d}\rho}{\mathrm{d}g} \tag{12}$$

and we obtain the extra condition that at a minimum ρ may differ from $\tilde{\rho}$ at the points where $\mathrm{d}\rho/\mathrm{d}g = 0$. The restrictions defined in the previous examples can also be introduced with the technique of the Lagrange multipliers and will be discussed later.

Besides constraints of the type $\rho = \rho(g)$, other types of electron density restrictions can be easily introduced. For example, if a protein structure is determined at a resolution which permits one to distinguish the molecule from the solvent region, then the constraint $\rho(r)$ = constant when r is in the solvent region should give an improved (better contrasted) electron density map.

The equation $\delta R(g)/\delta g(r) = 0$ is a condition for a stationary point rather than a condition for a minimum. A procedure to solve (9) that will ensure that a minimum of the functional will be attained is defined by the steepest-descent method. By considering infinitesimal steps, this method leads to the equation

$$\frac{\partial g(r,t)}{\partial t} = -\frac{\delta R}{\delta g(r,t)}, \tag{13}$$

where t is a parameter in iteration space. This is essentially the evolution equation for the relaxation of the thermodynamical system in the method of Khachaturyan *et al.* (1981).

From (12) and (13) we get

$$\frac{\partial g}{\partial t} = -\frac{\delta R}{\delta \rho} \frac{\mathrm{d}\rho}{\mathrm{d}g}, \tag{14}$$

which, together with the relationship $\rho = \rho(g)$, gives a differential equation in terms of ρ. For example, the positivity condition $\rho = g^2/4$ gives

$$\frac{\partial \rho}{\partial t} = -(\rho - \tilde{\rho})\,\rho. \tag{15}$$

The alternative choice $\rho = e^g$ yields

$$\frac{\partial \rho}{\partial t} = -(\rho - \tilde{\rho})\,\rho^2. \tag{16}$$

It is clear that (15) and (16) carry the same physical information and discrimination of one against the other should be based on numerical convenience. For the initial condition $\rho > 0$, differential equations (15) and (16) guarantee that the solutions also will be positive all along the pathway in iteration space.

The behaviour of (15) was explored by means of a few numerical one-dimensional test cases. The details of the calculations will be given in another section. The general pattern of the results can be summarized as follows:

(*a*) The minimization functional can be made virtually zero in almost all cases.

(*b*) Except for an origin translation and/or an enantiomorphic reversal (the test structures had symmetry $P1$) all peaks were obtained in the correct positions.

(*c*) The sizes of the spurious peaks were normally quite different from those of the correct ones.

(*d*) Spurious peaks, of sizes comparable with those of the correct peaks, were sometimes found.

(*e*) In general, recognition of the structure became more difficult as the number of atoms was increased.

It was clear from these examples that introduction of positivity alone was not restrictive enough to reconstruct the electron density function from scratch in linear problems with about six atoms. What is rather remarkable, however, is that limited success was obtained only with the positivity criterion without using the very powerful information of atomicity (form factors) which is essential in the use of the current direct-method procedures. This immediately suggests that the method might be considerably enhanced by the introduction of further information. A general way of doing so is discussed in the next section.

Use of constraints

The electron density function $\rho(r)$ possesses some general features which are more or less independent of the particular structure under consideration. For example, $\rho(r)$ must be everywhere positive and bounded and must also display atomicity, that is, $\rho(r)$ should be very small almost everywhere and should peak in small regions centered at the atomic positions. Whenever these general characteristics can be expressed in terms of equations they can be incorporated into the formalism to force some *modeling* on the functions sought. We have already discussed the particular cases of positivity and boundness in a previous section. A general and convenient way of dealing with model

equations is to incorporate them as constraints with the technique of the Lagrange multipliers.

In this section we shall give a few examples of this approach by discussing the introduction of some models and also by discussing the simple use of constraints that do not necessarily correspond to some general property of the electron density function.

A very useful technique employed to introduce the constraints of atomicity is to make the electron density function resemble its square. When this condition is translated into reciprocal space, the well known Sayre equation is obtained. Assuming equal atoms and the validity of Friedel's law, Sayre's equation can be written

$$S(h) = F(h) - \theta(h) \sum_k F(h+k)\, F^*(k) = 0 \qquad (17)$$

with

$$\theta(h) = f/(f * f), \qquad (18)$$

where f is the atomic form factor and the symbol $*$ denotes convolution. The particular form of (17) automatically ensures that Friedel's law is satisfied.

In the variational formalism the electron density can be forced to satisfy a set of Sayre's equations (17) by incorporating them into the minimization function with the technique of Lagrange multipliers.

Owing to the particular form of (17) this is better formulated in reciprocal space. Since Sayre's equation will not be the only type of constraint we shall be interested in, we shall consider first a general case of a set of complex constraint equations $C_{(H)} = 0$ in reciprocal space, where H belongs to some given domain H, in order to state a few general rules.

The variation of any functional $\mathsf{M}(F)$ of the structure factors F should be calculated in terms of the independent variations of the real and imaginary parts of F. That is, if $F = A + iB$,

$$\delta\mathsf{M} = \frac{\delta\mathsf{M}}{\delta A}\,\delta A + \frac{\delta\mathsf{M}}{\delta B}\,\delta B. \qquad (19)$$

It is convenient, however, to calculate the variations in terms of the two independent variables F and F^* defined through

$$\begin{bmatrix} F \\ F^* \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} A \\ B \end{bmatrix}, \qquad (20)$$

where the a's are complex constants. After all variations are calculated in terms of these new variables we make the identification $a_{11} = a_{21} = 1$, $a_{12} = -a_{22} = i$, which renders F^* as the complex conjugate of F. This has the advantage that the variations with respect to A and B are condensed in just a variation with respect to F, so that we have

$$\frac{\delta\mathsf{M}}{\delta F} = \frac{\delta\mathsf{M}}{\delta A} - i\frac{\delta\mathsf{M}}{\delta B} \qquad (21)$$

and also

$$\frac{\delta\mathsf{M}}{\delta F^*} = \left(\frac{\delta\mathsf{M}}{\delta F}\right)^*. \qquad (22)$$

Since the functional to be minimized must be real and $C(H)$ is generally complex, we define M in the form

$$\mathsf{M} = R + \sum_{H\in\mathsf{H}} \left[\mu_H \frac{C_{(H)} + C^*_{(H)}}{2} + \nu_H \frac{C_{(H)} - C^*_{(H)}}{2i}\right]. \qquad (23)$$

Denoting $2\lambda_H = \mu_H + i\nu_H$, we also have

$$\mathsf{M} = R + \sum_{H\in\mathsf{H}} (\lambda^*_H\, C_{(H)} + \lambda_H\, C^*_{(H)}). \qquad (24)$$

H denotes the set of the observed structure-factor magnitudes. The evolution equation is now

$$\frac{\partial F}{\partial t} = \frac{\delta\mathsf{M}}{\delta F^*} \qquad (25)$$

so that M yields a decreasing function of t:

$$\frac{\partial\mathsf{M}}{\partial t} = \sum_k \left[\frac{\delta\mathsf{M}}{\delta F(k)}\frac{\partial F(k)}{\partial t} + \frac{\delta\mathsf{M}}{\delta F^*(k)}\frac{\partial F^*(k)}{\partial t}\right] = -2\sum_k \left|\frac{\partial F(k)}{\partial t}\right|^2 \le 0. \qquad (26)$$

The multipliers λ_H can be set at some given pre-assigned values or may be determined by the introduction of further conditions. In the latter case one can, for instance, ask that constraint equations be satisified all along the iteration pathway, that is $\partial C_{(H)}/\partial t = 0$, $H \in \mathsf{H}$. Then

$$\frac{\partial C_{(H)}}{\partial t} = \sum_k \left[\frac{\delta C_{(H)}}{\delta F(k)}\frac{\partial F(k)}{\partial t} + \frac{\delta C_{(H)}}{\delta F^*(k)}\frac{\partial F^*(k)}{\partial t}\right] = 0. \qquad (27)$$

This is a homogeneous linear system for the real and imaginary parts of the shifts $\partial\mathsf{M}/\partial F$, which in general admits non-trivial solutions only if the number of unknowns is greater than the number of equations. As the number of constraint equations increases, the size of the shifts tends toward zero or, in other words, the system tends to be frozen by the imposition of too many constraints. Inserting the equations for $\partial F/\partial t$ in (27) we obtain a linear system for the Lagrange multipliers. It can be shown that if both H and $-H \in \mathsf{H}$ (as will be always assumed) then $\lambda_{-H} = \lambda^*_H$ when $F(H) = F^*(-H)$.

If the number of constraints is large the solution of the linear system (27) can be a serious handicap of the method. In such cases it would be preferable to fix the Lagrange multipliers at some convenient values. To do

so we must be sure that each constraint equation has a lower bound. If such is not the case a safe procedure is to use the functional

$$\mathsf{M} = R + \sum_{H\in\mathsf{H}} \lambda_H |C_{(H)}|^2 \tag{28}$$

instead of (24). When the domain H is extended to the whole reciprocal space and each constraint equation is given the same weight (all the λ's are equal) the functional (28) has the additional property of being symmetric in direct and reciprocal space because of Parseval's theorem.

We can now apply these rules to a few particular examples to illustrate the use of the formalism.

(*a*) *Sayre's equations as constraints*

We define the functional

$$\mathsf{M} = R + \sum_{H\in\mathsf{H}} [\lambda_H^* S_{(H)} + \lambda_H S_{(H)}^*] \tag{29}$$

from which we get

$$\frac{\partial \mathsf{M}}{F^*(h)} = F(h) - \tilde{F}^{\text{obs}}(h) + \sum_{H\in\mathsf{H}} \{\lambda_H\, \delta(h-H) - 2\theta(H)\, \lambda_H F(h-H)\}, \tag{30}$$

which, as indicated by the evolution equation (13), is proportional to the shift $\partial F/\partial t$. The equations satisfied by the Lagrange multipliers are

$$\begin{aligned}\sum_{H\in\mathsf{H}} [&\delta(K-H) + 2\{2\theta(K)\,\theta(H) \\ &\times \sum_h F(K-h)\, F(h-H) \\ &- [\theta(H) + \theta(K)] F(K-H)\}]\lambda_H \\ = &- \sum_h \{\delta(K-h) - 2\theta(K)\, F(K-h)\} \\ &\times [F(h - \tilde{F}^{\text{obs}}(h)].\end{aligned} \tag{31}$$

This may not be the best way of dealing with these constraints because the linear system (31) has to be solved at each step. The procedure of arbitrarily fixing the values of the multipliers is not feasible here because the constraint terms do not have a lower bound, though this can be introduced by further defining ρ as a convenient function of g [for example, expression (11)]. Writing the minimization function in the form (28)

$$\mathsf{M} = R + \sum_{H\in\mathsf{H}} \lambda_H |F(H) - \theta(H) \sum_k F(H+k)\, F^*(k)|^2, \tag{32}$$

we obtain for the shift

$$\frac{\delta \mathsf{M}}{\delta F^*(h)} = F(h) - \tilde{F}^{\text{obs}}(h) + \sum_{H\in\mathsf{H}} \lambda_H S_H[\delta(h-H) - 2\theta(H)\, F(h-H)]. \tag{33}$$

Note that in contrast to the sum over H, the sum over k in (32) runs in principle over all reciprocal space.

(*b*) *A subset of complex (modulus and phase) structure factors F'_H, $H\in\mathsf{H}$, is fixed*

This is a common problem in protein crystallography when phases in a low-resolution reciprocal-space region are determined, for example, by the method of isomorphous replacement, and phase extension to a larger region is required. We start from

$$\mathsf{M} = R + \sum_{H\in\mathsf{H}} \{\lambda_H^*[F(H) - F'(H)] + \lambda_H[F^*(H) - F'^*(H)]\}. \tag{34}$$

Then

$$\frac{\partial \rho}{\partial t} = -\{\rho - \tilde{\rho} + \sum_{H\in\mathsf{H}} \lambda_H^* \exp(2\pi i H r) + \lambda_H \exp(-2\pi i H r)\}\, \rho. \tag{35}$$

The conditions to determine the λ's are

$$\frac{1}{V}\int \frac{\partial \rho}{\partial t} \exp(-2\pi i K r)\, \mathrm{d}^3 r = 0; \quad K\in\mathsf{H}, \tag{36}$$

from which we get

$$\sum_{H\in\mathsf{H}} [\lambda_H^* F(H+K) + \lambda_H F^*(H-K)] = F\{(\rho - \tilde{\rho})\,\rho\}_K, \tag{37}$$

where $F\{\ \}_K$ stands for the Fourier transform operator and the relation $\rho = g^2/4$ has been used. Since we again assume that if $F(H)\in\mathsf{H}$ then also $F(-H)\in\mathsf{H}$,

$$\sum_{H\in\mathsf{H}} \lambda_H F(K-H) = \frac{1}{2} F\{(\rho - \tilde{\rho})\,\rho\}_K \tag{38}$$

or in matrix notation

$$K\boldsymbol{\lambda} = 1/2 F\{(\rho - \tilde{\rho})\,\rho\}, \tag{39}$$

where K is the Karle & Hauptman matrix (whose leading terms are the F's which belong to the set H). Once the multipliers are determined from this equation, they must be inserted into (34) to obtain the shifts to the electron density function. In particular, if we fix the value of $\int \rho \mathrm{d}^3 r$ [which coincides with $F(0)$], we get

$$\frac{\partial \rho}{\partial t} = -\{\rho - \tilde{\rho} - \int (\rho - \tilde{\rho})\,\rho \mathrm{d}^3 r / \int \rho \mathrm{d}^3 r\}\, \rho. \tag{40}$$

(c) The Patterson function from calculated F's should resemble the one from observed F^{obs}'s

This is another potentially useful way of modeling the electron density function. If, for example, we ask for the calculated Patterson function to be zero on a spherical shell around the origin peak, all along the refining trajectory, this will prevent electron density peaks from coming closer to each other than a certain limit.

Let us consider L points $\mathbf{a}_j$, $j = 1, \ldots, L$, possibly around the origin, at which the observed Patterson map is small:

$$P^{obs}(a_j) \simeq 0; \quad j = 1, \ldots, L. \tag{41}$$

We want the calculated Patterson function

$$P(a) = \sum_k F(k)\, F^*(k) \cos(2\pi k a) \tag{42}$$

to be zero at the same points all along the refining pathway. Starting from the functional

$$\mathsf{M} = R + \sum_{j=1}^{L} \lambda_j P(a_j) \tag{43}$$

we get the system of linear equations for the multipliers

$$\sum_j \{P(a_j + a_i) + P(a_i - a_j)\}\, \lambda_j = 2\{\tilde{P}(a_i) - P(a_i)\} \tag{44}$$

with

$$\tilde{P}(a) = \sum_h |F(h)|\, |F^{obs}(h)| \cos(2\pi h a). \tag{45}$$

Numerical results

Several one-dimensional tests were carried out in order to assess the numerical behavior of the method. We did not intend to perform an exhaustive evaluation of the many theoretical possibilities. We shall only report here a sample of some interesting results which illustrates the power of the procedure and indicates possible lines for further research. Numerical methods must be used because of the impossibility of solving analytically the differential equations. This introduces a problem of sampling which is independent of, and must not be confused with, the limitations due to experimental resolution. In particular, constraint equations of Sayre's type are only limited by numerical sampling and are independent of the experimental resolution which is revealed only on the R term of the minimizing functional. When referring to sampling we shall speak of grid resolution to differentiate it from the experimental resolution. In all examples the grid resolution was chosen to be such as to render a good atomic electron density function. Typically, a grid division of 0·2 Å allows calculations of structure factors with relative errors less than 10^{-3} from an electron density made of Gaussian atoms. More technically, the atomic form factors impose a finite band-width which in turn determines the grid resolution through Shannon's criterion of sampling (Brillouin, 1962). A greater grid resolution is useless since Shannon's sampling allows us to build the density function at any point in the unit cell. This limited band-width is in general a consequence of the built-in characteristic of the sought-after electron density function. For a domain of very general functions, like that of the functions satisfying only the requirement of positivity, there are no restrictions in the band-width and therefore no *a priori* limitations on the grid resolution.

Equations (15) and (30) were solved by using a second-order Runge–Kutta iteration procedure (which is somewhat related to the conjugate-gradient technique) starting from an arbitrary initial electron density consistent with the constraints. The step size was periodically adjusted after a few iterations to a large value satisfying the condition that the minimization function should decrease monotonically.

When the domain H contained more than a few tens of reflections, the system of equations (30)–(31) showed an extremely slow convergence, so that no solution was attained. On the other hand, if only a small number of constraint equations were introduced, the density function failed to resemble its square with the consequence of a proliferation of negative electron density regions. Although this method did not prove useful for *ab initio* phase calculations, perhaps it may be used in the final stages of phase refinement when the initial electron density is known with reasonably good accuracy.

The most interesting results were obtained by using Sayre's equations as constraints with all multipliers set equal to the same constant value λ. Since the R term and the constraint term have similar functional dependence on the F's, the λ value was set to unity in order to give them the same weight. First a minimization of the functional (32) was carried out by the method of steepest descent. This method was chosen because it is numerically more convenient and also because we are only interested in the limiting value of the density function when the iteration parameter tends to infinity rather than its value along the refinement pathway. A one-dimensional, $p1$, seven equal-atom structure was generated and a set of 128 structure-factor moduli was calculated from it. Starting from a random initial electron density distribution, after about ten iterations the crystallographic R factor stabilized at a value of 26%. The R term and the constraint term in (32) dropped from the initial values 0·26 and 2·30 to $4{\cdot}8 \times 10^{-4}$ and $9{\cdot}6 \times 10^{-4}$, respectively. The solution displays the main features of the structure but it is

somewhat obscured by the presence of some spurious peaks and small negative regions.

The general behavior of the procedure can be understood by analyzing the nature of the contribution to the shifts coming from the terms R and S in the minimization functional. This is more simply carried out for the particular case in which $\theta(h)$ is a constant. In this case, Sayre's equation corresponds to the condition

$$\rho - \rho^2 = 0, \tag{46}$$

which is satisfied if $\rho = 0$ or $\rho = 1$ throughout the unit cell. This constraint might be useful in, for example, the hypothetical case mentioned by Sayre (1980) in which the resolution is low enough for the structure to appear to be composed of only two different densities. Also, since it gives rise to faster iterative equations, it might be useful in the first stages of refinement. Again, minimizing the functional

$$\mathsf{M} = R + \lambda \frac{1}{2V} \int (\rho - \rho^2)^2 \, \mathrm{d}^3 r \tag{47}$$

could be interpreted as a least-squares fitting of a square peaked electron density function to the observed data. The shift corresponding to the constraint term in (47) is

$$\Delta\rho \simeq -\lambda\rho(1-\rho)(1-2\rho). \tag{48}$$

It follows that for

$$\begin{array}{llll} \rho < \theta & \text{then } \Delta\rho > 0 & \text{which makes} & \rho \to 0 \\ 0 < \rho < \frac{1}{2}, & \Delta\rho < 0 & & \rho \to 0 \\ \frac{1}{2} < \rho < 1, & \Delta\rho > 0 & & \rho \to 1 \\ \rho > 1, & \Delta\rho < 0 & & \rho \to 1. \end{array}$$

Obviously the constraint term in (47) behaves as a contrast device which tends to eliminate peaks smaller than $\frac{1}{2}$ and to scale to unity peaks of height greater than $\frac{1}{2}$. In the case where the electron density is small and there is a peak given by the difference term $\delta R/\delta\rho$, the constraint term tends to inhibit growth of the peak in that region.

Since this competition between terms may not always be desirable we tried a combined use of (15) and (33) in the following way. A few iterations were performed with (15) and the current density function was injected as input to (33), which after a few iterations was in turn re-injected into (15) and the process repeated a few more times. This switching scheme always gave the correct solution for any arbitrary initial electron density. A centrosymmetric solution was avoided either by choosing random initial phases or by arbitrarily setting to zero the electron density in a substantial fraction of the unit cell. A typical result gave an R factor of about 2% with a remarkable reproduction of the true electron density and the absence of negative regions and noise peaks.

Discussion

(a) *Sayre's equations are constraints*

Refinement techniques involving the solution of a set of Sayre's equations (17) have been proposed by Krabbendam & Kroon (1971) and Sayre (1972). Sayre's method amounts to a least-squares solution of (17) considered as functions of the phases and has been successfully applied to phase extension and refinement in rubredoxin (Sayre, 1974) and in insulin (Cutfield *et al.*, 1975). Because of the assumptions involved in deriving it, Sayre's equation is not expected to hold at low resolution [however, see the second footnote in the paper by Sayre (1975)], and phase refinement requires a set of F^{obs} to a resolution of 1·5 Å or better. Also, the starting set of phases led to false minima, in the case of rubredoxin, when it corresponded to a resolution lower than 2·5 Å (Sayre, 1974).

Minimizing the functional (32) is obviously related to the method of Sayre but it is fundamentally different in one important respect: the electron density with which we want to fit our experimental data does satisfy Sayre's equation, irrespective of the resolution of the F^{obs} set, because the constraint equation is forced upon the calculated F's, in which both phases and moduli are considered as variables. In other words the electron density sought is in our case an atomic ρ function which we wish to fit, in the least-squares sense, to the experimental data of whatever resolution at our disposal. In principle, at least, the method has no built-in experimental resolution limitation and might therefore prove to have a larger radius of convergence than that of Sayre.

It is interesting to notice the close relationship between the use of Sayre's equations as constraints and the empirical electron density modification procedure of Zwick *et al.* (1976). The total shift corresponding to the functional (47) is

$$\Delta\rho \simeq -(\rho - \tilde{\rho}) - \lambda\{\rho - (3\rho^2 - 2\rho^3)\}, \tag{49}$$

which can be interpreted as composed of two contributions, one from the difference map and the other from the filter function of Collins (1975). In fact Collins modifies the electron density according to

$$\rho + \Delta\rho = 3\rho^2 - 2\rho^3, \tag{50}$$

which gives for $\Delta\rho$ the expression between brackets in (49).

(b) *Applications of the principle of maximum entropy*

The principle of maximum entropy (PME) has been discussed by various authors in connection with the problem of obtaining the best electron density function consistent with a particular set of data.

Since the form in which it has been applied is amenable to a variational analysis, we consider a discussion of it to be of interest in the present context.

The entropy, or lack of information, associated with a continuous probability density $p(x_1, \ldots, x_N)$ of the random variables $x_1, \ldots, x_N$ is defined, except for an additive constant, by

$$\mathrm{S} = -K \int p(x_1, \ldots, x_N) \ln p(x_1, \ldots, x_n)\, \mathrm{d}x_1 \ldots \mathrm{d}x_n, \tag{51}$$

where K is a positive constant. The PME states that the probability distribution $p(x_1, \ldots, x_N)$ of the set of physical variables $x_1, \ldots, x_N$, not amenable to a complete experimental determination, which renders a maximum of entropy subjected to whatever is known, provides the most unbiased representation of the system (Shannon & Weaver, 1949).

Construction of the electron density function is a problem conditioned by lack of information, such as limited resolution, errors in the structure-factor magnitudes and phases, missing phase value, or some combination of these. The PME then provides the natural framework in which the problem should be treated.

Gull & Daniel (1978) have described a method for image reconstruction in radio astronomy which is also applicable in X-ray crystallography. The PME is applied in the following ways: owing to the lack of information one can think of many different electron density maps consistent with the data. Each of these maps is assigned an entropy value.

$$\mathrm{S} = -\frac{\lambda}{V} \int \rho(r) \ln \rho(r)\, \mathrm{d}^3 r, \tag{52}$$

as proposed by Frieden (1972) in connection with image-reconstruction techniques. Then, maximizing the entropy with the constraint that (1) should be as small as possible is equivalent to minimizing the functional

$$\mathrm{M} = -\mathrm{S} + R = \lambda \frac{1}{V} \int \rho(r) \ln \rho(r)\, \mathrm{d}^3 r + \frac{1}{2} \sum_h [|F(h)| - |F^{\mathrm{obs}}(h)|]^2. \tag{53}$$

This is a variational problem similar to the ones we have already described. From the point of view of our approach, S can be thought of as a constraint which excludes non-negative functions from the domain of argument functions and the question arises whether minimizing (53) renders a better ρ function than minimizing (1) with, for example, the condition $\rho = g^2$. This question is not immediately answered by the PME, as one would at first sight expect, because the expression (52) for the entropy implies that some model, by no means unique, has been assumed. To show this point we shall describe an alternative procedure from the one proposed by Frieden to generate possible maps.

Following Frieden we imagine the unit cell to be divided into J equal-sized smaller cells each centered at the points r_j. The electron density is then represented by the sequence of average ρ values $\rho(\tilde{r}_j) = \rho_j$, $j = 1, \ldots, J$. Further, the ρ_j values are expressed in terms of some elementary unit of electron density or quanta $\Delta\rho$ in the form

$$\rho_j = N_j \Delta\rho, \tag{54}$$

where N_j is a pure number, and we state the constraint that

$$\sum_{j=1}^{J} N_j = N. \tag{55}$$

To obtain the result of Frieden we imagine that each elementary cell is further subdivided into G subcells such that $G \gg N_j$ for each j. The number of ways a given sequence ρ_j can be realized is counted by putting every one of the N_j quanta into one of the G subdivisions, this being approximately G^{N_j} (because of the assumption $G \gg N_j$) and dividing by the number of configurations differing only by a permutation of the particles

$$\frac{G_j^{N_j}}{N_j!}. \tag{56}$$

Then, since the groups of particles are mutually independent, we obtain for the total number

$$W_0(N_1, \ldots, N_j) = \prod_{j=1}^{J} \frac{G^{N_j}}{N_j!} \tag{57}$$

The entropy associated with the sequence is then

$$\mathrm{S} \simeq \ln W_0 = -G \sum_{j=1}^{J} n_j \ln (n_j/e), \quad n_j = \frac{N_j}{G}, \tag{58}$$

in which Stirling's approximation $\ln N! \simeq N \ln (N/e)$ has been used. This result is equivalent to that of Frieden except for a multiplicative constant.

It should be recognized that this way of counting different configurations is equivalent to Boltzman counting in statistical mechanics. If the counting is performed under the different hypothesis $N_j \simeq G$ for all j, we obtain Fermi or Bose–Einstein statistics depending on whether each one of the G subcells is allowed to contain no more than one or any number of quanta. The expressions for the entropy are, respectively,

$$\mathrm{S} \simeq -G \sum_{j=1}^{J} [n_j \ln n_j + (1 - n_j) \ln (1 - n_j)] \tag{59}$$

(Fermi), or

$$\mathrm{S} \simeq -G \sum_{j=1}^{J} [n_j \ln n_j - (1 + n_j) \ln (1 + n_j)] \tag{60}$$

Bose–Einstein.

Clearly there is no *a priori* physical means to determine which of the statistics gives the most appropriate model to calculate the entropy associated

with electron density maps and preference of a particular one should be based on practical convenience.

Moreover it is clear that the three hypotheses from which (58)–(60) are obtained are by no means the only possible ones.

Formulas (58)–(60) can be reinterpreted in the light of our previous formalism. As was mentioned before, rather than maximizing the entropy with the constraint that the R term should be as small as possible, we consider the functional

$$\mathsf{M} = R - \lambda \mathsf{S}, \tag{61}$$

where S is any of the entropy expressions (58)–(60), as one which incorporates a constraint equation S to the fundamental R which we want to maximize. The Lagrange multiplier λ is then a measure of the allowed spread in the error $\rho - \tilde{\rho}$.

Expression (58) introduces the constraint of positivity. The function ρ which minimizes (61) satisfies

$$\rho = \langle\rho\rangle \exp\left[-(\rho - \tilde{\rho})/\lambda\right]; \quad \ln\langle\rho\rangle = \frac{1}{V}\int \ln\rho \,\mathrm{d}^3 r. \tag{62}$$

For small values of λ, ρ is approximately $\tilde{\rho}$ wherever $\tilde{\rho} > 0$ and takes small positive values when $\tilde{\rho} < 0$. For $\lambda \gg 1$, ρ tends to a constant, a result which holds for any functional of ρ of the form

$$\int f(\rho)\,\mathrm{d}^3 r \tag{63}$$

with a normalizing condition

$$\frac{1}{V}\int \rho \,\mathrm{d}^3 r = \text{constant} \tag{64}$$

if ρ belongs to the domain of continuous functions.

An expression similar to (59) was obtained by Khachaturyan *et al.* (1981) from thermodynamic analogies. Equation (59) seems particularly attractive because it puts a bound to the maximum possible value of a positive ρ, which is a physically meaningful constraint.

The last expression for the entropy (60) ensures positivity and, as is known from quantum statistics, allows for the effect of condensation, a property whose implications in the present context are not apparent.

Three-dimensional calculations were performed to compare the results of minimizing the functional (1) with the positivity condition $\rho = g^2/4$, with the results of minimizing a functional containing an entropy term as in (53).

A thirteen equal-atom structure with chemically realistic bond distances and angles was generated in space group $P1$ and the moduli of the 1183 structure factors in the box $-6 \le h \le 6, -6 \le k \le 6, 0 \le l \le 6$ were calculated. The electron density was sampled at thirteen sections of 13 × 14 points, thus providing a grid resolution of about 0·4 Å. The initial calculated electron density consisted of the first section of the map obtained from the model and the twelve other sections with the electron density set at an arbitrary constant value close to zero. The first section contained fractions of two peaks of the model which permitted the automatic fixation of the unit-cell origin.

Firstly, (15) was solved by iteration. It was rewritten in the form

$$-\Delta\rho = \alpha\rho(\rho - \tilde{\rho})$$

and α was estimated at each step so as to produce the maximum decrease in the minimization function (1). Calculation of α was performed by linear search along the search direction $\Delta\rho$, as described by Agarwal (1978). In fifteen cycles the crystallographic R factor dropped from 70 to 12% and the final calculated electron density showed the thirteen peaks of the structure in their correct positions. The height of the peaks was generally incorrect (enhanced for the two partially contained in the first section and depressed for all others) and there appeared a few small spurious peaks, which, however, did not jeopardize the interpretation of the map.

Secondly, a similar procedure was carried out minimizing M as given by (53). The shifts are in this case

$$\Delta\rho = -\alpha\frac{\delta\mathsf{M}}{\delta\rho} = -\alpha\{(\rho - \tilde{\rho}) - \lambda(1 - \ln\rho)\}.$$

Several trials, each for a different value of λ in the range between 10 and 10^{-1}, were performed. For $\lambda > 1$, low-contrast solutions were obtained. In particular, for $\lambda = 10$, the final electron density was flat and featureless. On the other hand, very noisy solutions corresponded to $\lambda = 0{\cdot}1$, which made them uninterpretable in terms of the original thirteen-atom model. For $\lambda = 1$, the final electron density function was of quality similar to the one obtained by minimizing (15), as was also the computational effort (less than 1 s per cycle on an IBM 370/125 computer) and the final R factor. In this example there was no advantage in using the entropy-containing functional (53) rather than just (1) with the constraint of positivity. In the light of our previous discussion this seems to indicate that the entropy term in (53) corresponds to a model which merely enforces positivity on the calculated electron density function.

Three-dimensional tests with realistic macromolecular problems are in progress.

References

AGARWAL, R. C. (1978). *Acta Cryst.* A**34**, 791–809.

BRILLOUIN, L. (1962). *Science and Information Theory*. New York: Academic Press.

BRITTEN, P. L. & COLLINS, D. M. (1982). *Acta Cryst.* A**38**, 129–132.

COLLINS, D. M. (1975). *Acta Cryst.* A**31**, 388–389.
COLLINS, D. M. (1982). *Nature (London)* **298**, 49–51.
CUTFIELD, J. F., DODSON, E. J., DODSON, G. G., HODGKIN, D. C., ISAACS, N. W., SAKABE, K. & SAKABE, N. (1975). *Acta Cryst.* A**31**, S21.
FRIEDEN, R. (1972). *J. Opt. Soc. Am.* **62**, 511–518.
GULL, S. F. & DANIEL, G. J. (1978). *Nature (London)*, **272**, 686–690.
KHACHATURYAN, A., SEMENOVSKAYA, S. & VAINSHTEIN, B. (1981). *Acta Cryst.* A**37**, 742–754.
KRABBENDAM, H. & KROON, J. (1971). *Acta Cryst.* A**27**, 48–53.
NARAYAN, R. & NITYANANDA, R. (1981). *Acta Cryst.* A**38**, 122–128.
PIRO, O. (1983). *Acta Cryst.* A**39**, 61–68.
SAYRE, D. (1972). *Acta Cryst.* A**28**, 210–212.
SAYRE, D. (1974). *Acta Cryst.* A**30**, 180–184.
SAYRE, D. (1975). In *Crystallographic Computing Techniques*, edited by F. R. AHMED, pp. 322–327. Copenhagen: Munksgaard.
SAYRE, D. (1980). *Theory and Practice of Direct Methods in Crystallography*, edited by M. F. C. LADD & R. A. PALMER, pp. 271–286. New York: Plenum Press.
SHANNON, C. E. & WEAVER, W. (1949). *The Mathematical Theory of Communication*. Urbana: Univ. of Illinois Press.
ZWICK, M., BANTZ, D. & HUGHES, J. (1976). *Ultramicroscopy*, **1**, 275–277.

Acta Cryst. (1983). A**39**, 631–641

Willis Formalism of Anharmonic Temperature Factors for a General Potential and its Application in the Least-Squares Method

BY KIYOAKI TANAKA AND FUMIYUKI MARUMO

The Research Laboratory of Engineering Materials, Tokyo Institute of Technology, Nagatsuta 4259, *Midori-ku, Yokohama* 227, *Japan*

(*Received* 18 *September* 1982; *accepted* 7 *March* 1983)

Abstract

Willis treatment of anharmonic temperature factors including up to fourth-order terms has been generalized and incorporated into a conventional full-matrix least-squares program. The temperature factor $T(\mathbf{S})$ including the anharmonic vibration effect is formulated in the general case using Willis's method. $T(\mathbf{S})$ is based on the Cartesian coordinates defined by the three principal axes of the harmonic thermal ellipsoid. The simultaneous refinement of the parameters in $T(\mathbf{S})$ with the conventional parameters in crystallography, which are based on the crystal lattice system, is possible. In order to introduce $T(\mathbf{S})$ into conventional full-matrix least-squares programs, some other relations were also derived, such as that between crystallographic symmetry and $T(\mathbf{S})$, and that among parameters due to point symmetry of the atom, and so on. The present method was applied to the K and two F atoms in $KCuF_3$ crystals at 296 K with the point symmetries of the sites 422, $\bar{4}2m$ and $mm2$, respectively, and Al and O atoms in α-Al_2O_3 crystals at 2170 K with the point symmetries 3 and 2, respectively. The features of the potentials of atoms in $KCuF_3$ crystals correspond very well to the peaks on the difference-Fourier maps. After the correction for anharmonic vibration, the difference-Fourier map around each atom became flat. It indicates that in an accurate electron density study the anharmonic vibration effect is not negligible and the Willis method works effectively.

0108-7673/83/050631-11$01.50

Introduction

Studies on anharmonic vibration have been mainly performed by neutron diffraction, since the constant value of the neutron cross section prevents steep diminution in intensities of high-angle reflexions, and since no interaction between aspherical thermal vibrations of nuclei and aspherical electron distribution favors the neutron diffraction study. However, recent advances in X-ray diffraction have made it possible to observe a large number of high-angle reflexions accurately. In the recent X-ray study of $KCuF_3$ (Tanaka & Marumo, 1982), it was shown that the interaction between the two asphericities was not so severe. And the peaks of the anharmonic vibration appeared on a difference-Fourier map after the removal of those of the aspherical electron distribution. Thus, the main difficulties in the X-ray study of anharmonic vibration in crystals where covalency of bonds does not play a significant role are believed to be overcome. It has become highly necessary and desirable to modify the harmonic temperature-factor formalism to that of the general temperature factor (GTF), which includes the effects of anharmonic vibration of atoms.

Assuming a crystal to be an assembly of independent oscillators, several authors formulated the GTF by taking the ensemble average of the one-particle potential (OPP). Willis (1969) expressed the OPP as a power-series expansion and formulated the GTF for atoms with cubic point symmetries. Kurki-Suonio,