Methods of Minimization and their Implications

D. E. Tronrud

Howard Hughes Medical Institute
and
Institute of Molecular Biology
University of Oregon
Eugene, OR 97403

January 4, 1994

Abstract

The process of refinement is a large problem in function minimization. To reduce the amount of computation the methods chosen to minimize the function incorporate a number of assumptions. When these assumptions break down special procedures must be used.

Most of these procedures are commonly known, such as rigid body refinement, but an understanding of the details of the methods themselves allows one to know when and what procedure to apply.

1 Introduction

After the crude model of a protein is constructed we enter the stage of refinement. The parameters of the model are altered to improve the agreement between the model and the experimental observations. If we construct a

function which reflects the discrepancy between the these two, refinement becomes the minimization of this function.

All are familiar with fitting models to data in this fashion. Finding the "least-squares" line through a collection of points is the classic example. However line fitting is easy and refinement is hard — The difference lies in the relationship between the model and the experimental observations. This paper will discuss how the particulars of our structural model limits our ability to interpret diffraction data.

1.1 Method vs. Function

There is one distinction which must be clearly made, but is usually treated in an ambiguous fashion. This is the difference between the two choices to be made. First the function which describes the difference between the observations and the predictions of the model. The second is the choice of the method by which this function will be minimized.

There are several methods of minimization commonly used today. Most are described in detail below. Each of them can be used to minimize any function.

In crystallography three functions commonly used. They are the least-squares residual, the empirical energy function, and the correlation coefficient.

The least-squares residual function is

$$f(\mathbf{p}) = \sum_{i}^{\text{all data}} \frac{1}{\sigma(i)^2}(Q_o(i) - Q_c(i,\mathbf{p}))^2, \tag{1}$$

where $Q_o(i)$ and $\sigma(i)$ are the value and standard deviation for observation number i. $Q_c(i,\mathbf{p})$ is the model's prediction for observation i using the set of model parameters $\mathbf{p}$. The values of the parameters found by minimizing this function are those which have the smallest individual standard deviation, or the smallest probable error[4].

The justification for refining against an empirical energy function is the belief that the true protein structure should be at an energy minimum as well as a best fit to the crystallographic observations. While this is undoubtedly correct in the absence of errors in the measured intensities and energy parameters, an analysis of the effect of the presence of such errors has not

been done. In practice, usually the parameters of the energy function are chosen in a fashion to allow the energy to mimic the least-squares residual. Confusion can result if the value of such an "energy" function is interpreted as an energy.

The correlation coefficient is a different measure of the agreement between the model and the observations. In statistics it is used to judge whether there is any agreement at all. This makes it very sensitive to changes in the model when the agreement between the model and the observations is only barely detectable. The correlation coefficient is commonly used in the solution of rotation functions, but has not been used commonly in individual atom refinement.

To describe a refinement protocol it is not sufficient to state one or the other of these choices. One can not meaningfully state that a model was refined with "least-squares". Both the function and the method must be stated.

2 Minimization Methods

Function minimization methods fall on a continuum. The distinguishing characteristic is the amount of information about the function which must be explicitly calculated and supplied for the algorithm. All methods require the ability to calculate the value of the function given a particular set of values for the parameters of the model. There are methods which require only the function values (Simulated Annealing is such a method, it uses the gradient of the function only incidentally in generating new sets of parameters.). Some methods require gradient of the function as well. These methods, as a class, are called Gradient Descent methods.

The method of minimization which uses the gradient and all of the second derivative (or curvature) information is call the "Full-Matrix" method. The Full-Matrix method is quite powerful but the requirements of memory and computations for its implementation are beyond current computer technology except for small molecules and smaller proteins. Also, for reasons to be discussed, this algorithm can only be used when the model is very close to the minimum — closer than most "completely" refined protein models. For proteins, it has only been applied to cases where the molecule is small (< 1000 atoms) which diffract to high resolution and have previously been

exhaustively refined with gradient descent methods.

2.1 The Full-Matrix Method

An analysis of the Full-Matrix method, and all gradient descent methods begins with the Taylor's series expansion of the function being minimized. For a generic function ($f(\mathbf{p})$) the Taylor's expansion is

$$f(\mathbf{p}) = f(\mathbf{p}_0) + \left|\frac{df(\mathbf{p})}{d\mathbf{p}}\right|^t_{\mathbf{p}=\mathbf{p}_0} (\mathbf{p}-\mathbf{p}_0) + \frac{1}{2}(\mathbf{p}-\mathbf{p}_0)^t \left|\frac{d^2 f(\mathbf{p})}{d\mathbf{p}^2}\right|_{\mathbf{p}=\mathbf{p}_0} (\mathbf{p}-\mathbf{p}_0) + \cdots, \tag{2}$$

where $\mathbf{p}_0$ is the current set of parameters of the model. In all cases the additional terms (represented by "$\cdots$") are ignored. This assumption has considerable consequences which will be discussed later.

We can change the nomenclature used in equation 2 to more closely match those in refinement by defining $\mathbf{p}_0$ to be the parameters of the current model and $\mathbf{s}$ to be a "shift vector" which we want to add to $\mathbf{p}_0$. $\mathbf{s}$ is equal to $\mathbf{p}-\mathbf{p}_0$. The new version of Equation 2 is

$$f(\mathbf{p}_0+\mathbf{s}) = f(\mathbf{p}_0) + \left|\frac{df(\mathbf{p})}{d\mathbf{p}}\right|^t_{\mathbf{p}=\mathbf{p}_0} \mathbf{s} + \frac{1}{2}\mathbf{s}^t \left|\frac{d^2 f(\mathbf{p})}{d\mathbf{p}^2}\right|_{\mathbf{p}=\mathbf{p}_0} \mathbf{s} \tag{3}$$

and its derivative is

$$\left|\frac{df(\mathbf{p})}{d\mathbf{p}}\right|_{(\mathbf{p}=\mathbf{p}_0+\mathbf{s})} = \left|\frac{df(\mathbf{p})}{d\mathbf{p}}\right|_{\mathbf{p}=\mathbf{p}_0} + \left|\frac{d^2 f(\mathbf{p})}{d\mathbf{p}^2}\right|^t_{\mathbf{p}=\mathbf{p}_0} \mathbf{s}. \tag{4}$$

Since the first and second derivatives can be calculated given any particular value for $\mathbf{p}_0$ this equation allows the gradient of the function to be calculated given any shift vector. In addition the equation can be inverted to allow the shift vector to be calculated given the gradient of the function.

At the minimum (or maximum) of a function all components of the gradient are zero. Therefore we should be able to calculate the shift vector between the current model ($\mathbf{p}_0$) and the minimum. The equation for this is simple —

$$\mathbf{s} = -\left|\frac{d^2 f(\mathbf{p})}{d\mathbf{p}^2}\right|^{-1}_{\mathbf{p}=\mathbf{p}_0} \left|\frac{df(\mathbf{p})}{d\mathbf{p}}\right|_{\mathbf{p}=\mathbf{p}_0}. \tag{5}$$

The Full-Matrix method is to use this equation, evaluated with the current parameters, to calculate $\mathbf{s}$. $\mathbf{s}$ is then added to $\mathbf{p}_0$ to give the set of parameters which cause the function to be minimal, and in the case of refinement the best fit to the observations.

This method sounds great. One calculates a single expression and the minimum is discovered. When fitting a line to a set of points this is exactly what is done. In refinement something is obviously different. The difference arrises from the "$\cdots$" which we choose to ignore. In the case of fitting a line to points the terms represented by "$\cdots$" in fact are zero. The truncated Taylor's series is exact and the shift vector is also exact. In refinement these terms are not equal to zero resulting in the shift vector giving only the approximate location of the minimum.

The quality of the estimate is limited by the size of the terms which are ignored. The terms of the Taylor's series have increasing powers of $\mathbf{s}$. The first term ignored is multiplied by $\mathbf{s}^3$ and the higher order terms are multiplied by ever higher powers. If $\mathbf{s}$ is small these higher order terms become quite small also. Therefore the closer $\mathbf{p}_0$ is to the minimum the better estimate $\mathbf{s}$ becomes.

The Full-Matrix method, and all the gradient descent methods which are derived from it, becomes a series of successive approximations. An initial guess for the parameters of the model ($\mathbf{p}_0$) is manufactured somehow. For the shift vector to actually give an improved set of parameters the guess must be sufficiently close to the minimum. The "sufficiently close" criteria is rather difficult to calculate exactly.

The distance from the minimum at which a minimization method breakdown is called the "radius of convergence". It is clear is that the Full-Matrix method is much more restrictive than the gradient descent methods, and the gradient descent methods are more restrictive than simulated annealing, Metropolis, and Monte Carlo methods. Basically the less information about the function calculated at a particular point the larger the radius of convergence will be.

The property of the Full Matrix method which compensates for its restricted radius of convergence is its "power of convergence". If the starting model is within the radius of the Full Matrix method that method will be able to bring the model to the minimum quicker than any other method.

2.1.1 The Normal Matrix

The aspect of the Full-Matrix minimization method which prevents it being used in common refinement is the difficulty in calculating the term

$$\left| \frac{d^2 f(\mathbf{p})}{d\mathbf{p}^2} \right|^{-1}_{\mathbf{p}=\mathbf{p}_0} . \tag{6}$$

This matrix written out in full is

$$\begin{pmatrix} \frac{\partial^2 f(\mathbf{p})}{\partial p_1^2} & \frac{\partial^2 f(\mathbf{p})}{\partial p_2 \, \partial p_1} & \cdots & \frac{\partial^2 f(\mathbf{p})}{\partial p_n \, \partial p_1} \\ \frac{\partial^2 f(\mathbf{p})}{\partial p_1 \, \partial p_2} & \frac{\partial^2 f(\mathbf{p})}{\partial p_2^2} & \cdots & \frac{\partial^2 f(\mathbf{p})}{\partial p_n \, \partial p_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\mathbf{p})}{\partial p_1 \, \partial p_n} & \frac{\partial^2 f(\mathbf{p})}{\partial p_2 \, \partial p_n} & \cdots & \frac{\partial^2 f(\mathbf{p})}{\partial p_n^2} \end{pmatrix}^{-1} . \tag{7}$$

This matrix contains $n \times n$ elements, where n is the number of parameters in the model. In a typical case n will be on the order of 10,000. The number of elements in the second derivative matrix, often called the Normal matrix, would be 100,000,000. It would take a lot of computer time to calculate it, a lot of memory to store it, and a lot more computer time to invert it. The gradient descent methods make various assumptions about the importance of different parts of the Normal matrix to reduce these requirements.

To understand the relative importance of the different elements of the Normal matrix we need to understand the meanings of each part. The most important classification of the elements is the distinction between the elements on the diagonal and those off it. The elements on the diagonal are affected by a single parameter and are therefore somewhat easier to analyse. The off-diagonal elements are affected jointly by two parameters.

The information contained in the off-diagonal elements described how the effect on the function of changing parameter a is affected by changes in parameter b. In essence it is related to the correlation of the two parameters. If one considers the simple case where each parameter is varied in turn. Parameter a is moved to the value where the function in minimized. Then parameter b is changed. If the off-diagonal element for a and b is nonzero than parameter a will have to be readjusted, and the larger the value the greater the adjustment required.

The diagonal elements contain information about the affect of a parameter's value on its own affect on the function. This, of course, will always be

large. (If the diagonal element is zero than any value for that parameter will be equivalent: a property which is usually undesirable in a parameter.)

2.2 Sparce Matrix Method

One can examine the relationship between the parameters in the model to determine which pairs will have significant off-diagonal elements in the normal matrix. The pairs whose off-diagonal elements are predicted to be small can then be ignored. Such selective attention only pays off when the vast majority of elements are removed.

With some functions all the off-diagonal elements may be ignored where other functions do not allow any. One must treat functions on a case by case basis to determine which elements to use. An analysis of the residual function for x-ray diffraction shows that the size of the off-diagonal elements is related to the extent of electron density overlap of the two atoms. Since atoms are fairly compact all off-diagonal terms between parameters in atoms are negligible except for atoms bonded to one another, and the terms for those pairs are small. Since an atom has a large overlap with its own electrons the diagonal elements are very large compared to any off-diagonal ones.

The stereochemical restraints commonly used in protein refinement have a different pattern. Here the parameters of atoms connected by a bond distance or angle have strong correlation. Atoms not restrained to one another have no correlation at all. The off-diagonal terms which are nonzero are as significant as the diagonal ones.

This knowledge allows one to calculate the normal matrix as a sparse matrix, the vast majority of the off-diagonal elements are never calculated or even have computer memory allocated for their storage. The only elements calculated are the diagonal ones (including contributions from both the crystallographic and stereochemical restraints) and the off-diagonal elements for parameters from atoms directly connected by geometric restraints.

Even with the simplification of the normal matrix introduced by the sparse approximation the problem of inverting the matrix is difficult. There are a number of methods available for generating an approximation to the inverse of a sparse matrix. A discussion of these methods is beyond the scope of this paper. However it is important to note that each of them includes assumptions and approximations which should be understood when they are used.

The refinement program PROLSQ[2] uses the sparse matrix approximation to the normal matrix. PROLSQ inverts the matrix using a method called "Conjugate Gradient" which is unrelated to the Conjugate Gradient method used to minimize functions. It is a sign of confusion to state that X-PLOR[1] and PROLSQ both use the Conjugate Gradient method.

It is quite difficult to calculate the proper values for the elements of the normal matrix. To simplify these calculations Konnert and Hendrickson decided to implement all stereochemical restraints as distances. While this restructuring of the restraints does simplify the normal matrix it makes the restraints more difficult for the user to visualize and prevents the minimization method from seeing the true, underlying nature of the restraints.

While the minimization method used in PROLSQ is the most powerful of those used in large molecule refinement (and therefore the smallest radius of convergence) in practice it does not seem to work any better than simple the Conjugate Gradient method. Its limitations arise from the approximations made in the calculation of the normal matrix elements and the way the space matrix is inverted.

2.3 Diagonal Matrix

A further step in simplifying the normal matrix is made by ignoring all off-diagonal elements. The normal matrix becomes a diagonal matrix, which is inverted by simply inverting each diagonal element in turn. In essence working with the matrix has become a one-dimensional problem. Since any correlation between parameters has been assumed away the shift for a particular parameter can be calculated in isolation from the shifts of all other parameters. The Full Matrix equation 5 becomes

$$s_i = - \left| \frac{\partial f(\mathbf{p})}{\partial p_i} \right|_{\mathbf{p}=\mathbf{p}_0} \Bigg/ \left| \frac{\partial^2 f(\mathbf{p})}{\partial p_i^2} \right|_{\mathbf{p}=\mathbf{p}_0} . \tag{8}$$

2.4 Steepest Descent

A further simplification can be made if all the diagonal elements of the normal matrix have the same value. If this is true none of them need to be calculated at all. The average value can be estimated from the behaver of the function value as the parameters are shifted. The shift for a particular parameter is

simply

$$s_i = -\left|\frac{\partial f(\mathbf{p})}{\partial p_i}\right|_{\mathbf{p}=\mathbf{p}_0}. \tag{9}$$

The Steepest Descent method is far from Full Matrix. However is has the advantage of a large radius of convergence. Since the gradient of a function points in the steepest direction up hill, the Steepest Descent method simply shifts the parameters in the steepest direction down hill. It is guaranteed to reach the local minimum, given enough time. Any method which actually divides by the second derivative is subject to problems in the curvature is negative, or worst yet zero. Near a minimum all second derivatives must be positive. Near a maximum they are all negative. As one moves away from the minimum the normal matrix elements tend toward zero. The curvature becomes zero at the inflection point which surrounds each local minimum. The Full Matrix becomes unstable somewhere between the minimum and the inflection point. The Diagonal Approximation method has a similar radius of convergence.

However, the Steepest Descent method simply moves the parameters to decrease the function value. It will move toward the minimum when the starting point is anywhere within the ridge of hills surrounding the minimum.

2.5 Conjugate Gradient

The Steepest Descent method is very robust. It will smoothly converge to the local minimum whatever the starting parameters are. However it will require a great deal of time to do so. One would like a method which would reach the minimum quicker.

The problem with Steepest Descent is that no information about the normal matrix is used to calculate the shifts to the parameters. Where ever the assumptions break down (the parameters have correlation and have different diagonal elements) the shifts generated will be inefficient.

Just as one can calculate an estimate for the slope of a function by looking at the function value at two nearby points, one can estimate the curvature of a function by looking at the change in the function's gradient at two similar points. This experiment is routinely performed in Steepest Descent refinement. The gradient is calculated, the parameters shifted a little, and the gradient calculated again. In Steepest Descent the two gradients are never

compared but if they were a bit of information about the normal matrix could be learned.

The Conjugate Gradient method[3] does just this. The analysis of Fletcher and Reeves showed that the Steepest Descent shift vector can be improved by adding a well defined fraction of the shift vector of the previous cycle. Each cycle essentially "learns" about one dimension of curvature in the n dimensional refinement space. Therefore after n cycles everything is known about the normal matrix and the minimum is found.

The shift vector for cycle $k+1$ using Conjugate Gradient is

$$\mathbf{s}_{k+1} = -\left|\frac{df(\mathbf{p})}{d\mathbf{p}}\right|_{\mathbf{p}=\mathbf{p}_k} + \beta_{k+1}\mathbf{s}_k, \tag{10}$$

where β_{k+1} is the ratio of the length of the function's present gradient to its previous length. During the first cycle there is no previous cycle. The first cycle must be Steepest Descent.

The fundamental limitation of the Conjugate Gradient method is that it is guaranteed to reach the minimum in n cycles only if the Taylor's series does indeed terminate, as assumed in equation 3. If there are higher order terms, and there are in crystallographic refinement, then n cycles will only get the model nearer to the minimum. One should start over with a new run of n cycles to get the model even closer.

Even n cycles is a lot in crystallography. No one runs thousands of cycles of Conjugate Gradient refinement, nor can they be run with current software. The shifts become too small to be represented with the precision of current computers. Small shifts are not necessary unimportant ones. These small shifts add up to significant changes in the model, but we cannot calculate them.

2.6 Conjugate Direction

The Conjugate Gradient method is better than the Steepest Descent method because it uses some information about the normal matrix to improve the quality of the shift vector. It would seems reasonable to believe that the shift vector could be improved further if additional information were added. For instance, we can calculate the diagonal elements of the normal matrix directly, and quickly.

All this information is combined together in the Conjugate Direction method[5]. This method operates like the Conjugate Gradient method except it uses the shifts from the Diagonal Matrix method for its first cycle instead of the Steepest Descent method's. The shift vector in Conjugate Direction is

$$\mathbf{s}_{k+1} = -\left|\frac{df(\mathbf{p})}{d\mathbf{p}}\right|_{\mathbf{p}=\mathbf{p}_k} \Bigg/ \left|\frac{d^2 f(\mathbf{p})}{d\mathbf{p}_i^2}\right|_{\mathbf{p}=\mathbf{p}_k} + \beta'_{k+1}\mathbf{s}_k \ , \tag{11}$$

where the trick is calculating β'_{k+1} correctly. This matter is discussed in detail in [5].

3 What Does This Mean?

You have now read more than you ever wanted to know about function minimization methods. The basic facts which have been presented are that the methods commonly used find difficult cases where there are correlation between parameters and where the diagonal elements for some parameters differ from the average. These two limitations are the source of most problems in refinement.

3.1 Rigid Body

If you have a model which is in error because of an overall rotation and translation, or even if only one domain has such an error, none of the refinement packages will be able to correct this automatically. To correct this error a concerted shift of a large number of atoms must be made. The indication that such a shift is required in located in the off-diagonal elements of the normal matrix, which has been discarded. For this problem to be corrected you will have to resort to rigid body refinement.

In rigid body refinement the molecule is defined to contain one or more groups within which the atoms cannot move relative to one another. Basically the parameters of the model are recast in a form in which there no longer are correlations between the parameters.

While it is unlikely that errors of this type will arise when solving a problem with the MIR method it is quite common in MR and even Molecular Substitution (Isomorphous inhibitor or mutant structures). Refinement in these latter cases should always be started with overall rigid body refinement

and refinement with each domain as a separate body. There have been a number of cases where the refinement has "hung up" in the mid to upper 20's (percent R-factor) where the problem was eventually traced to a small unguessed domain shift.

3.2 Separate XYZ and B

Whenever no part of the normal matrix is directly calculated, as in Steepest Descent and Conjugate Gradient, the method tends to minimize the function by shifting only those parameters which have large diagonal elements. Because the diagonal elements are larger for positional parameters (like x, y, and z) than the thermal factors, the B values will not be shifted to their correct values. This is why routinely these classes of parameters are refined in separate cycles. One first refines the positional parameters holding the B values fixed and then refines the thermal factors holding the positions of the atoms constant.

However because these parameters are correlated to one another it is difficult for both types of parameters to reach their optimal values. One must repeat the cycle many times for the parameters to settle down, more cycles than are usually done.

All parameters may be varied simultaneously when the diagonal elements of the normal matrix are explicitly included in the calculation of the shifts. This is one reason why Conjugate Direction refinement requires fewer cycles.

3.3 Heavy Atoms

When refining with the Conjugate Gradient method the assumption is made that all the diagonal elements are equal. For the positional parameters these elements usually are similar enough that the real differences can be accommodated with the usual number of cycles. However one factor which contributes to the magnitude of the diagonal elements is the number of electrons surrounding the atom. Heavy atoms such as iron, calcium, and chlorine have much larger diagonal elements and will be shifted much larger distances. In fact, with Conjugate Gradient they will be shifted too far.

This is why heavy atoms tend to oscillate in refinement. The amount of shift is determined based on the average atom, which is about the size of carbon, so the heavy atoms will be over shifted. The next cycle of refinement

will attempt to correct their position but again will over shift. The atoms will slip back and forth from cycle to cycle. This problem occurs for all parameters of the heavy atom, both positional and thermal.

The problem can easily be overlooked if only the overall statistics are monitored. The mean and rms shift may be very small even though one or two atoms continue to move a great deal in each cycle. One must monitor the atoms with the largest shifts and try to understand why they continue to shift.

A minimization method which uses at least the diagonal elements will correct the problem.

3.4 High B's

Another factor which contributes to the magnitude of the diagonal elements is the size of the B value of the atom. Atoms with high B values have smaller diagonal elements for all their parameters. These atoms will be under shifted in Conjugate Gradient refinement. Since usually atoms are created with small B values and are under shifted in each cycle they will never reach their correct location nor will their thermal parameters become as large as required by the observations.

Atoms with large B values almost certainly should have even larger ones. Again, a method which uses the diagonal elements will not exhibit this problem. However, most programs which do utilize the diagonal elements of the normal matrix calculate them with the approximation that only the element type of the atom is important. They ignore the contribution of the B value to these terms. This is done for historical reasons. In small molecule structures, where refinement was originally developed, there usually are no atoms with particularly large B values, and the assumption is good. The early protein refinement packages made the same assumption without reconsidering its validity.

Therefore some programs will underestimate the thermal factors even when using the diagonal elements. To learn if this is a problem in your refinement you must discover exactly how these matrix elements are calculated. Usually this is not an easy task. The program's source code must be examined.

3.5 Local Minima

The last problem which must be considered is that we can never reach the local minimum. Often it has been said that refinement was continued until convergence at the local minimum. Even in a perfect case, where our refinement residual was quadratic, both Conjugate Gradient and Conjugate Direction would require n cycles where n is at least four times the number of atoms in the model. No one has ever ran that many cycles.

This means that no one has ever been "trapped in a local minimum". They have never reached a local minimum.

4 Summary

The fact that we cannot include all the information about our residual function into our refinement results in some parameters of the model oscillating, other becoming stuck, and the requirement that we run many, many cycles. Until more powerful methods of minimization become available the crystallographer must be on guard.

References

[1] Brunger, A.T., Kuriyan, K. & Karplus, M. (1987). Crystallographic R factor refinement by molecular dynamics. *Science* **235**, 458–460.

[2] Diamond, R., Ramaseshan, S. & Venkatesan, K., editors. *Computing in Crystallography*, chapter 13, pages 13.01–13.25. Indian Academy of Sciences, Bangalore, (1980).

[3] Fletcher, R. & Reeves, C. (1964). Function minimization by conjugate gradients. *Computer Journal* **7**, 81–84.

[4] Mandel, John. *The Statistical Analysis of Experimental Data.* Dover Publications, Inc., New York, (1984).

[5] Tronrud, D.E. (1992). Conjugate-direction minimization – An improved method for the refinement of macromolecules. *Acta Crystallogr A* **48**, 912–916.