Acta Cryst. (1998). A**54**, 543–562

Use of Statistical Information in X-ray Crystallography with Application to the Holographic Method

Abraham Szöke

Lawrence Livermore National Laboratory, Livermore, CA 94550, USA. E-mail: szoke1@llnl.gov

(*Received 17 July 1997; accepted 11 December 1997*)

Abstract

The holographic method solves the crystallographic inverse problem in real space. In addition to the measured structure-factor amplitudes, it uses other available information such as the positivity of the electron density, knowledge of part of the structure as well as MIR and/or MAD data. In the present paper, the range of useful information is extended to include knowledge that is statistical in nature. For example, it is known that the distribution of the structure-factor amplitudes of large molecules is described by Wilson statistics. Bayesian methods are used to optimize the signal-to-noise ratio of experimental measurements, to estimate missing reflections and to extrapolate measured data to higher resolution. In a similar vein, the cost function in the holographic algorithm is modified to account for the uncertainties of the measured structure factors. It is also shown how statistical knowledge about the unsolved part of the molecule may be utilized.

1. Introduction

In previous publications of the author and his collaborators [Szöke (1993, paper II), Maalouf *et al.* (1993, paper III), Somoza *et al.* (1995, paper IV) and Szöke *et al.* (1997, paper V)], X-ray crystallographic computations of macromolecules were re-examined starting from the analogy between X-ray diffraction and holography. When part of the molecule is known, the waves diffracted from that part are analogous to the reference wave in holography. The waves diffracted from the unknown part of the molecule are analogous to the object wave in holography. Their interference provides phase information that helps to decipher the X-ray diffraction pattern. The results of these considerations were called the holographic method. Our method reawakens a long dormant point of view in X-ray crystallography (Bragg, 1950; Tollin *et al.*, 1966).

Our approach can be alternatively described as a real-space method, based on a representation of the electron density in terms of basis functions with some overlap. The approach has several advantages. One of them is that various kinds of information about the electron density of the molecule can be incorporated consistently and transparently. Another one is that it shows the close analogy of the solution of the crystal structure to the reconstruction of an unknown object from a hologram and to the recovery of a three-dimensional image from a blurred picture. Both these subjects have an extensive literature and have been thoroughly analyzed mathematically. The analogy, therefore, gives good guidance for crystallographic computations.

In our previous papers on the holographic method, we did not consider the uncertainty in experimental data nor did we utilize information that can be expressed only in statistical terms, although statistical considerations have a long history in X-ray crystallography and have been very successful in improving the quality of the solutions of crystals. We refer the reader to articles in a recent book on macromolecular crystallography edited by Carter & Sweet (1997). This paper is our first direct contact with these considerations. Some of our derivations repeat earlier work; our excuse for doing so is that we put the earlier work into a new context.

In this paper, we start with a brief summary of our notation in §2. This is followed in §3 by a discussion on how to obtain the statistical distribution of the estimated magnitudes of X-ray structure factors from experimental measurements. The speed of modern computers makes it possible to use sophisticated statistical methods without simplifying assumptions and thereby get the best signal-to-noise ratio that the measurement can provide. The holographic method can then use the detailed statistics of the structure factors. In §4, we show how to use the statistics of the measured structure factors in the holographic method. We show in §5 how to use prior knowledge of missing data in terms of Wilson statistics. In §6, we go one step further: we show how to use statistical knowledge of the missing part of the molecule in order to lift the phase degeneracy of the cost function used by the holographic method. In §7, we make a few remarks on the relation of maximum-entropy reconstruction to our method and we present our summary in §8. Only the formulas in §4 of this paper have been incorporated so far into the computer program *EDEN*.

© 1998 International Union of Crystallography
Printed in Great Britain – all rights reserved

ISSN 0108-7673 © 1998

2. Brief summary of the theory

The notation in this paper is the same as in our previous papers (Szöke, 1993; Szöke *et al.*, 1997). The electron density in the unit cell of a crystal is divided into a known and an unknown part. The structure factors of the known part are denoted by $R(\mathbf{h})$. They are given by

$$R(\mathbf{h}) = \int_{\text{unit cell}} \rho_{\text{known}}(\mathbf{r}) \exp(2\pi i \mathbf{h} \cdot \mathcal{F}\mathbf{r})\, \mathrm{d}\mathbf{r}, \qquad (1)$$

where we use standard crystallographic notation. The unknown part of the electron density is described as a sum of Gaussian basis functions of equal widths, centered on a grid that divides the unit cell into P_a, P_b, P_c equal parts along the crystallographic axes a, b, c, respectively. The grid points are denoted by $\mathbf{r}_p$; $p = 1, \ldots, P$, where $P = P_a P_b P_c$. Each Gaussian blob (voxel) is assumed to contain an unknown number of electrons, $n(p)$:

$$\rho_{\text{unknown}}(\mathbf{r}) \simeq (\pi\eta\Delta r^2)^{-3/2} \sum_{p=1}^{P} n(p) \exp(-|\mathbf{r} - \mathbf{r}_p|^2/\eta\Delta r^2), \qquad (2)$$

where Δr is the mean grid spacing and η determines the width of the Gaussians relative to the grid spacing. If the grid spacing is sufficiently fine, the electron density of the unknown part of the molecule can be well approximated by such a superposition of Gaussians. When (2) is extended periodically over the repetitions of the unit cell, the structure factors of the unknown part, $O(\mathbf{h})$, can be expressed as

$$O(\mathbf{h}) = \exp[-\eta(\pi\Delta r|\mathcal{F}^T\mathbf{h}|)^2] \sum_{p=1}^{P} n(p) \exp(2\pi i \mathbf{h} \cdot \mathcal{F}\mathbf{r}_p). \qquad (3)$$

The notation $R(\mathbf{h})$ for the structure factors of the known part of the structure and $O(\mathbf{h})$ for those of the unknown part of the structure is adopted from holographic theory, where $R(\mathbf{h})$ and $O(\mathbf{h})$ denote the reference and object wave, respectively. The square of the absolute magnitude of the structure factors of the crystal, $|F(\mathbf{h})|^2$, then satisfies the equation

$$\begin{aligned} |F(\mathbf{h})|^2 &= |R(\mathbf{h}) + O(\mathbf{h})|^2 \\ &= |R(\mathbf{h})|^2 + R(\mathbf{h})O^*(\mathbf{h}) + R^*(\mathbf{h})O(\mathbf{h}) + |O(\mathbf{h})|^2. \end{aligned} \qquad (4)$$

When the representation of the unknown density is substituted from (3), (4) becomes a set of quadratic equations in the unknowns, $n(p)$. The number of equations, N_h, is usually not equal to the number of unknowns, P. The equations may contain inconsistent information, *e.g.* due to experimental errors, or lack of isomorphism in MIR, or incomplete non-crystallographic symmetry. The equations are also ill conditioned and therefore their solutions are extremely sensitive to noise in the data. Under these conditions, the equations may have many solutions or no solution at all. Our way of circumventing these problems is to obtain a 'quasi-solution' of (4) by minimizing the discrepancy or cost function (see *e.g.* Dainty & Fienup, 1987):

$$f_{\text{eden}} = \tfrac{1}{2} \sum_{\mathbf{h}} w'(\mathbf{h})^2 [|R'(\mathbf{h}) + O(\mathbf{h})| - |F'(\mathbf{h})|]^2, \qquad (5)$$

where $R'(\mathbf{h})$ and $F'(\mathbf{h})$ are apodized (smeared) versions of $R(\mathbf{h})$ and $F(\mathbf{h})$. The need for apodization and the correct ways to do it were discussed in papers II and V.

The solution of equation (5) is not unique: this is an expression of the well known phase problem of crystallography. The equivalent mathematical statement is that an arbitrary element of the null space of the 'holographic encoding operator' can be added to any vector $n(p)$ that minimizes the cost function (5) (see paper II). A simple geometric representation of this lack of uniqueness for acentric reflections, in the plane of complex numbers, is shown in Fig. 1. Since the phase of $|F(\mathbf{h})|$ is unknown, any $O(\mathbf{h})$ that connects the tip of $R(\mathbf{h})$ to any point on the circle with radius $|F(\mathbf{h})|$ results in the same value of the cost function. Thus, the difference between any two of the vectors $O(\mathbf{h})$ that satisfies this condition belongs to the null space of the encoding operator. However, additional information in the form of constraints reduces the arbitrariness of the solution.

As discussed in our previous papers, non-negativity of the electron density is always enforced. When part of the

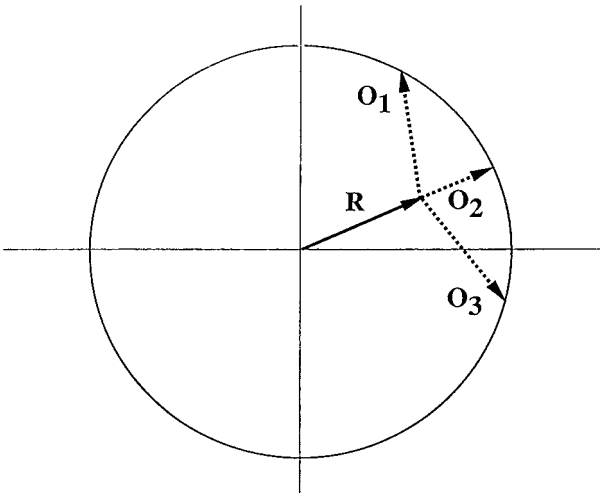

Fig. 1. Geometric representation of equation (4) in the complex plane for acentric reflections. $R(\mathbf{h})$ is the vector representing the structure factor of the known part of the electron density. The circle around the origin has the radius $|F(\mathbf{h})|$. Any vector $O(\mathbf{h})$ that starts at the tip of $R(\mathbf{h})$ and ends on the circle satisfies equation (4). The unweighted difference Fourier solution is $O_2(\mathbf{h})$. If the 'correct' solution is $O_1(\mathbf{h})$, the dual image is represented by $O_3(\mathbf{h})$. The Sim-weighted difference Fourier solution is the midpoint between the tips of $O_1(\mathbf{h})$ and $O_3(\mathbf{h})$.

molecule is known, there is statistical knowledge of the structure factors of the rest of the molecule (its unknown part). This knowledge will be used in §6 to lift the degeneracy of the cost function on the circle $|F(\mathbf{h})|$ in Fig. 1.

Sometimes there is also partial knowledge of some parts of the electron density of the molecule. Such knowledge is used by *EDEN* in terms of a 'target' density, denoted by $n(p)_{\text{target}}$. The electron density of the crystal is urged to agree with the target density by minimizing a real-space cost function, f_{space},

$$f_{\text{space}} = \tfrac{1}{2}\lambda_{\text{space}} P \sum_{p=1}^{P} \tilde{w}(p)^2 [n(p) - n(p)_{\text{target}}]^2. \quad (6)$$

The relative weight, λ_{space}, and the weights, $\tilde{w}(p)^2 \leq 1$, express the 'strength of our belief' in the correctness of the target density: the weights, $\tilde{w}(p)$, may be used to emphasize or de-emphasize different regions of the target density (although generally they are set to 1 or 0), while λ_{space} determines the relative importance of f_{space} with respect to f_{eden}. Several variants of target cost functions will be the subject of a forthcoming paper.

3. Statistical distribution of the 'true' values of structure-factor amplitudes

The amplitudes of the structure factors, $|F(\mathbf{h})|$, that appear in (4) are related to the measured intensities of the diffracted beam from the crystal. This section will concentrate on the statistical distribution of the 'true' values of the structure factors when they are estimated from (necessarily imperfect) experimental measurements. We will discuss one particular aspect of data reduction by trying to answer the following questions. Is there an optimal strategy for estimating the 'true' intensity of a Bragg reflection? Can we estimate the distribution of its likelihood, *i.e.* the probability density of the 'true' value of $|F(\mathbf{h})|^2$? Can we reliably identify 'outliers' or 'glitches', *i.e.* measurements where something went wrong with the experiment? How do we 'merge' data, *i.e.* what do we do when the intensities of two reflections that are supposed to be the same turn out to be different? We will answer these questions using Bayesian statistics.

We assume that an area detector was used in the experiment. We will skip over many of the problems faced by an experimenter: measurements of the reflected X-ray intensities have to be corrected for lack of uniformity, for possible nonlinearity and for localized flaws in the detector. Together with Bragg reflections, the detector also sees a more or less structureless background that is caused by X-ray scattering from disorder and thermal motion in the crystal as well as by the solvent in and around the crystal, by parts of the apparatus and even by the air path in front of the detector. Frames of data, which are taken sequentially, have to be put onto a common scale and possibly corrected for radiation damage to the crystal. Experimental results also have to be corrected for the absorption of X-rays in the crystal, for the polarization of the source and of the reflected beam, for possible extinction and for the rotation speed of the crystal. All these subjects are well summarized in standard textbooks (see *e.g.* Giacovazzo, 1992). There are also good computer programs that deal with these matters. Some aspects of the above problems are further elaborated on in Appendix *A*.

We assume that there are well defined regions of the area detector that are outside the Bragg peaks. These areas can be used to determine the intensity of the background. The simplest way to estimate the intensity of a Bragg reflection is by summing the measured photoelectrons in an area of the photodetector that encloses the reflection in question and by subtracting the estimated background counts from it. Our approach to the estimation of the statistical distribution of the true values of the reflected intensities follows closely that of French & Wilson (1978). Our excuse for repeating some of their work in §3.1 is to present some new results and to set the stage for the more sophisticated discussion in §3.2. In that section, we use the areas of the detector that are outside the Bragg peaks to estimate the spatial distribution of the background. In addition, we use an optimal (invariant three-dimensional) model for the shape of the Bragg reflections from the crystal. Then we optimally estimate the intensity of a reflection as well as the statistical distribution of this estimate. In §3.3, we give the likelihood that the measurement had a 'glitch' and discuss how to merge data from reflections that are supposed to have the same intensity. Both the first and the second methods are used in widely available data-processing programs. The first one is usually called data integration with background subtraction, or background–signal–background method. The second one is usually called profile fitting. For a recent review, we refer to Gilmore (1996).

3.1. *Data integration with background subtraction*

Let us assume that the integrated intensity of a certain Bragg reflection plus its background has been measured in units of photon counts by summing up the photon counts in a small area that contains 'just' the reflection. Let us denote the result by N. The background intensity in this area of the reflection can be estimated independently by adding up the photon counts in another area that surrounds the first one and scaling it by the ratio of the two areas. Let us denote the background intensity thus estimated by $\langle N_B \rangle$. Naturally, the estimated value of the background has an error associated with it. In addition to the experimental data, we will use the *a priori* (Wilson) statistics of the structure-factor amplitudes of the molecules in the crystal and the photocount

distribution of the detector. We want to know the 'true' value of $|F(\mathbf{h})|^2$ in the same units under these circumstances. This will be performed, following French & Wilson (1978), by finding the posterior distribution of the value of the integrated reflection intensity.

The 'true' value of $|F(\mathbf{h})|^2$ will be denoted by N_R. The (posterior) probability density of N_R, given the measurements N and $\langle N_B\rangle$, will be denoted by $p(N_R|N, \langle N_B\rangle)$. (Appendix B summarizes our statistical definitions and notation.) The probability distribution $p(N_R|N, \langle N_B\rangle)$ can be perceived as a conditional probability or likelihood density that expresses our ignorance of the true value of the Bragg reflection amplitude, given our knowledge of the measurement of that reflection (plus its background) and also given our best estimate of the background. It can also be thought of as a formula that determines the odds for betting on the outcome of a future experiment that would measure the same reflection intensity to very high accuracy (*e.g.* by counting for much longer).

We will now estimate $p(N_R|N, \langle N_B\rangle)$. First we discuss its 'ingredients': the uncertainty in N_B, the photocount distribution of N and the prior statistical knowledge of N_R. Then we use Bayes's law to put it all together.

The likelihood distribution of the true value of the background, N_B, given its estimated value, $\langle N_B\rangle$, will be assumed to be Gaussian,

$$p(N_B|\langle N_B\rangle) = (2\pi S^2)^{-1/2} \exp[-(N_B - \langle N_B\rangle)^2/2S^2], \quad 0 \le N_B, \langle N_B\rangle. \tag{7}$$

Equation (7) should be a good approximation when the uncertainty in the background measurement is small, $S \ll \langle N_B\rangle$. Thus, S is interpreted as the root mean square error of the estimate, expressing our ignorance of the true value of N_B.

Even though the true values of N_R and N_B are determined uniquely by the crystal and the X-ray apparatus, the actual number of counts measured in any experiment fluctuates statistically. On the most fundamental level, this follows from the quantum nature of X-rays and the limited quantum efficiency of the detector (Goodman, 1985), but imperfections of the measuring apparatus and even the judgment of the experimenter contribute to it. Although theories of photocount distributions are well founded and well tested, it is highly desirable to determine the actual statistical distribution of the signal when the detector is exposed to a constant X-ray intensity for best results (Appendix A). Such measurements can be made during the routine calibration of detectors. For illustration, we will use the highly simplified Poisson photocount distribution,

$$p(N|N_R, N_B) = [(N_R + N_B)^N/N!] \exp[-(N_R + N_B)], \quad 0 \le N, N_R, N_B. \tag{8}$$

Such a distribution would be obtained from a perfectly monochromatic nonfluctuating source and a noiseless X-ray detector. More general distributions are also described by Goodman (1985).

The additional information we will use is the nature of the crystal itself. We can assume that the dimensions of the unit cell as well as the chemical composition of the protein are known. While the chemical composition of the molecule can determine the absolute scale of the reflections as well as the solvent content of the unit cell, it is not needed at this stage. At medium to high resolution (greater than 4 Å), the structure factors of proteins are the Fourier transforms of the scattering factors of all the protein atoms, while the solvent contributes very little. Under these conditions, the amplitudes of the reflections have a Wilson distribution, which we use for our prior knowledge.

The specific Wilson distributions for centric and acentric reflections are, from (71c) and (71a) in Appendix $C1$:

$$p_c(N_R) = (2\pi\langle N_{\rm ave}\rangle N_R)^{-1/2} \exp(-N_R/2\langle N_{\rm ave}\rangle) \quad \text{(centric)} \tag{9c}$$

$$p_a(N_R) = \langle N_{\rm ave}\rangle^{-1} \exp(-N_R/\langle N_{\rm ave}\rangle) \quad \text{(acentric)}, \tag{9a}$$

where $\langle N_{\rm ave}\rangle$ is the average of the measured reflection intensities within a resolution shell. When the reflections are measured, their scale is unknown. It can be estimated from a simple-minded average and its magnitude can be corrected iteratively. In Bayesian statistics, these probabilities are called priors.

In order to estimate $p(N_R|N, \langle N_B\rangle)$, we will use the combination law of probabilities (67) and Bayes's rule (68):

$$p(N_R|N, \langle N_B\rangle) = \int_0^\infty p(N_R|N, N_B)p(N_B|\langle N_B\rangle)\, dN_B, \tag{10}$$

$$p(N_R|N, N_B) = p(N|N_R, N_B)p(N_R, N_B)/p(N, N_B). \tag{11}$$

As the intensities of the Bragg reflection and that of the background are statistically independent, the joint probabilities can be simplified:

$$p(N_R, N_B) = p(N_R)p(N_B) \tag{12}$$

$$p(N, N_B) = p(N|N_B)p(N_B). \tag{13}$$

The conditional probability, $p(N|N_B)$, is calculated in Appendix $C2$. Finally, substituting from (12) and (13), we get

$$p(N_R|N, \langle N_B\rangle) = p(N_R) \int_0^\infty p(N|N_R, N_B)[p(N_B|\langle N_B\rangle)/p(N|N_B)]\, dN_B, \tag{14}$$

where for centric and acentric reflections we should use $p_c(N_R)$ and $p_a(N_R)$, respectively, from (9c) and (9a) as

well as $p_c(N|N_B)$ and $p_a(N|N_B)$ from (75*c*) and (75*a*). For the simplified assumptions we made above, we can substitute from (8) for $p(N|N_R, N_B)$ and from (7) for $p(N_B|\langle N_B\rangle)$. For the most reliable results, experimental distributions for $p(N|N_R, N_B)$ and $p(N_R)$ should be used and (14), as well as (74*a*) and (74*c*), integrated numerically. Note that N, N_R, N_B and $\langle N_B\rangle$ are restricted to be non-negative.

When the uncertainty in the background measurement is small compared to N, $\langle N_B\rangle$ and also compared to $|N - \langle N_B\rangle|$, the integral can be carried out. The result is the substitution of $\langle N_B\rangle$ for N_B everywhere. If, in addition, the number of photons within the peak is large, $N \gg 1$, the Poisson distribution can be well approximated by a Gaussian distribution with a peak at N and a width $N^{1/2}$. The maximum of the likelihood distribution for centric reflections, p_c, is at

$$\begin{aligned} N_R &= N - \langle N_B\rangle - N/2\langle N_{\text{ave}}\rangle \\ &= N(1 - 1/2\langle N_{\text{ave}}\rangle) - \langle N_B\rangle, \end{aligned} \tag{15}$$

while for an acentric reflection the maximum of p_a is at

$$\begin{aligned} N_R &= N - \langle N_B\rangle - N/\langle N_{\text{ave}}\rangle \\ &= N(1 - 1/\langle N_{\text{ave}}\rangle) - \langle N_B\rangle. \end{aligned} \tag{16}$$

The likelihood distribution of N_R is itself a Gaussian of width $N^{1/2}$, so the σ value of $|F(\mathbf{h})|^2$ (in units of N_R) is $N^{1/2}$. The terms subtracted, $N/2\langle N_{\text{ave}}\rangle$ for centric and $N/\langle N_{\text{ave}}\rangle$ for acentric reflections, are corrections for our prior knowledge of the magnitudes of the reflections. It should be stressed that (15) and (16) are for the high signal-to-noise case only. Therefore, $\langle N_{\text{ave}}\rangle \gg 1$ and the corrections to N are relatively small. In the low signal-to-noise regime, where quantum fluctuations are important, the full equations (14) have to be used.

3.2. *'Profile fitting'*

Let us assume that most Bragg reflections are well separated on the area detector. Let us also assume that the positions of the reflections are predicted quite accurately by data-acquisition programs. Under such conditions, 'fitting the profile' of the reflections can improve the statistical uncertainty of the measurements significantly. This section will borrow ideas from Oatley & French (1982), as well as from Diamond (1969), Rossmann (1979), Rossmann *et al.* (1979), Kabsch (1988) and Otwinowski (1993).

A simple way to understand the possibility of improving the signal-to-noise ratio, compared to the simple integration introduced in the previous section, is by asking the question: how big an area should we integrate around a reflection in order to obtain the smallest statistical uncertainty? If we integrate a very large area, we encompass all the photon counts from the Bragg reflection but we include too many background counts. As the uncertainty is proportional to $N^{1/2}$ and increasing the integrated area hardly increases the counts from the reflection, we clearly increase our overall uncertainty. If we integrate too small an area, we lose part of the 'good' counts from the reflection itself, increasing the uncertainty of the measurement again. Thus there is clearly an optimum area to be integrated. [A computer program that optimizes the area was published by Bolotovsky *et al.* (1995).] One should do even better by a weighted integration of the counts in different pixels within the optimum area. The weights should be large where the ratio of the signal to the background is large and they should be small where the signal is small compared to the background. This is the procedure followed in *DENZO* (Otwinowski, 1993) and *MOSFLM* (CCP4, 1994). We should do even better if we had a good independent estimate of the background. All these considerations can be included properly in Bayesian estimation. The formalism in this section is a straightforward extension of our treatment in the previous section.

In Appendix *A*, we discuss the procedure to obtain estimates for the background intensity and its standard error as well as for the profiles of the reflection spots and their error distribution. In this section, we will assume that those profiles have been obtained, parametrized and their correct positions have been predicted by a suitable indexing program. For ease of presentation, we will assume two-dimensional Gaussian reflection profiles, no error in their widths and a two-dimensional Gaussian distribution of the error in their predicted position. The formalism is independent of these assumptions but their use gives relatively simple analytic results.

We are given the following data: (*a*) A set of measured photon counts $\{N(i)\}$, where i denotes the two-dimensional pixel coordinates of the detector in a generous vicinity of a Bragg reflection and where the braces $\{\ldots\}$ signify that we deal with the whole set of measurements, $\{N(i), i = 1, \ldots, k\}$. (*b*) An estimate of the center position of the reflection $\langle i_o\rangle$ and the probability distribution of the error in this estimate. (*c*). An estimate of the shape of the reflection, $\{\Pi_\alpha(i - i_o)\}$, expressed as an analytic function with some parameters and the probability distribution of the errors in those parameters. (*d*) An estimate of the background $\{\langle N_B(i)\rangle\}$ as a function of position and the standard error of that estimate. (*e*) The *a priori* (Wilson) statistics of the structure-factor amplitudes of the molecules. (*f*) The photocount distribution of the detector. We want to know the posterior distribution of the value of the integrated reflection intensity N_R under these circumstances. The distribution we seek will be denoted by $p(N_R|\{N(i)\}, \{\langle N_B(i)\rangle\}, \langle i_o\rangle, \{\Pi_\alpha(i - i_o)\})$, where the symbols will be defined more accurately below. As in the previous section, we will first discuss the separate ingredients of the distribution, then put them together using Bayes's rule.

We denote by $\langle i_o \rangle$ our best estimate of the center position of the reflection; let the probability distribution of its true center, i_o, be given by

$$p(i_o|\langle i_o\rangle) = [1/2\pi(\Delta i)^2] \exp[-|i_o - \langle i_o\rangle|^2/2(\Delta i)^2], \quad (17)$$

where Δi is the uncertainty in the estimated position of the reflection. Our best estimate of the reflection profile was denoted by $\{\Pi_\alpha(i - i_o)\}$, where the symbol α denotes a set of parameters to describe its actual shape. For definiteness, we will use a Gaussian shape of width $\Delta\Pi$. What this means is that we can calculate the 'true' value of the contribution of the Bragg reflection at point i from its 'true' integrated value, N_R, using

$$\begin{aligned} N_R(i) &= N_R \Pi_\alpha(i - i_o) \\ &= N_R[1/2\pi(\Delta\Pi)^2] \exp[-|i - i_o|^2/2(\Delta\Pi)^2]. \end{aligned} \quad (18)$$

For simplicity, we will forego the probability distribution in the value of $\Delta\Pi$. The last two equations are actually oversimplified. As discussed in Appendix A, we should be able to establish a good 'invariant' reciprocal-space profile for the crystal. We can then use our knowledge of the experimental geometry to transform the invariant profile into detector coordinates. Even if the three-dimensional invariant profile is Gaussian, it is not assured that (18) will describe $\{N_R(i)\}$ correctly or that the uncertainty in the estimated position of the center of the reflection will be given by (17). The good news is that, by doing the computations properly, correct formulas can be obtained and used.

We will make two alternative assumptions about the uncertainty in the estimation of the background intensity. First, we will assume that it is independent of position; in that case, (7) will be modified to

$$\begin{aligned} &p(N_B(i)|\langle N_B(i)\rangle) \\ &\quad = (2\pi S^2)^{-1/2} \exp\{-[N_B(i) - \langle N_B(i)\rangle]^2/2S^2\}, \end{aligned} \quad (19a)$$

with the conditions $0 \le N_B(i), \langle N_B(i)\rangle$. An alternative assumption is that the uncertainty in the background is proportional to the value of $\langle N_B(i)\rangle$. This means that the uncertainty is essentially a scale factor

$$\begin{aligned} &p(N_B(i)|\langle N_B(i)\rangle) \\ &\quad = [2\pi\Sigma^2\langle N_B(i)\rangle^2]^{-1/2} \\ &\qquad \times \exp\{-[N_B(i) - \langle N_B(i)\rangle]^2/2\Sigma^2\langle N_B(i)\rangle^2\}, \end{aligned} \quad (19b)$$

where $S = \Sigma\langle N_B\rangle$ and $\langle N_B\rangle$ denotes an average background value. The photocount distribution of each pixel is the same and the photon counts are not correlated in different pixels. Therefore, (8) can also be generalized to

$$\begin{aligned} &p(N(i)|N_R, N_B(i)) \\ &\quad = \{[N_R(i) + N_B(i)]^{N(i)}/N(i)!\} \exp\{-[N_R(i) + N_B(i)]\}, \end{aligned} \quad (20)$$

with the conditions $0 \le N(i), N_R(i), N_B(i)$. We have to use (18) for the actual values of $N_R(i)$, given N_R.

We can now extend (20) to calculate the joint probability of the measurements $\{N(i)\}$ by using the independence of the photocount statistics of different pixels.

$$\begin{aligned} p(\{N(i)\}|N_R, \{N_B(i)\}) &= \prod_{i=1}^{k} p(N(i)|N_R(i), N_B(i)) \\ &= \prod_{i=1}^{k} \{[N_R(i) + N_B(i)]^{N(i)}/N(i)!\} \\ &\quad \times \exp\{-[N_R(i) + N_B(i)]\}, \end{aligned} \quad (21)$$

where again we have to use (18) for the actual values of $N_R(i)$, given N_R.

We can now extend the arguments leading to equations (11)–(14). Let us start with Bayes's rule (68) and apply it to (21):

$$\begin{aligned} &p(N_R|\{N(i)\}, \{N_B(i)\}) \\ &\quad = p(\{N(i)\}|N_R, \{N_B(i)\}) \\ &\qquad \times p(N_R, \{N_B(i)\})/p(\{N(i)\}, \{N_B(i)\}). \end{aligned} \quad (22)$$

As in (12) and (13), we use the statistical independence of N_R and $N_B(i)$ and use

$$p(N_R, \{N_B(i)\}) = p(N_R)p(\{N_B(i)\}) \quad (23)$$

$$p(\{N(i)\}, \{N_B(i)\}) = p(\{N(i)\}|\{N_B(i)\})p(\{N_B(i)\}) \quad (24)$$

$$\frac{p(N_R, \{N_B(i)\})}{p(\{N(i)\}, \{N_B(i)\})} = \frac{p(N_R)}{p(\{N(i)\}|\{N_B(i)\})}. \quad (25)$$

Finally, we can put all our formulas together. The required probability distribution will be interpreted as

$$\begin{aligned} &p(N_R|\{N(i)\}, \{\langle N_B(i)\rangle\}, \langle i_o\rangle, \{\Pi_\alpha(i - i_o)\}) \\ &\quad = \int p(i_o|\langle i_o\rangle)\, \mathrm{d}i_o \int_0^\infty p(\{N_B(i)\}|\{\langle N_B(i)\rangle\})\, \mathrm{d}\{N_B(i)\} \\ &\qquad \times p(N_R|\{N(i)\}, \{N_B(i)\}), \end{aligned} \quad (26)$$

where the shorthand $p(\{N_B(i)\}|\{\langle N_B(i)\rangle\})$ denotes the product of independent probabilities of the background distribution, (19a) or (19b), and $\mathrm{d}\{N_B(i)\}$ is the k-dimensional integration over them. Substituting from (22) and (23), we get the general formula for 'profile fitting',

$$\begin{aligned} &p(N_R|\{N(i)\}, \{\langle N_B(i)\rangle\}, \langle i_o\rangle, \{\Pi_\alpha(i - i_o)\}) \\ &\quad = \int \mathrm{d}i_o \int_0^\infty \mathrm{d}\{N_B(i)\}\, p(\{N(i)\}|N_R, \{N_B(i)\})p(N_R) \\ &\qquad \times p(i_o|\langle i_o\rangle)p(\{N_B(i)\}|\{\langle N_B(i)\rangle\}/p(\{N(i)\}|\{N_B(i)\}). \end{aligned} \quad (27)$$

The factors in the integrand of this complicated formula were defined elsewhere: $p(\{N(i)\}|N_R, \{N_B(i)\})$ is defined in (21); $p(N_R)$ is defined in (9c) and (9a) for centric and acentric reflections, respectively; $p(i_o|\langle i_o\rangle)$ is defined in

(17); $p(N_B(i)|\langle N_B(i)\rangle)$ is defined in (19*a*) or (19*b*); $p(\{N(i)\}|\{N_B(i)\})$ is calculated in Appendix *C*2, equations (76*c*) and (76*a*), and we neglect the estimated error in the shape of the profile everywhere.

We will now explore some consequences of this formula. Let us first neglect the uncertainty in the position and the shape of the reflection as well as the uncertainty in the estimate of the background. Let us also assume that the *a priori* distribution of the reflection intensities is uniform, *i.e.* $p(N_R)$ is constant, and that there are many photon counts in each pixel so the Poisson distribution, (21), can be approximated by a Gaussian of mean $[N_R(i)+N_B(i)]$ and of standard uncertainty $[N_R(i)+N_B(i)]^{1/2}$. Then,

$$p(N_R|\{N(i)\},\{\langle N_B(i)\rangle\},\langle i_o\rangle,\{\Pi_\alpha(i-i_o)\}) \simeq \prod_{i=1}^{k}\{2\pi[N_R(i)+N_B(i)]\}^{-1/2} \times \exp\{[N(i)-N_R(i)-N_B(i)]^2/2[N_R(i)+N_B(i)]\}. \quad (28)$$

The maximum-likelihood estimate is obtained by minimizing the exponent,

$$\sum_{i=1}^{k}[N(i)-N_R(i)-N_B(i)]^2/2[N_R(i)+N_B(i)] = \min. \quad (29)$$

Substituting from (18) for the reflection profile and estimating $[N_R(i)+N_B(i)]$ in the denominator by $N(i)$, we get the formulas used in *DENZO* (Otwinowski, 1993) and *MOSFLM* (CCP4, 1994). It is clear therefore that our equations generalize those for the case when there are errors in the position and the shape of the reflection as well as in the background and when there is *a priori* knowledge of Wilson statistics. We can also calculate the mean square error in our estimate of N_R to be

$$(\Delta N_R)^2 = \left[2\sum_i \Pi_\alpha(i-i_o)^2/2N(i)\right]^{-1}. \quad (30)$$

3.3. *Combination of several measurements and check of their consistency*

In the 'profile fitting' method of §3.2, the value of N_R is obtained from a large set of measurements, $\{N(i)\}$, $\langle N_B(i)\rangle$ on a single Bragg reflection. If the uncertainty in each measurement, (20), can be approximated by a Gaussian, we can ask the question whether the distribution of photon counts in individual pixels is compatible with our premise of a single reflection with a well estimated shape. The simplest statistical test to answer this question is the χ^2 test:

$$\chi^2 = \sum_{i=1}^{k}[N(i)-N_R(i)-N_B(i)]^2/2[N_R(i)+N_B(i)], \quad (31)$$

where we should compare it with tables for $k-1$ degrees of freedom. If our estimate of the invariant profile of the reflection is correct, the values of χ^2 for most reflections should fall within the limits $(k-1)\pm[2(k-1)]^{1/2}$. If most of the values are significantly larger, this is a sign that we did not estimate the profile well or that there is some external variability of the measurement that was not accounted for. In either case, we should take remedial action. If there are 'glitches' in the measurement of any reflection, owing to a cosmic ray hitting the detector or some stray radiation reflected from a part of the apparatus, this will show up as a particularly large value of χ^2 for a very few reflections.

In addition, most reflections are measured repeatedly in equivalent crystallographic positions. In present practice, their values are combined with appropriate weights to give an experimental value and variance for the 'unique' reflection chosen to represent the group of equivalent reflections. If the differences among the individual measurements are too large, some of the reflections are not included in the average. We propose to combine reflections using the ideas of Press (1997). First, as written above, we have an estimate for the variances of N_R and the various measurements of the equivalent reflections should be reasonably within the required statistical distribution. Second, we may use Press's method to assign a probability to the 'truth' of the estimate of the variances of the equivalent reflections and combine them that way.

In summary, measured reflection data should be examined on two different levels. First, the counts in each pixel within an individual reflection should be statistically compatible. Second, reflections that are supposed to be the same (like $\mathbf{h}$ and $-\mathbf{h}$ in the absence of anomalous dispersion) are measured separately and their values are combined. Knowing the standard errors in the estimates of each reflection allows a statistical comparison and the use of Press's (1997) technique.

4. *EDEN* cost function in the presence of experimental errors

The holographic method is based on the solution of the set of algebraic equations (4). A mathematically stable solution is provided by the minimization of the basic cost function in *EDEN*, (5). It was implicitly assumed in the derivation that the values of the structure factors, $|F(\mathbf{h})|$, are perfectly known. In view of the previous section, we should modify the cost function in order to minimize its expected value. The likelihood distribution of the 'true' structure-factor amplitudes, N_R, was given by (14) for simple integration with background subtraction, and it was given by the much more

complicated expression, (27), for the 'profile-fitting' case. In order to simplify matters, we will assume that one or other formula was reduced to the probability distribution of the structure-factor amplitudes, $p(|F'(\mathbf{h})|\big||F'(\mathbf{h})_{\rm meas}|)$, where $|F'(\mathbf{h})_{\rm meas}|$ are the 'measured' structure-factor amplitudes as they appear in the files of experimental values and (possibly) apodized by *EDEN*. The expected value of the cost function, $\langle f_{\rm eden}\rangle$, is therefore given by

$$\langle f_{\rm eden}\rangle = \tfrac{1}{2}\sum_{\mathbf{h}} w'(\mathbf{h})^2 \int [|R'(\mathbf{h}) + O(\mathbf{h})| - |F'(\mathbf{h})|]^2 \times p(|F'(\mathbf{h})|\big||F'(\mathbf{h})_{\rm meas}|)\, {\rm d}|F'(\mathbf{h})|. \quad (32)$$

This formula applies to an arbitrary statistical distribution of the measurement. In particular, it is known from French & Wilson (1978) that for small measured values the distribution $p(|F'(\mathbf{h})|\big||F'(\mathbf{h})_{\rm meas}|)$ is not Gaussian.

We will now show that if $p(|F'(\mathbf{h})|\big||F'(\mathbf{h})_{\rm meas}|)$ is approximately Gaussian, equation (32) reduces to the usual least-squares prescription.† For a Gaussian distribution of width σ, the integration can be carried out; the result is

$$\langle f_{\rm eden}\rangle = \tfrac{1}{2}\sum_{\mathbf{h}} w'(\mathbf{h})^2 \{[|R'(\mathbf{h}) + O(\mathbf{h})| - |F'(\mathbf{h})_{\rm meas}|]^2 + \sigma(\mathbf{h})^2\}. \quad (33)$$

Further progress can be made by using the discrepancy principle (Morozov, 1966). The principle says that the solution of a crystal structure should not be better than the quality of the data. It translates to a criterion that each term in (33) should not be smaller than $\sim \sigma(\mathbf{h})^2$. When this criterion is satisfied, the derivative of each term in (33) is proportional to $\sigma(\mathbf{h})$, so the use of $\langle f_{\rm eden}\rangle$ would have the tendency of optimizing the less well known structure factors relatively better than the better known ones. In order to equalize their uncertainties, each term should be weighted by $1/\sigma(\mathbf{h})^2$. Then the last term in the curly brackets in (33) becomes a constant additive term and can be neglected. The result has to be normalized to make it independent of the values of the experimental uncertainty, $\sigma(\mathbf{h})^2$.

$$\langle f_{\rm eden}\rangle = \tfrac{1}{2}(1/\mathcal{N})\sum_{\mathbf{h}}[w'(\mathbf{h})^2/\sigma(\mathbf{h})^2] \times \{[|R'(\mathbf{h}) + O(\mathbf{h})| - |F'(\mathbf{h})_{\rm meas}|]^2\}, \quad (34)$$

where $\mathcal{N}$ is the average weight in the unit cell,

$$\mathcal{N} = (1/N_{hkl})\sum_{\mathbf{h}} w'(\mathbf{h})^2/\sigma(\mathbf{h})^2, \quad (35)$$

and N_{hkl} is the number of terms in the summation in (34). The discrepancy principle also determines a stopping criterion for the solver: the average value of $[|R'(\mathbf{h}) + O(\mathbf{h})| - |F'(\mathbf{h})_{\rm meas}|]^2$ should be $\sigma(\mathbf{h})^2$. This translates to

$$\langle f_{\rm eden}\rangle_{\rm min} = \tfrac{1}{2}(1/\mathcal{N})\sum_{\mathbf{h}} w'(\mathbf{h})^2. \quad (36)$$

Equations (34) and (36) are a very satisfactory result. Inverse σ^2 weighting has been used in crystallography for a long time and its most sophisticated justifications have used a maximum log-likelihood criterion (see *e.g.* Read, 1986, 1990) or a maximum-entropy argument (Gilmore, 1996). We have shown that, when the crystallographic reconstruction is cast in the language of inverse problems, the same criterion results from the statistical uncertainty of the data (at least under conditions of a Gaussian distribution of the experimental errors).

It should be noted that in equation (34) the minimum of the cost function occurs when the calculated structure factors are equal to the measured ones. The inverse σ^2 weighting gives relatively more weight to well measured reflections. In usual practice, an inverse σ^2-weighted map is calculated by multiplying the measured structure factors by $1/\sigma^2$ so the resulting maps are necessarily distorted. The above criticism does not apply to the use of $1/\sigma^2$ weights in crystallographic refinement. It should also be noted that our use of $1/\sigma^2$ weights in MIR or MAD problems should obviate the separate use of figures of merit. The above formulas were incorporated in *EDEN*, using $w'(\mathbf{h})^2 = 1$. For the record, the detailed formulas are listed in Appendix *D*. As expected, the inverse σ^2 weighting improved some electron-density maps and the use of (36) as a stopping criterion has reduced the extent to which *EDEN* 'churns' beyond a reasonable stopping point. Details of our experience will be presented in a forthcoming paper.

5. Missing data

In the holographic method, we take great care to distinguish between missing data and data with the value zero. In all cost functions in reciprocal space, *e.g.* equations (5) and (34), (35), only measured reflections appear. (Reflections forbidden by symmetry are automatically added by *EDEN*.) Similarly, the computation of the gradient (see Appendix *A* of paper V) involves only transforms from real space to reciprocal space. When the electron density that corresponds to a set of calculated structure factors is needed, we still only use Fourier transformation from real space to reciprocal space and never in the opposite direction. The common thread in all these methods is that reflections that are not measured are free to take any value.

This procedure, although clearly superior to substituting a zero value for missing reflections, can be criticized on two grounds. First, leaving the values of the unknown reflections completely free may cause numerical instability. In particular, if the gridding resolution is

† For a skewed distribution, the minimum of $\langle f_{\rm eden}\rangle$ contains additional terms and it does not reduce to the familiar least-squares formula.

higher than the measured data resolution, the electron density can develop spurious high-frequency oscillations. The traditional way of dealing with such phenomena is to introduce regularization methods (Morozov, 1966) or use maximum-entropy reconstruction (Gilmore, 1996). We deal with similar problems partly by apodization. Second, more importantly, we do have statistical information on the missing data that we have not used yet. This will be discussed below.

If most of the reflections within a resolution shell are measured, *and all measured reflections are kept*,† we can calculate the average of the squares of the structure-factor amplitudes in a properly chosen resolution shell, $\Sigma = \langle |F(\mathbf{h})|^2 \rangle$. The *a priori* distribution of the magnitudes of the structure factors within a resolution shell will be assumed to have the Wilson distribution (Appendix *C*1)

$$p_c(|F(\mathbf{h})|) = (2/\pi\Sigma)^{1/2} \exp[-|F(\mathbf{h})|^2/2\Sigma] \quad \text{(centric)}, \tag{37}$$

$$p_a(|F(\mathbf{h})|) = [2|F(\mathbf{h})|/\Sigma] \exp[-|F(\mathbf{h})|^2/\Sigma] \quad \text{(acentric)}. \tag{38}$$

These results can be used in *EDEN* by adding a cost function that is small when $p_c(|F(\mathbf{h})|)$ and $p_a(|F(\mathbf{h})|)$ are large and *vice versa*. The relative weight λ_{Wilson} of this cost function can be determined *a posteriori* by the resulting distribution of $|F(\mathbf{h})|$: if the distribution is too narrow, λ_{Wilson} has to be lowered and *vice versa*.

Our gridding resolution is usually higher than the extent of the measured reflections. If the measured resolution is higher than ~3 Å, we can safely use the theoretical value,

$$\Sigma = \langle |F(\mathbf{h})|^2 \rangle = \exp[-B(|\mathcal{F}^T\mathbf{h}|)^2/4] \sum_n f_n^2, \tag{39}$$

for the average magnitude of $|F(\mathbf{h})|^2$; we can use (37) and (38) for the statistical distribution of the centric and acentric reflections, respectively. The resulting cost function is expected to be a good regularizer at high resolution.

6. *EDEN* cost function for the unknown part of the unit cell

From Fig. 1, we see that the structure factors arising from the missing part of the molecule satisfy the inequality $||F(\mathbf{h})| - |R(\mathbf{h})|| \le |O(\mathbf{h})| \le ||F(\mathbf{h})| + |R(\mathbf{h})||$ but, in the absence of additional information, any point on the circle in Fig. 1 is equally probable as the end point of $O(\mathbf{h})$. Accordingly, the basic cost function of *EDEN*, (5), is degenerate on the circle: it has the same value (zero) at all points. In fact, we usually know the chemical composition of the missing part of the molecule. That information determines the statistical distribution of $|O(\mathbf{h})|$. Therefore, not all points on the circle in Fig. 1 should be equally probable. In this section, we work out the mathematics.

At reasonably high resolution, the solvent region does not contribute to the structure factors significantly; thus it will be ignored in the first approximation. The number of atoms in the unknown part and their atomic number are known. Under the usual statistical assumptions (see also Appendix *C*1), we get the Wilson distribution for the magnitudes of the structure factors of the unknown part, $|O(\mathbf{h})|$:

$$p_c(|O(\mathbf{h})|) = (2/\pi\Sigma)^{1/2} \exp[-|O(\mathbf{h})|^2/2\Sigma] \quad \text{(centric)}, \tag{40}$$

$$p_a(|O(\mathbf{h})|) = [2|O(\mathbf{h})|/\Sigma] \exp[-|O(\mathbf{h})|^2/\Sigma] \quad \text{(acentric)}, \tag{41}$$

where Σ is the sum of the squared scattering factors, f_n^2, of the missing atoms only, taking into account their average thermal motion

$$\Sigma = \exp[-B(|\mathcal{F}^T\mathbf{h}|)^2/4] \sum_n f_n^2. \tag{42}$$

Alternatively, we can assume that the missing part of the structure is similar to the known part, and scale the measured structure factors by the ratio of the number of known to the number of unknown electrons of the molecule. The above considerations are very similar to those presented by Bricogne (1984) and Read (1986).

6.1. *No 'solvent mask'*

The calculations in this section will be similar to those in §3. We want to calculate the (posterior) probability distribution $p(|O(\mathbf{h})|\,\big|\,|R(\mathbf{h})|, |F(\mathbf{h})_{\text{meas}}|)$. In order to simplify the notation, we will omit the index in reciprocal space and write $p(|O|\,\big|\,|R|, |F_{\text{meas}}|)$. Taking into account the experimental uncertainties in the 'true' value of $|F|$, as discussed at length in §3, we get from the composition law of probabilities, (67),

$$p(|O|\,\big|\,|R|, |F_{\text{meas}}|) = \int p(|O|\,\big|\,|R|, |F|) p(|F|\,\big|\,|F_{\text{meas}}|)\, \mathrm{d}|F|. \tag{43}$$

Using Bayes's law, a calculation similar to (11) gives

$$p(|O|\,\big|\,|F|, |R|) = p(|F|\,\big|\,|R|, |O|) p(|R|, |O|)/p(|F|, |R|), \tag{44}$$

where the magnitude of $|O|$ on the left-hand side is bounded by

$$\big||F| - |R|\big| \le |O| \le |F| + |R|. \tag{45}$$

† In §3, we outlined an optimal method for deducing the 'true' amplitudes of the reflections. We strongly disagree with the practice of disregarding measured reflections only because their amplitudes are small.

As $|R|$ and $|O|$ refer to different parts of the molecule, their magnitudes are statistically independent. Therefore,

$$p(|R|, |O|) = p(|R|)p(|O|). \tag{46}$$

The magnitudes of $|F|$ and $|R|$ are not statistically independent. For brevity, we will present below the calculations for acentric reflections. (In Appendix *C*3, centric reflections are treated as well.) We know that $F = R + O$, where all three denote complex quantities. We introduce, as auxiliary variables, the magnitude of $|O|$ and the relative phase between R and O, which we will denote by φ. Then,

$$\begin{aligned} p_a(|F|, R) &= p_a(|F|\big|R)p_a(R) \\ &= p_a(|R|) \int p_a(|F|\big|R, |O|, \varphi) \\ &\quad \times p_a(|O|, \varphi)\, \mathrm{d}|O|\, \mathrm{d}\varphi. \end{aligned} \tag{47}$$

The assumptions that lead to (47) are that the phases of R are uniformly distributed and that the relative phase between R and O, denoted by φ, is independent of the absolute phase of R. We will make now the more restrictive assumption that the magnitude and the phase of O are statistically independent, $p_a(|O|, \varphi) = p_a(|O|)p_a(\varphi)$.

A similar substitution into the numerator of (44) gives

$$\begin{aligned} &p_a(|F|\big|R, |O|)p(|O|) \\ &\quad = \int p_a(|F|\big|R, |O|, \varphi)p_a(|O|)p_a(\varphi)\, \mathrm{d}\varphi. \end{aligned} \tag{48}$$

Finally, substituting (47) and (48) into (44), we get

$$p_a(|O|\big||F|, R) = \frac{\int p_a(|F|\big|R, |O|, \varphi)p_a(|O|)p_a(\varphi)\, \mathrm{d}\varphi}{\iint p_a(|F|\big|R, |O|, \varphi)p_a(|O|)p_a(\varphi)\, \mathrm{d}|O|\, \mathrm{d}\varphi}. \tag{49}$$

Equation (49) contains very similar expressions in the numerator and the denominator. The calculations are carried out in detail in Appendix *C*3 both for acentric and for centric reflections. The results for acentric reflections are

$$\begin{aligned} p_a(|O|\big||F|, R) &= |O| \exp[-|O|^2/\Sigma] \\ &\quad \times \left(|F|\{[|F|^2 - (|R| - |O|)^2] \right. \\ &\quad \times \left. [(|R| + |O|)^2 - |F|^2]\}^{1/2}\right)^{-1} \\ &\quad \times \{\exp[-(|F|^2 + |R|^2)/\Sigma] \\ &\quad \times I_0[2|F||R|/\Sigma]\}^{-1} \\ &\qquad\qquad \text{(acentric)}, \end{aligned} \tag{50a}$$

where I_0 is the modified Bessel function of the first kind, zero order (Abramowitz & Stegun, 1972). The result agrees with Goodman (1985, equations 2.9–20).

The results for centric reflections are

$$\begin{aligned} p_c(|O|\big||F|, R) &= \{\delta(|O| - ||R| - |F||) \\ &\quad \times \exp[-(|R| - |F|)^2/2\Sigma]\} \\ &\quad \times \{\exp[-(|R| - |F|)^2/2\Sigma] \\ &\quad + \exp[-(|R| + |F|)^2/2\Sigma]\}^{-1} \\ &\quad + \{\delta(|O| - ||R| + |F||) \\ &\quad \times \exp[-(|R| + |F|)^2/2\Sigma]\} \\ &\quad \times \{\exp[-(|R| - |F|)^2/2\Sigma] \\ &\quad + \exp[-(|R| + |F|)^2/2\Sigma]\}^{-1} \\ &\qquad\qquad \text{(centric)}. \end{aligned} \tag{50c}$$

The reader is reminded that in both formulas (50*a*) and (50*c*) the limits on $|O|$ satisfy (45). Finally, (50*a*) or (50*c*) have to be substituted into (43). The integrals have to be evaluated numerically.

These results can be used in *EDEN* by adding a cost function in reciprocal space that is small when $p(|O|\big||R|, |F_{\mathrm{meas}}|)$ is large and *vice versa*. The relative weight $\lambda_{\mathrm{missing}}$ of this cost function can be determined *a posteriori* by the resulting distribution of the amplitudes of the structure factors of the unknown part, $|O|$: if the distribution is too narrow, $\lambda_{\mathrm{missing}}$ has to be lowered and *vice versa*.

6.2. *Spatial knowledge about the unknown part of the molecule*

Sometimes there is knowledge about the approximate location of the missing part of the molecule. This can happen if there is a known solvent region or there are parts of the molecule that have been satisfactorily solved and then we know that the missing part is somewhere else. There are two consequences of such knowledge. First, the prior distribution of the density of the unknown part is not uniform. This manifests itself in a prior probability distribution that is different from those in (40) and (41). Second, knowledge of the whereabouts of the missing part can be incorporated directly into *EDEN* by a spatial target function.

The influence of spatial nonuniformity of the molecule on the statistics of its structure factors was derived by Bricogne (1984) and summarized on p. 382 of his article in Carter & Sweet (1997). The simplest prior distribution of the unknown part of the structure is a uniform one in the allowed region. For our purposes, we can then use a Gaussian approximation to the structure-factor amplitudes, whose first and second moments are given explicitly by Bricogne's formulas. The resulting distribution can be substituted into equation (48). Knowledge of the assumed location of the unknown part of the molecule should be used in one more way: the calculated structure factors of the unknown part $O'(\mathbf{h})$ are to be calculated only from the part of the molecule that is in the unknown region. We will elaborate on this

topic in a forthcoming paper that deals primarily with spatial target functions.

7. Maximum-entropy reconstruction

The essence of the holographic method is that it obtains an electron-density map that simultaneously agrees with all available knowledge both in real and in reciprocal space. However, there is often a range of maps that meet the available information. It is probably best to choose the simplest (smoothest) map among them. Selecting this map should stabilize the solution and should minimize any bias in the map by removing features that are not needed to explain the available data. From the point of view of information theory, this selection should be equivalent to selecting the map with the maximum entropy that satisfies the supplied constraints (Papoulis, 1991; Gilmore, 1996). An important 'side effect' of entropy maximization is the positivity of the resultant electron density.†

We define a cost function that encourages smoothness by maximizing the entropy of the electron-density map. The entropy of a given distribution of electrons, $\{n(p)\}$, is

$$S = -\sum_{p=1}^{P} n(p)\ln[n(p)]. \tag{51}$$

It should be maximized subject to the magnitudes of the calculated structure factors being equal to the measured ones, as expressed *e.g.* in equation (4). Explicitly, the constraints are

$$|F(\mathbf{h})| = \left|\exp[-\eta(\pi\Delta r|\mathcal{F}^T\mathbf{h}|)^2]\sum_{p=1}^{P} n(p)\exp(2\pi i\mathbf{h}\cdot\mathcal{F}\mathbf{r}_p)\right|, \tag{52}$$

where all symbols have been defined in §2. In *EDEN*, we use cost functions that are minimized, therefore we define the maximum-entropy cost function, $f_{\rm maxent}$, to be proportional to $-S$. The usual way to find such a constrained minimum is by adding to (51) the differences between the two sides of the (equality) constraints, (52), multiplied by Lagrange multipliers, $\lambda(\mathbf{h})$, that are themselves unknown:

$$\begin{aligned} f_{\rm maxent} = {} & C\sum_{p=1}^{P} n(p)\ln[n(p)] \\ & + \lambda(\mathbf{h})\left\{\left|\exp[-\eta(\pi\Delta r|\mathcal{F}^T\mathbf{h}|)^2]\right.\right. \\ & \left.\left.\times\sum_{p=1}^{P} n(p)\exp(2\pi\mathbf{h}\cdot\mathcal{F}\mathbf{r}_p)\right| - |F(\mathbf{h})|\right\}, \end{aligned} \tag{53}$$

where C is an appropriate constant. Then the unconstrained minimum of $f_{\rm maxent}$ and the values of the Lagrange multipliers, $\lambda(\mathbf{h})$, are found so that the constraint equations are satisfied. This results in a set of electrons per voxel, $\{n_{\rm maxent}(p)\}$, that corresponds to the maximum-entropy density, $q^{\rm ME}(r)$, in equation (ME1) of Bricogne (1984), or equation (6) of Prince (1993), or equation (3.10) on p. 384 of Carter & Sweet (1997).

Let us suppose that we are able to solve the maximum-entropy equations.‡ A fundamental result of constrained optimization theory is that the values of the Lagrange multipliers exactly compensate for the gradient of the entropy function at the point of the constrained minimum of $f_{\rm maxent}$. [For exact conditions see *e.g.* Luenberger (1984).] For small deviations from the maximum, we can therefore approximate the entropy term, (51), using only the second-order term in the expansion of the logarithm around $\{n_{\rm maxent}(p)\}$,

$$f_{\rm maxent} \simeq \tfrac{1}{2}C\sum_{p=1}^{P}[n(p) - n_{\rm maxent}(p)]^2/n_{\rm maxent}(p) \quad \text{(entropy term).} \tag{54}$$

We see that, apart from numerical factors and the denominator $n_{\rm maxent}(p)$, this cost function is very similar to the one in equation (6). The second-order expansion of the structure-factor part of equation (52) gives terms that are similar to the 'basic' cost function (5). Displaying those terms gives formulas similar to those of Collins (1982). We can obviously extend (54) to satisfy more constraints by defining a cost function that is even more similar to equation (6),

$$f_{\rm maxent} = \tfrac{1}{2}\lambda_{\rm maxent}\,P\sum_{p=1}^{P}\tilde{w}(p)^2[n(p) - n_{\rm maxent}(p)]^2, \tag{55}$$

where the weights should default to $\tilde{w}(p)^2 = \langle n\rangle/n_{\rm maxent}(p)$, where $\langle n\rangle = N/P$, the average number of electrons per voxel. If there are no external constraints, $n_{\rm maxent}(p) = \langle n\rangle$. We then get

$$f_{\rm maxent} = \tfrac{1}{2}\lambda_{\rm maxent}\,P\sum_{p=1}^{P}\tilde{w}(p)^2[n(p) - \langle n\rangle]^2. \tag{56}$$

It is clear that, by assigning different weights for different parts of the unit cell, we can selectively maximize the entropy of some unknown region while ignoring other regions whose electron densities are presumably known.

8. Discussion and conclusions

In this paper, we have shown some ways to incorporate statistical knowledge into the holographic method. We started in §3 with a rather lengthy discussion of the

† In the holographic method non-negative electron densities are always assured. Therefore this side effect is not important.

‡ It is not obvious that solving those equations is inherently easier than solving the crystal structure. In fact, when the structure is solvable by direct methods, the maximum-entropy density is the solution of the crystal structure. The advantage of the maximum-entropy equations is that, given the set of constraints, they always have a unique and stable solution.

statistical distribution of the estimated structure factors. Measurements of integrated intensities of Bragg reflections have inherent errors that lead to uncertainties of the estimates of their true values. The goal of our derivation was to reduce the errors as much as possible by introducing additional information. First, the structure factors are known *a priori* to satisfy Wilson statistics. Second, the position as well as the shape of the Bragg reflections can be predicted within some accuracy. Third, the X-ray background intensity can also be estimated fairly accurately. The pivotal point of our derivation is that, if the maximum amount of statistical information (and no more) is incorporated into Bayes's law, the distribution of the estimated structure-factor amplitudes is correctly described by their posterior probability distribution. As such, it is trivially optimal. In particular, 'profile-fitted' intensities should always give better estimates of the 'true' intensities and smaller error bars than 'data integration with background subtraction'. Our discussion was based on the work of Oatley & French (1982), but went beyond it. We tried to give a comprehensive enough discussion that current data-processing programs could benefit from it.

Our computer program, *EDEN*, is able to use (in principle) the full statistical distribution of the estimated amplitudes of the structure factors. The most straightforward use of this distribution is to demand that the program minimize the expected value of the squared differences of the calculated and the 'true' structure-factor amplitudes. For a Gaussian distribution of errors, this led us to a $1/\sigma^2$ weighting scheme for the reciprocal-space cost function (34) and to a stopping criterion, (36), that prevents overfitting. These are familiar and thus very satisfactory results. Several remarks are in order. First, the $1/\sigma^2$ weighting scheme is usually derived by maximizing the log-likelihood estimate and not by minimizing the expected value of the cost function. Second, minimizing the expectation value of the cost function and the subsequent weighting scheme is an equally valid procedure for a more complicated probability distribution, *e.g.* equation (27) or (29). Third, our derivation is valid for a general 'omit map'. In usual practice, analogous weights (Sim weights) are used to multiply the magnitudes of the measured structure factors (see paper V). Therefore, the calculated structure factors cannot agree with the 'true' ones even in the best case, while in our procedure they do. Therefore, the maps of *EDEN* are intrinsically more accurate. In computer programs for crystallographic refinement, weights are used correctly. Our formula (34) is similar to that used in 'least-squares' refinement and maximum-likelihood refinement is similar to our formulas in §6. Fourth, the use of the stopping criterion should prevent overfitting the map just as cross validation or free R factor (Brünger, 1992) does. Fifth, using the $1/\sigma^2$ weighting scheme for the native and each derivative in MIR and for each wavelength in MAD automatically gives the correct distribution of phase errors in MIR and MAD maps. Therefore, it is equivalent to figures of merit and to more elaborate descriptions of phase errors. (Note that the latter are consequences of exactly the same weighting scheme for the individual reflection amplitudes.)

In §5, we dealt with two categories of missing data. The *a priori* knowledge that was brought to bear on this problem was just the Wilson distribution, but the formulas are equally valid for more elaborate distributions. In §6, we went one step further: we used the statistical distribution of the structure factors of the unknown part of the molecule to derive the probability distribution of the relative phases of the known and the unknown part, given the measured structure factors for the whole molecule. Our results are essentially equivalent to those of Bricogne (1984) and Read (1986). The consequence of this statistical knowledge is that the cost function, (5) or (34), ceases to be degenerate on the circle of Fig. 1. One additional step, briefly discussed in §6.2, is that we can easily restrict the spatial extent of the unknown part of the structure. This can be important *e.g.* for molecular replacement.

Our final comments are about the relation of the holographic method to maximum-entropy methods used for recovering the electron-density distribution in crystals. Naturally, the most important information for solving the crystal structure comes from available data: diffraction amplitudes, MIR, MAD, NCS, the location and density of known parts of the molecule and of solvent regions. The alert reader might have noticed that, in addition, some of the 'side effects' of the maximum-entropy method were already incorporated into the holographic method. More specifically, we always ensure the positivity of the electron density. We are also able to use available statistical information to fill in missing data and we are able to use the statistical distribution of structure-factor amplitudes as well as the presumed location of the missing electron density. Usually, the resultant electron density is still not completely determined; in fact, there is often a large range of uncertainty. The maximum-entropy map introduces the least amount of unwarranted information (bias).

We repeat our apologia from paper V. On the theoretical side, we scrupulously distinguish between lack of information and tacitly assumed information. We try to abide by the dictum of Lánczos: use all available information and no more. In principle, given a sufficient amount of information, it is possible to recover the crystal perfectly. However, different algorithms may have very different convergence properties and may have very different sensitivity to imperfections in the data. In our opinion, this last point alone is sufficiently important to justify the development of new methods for crystallographic computations.

APPENDIX *A*

In this Appendix, we discuss some considerations on the processing of experimental data. It is self evident that the solution of a crystal structure can only be as good as the data. In other words, experiments have to be performed well and data processed carefully for best results.

We will concentrate on data taken with area detectors, be they image plates, charge-coupled-device arrays (CCD) or silicon diode arrays. We assume that the following preliminary measurements have already been carried out. First, the detector was illuminated by a uniform X-ray flux for several well measured time intervals. This procedure established the relative sensitivity of each image element (pixel), its linearity and saturation. In particular, 'dead' pixels were noted. In addition, we assume that the actual counting statistics of the detector were established as a function of the total number of detected photons. Probably, it is good enough to assume that all 'live' pixels have the same photocount statistics so this can be extracted from the previous measurement with uniform illumination. If the validity of this assumption is in doubt, the detector can be illuminated by the same total flux many times and the statistics of each pixel measured separately. Next, we assume that the geometrical distortion of the detector and its spatial resolution have been established by illuminating it through a well defined array of small pinholes. In brief, we will assume that we can correct all distortions of the detector and obtain valid photocounts at the correct locations. If the detector is stable, such calibrations need to be performed infrequently.

We now discuss the expected shapes of the Bragg reflections in biological macromolecules. We follow the notation of Jagodzinski & Frey (1992) and of Willis (1992). James (1982) and Cowley (1981) present essentially the same formulas in different notation. We restrict ourselves to the kinematic theory of X-ray diffraction, *i.e.* to the first Born approximation. This is the usual approximation used in macromolecular crystallography. It should be augmented by absorption corrections and corrected for primary and secondary extinction for the strongest reflections. We define an invariant reflection profile to be the intensity distribution of the Bragg spots reflected from a crystal described as a function of the (continuous) reciprocal-lattice vector, **H**. As we have not found a clear and unequivocal definition of this term in the literature (see *e.g.* Kabsch, 1988), we attempt to define it in a way that is independent of the experimental arrangement and that will be useful for our purposes.

We start with the general expression for X-ray scattering of a crystal in the first Born approximation. Both the source of the incident X-ray beam and the detector are assumed to be at a very large distance compared with the crystal size. The cross section for the scattered X-ray intensity per unit solid angle per unit frequency interval (the so-called doubly differential cross section) for the whole crystal is given by Jagodzinski & Frey (1992), from the general expression of linear response theory

$$\mathrm{d}^2\sigma_{\mathrm{coh}}/\mathrm{d}\Omega\,\mathrm{d}(h\nu) = (|\mathbf{k}|/|\mathbf{k}_o|)r_c^2 S_{\mathrm{coh}}(\mathbf{H}, \nu), \tag{57}$$

where $\mathbf{k}_o$ and $\mathbf{k}$ are the incident and scattered wave vectors of the X-ray beam. The magnitudes of the wave vectors are $|\mathbf{k}_o| \simeq |\mathbf{k}| \simeq 2\pi/\lambda$ and $\mathbf{H} = (\mathbf{k} - \mathbf{k}_o)/2\pi$ is the momentum transfer vector. The space–time Fourier transform of the van Hove autocorrelation function, $S_{\mathrm{coh}}(\mathbf{H}, n\nu)$, is defined below. (We write capital **H** in order to emphasize that it is a continuous variable, as opposed to **h**, which will be restricted to integer values of the reciprocal lattice.) The scattering length for electrons is their classical radius

$$r_c = e^2/mc^2. \tag{58}$$

In notation familiar to crystallographers, $S_{\mathrm{coh}}(\mathbf{H}, n\nu)$ is the Fourier transform of the generalized (space- and time-dependent) electron-density autocorrelation function, or Patterson function.

$$S_{\mathrm{coh}}(\mathbf{H}, n\nu) = (1/2\pi) \int\!\!\int_{\mathbf{r}\,t} P(\mathbf{r}, t) \exp\{2\pi i(\mathbf{H}\cdot\mathbf{r} - \nu t)\}\,\mathrm{d}\mathbf{r}\,\mathrm{d}t, \tag{59}$$

where

$$P(\mathbf{r}, t) = 1/|t_o| \int_{\mathbf{r}'} \int_{t'=-t_o}^{0} \rho(\mathbf{r}', t')\rho(\mathbf{r}' + \mathbf{r}, t' + t)\,\mathrm{d}\mathbf{r}'\,\mathrm{d}t', \tag{60}$$

and $\rho(\mathbf{r}', t')$ is the time-dependent electron density in the crystal in units of electrons per unit volume. The limits of integration over t' ensure that the crystal was actually illuminated by the X-ray beam (which was turned on at $-t_o$) and that causality is obeyed. The $\mathbf{r}'$ integration is over the illuminated (and observed) volume of the crystal; in fact both $\mathbf{r}'$ and $\mathbf{r}' + \mathbf{r}$ have to be illuminated. The electron densities (multiplied by r_c) are closely related to the atomic scattering factors, f; in fact, the usual X-ray structure factors are defined as Fourier transforms of either of them.

Usually, the length of the X-ray exposure is much longer than the correlation time of internal motions in the crystal. In such cases, we can define a time-averaged density autocorrelation function as the limit for large T of the expression

$$\langle P(\mathbf{r}, t)\rangle = 1/T \int_{\mathbf{r}'} \int_{t'=-T}^{0} \rho(\mathbf{r}', t')\rho(\mathbf{r}' + \mathbf{r}, t' + t)\,\mathrm{d}\mathbf{r}'\,\mathrm{d}t'. \tag{61}$$

The resulting formula is

$$S_{\rm coh}(\mathbf{H}, \nu) = 1/2\pi \int\int_{\mathbf{r}\,t} \langle P(\mathbf{r}, t)\rangle \exp\{2\pi i(\mathbf{H}\cdot\mathbf{r} - \nu t)\}\,\mathrm{d}\mathbf{r}\,\mathrm{d}t, \tag{62}$$

where the $\mathbf{r}$ integration is again over the illuminated (and observed) volume of the crystal.

The frequency width (spectral width) of the scattered X-rays is comparable to the inverse of the time scales of the internal motions of the atoms in the crystal. X-ray detectors used in crystallography cannot resolve the spectrum but they measure the integrated intensity. Accordingly, we define a frequency integrated cross section

$$\mathrm{d}^2\sigma_{\rm coh}/\mathrm{d}\Omega = (|\mathbf{k}|/|\mathbf{k}_o|)r_c^2 \int S_{\rm coh}(\mathbf{H}, \nu)\,\mathrm{d}\nu. \tag{63}$$

The frequency integration can be carried out in equation (59) or (62), giving

$$\int S_{\rm coh}(\mathbf{H}, n\nu)\,\mathrm{d}\nu = 1/2\pi \int_{\mathbf{r}} \langle P(\mathbf{r}, 0)\rangle \exp\{2\pi i\mathbf{H}\cdot\mathbf{r}\}\,\mathrm{d}\mathbf{r}. \tag{64}$$

For completeness, we repeat the definition of $\langle P(\mathbf{r}, 0)\rangle$

$$\langle P(\mathbf{r}, 0)\rangle = 1/T \int_{\mathbf{r}'} \int_{t'=-T}^{0} \rho(\mathbf{r}', t')\rho(\mathbf{r}' + \mathbf{r}, t')\,\mathrm{d}\mathbf{r}'\,\mathrm{d}t'. \tag{65}$$

Thus, it can be seen that the directional distribution of the frequency-integrated X-ray scattering intensity is related to the space–time average of the instantaneous autocorrelation function of the electron density over the whole crystal.

The first general conclusion that can be drawn from this formula is that the shape of the X-ray scattering pattern, including Bragg peaks and background, depends only on the momentum-transfer vector $\mathbf{H}$. This conclusion assumes that the illuminated volume of the crystal does not depend on the crystal orientation and that the kinematic approximation is valid.

The electron density of crystals has a repetitive pattern over some distance but real crystals have various degrees of imperfection. The exact shapes of the diffraction spots (in reciprocal space) depend on the exact nature of the disorder in the crystal. From equations (63), (64) and (65), we see that only the time-averaged part of the crystal density contributes to the Bragg spots.

In order to make some progress, we now discuss some simple models of imperfect crystals. The reflections from a small but perfect crystal of $N_1 \times N_2 \times N_3$ unit cells has a width of $1/N_1 \times 1/N_2 \times 1/N_3$ along the reciprocal-lattice vectors. This width is independent of the indices h, k, l of the centers of the reflections. From equation (65), we see that this statement is approximately true even when the crystalline order has a correlation function that is not a 'box function' as above. For example, if the imperfect crystal is a collection of incoherently scattering crystallites, all having exactly the same direction and lattice spacing, the shapes of the reflections are given by the convolution of the Fourier transform of the autocorrelation function of the statistical ensemble of the crystallites with the reciprocal-lattice of each crystal. Therefore, the shape of the reflections for such a collection of incoherently scattering crystallites can be described by a broadening of the integral points of the reciprocal lattice by $\sim 1/N_1 \times 1/N_2 \times 1/N_3$ for the average crystallite along the reciprocal-lattice vectors. Now we can generalize to a collection of slightly misoriented crystallites (the so-called mosaic crystal.) We can describe the statistics of misorientation by some probability distribution as a function of Euler angles by which the crystallites are rotated. For small rotations, such misorientation can be described as small rotations around three orthogonal axes. As the reciprocal lattice of each crystal is rigidly tied to its real lattice, the mosaic crystal can be described as an ensemble of slightly rotated reciprocal lattices whose points are broadened by the correlation function, as described above. The resulting qualitative picture is a reciprocal lattice at the 'mean' crystal position, whose points are broadened by a convolution of two terms: a constant term and another one that has a pancake shape along a spherical shell with a given angular distribution, therefore with a width in reciprocal space that is proportional to the magnitude of the reciprocal-lattice vector. Such a broadening will eventually make the reflections overlap at high resolution. In addition, the incident X-ray beam can be described by its angular distribution and by its spectral distribution. These can also be translated into a three-dimensional distribution of the incident $\mathbf{k}_o$. Yet another source of broadening is the finite resolution of the detector. A discussion along these lines was given by Roth (1986), who gave explicit formulae for Gaussian distributions in all the above variables. A recent more detailed calculation along similar lines was presented by Bolotovsky & Coppens (1997).

In usual practice, the shapes of the diffraction peaks, as they appear on the detector, are deduced empirically from the observed shapes of a number of strong reflections. Most programs assume that the shapes of the diffraction spots vary slowly as a function of their position on the detector and as a function of crystal rotation. They 'learn' the actual shapes of the strong reflections in some areas (neighborhoods) of the detector and then use interpolation to determine the expected shapes of other reflections. We propose, along the lines of Roth (1986), that an 'invariant' three-dimensional profile of the reflections in reciprocal space be established, characterized by a small number of parameters and used for all the reflections. We suggest the following procedure. First, the parameters of the incident beam have to be established, *i.e.* one should measure its angular distribution and its spectral distribution. This can be performed as part of the calibration of the X-ray source. Second, one should record a few

frames of diffraction patterns of the molecule being studied, using the intended method, *e.g.* small-angle oscillation. After indexing the diffraction pattern, one should take a few static diffraction images with very small rotations between the pictures. The profiles of selected strong reflections can be learned by back-transforming their shapes into reciprocal-space coordinates. This can help to parametrize the 'invariant' shapes of the reflections in reciprocal space. The strong reflections whose shapes are taken should sample the reciprocal space relatively uniformly. After this procedure, one should go ahead and take all the diffraction images of the crystal and refine the parameters of the shapes of the diffraction spots. From these patterns, one should try to find the shapes' slowly varying component – in other words, how many parameters are needed to describe the whole pattern.

We add a few words on diffuse scattering. From the data-processing point of view, diffuse scattering has a component that is centered on the incident beam (the short-range-order part) and a component around each diffraction spot. Thermal motions of atoms attenuate the integrated intensity of the Bragg spots by the factor $\exp(-B|\mathcal{F}^T\mathbf{h}|^2/4)$. The missing intensity is transferred to the vicinity of the Bragg reflection as diffuse scattering (Willis, 1992). The intensity distributions of the above diffuse intensity peaks around the Bragg spots (for an ideal crystal) are proportional to $1/|\mathbf{q}|^2$, where $\mathbf{q} = \mathbf{H} - \mathbf{h}$ is the displacement from the exact Bragg condition in reciprocal space. In real crystals, the distribution of crystal orientations smears out this distribution of the diffuse spots. It is important to notice that the integrated diffuse intensity is proportional to the Bragg intensity itself and its magnitude increases with resolution. Therefore, if a well defined (and resolution-independent) fraction of the diffuse scattering is mistakenly identified as belonging to the Bragg peak, the net result of this error is that the thermal motion of all atoms is mistakenly reduced by the same factor but the relative magnitude of the thermal motion (B value) for each individual atom is not changed. Therefore, it is not too critical to put all the thermal scattering into the background as long as we cut a constant fraction of it.

We propose the following procedure to find the background as a function of diffraction angle and to establish the shapes of the reflections in reciprocal space. Let us define the coordinates of the detector as pixel space. This is the space in which data were recorded. After the required geometrical corrections, the exact position and (lack of) flatness of the detector have been established, we are able to transform from pixel coordinates to angle coordinates at the crystal. All data-processing programs perform this transformation one way or another.

A single exposure of the diffraction pattern yields a set of 'photocounts', $I_o(\mathbf{x})$, on the detector as a function of pixel coordinates, $\mathbf{x}$. This set of counts has first to be corrected for the sensitivity of each pixel as established by the calibration procedure, yielding a corrected image, $I_c(\mathbf{x})$. Next, the corrected image has to be 'filled in' for missing data. By missing data, we mean dead pixels and the image of the beam stop. Dead pixels are flagged during calibration and the image of the beam stop can be established by looking for a large region of very low counts surrounded by high counts. These regions should also be flagged for future use and they should be filled in using a low-order spline interpolation in two dimensions. Also, the detector frame should be surrounded by a larger frame (of appropriate magnitude) and filled in so that the boundaries of the outer frame satisfy periodic boundary conditions with continuous derivatives. We denote the filled-in (augmented) picture by $I_a(\mathbf{x})$. The image of the beam stop, the dead pixels and the coordinates of the larger frame change very little from exposure to exposure; therefore they can be learned and used later (with slight corrections if needed). We will call them by the generic name dead pixels.

As we intend to Fourier transform $I_a(\mathbf{x})$, the 'filling in' was performed in order to eliminate spurious oscillations and aliasing in the Fourier transform due to sudden jumps in the data and due to nonperiodic boundaries. As discussed above, the $1/|\mathbf{q}|^2$ dependence of thermal diffuse scattering gives relatively high frequency components, which may present a problem how to decompose $I_a(\nu)$ correctly, but this problem is mitigated in real crystals with sizable mosaic spread.

Now we are ready to find an 'optimum' background. Our basic premise is that the two-dimensional 'augmented' count distribution $I_a(\mathbf{x})$ is a sum of a relatively smooth background and of prominent peaks. Let us first Fourier transform $I_a(\mathbf{x})$ in two dimensions, giving $I_a(\nu) = \mathrm{DFT}[I_a(\mathbf{x})]$. The expected shape of the magni-

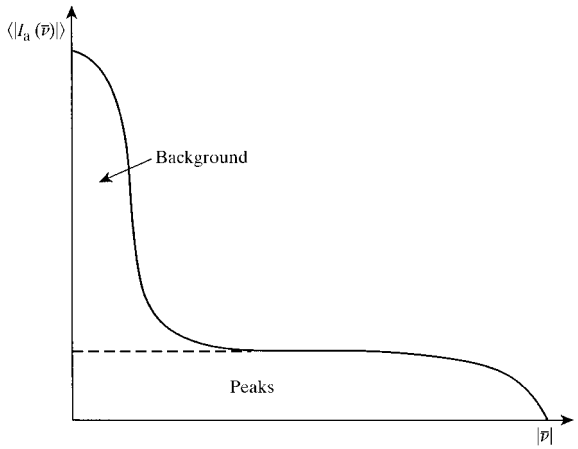

Fig. 2. $\langle|I_a(\nu)|\rangle$ is the Fourier transform of the X-ray diffraction image, in pixel coordinates, averaged within resolution shells, plotted *vs* the resolution $|\nu|$. Its decomposition into contributions from Bragg peaks and background is discussed in Appendix *A*.

tude of $I_a(\nu)$ is shown in a schematic way in Fig. 2, where we plot $\langle|I_a(\nu)|\rangle$ averaged over a resolution shell *vs* $|\nu|$. We will now disentangle the spectrum into its two components, $I_a(\nu) = I_p(\nu) + I_b(\nu)$, where $I_p(\nu)$ and $I_b(\nu)$ are the Fourier transforms of the peaks and the background, respectively, as indicated in Fig. 2. This is a nonlinear problem because $I_a(\nu)$, $I_p(\nu)$ and $I_b(\nu)$ are complex. The decomposition can be solved iteratively. We divide $\langle|I_a(\nu)|\rangle$ into two fractions as outlined in Fig. 2 by the broken line. We now divide each $|I_a(\nu)|$ itself into the same two fractions as $\langle|I_a(\nu)|\rangle$ was divided into, in order to get $|I_p(\nu)|$ and $|I_b(\nu)|$ and we set the phases of $I_p(\nu)$ and $I_b(\nu)$ to be those of $I_a(\nu)$. Let us call these components $I_p^{(1)}(\nu)$ and $I_b^{(1)}(\nu)$, respectively. Back transforming them into pixel space gives $I_p^{(1)}(\mathbf{x})$ and $I_b^{(1)}(\mathbf{x})$ (both real quantities). Let us subtract the background from the peaks, $I_{\rm net}^{(1)}(\mathbf{x}) = I_p^{(1)}(\mathbf{x}) - I_b^{(1)}(\mathbf{x})$. If there were no phase error in our procedure and no noise in the measurement, the peaks would appear essentially with no background. Because of both errors, the space between the peaks is noisy. Next we use 'soft thresholding', *i.e.* we set all the values of $I_{\rm net}^{(1)}(\mathbf{x}) < I_{\rm thresh}$ to zero and leave all values of $I_{\rm net}^{(1)}(\mathbf{x}) \geq I_{\rm thresh}$ unchanged. Let us call the result $I_p^{(2)}(\mathbf{x})$ and its Fourier transform $I_p^{(2)}(\nu)$. We can form the second approximation of the background by forming the complex difference $I_b^{(2)}(\nu) = I_a(\nu) - I_p^{(2)}(\nu)$. As this contains all the high-frequency noise of the background counts, it should be low-pass filtered by multiplying it with an appropriate filter, $\mathrm{Lp}(\nu)$. The accuracy of the filled-in (augmented) image can be improved by replacing the dead pixel values by $I_b^{(2)}(\mathbf{x})$. Finally, at this point we can iterate in an obvious way. We expect that the result will converge in very few iterations.

For purposes of data compression, the differences $D(\mathbf{x}) = I_c(\mathbf{x}) - I_b(\mathbf{x})$ can be used, together with $I_b(\nu)$. There are relatively few components in the Fourier-space part and we expect that $|D(\mathbf{x})| \ll I_c(\mathbf{x})$ for the live pixels. Note that by our manipulations so far we have not lost any data. One can also refine the procedure by finding the beam-stop pixel coordinates as those with large negative values of $D(\mathbf{x})$ then setting the values of all the 'dead' pixels to be $I_b(\mathbf{x})$ and iterating.

It should be stressed that this procedure provides a background that is not constrained by preconceptions about its shape. It is optimal in the sense that it is the best smooth surface. We retain $I_b(\mathbf{x})$ from the procedure as our estimated background and we can use the converged values of the peaks, $I_p(\mathbf{x})$, in order to estimate the parameters of the peaks in reciprocal space as discussed above. The standard error of the background can be estimated in the following way. We can determine the variance of the noise in the background region, V_b, by calculating the mean deviation of the actual values of $I_c(\mathbf{x})$ in that region from the smoothed background, $I_b(\mathbf{x})$. The estimated value of the variance should not be significantly different from that obtained during the calibration of the detector. (This can be used to check the consistency of the procedure and as a check of the health of the experiment.) If the average bandwidth of $I_b(\nu)$ is $1/\Delta$ pixels, the standard error of the estimation of the background is V_b/Δ^2. The background intensity distribution should be analyzed for the penumbra of the beam stop and Bragg reflection intensities affected by it should be corrected.

As a final note, we would like to point out some additional needs and possibilities. We proposed finding the background and the shapes of the reflections using pixel coordinates. This is surely the easiest system to use for computations. Eventually, we have to transform out the distortions of the detector using our calibration. We expect very little distortion when image plates or silicon diode arrays are used but in optical-fiber-coupled CCD arrays the distortion can be considerable. In the latter case, if needed, the transformation to undistorted coordinates can be performed before finding the shapes of the background and the Bragg spots. By using methods of wavelet expansion or fast Fourier transforms of non-equispaced data, the proposed procedure can be carried out to the needed precision in negligible computer time even after the transformation to undistorted coordinates. We did not specify the optimum algorithm for finding the shapes of the Bragg reflections in reciprocal-space either. The procedures are not very different from those practiced in existing data-processing programs. We stressed the need for a consistent parametrization of the reflection shapes. In §3.2, we assumed a Gaussian shape of the reflections in reciprocal space and a Gaussian distribution of its width and of its predicted center. The central point is that not only the best parametrization but also its distribution can be found from experiment.

The purpose of the above paragraphs was to outline one data-processing procedure that is consistent with modern statistical practice. We are aware of many good data-processing programs that do some parts of the procedure very well and we do not claim that this is the only method available.

APPENDIX *B*

Some basic statistical notation

We use the standard axiomatic definition of probabilities, as appear *e.g.* in Goodman (1985) or Papoulis (1991). The probability of an event A will be denoted by $P(A)$. The probability distribution function, $F_A(A_o) = P(A \leq A_o)$, expresses the probability that the random variable A has a value $\leq A_o$. $F_A(A_o)$ has to be non-decreasing and it has to have the limits 0 for $A_o \rightarrow -\infty$ and 1 for $A_o \rightarrow \infty$. If the probability distribution function has a finite derivative, the probability density $p(A_o)$ can be identified with the probability of A being between A and $A + \mathrm{d}A$, by using the formula

$dF_A(A_o) = P(A_o \le A \le A_o + dA) = p(A_o)dA$. Otherwise, $p(A)$ has to be defined in terms of distributions, meaning that it may also have Dirac-delta-function components. In that case, all integrations below are Lebesgue integrals. The normalization, $\int p(A)\,dA = 1$ follows from the definition of $p(A)$. The joint probability density of two events A *and* B will be denoted by $p(A, B)$. It satisfies $\int p(A, B)\,dA\,dB = 1$. The conditional probability density, A *if* B, is denoted by $p(A|B)$. The two composition laws needed are

$$p(A, B) = p(A|B)p(B) = p(B|A)p(A) \tag{66}$$

$$p(A) = \int p(A, B)\,dB = \int p(A|B)p(B)\,dB. \tag{67}$$

Equation (67) defines what is called marginal probability. From the second equality in equation (66), one can deduce Bayes's rule,

$$p(A|B) = p(B|A)p(A)/p(B). \tag{68}$$

If A and B are legitimate random events, there is no difficulty in interpreting Bayes's rule. In our use, we will have to be more cautious with the interpretation of Bayes's rule because we apply it to cases where A refers to the likelihood of a hypothesis that a variable has the value A.† The most satisfactory interpretation the author knows is that $p(A|B)$ describes the odds for betting on the outcome of a hypothetical experiment on A if the value B has already been measured. Note that the value of A is given by nature and it does not necessarily have a probability distribution.

The expected value of a function of the random variable is its integral over its probability distribution

$$E(g(A)) = \int g(A)p(A)\,dA. \tag{69}$$

APPENDIX *C*

*C*1. *The Wilson distribution*

For completeness, we list the formulas for the simplest statistical distribution, called the Wilson distribution, of the magnitudes of the structure factors in large molecules. For our purposes, we need only prior distributions: those correspond to the atoms being randomly distributed in the asymmetric unit. Those were called fixed-index distributions by Shmueli & Wilson (1992). More detailed and elaborate distributions are found in Giacovazzo (1992), French & Wilson (1978), Sivia & David (1994), Castlenden & Fortier (1994, and references therein), Shmueli & Weiss (1994, and references therein) and Bricogne (1997). Although the Wilson distribution is usually given in terms of $|F|$, we will need it also in terms of $N_R = |F|^2$. Using the basic definition of the probability density in Appendix B,

$$dP = p(|F(\mathbf{h})|)\,d|F(\mathbf{h})| = p(N_R)\,dN_R. \tag{70}$$

† We do not use the likelihood function $\Lambda(A|B) = p(B|A)$ explicitly.

The specific Wilson distributions for centric and acentric reflections are

$$\begin{aligned} dP_c &= [2/(\pi\Sigma)^{1/2}]\exp[-|F(\mathbf{h})|^2/2\Sigma]\,d|F(\mathbf{h})| \\ &= [1/(2\pi\Sigma N_R)^{1/2}]\exp(-N_R/2\Sigma)\,dN_R, \\ &\qquad 0 \le N_R, \Sigma \quad \text{(centric)} \end{aligned} \tag{71c}$$

$$\begin{aligned} dP_a &= [2|F(\mathbf{h})|/\Sigma]\exp[-|F(\mathbf{h})|^2/\Sigma]\,d|F(\mathbf{h})| \\ &= (1/\Sigma)\exp(-N_R/\Sigma)\,dN_R, \\ &\qquad 0 \le N_R, \Sigma \quad \text{(acentric)}. \end{aligned} \tag{71a}$$

If the average magnitudes of the structure factors within a resolution shell are put onto an absolute scale, $\langle|F(\mathbf{h})|^2\rangle$ is well approximated at high resolution (> 3 Å) by

$$\Sigma = \langle|F(\mathbf{h})|^2\rangle = \exp[-B(|\mathcal{F}^T\mathbf{h}|)^2/4]\sum_n f_n^2, \tag{72}$$

where f_n^2 are the squares of the scattering factors of the atoms in the molecule and B characterizes the average thermal motion of the atoms. Note that Σ depends on the resolution shell, $1/d^2 = |\mathcal{F}^T\mathbf{h}|^2$, and it can be estimated from the number and kind of atoms and their average thermal motion. At low resolution, where the solvent contributes significantly, there are large deviations from (72). Nevertheless, within each resolution shell, we will assume a Wilson distribution of the structure-factor amplitudes and use the actual average intensity within that shell, $\langle N_{\rm ave}\rangle$, in the relevant formula. The latter has to be estimated from the measured reflections themselves. Such an estimation may have to be performed iteratively. French & Wilson (1978) investigated the validity of this assumption and found it to be satisfied. They have also discussed a successive approximation procedure to find the values of $\langle N_{\rm ave}\rangle$ in each resolution shell. Recently, Bricogne (1997) has examined the applicability of the above assumptions.

*C*2. *Calculation of the integrals in §3*

We will compute the conditional probabilities, $p(N|N_B)$, for centric and acentric reflections. From the combination rule (67),

$$p(N|N_B) = \int_0^\infty p(N|N_R, N_B)p(N_R)\,dN_R. \tag{73}$$

For $p(N_R)$, we use the Wilson distribution, (71*c*) and (71*a*), respectively, and, for $p(N|N_R, N_B)$, we use the Poisson distribution (8). The formulas are

$$p_c(N|N_B) = \int_0^\infty \mathrm{d}N_R\,[(N_R+N_B)^N/N!]\exp[-(N_R+N_B)] \times [1/(2\pi\langle N_{\rm ave}\rangle N_R)^{1/2}]\exp(-N_R/2\langle N_{\rm ave}\rangle) \quad \text{(centric)} \qquad (74c)$$

$$p_a(N|N_B) = \int_0^\infty \mathrm{d}N_R\,[(N_R+N_B)^N/N!]\exp[-(N_R+N_B)] \times (1/\langle N_{\rm ave}\rangle)\exp(-N_R/\langle N_{\rm ave}\rangle) \quad \text{(acentric)}. \qquad (74a)$$

The integrals can be performed analytically for integer N, by using the binomial expansion for $(N_R+N_B)^N$ (Kittel & Kroemer, 1980), then using formulas 3.351-3 and 3.371 from Gradshteyn & Rizhik (1980). The results are, using $\alpha = 1 + 1/(2\langle N_{\rm ave}\rangle)$,

$$\begin{aligned} p_c(N|N_B) &= [\exp(-N_B)/(2\langle N_{\rm ave}\rangle)^{1/2}\alpha^{N+1/2}] \\ &\quad \times \sum_{n=0}^{N}[(2n-1)!!/(2n)!!][(\alpha N_B)^{N-n}/(N-n)!] \\ &\simeq [(2\langle N_{\rm ave}\rangle)^N/(1+2\langle N_{\rm ave}\rangle)^{N+1/2}] \\ &\quad \times \exp(N_B/2\langle N_{\rm ave}\rangle), \end{aligned} \qquad (75c)$$

where $(2n-1)!! = 1\times 3\times\ldots\times(2n-1)$ and, using $\beta = 1+1/\langle N_{\rm ave}\rangle$, we get

$$\begin{aligned} p_a(N|N_B) &= [\exp(-N_B)/\langle N_{\rm ave}\rangle\beta^{N+1}]\sum_{n=0}^{N}(\beta N_B)^k/k! \\ &\simeq [(\langle N_{\rm ave}\rangle)^N/(1+\langle N_{\rm ave}\rangle)^{N+1}]\exp(N_B/\langle N_{\rm ave}\rangle). \end{aligned} \qquad (75a)$$

In the profile-fitting case, the integrals can also be performed by a close analogy to the previous ones. The results are

$$\begin{aligned} &p_c(\{N(i)\}\big|\{N_B(i)\}) \\ &\quad \simeq \prod_{i=1}^{k}[(2\langle N_{\rm ave}\rangle)^{N(i)}/(1+2\langle N_{\rm ave}\rangle)^{N(i)+1/2}] \\ &\qquad \times \exp[N_B(i)/2\langle N_{\rm ave}\rangle], \end{aligned} \qquad (76c)$$

$$\begin{aligned} &p_a(\{N(i)\}\big|\{N_B(i)\}) \\ &\quad \simeq \prod_{i=1}^{k}[(\langle N_{\rm ave}\rangle)^{N(i)}/(1+\langle N_{\rm ave}\rangle)^{N(i)+1}] \\ &\qquad \times \exp[N_B(i)/\langle N_{\rm ave}\rangle]. \end{aligned} \qquad (76a)$$

*C*3. *Calculation of the integrals in §6*

For completeness, in this Appendix we present the detailed calculation of the integrals in (48), (47). Although similar calculations have been presented many times (Goodman, 1985; Srinivasan & Ramachandran, 1965), the calculations are somewhat subtle. We will use the intermediary of the probability distribution function. Let us define

$$P(|F|\le F_o\big|R,|O|) = \int_0^{2\pi} P(|F|\le|F_o|\big|R,|O|,\varphi)p(\varphi)\,\mathrm{d}\varphi. \qquad (77)$$

The structure factor of the whole molecule is the sum of that for the known part and the unknown part, $F = R + O$, where all three are complex quantities. For their magnitudes, the cosine theorem of trigonometry gives

$$|F|^2 = |R|^2 + |O|^2 + 2|R||O|\cos\varphi, \qquad (78)$$

where φ is the angle between R and O. (It is the external angle of the *ROF* triangle.) We can deduce that $|F|\le|F_o|$ if $|R|^2+|O|^2+2|R||O|\cos\varphi \le |F_o|^2$, or

$$\cos\varphi \le (|F_o|^2 - |R|^2 - |O|^2)/2|R||O|. \qquad (79)$$

From $|\cos\varphi|\le 1$, it follows that $(|R|-|O|)^2 \le |F_o|^2 \le (|R|+|O|)^2$. Thus, we can define the raw probability,

$$P(|F|\le|F_o|\big|R,|O|,\varphi) = \begin{cases} 0, & |F_o|^2 < |R|^2+|O|^2+2|R||O|\cos\varphi \\ 1, & |F_o|^2 \ge |R|^2+|O|^2+2|R||O|\cos\varphi. \end{cases} \qquad (80)$$

At this point, we have to carry out the calculation separately for acentric and centric reflections.

Let us start with the acentric reflections. We will assume that the relative phases of R and O are distributed uniformly, $p(\varphi) = 1/2\pi$. The integral in (77) can be carried out, the result is

$$P(|F|\le|F_o|\big|R,|O|) = \begin{cases} 0, & |F_o|^2 < (|R|-|O|)^2 \\ 1-(1/\pi)\cos^{-1}\{(|F_o|^2-|R|^2-|O|^2)/2|R||O|\}, & \\ \qquad (|R|-|O|)^2 \le |F_o|^2 < (|R|+|O|)^2 & \\ 1, & (|R|+|O|)^2 \le |F_o|^2. \end{cases} \qquad (81)$$

From the fundamental definition of the probability density,

$$p(|F|\big|R,|O|) = \partial P(|F|\le F_o|\big|R,|O|)/\partial|F_o| \quad \text{at} \quad |F_o| = |F|, \qquad (82)$$

We will use $\mathrm{d}(\cos^{-1}x)/\mathrm{d}x = -(1-x^2)^{-1/2}$. Also, from (78), $\cos\varphi = (|F|^2-|R|^2-|O|^2)/2|R||O|$, therefore $\mathrm{d}\cos\varphi/\mathrm{d}|F| = |F|/|R||O|$. Finally, after some algebra,

$$p(|F|\big|R,|O|) = -(2|F|/\pi)\{[|F|^2-(|R|-|O|)^2] \times [(|R|+|O|)^2-|F|^2]\}^{-1/2}. \qquad (83)$$

The denominator of (49) is the integral of (83) over $|O|$. The inequality constraints (45) provide the limits of the integral. Thus, the final formula for (49), under the

assumptions of Wilson statistics and uniform phase distribution of the relative angle of R and O, is

$$p(|O|\big||F|,R) = \{[|F|^2-(|R|-|O|)^2][(|R|+|O|)^2-|F|^2]\}^{-1/2} \times p_a(|O|)\left(\int_{||F|-|R||}^{|F|+|R|}\{[|F|^2-(|R|-|O|)^2] \times [(|R|+|O|)^2-|F|^2]\}^{-1/2}p_a(|O|)\,\mathrm{d}|O|\right)^{-1}. \tag{84}$$

In this expression, we have to substitute from (41) and (42):

$$p_a(|O|) = (2|O|/\Sigma)\exp(-|O|^2/\Sigma), \tag{41}$$

where

$$\Sigma = \sum_n f_n^2. \tag{42}$$

In order to carry out the integral in the denominator, we used the formula (9.6.18) in Abramowitz & Stegun (1972) for the integral representation of the modified Bessel function,

$$I_o(z) = (1/\pi)\int_{-1}^{1}(1-t^2)^{-1/2}\exp(-zt)\,\mathrm{d}t. \tag{85}$$

When the substitution $t = (|O|^2-|R|^2-|F|^2)/2|R||F|$ is made, after some algebra, equation (50*a*) is obtained.

The calculation for the centric reflection is carried out next. The basic relations (77), (79) and (80) are valid for centric reflections too. The angles φ are restricted to 0 and 180°. This makes the probability density $p(\varphi) = 1/2[\delta(\varphi-0)+\delta(\varphi-\pi)]$. The integral in (77) can be carried out again, the result is

$$P(|F|\le|F_o|\big|R,|O|) = \begin{cases} 0, & |F_o|^2 < (|R|-|O|)^2 \\ 1/2, & (|R|-|O|)^2 \le |F_o|^2 < (|R|+|O|)^2 \\ 1, & (|R|+|O|)^2 \le |F_o|^2. \end{cases} \tag{86}$$

The probability density, from (82), is a distribution, it is

$$p(|F|\big|R,|O|) = 1/2[\delta(|F|-||R|-|O||) + \delta(|F|-||R|+|O||)]. \tag{87}$$

The integral in the denominator of (49) can be carried out explicitly for the probability density from (40):

$$p_c(|O|) = (2/\pi\Sigma)^{1/2}\exp(-|O|^2/2\Sigma), \tag{40}$$

giving the result

$$p(|O|\big||F|,R) = \{[\delta(|F|-||R|-|O||) + \delta(|F|-||R|+|O||)]\exp(-|O|^2/2\Sigma)\} \times\{\exp[-(|R|-|O|)^2/2\Sigma] + \exp[-(|R|+|O|)^2/2\Sigma]\}^{-1}. \tag{88}$$

The variables can be changed in this expression, giving

$$p(|O|\big||F|,R) = \delta(|O|-||R|-|F||) \times\exp[-(|R|-|F|)^2/2\Sigma] \times\{\exp[-(|R|-|F|)^2/2\Sigma] + \exp[-(|R|+|F|)^2/2\Sigma]\}^{-1} + \{\delta(|O|-||R|+|F||) \times\exp[-(|R|+|F|)^2/2\Sigma]\} \times\{\exp[-(|R|-|F|)^2/2\Sigma] + \exp[-(|R|+|F|)^2/2\Sigma]\}^{-1}. \tag{89}$$

APPENDIX *D*
Inverse σ^2 weighted cost function with MIR/MAD

In this Appendix, we expand the formulas in §4 to include MIR/MAD. We start from equation (30) of paper V:

$$f_{\rm eden} = \tfrac{1}{2}\sum_{m=0}^{M}\lambda_m\sum_{\mathbf{h}}w'_m(\mathbf{h})^2[|R_m(\mathbf{h})+O_m(\mathbf{h})|-|F'_m(\mathbf{h})|]^2, \tag{90}$$

which is the obvious generalization of (5). For a Gaussian distribution of the true intensities, the generalization of (33) is

$$\langle f_{\rm eden}\rangle = \tfrac{1}{2}\sum_{m=0}^{M}\lambda_m\sum_{\mathbf{h}}w'_m(\mathbf{h})^2\{[|R'_m(\mathbf{h})+O_m(\mathbf{h})| - |F'_m(\mathbf{h})|]^2+\sigma_m(\mathbf{h})^2\}. \tag{91}$$

Dividing each term by $\sigma_m(\mathbf{h})^2$ and normalizing the sum to be independent of the absolute value of the errors gives, as in (34),

$$\langle f_{\rm eden}\rangle = \tfrac{1}{2}(1/\mathcal{N})\sum_{m=0}^{M}\lambda_m\sum_{\mathbf{h}}[w'_m(\mathbf{h})^2/\sigma_m(\mathbf{h})^2] \times[|R'_m(\mathbf{h})+O_m(\mathbf{h})|-|F'_m(\mathbf{h})|]^2, \tag{92}$$

where $\mathcal{N}$ is the average weight in the unit cell,

$$\mathcal{N} = \sum_{m=0}^{M}\lambda_m\sum_{\mathbf{h}}[w'_m(\mathbf{h})^2/\sigma_m(\mathbf{h})^2]\Big/\Big(\sum_{m=0}^{M}\lambda_m N_m\Big), \tag{93}$$

and where N_m is the number of measured structure factors in each derivative. The discrepancy principle also determines a stopping criterion for the solver: the average value of $[|R'_m(\mathbf{h})+O_m(\mathbf{h})|-|F'_m(\mathbf{h})|]^2$ should be $\sigma_m(\mathbf{h})^2$. This translates to

$$\langle f_{\rm eden}\rangle_{\rm min} = \tfrac{1}{2}(1/\mathcal{N}) \sum_{m=0}^{M} \lambda_m \sum_{\mathbf{h}} w'_m(\mathbf{h})^2. \qquad (94)$$

I would like to thank John Somoza and Hanna Szöke, the other two members of the *EDEN* team, for help and stimulation. I also thank Lawrence Livermore National Laboratory for providing me with their physical and computational resources, and especially Mark Eckart, Richard More and the Physics Department for their support. Part of this work was performed under the auspices of the US Department of Energy under Contract No. W-7405-ENG-48.

References

Abramowitz, M. & Stegun, I. A. (1972). *Handbook of Mathematical Functions.* New York: Dover.
Bolotovsky, R. & Coppens, P. (1997). *J. Appl. Cryst.* **30**, 65–70.
Bolotovsky, R., White, M. A., Darowsky, A. & Coppens, P. (1995). *J. Appl. Cryst.* **28**, 86–95.
Bragg, W. L. (1950). *Nature (London)*, **166**, 399–400.
Bricogne, G. (1984). *Acta Cryst.* A**40**, 410–445.
Bricogne, G. (1997). *Macromolecular Crystallography*, edited by C. W. Carter Jr & R. M. Sweet, Part B, *Methods in Enzymology*, Vol. 277. San Diego: Academic Press.
Brünger, A. T. (1992). *Nature (London)*, **355**, 472–474.
Carter, C. W. Jr & Sweet, R. M. (1997). Editors, *Macromolecular Crystallography*, Part A, *Methods in Enzymology*, Vol. 276. San Diego: Academic Press.
Castlenden, I. R. & Fortier, S. (1994). *Acta Cryst.* A**50**, 9–17.
CCP4 (1994). The CCP4 Project, Daresbury Laboratory, Warrington WA4 4AD, England.
Collins, D. M. (1982). *Nature (London)*, **298**, 49–51.
Cowley, J. M. (1981). *Diffraction Physics*, 2nd ed. Amsterdam: North Holland.
Dainty, J. C. & Fienup, J. R. (1987). *Image Recovery: Theory and Applications*, edited by H. Stark, ch. 7. Orlando: Academic Press.
Diamond, R. (1969). *Acta Cryst.* A**25**, 43–55.
French, S. & Wilson, K. (1978). *Acta Cryst.* A**34**, 517–525.
Giacovazzo, C. (1992). Editor. *Fundamentals of Crystallography.* International Union of Crystallography/Oxford University Press.
Gilmore, C. J. (1996). *Acta Cryst.* A**52**, 561–589.
Goodman, J. W. (1985). *Statistical Optics.* New York: Wiley.
Gradshteyn, L. S. & Rizhik, I. M. (1980). *Table of Integrals, Series and Products.* Orlando: Academic Press.
Jagodzinski, H. & Frey, F. (1992). *International Tables for Crystallography*, Vol. B, edited by U. Shmueli. Dordrecht: Kluwer Academic Publishers.
James, R. W. (1982). *The Optical Principles of the Diffraction of X-rays.* Reprinted by Ox Bow Press, Woodbridge.
Kabsch, W. (1988). *J. Appl. Cryst.* **21**, 916–924.
Kittel, C. & Kroemer, H. (1980). *Thermal Physics*. New York: Freeman.
Luenberger, D. G. (1984). *Linear and Nonlinear Programming.* Reading, MA: Addison-Wesley.
Maalouf, G. J., Hoch, J. C., Stern, A. S., Szöke, H. & Szöke, A. (1993). *Acta Cryst.* A**49**, 866–871.
Morozov, V. A. (1966). *Sov. Math. Dokl.* **7**, 414–417.
Oatley, S. & French, S. (1982). *Acta Cryst.* A**38**, 537–549.
Otwinowski, Z. (1993). In *Data Collection and Processing*, Proceedings of the CCP4 Study Weekend, Report DL/Sci/R34. SERC Daresbury Laboratory, Warrington, England.
Papoulis, A. (1991). *Probability, Random Variables and Stochastic Processes.* New York: Mc-Graw Hill.
Press, W. H. (1997). In *Unsolved Problems in Astrophysics*, edited by J. N. Bahcall & J. P. Ostriker. Princeton University Press.
Prince, E. (1993). *Acta Cryst.* D**49**, 61–65.
Read, R. J. (1986). *Acta Cryst.* A**42**, 140–149.
Read, R. J. (1990). *Acta Cryst.* A**46**, 900–912.
Rossmann, M. G. (1979). *J. Appl. Cryst.* **12**, 225–238.
Rossmann, M. G., Leslie, A. G., Abdel-Meguid, S. S. & Tsukihara, T. (1979). *J. Appl. Cryst.* **12**, 570–581.
Roth, M. (1986). Proceedings of the EEC Cooperative Workshop on Position Sensitive Detector Software, LURE, Université Paris-Sud, Paris, France.
Shmueli, U. & Weiss, G. H. (1994). *Acta Cryst.* A**50**, 605–608.
Shmueli, U. & Wilson, A. J. C. (1992). In *International Tables for Crystallography*, Vol. B, edited by U. Shmueli. Dordrecht: Kluwer Academic Publishers.
Sivia, D. S. & David, W. I. F. (1994). *Acta Cryst.* A**50**, 703–714.
Somoza, J. R., Szöke, H., Goodman, D. M., Béran, P., Truckses, D., Kim, S.-H. & Szöke, A. (1995). *Acta Cryst.* A**51**, 691–708.
Srinivasan, R. & Ramachandran, G. N. (1965). *Acta Cryst.* **19**, 1008–1014.
Szöke, A. (1993). *Acta Cryst.* A**49**, 853–866.
Szöke, A., Szöke, H. & Somoza, J. R. (1997). *Acta Cryst.* A**53**, 291–313.
Tollin, P., Main, P., Rossmann, M. G., Stroke, G. W. & Restrick, R. C. (1966). *Nature (London)*, **209**, 603–604.
Willis, B. T. M. (1992). In *International Tables for Crystallography*, Vol. B, edited by U. Shmueli. Dordrecht: Kluwer Academic Publishers.