Journal of
Applied Crystallography
ISSN 0021-8898

Received 11 July 2008
Accepted 7 November 2008

Post-processing intensity measurements at favourable dose values

Kay Diederichs[a]* and Michael Junk[b]

[a]Fachbereich Biologie, Universität Konstanz, 78451 Konstanz, Germany, and [b]Fachbereich Mathematik und Statistik, Universität Konstanz, 78451 Konstanz, Germany. Correspondence e-mail: kay.diederichs@uni-konstanz.de

In a macromolecular X-ray experiment, many sets of intensity measurements are collected, with each measurement corresponding to the intensity of a unique reflection at a different X-ray dose. The computational correction of radiation damage, which occurs as a function of dose during the experiment, is a concept suggesting the approximation of each set of measured intensities with a smooth function. The value of the approximating function at a user-defined point common to all unique reflections is then used as an interpolated snapshot of the true intensities at this specific dose. It is shown here that, under realistic assumptions, interpolation with a linear function has the smallest amount of error at or near two well defined points in the dose interval. This result is a special case from a mathematical analysis of polynomial approximations which proves that the points of minimum error in the approximation of a polynomial of order n by a polynomial of order $n-1$ are independent of the function values. Conditions are formulated under which better intensities are obtained from linear interpolation than from the usual averaging of observations.

1. Introduction

The effects of radiation damage on crystals of macromolecules were described by Blake & Phillips (1962) more than 40 years ago. Experience at third-generation synchrotron beamlines has shown that serious radiation damage occurs within 100–500 s of irradiation of crystals with the unattenuated beam, even if they are cryocooled (Burmeister, 2000; Weik *et al.*, 2000; Ravelli & McSweeney, 2000).

Radiation damage results in non-isomorphism between successive sets of data measured from the same crystal. This is caused by specific and nonspecific changes (affecting all atoms) of the structure. Nonspecific changes, producing many small and random changes of electron density and arrangement of atoms, lead to random effects and an overall reduced scattering of the crystal, and thus to data with reduced resolution. Another nonspecific effect is an increase in the mosaicity of the crystal, often due to a variation in cell parameters throughout the crystal, which results in systematic differences between successive data sets from the same crystal. Unit-cell changes lead to different sampling of the molecular transform, which influences intensities more strongly at high resolution.

Specific changes, such as the breakage of disulfide bonds, decarboxylation of acids (Weik *et al.*, 2000) or rupture of covalent bonds to heavier atoms, are confined to particularly susceptible sites of the macromolecule. As the electron density of the crystal changes as a function of dose, intensity measurements from the crystal, which correspond to the Fourier transform of the unit cell sampled at reciprocal lattice points, are dose-dependent. This fact alone may reduce the phasing signal expected from data collected successively at the different wavelengths of a MAD experiment. The reduction of phasing power is even more serious because those sites that provide the phasing signal, the heavy atoms, are worst affected by radiation damage.

Usually, this effect is partly accounted for using a smooth correction function which depends on the scattering angle and the dose. It is important to realize that, even after accounting for average changes of intensities in resolution shells using this correction function, about half of the reflections increase in intensity relative to their ideal value, whereas the other half decrease.

A macromolecular X-ray data set has many observations belonging to a set of unique reflections. The ratio of observations to unique reflections is called 'multiplicity' (sometimes less appropriately called 'redundancy'); this ratio is, in most practical experimental situations, in the range 2–16. In the presence of significant radiation damage, increased multiplicity of measurements, if these are combined to form a weighted average, may in effect lead to the deterioration of the overall data quality instead of its improvement.

Computational correction of radiation damage has been shown to be effective at the level of the raw unmerged intensity data, thus exploiting the multiplicity of observations of the unique reflections during data reduction and scaling (Diederichs *et al.*, 2003). However, it has remained unclear how much multiplicity is required to be able to extrapolate the

intensity of each reflection to zero dose, or to obtain a more accurate interpolated value of the intensity in the dose interval that is covered by the data set.

Rather than adding evidence by the analysis of specific data sets, we have investigated the general properties of the approximation of noisy functions by low-order polynomials. Our analysis reveals that, depending on the order of the approximating polynomial, interpolation may yield more accurate estimates of the underlying function value than mere weighted averaging.

2. Favourable points

In this section, we present a general idea of how to construct suitable locations in the dose interval where the evaluation of the polynomially approximated intensity values leads to small errors. First, our idea is illustrated in the hypothetical case where all values of the measured function are available without error. The approach is then generalized to the more realistic scenario when only finitely many noisy data points are available.

2.1. Basic idea

The starting point of our observation is a rule suggested by the examples presented in Fig. 1: whenever a linear fit to a nonlinear function on [0, 1] is reasonable, the approximation error seems to be minimum close to $x \simeq 0.2$ and $x \simeq 0.8$.

To illustrate the mathematical reason underlying this rule, let us consider the example where the nonlinear function is a general parabola $Q(x) = \alpha + \beta x + \gamma x^2$ with $\gamma \neq 0$. To find the parameters a, b of the linear fit polynomial $P(x) = a + bx$, we minimize the squared summed distance

$$D(a, b) = \int_0^1 [Q(x) - P(x)]^2 \, \mathrm{d}x \tag{1}$$

with respect to a and b. Equating the partial derivatives

$$\begin{aligned} \frac{\partial D}{\partial a} &= -2\int_0^1 Q(x) - P(x) \, \mathrm{d}x, \\ \frac{\partial D}{\partial b} &= -2\int_0^1 [Q(x) - P(x)]x \, \mathrm{d}x \end{aligned} \tag{2}$$

to zero, we obtain two linear conditions on the optimum parameters $\overline{a}$, $\overline{b}$:

$$\int_0^1 \overline{a} + \overline{b}x \, \mathrm{d}x = \int_0^1 Q(x) \, \mathrm{d}x, \quad \int_0^1 \overline{a}x + \overline{b}x^2 \, \mathrm{d}x = \int_0^1 xQ(x) \, \mathrm{d}x, \tag{3}$$

which finally yield

$$\overline{a} = \alpha - (1/6)\gamma, \quad \overline{b} = \beta + \gamma. \tag{4}$$

Obviously, the approximation error $|Q(x) - P(x)|$ vanishes (and thus is minimum) at those points $\overline{x}$ where the linear function P coincides with the original function Q. To compute these points, we have to solve the quadratic equation

$$P(\overline{x}) = \alpha - (1/6)\gamma + (\beta + \gamma)\overline{x} = \alpha + \beta\overline{x} + \gamma\overline{x}^2 = Q(\overline{x}). \tag{5}$$

Surprisingly, the parameters α, β, γ of the original function vanish from this equation, giving rise to the condition

$$\overline{x}^2 - \overline{x} + 1/6 = 0 \tag{6}$$

with solutions

$$\overline{x} \in \left\{(3 - 3^{1/2})/6, \ (3 + 3^{1/2})/6\right\} \simeq \{0.2113, \ 0.7887\}. \tag{7}$$

In other words, for all parabolae, the minimum approximation error of the linear least-squares fit occurs at the same points $x \simeq 0.21$ and $x \simeq 0.79$, supporting the rule formulated at the beginning of this section. However, the result for quadratic polynomials does not carry over to all functions in this universal form. In fact, the approximation error can be far from minimum at the points characterized in equation (7) when the graph is not close to a parabola, as shown in Fig. 2(*a*).

In order to quantify how well a function f can be approximated by polynomials of degree 2, we first introduce the maximum discrepancy between f and a general quadratic polynomial Q, *i.e.*

$$\|f - Q\|_\infty := \max_{x\in[0,1]} |f(x) - Q(x)|. \tag{8}$$

The closest parabola to f is then the polynomial Q^* which minimizes this distance in the set Π_2 of all possible quadratic polynomials

$$d_2(f) := \|f - Q^*\|_\infty = \min_{Q\in\Pi_2} \|f - Q\|_\infty. \tag{9}$$

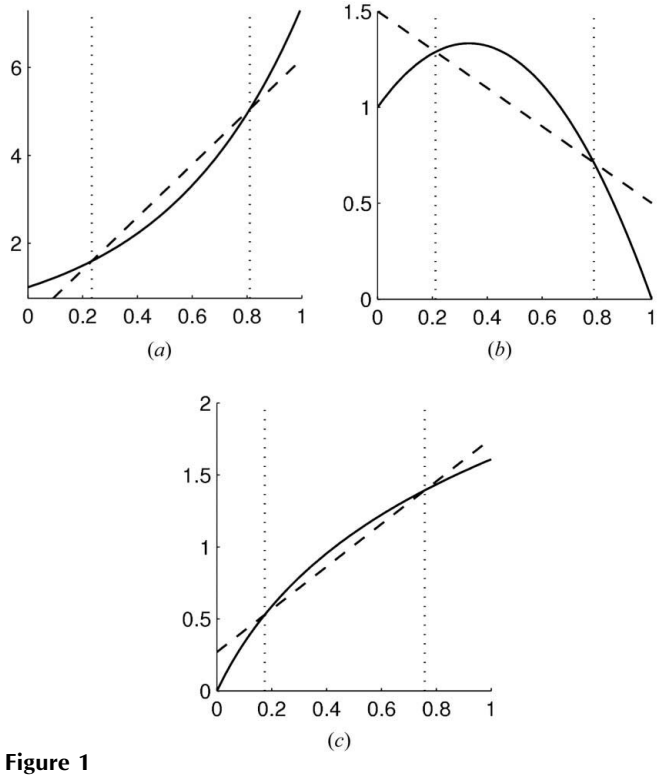

Figure 1
Various graphs (solid lines) with corresponding linear least-squares fits (dashed lines). The graphs are given by (*a*) $\exp(2x)$, (*b*) $1 + 2x - 3x^2$ and (*c*) $\ln(1 + 4x)$. The minimum approximation error occurs at $x \simeq 0.2$ and $x \simeq 0.8$.

With this notation, we can translate the assumption included in our rule that the linear fit should be reasonable into the quantitative statement that the minimum distance $d_2(f)$ of f from the quadratic polynomials should be small. Equipped with this measure, we can now cast the rule into a strict mathematical theorem, the proof of which can be found in Appendix *A*.

Theorem 1. Let $f : [0, 1] \to \mathbb{C}$ be square-integrable and let P be its linear least-squares fit. The approximation error at $\overline{x} = (3 \pm 3^{1/2})/6$ can then be estimated by

$$|f(\overline{x}) - P(\overline{x})| \leq 3d_2(f). \tag{10}$$

In other words, the linear fit at the favourable points $\overline{x}$ is almost as accurate as the best quadratic approximation of *f*.

2.2. Main result

In order to increase the applicability of Theorem 1, we modify it in two respects. First, we relax the assumption of the linearity of the fitted polynomial, allowing for polynomials of arbitrary degree and thus increasing the range of functions it is possible to fit (see, for example, Fig. 2*b*). The second and more important modification concerns the restriction of available information on the function *f*. We assume that only finitely many measurements $Y(x_i)$, $i = 1, \ldots, N$ are given, where x_i are pairwise different and $Y(x_i)$ are of the form

$$Y(x_i) = f(x_i) + \varepsilon_i \tag{11},$$

with independent and identically distributed measurement errors ε_i having mean 0 and variance σ^2.

It turns out that these modifications influence the number and location of the favourable points $\overline{x}$. In order to give a precise definition in this more general situation, we introduce the set $\mathbb{L}^2(X)$ of square summable functions on $X = \{x_1, \ldots, x_N\}$ with the scalar product

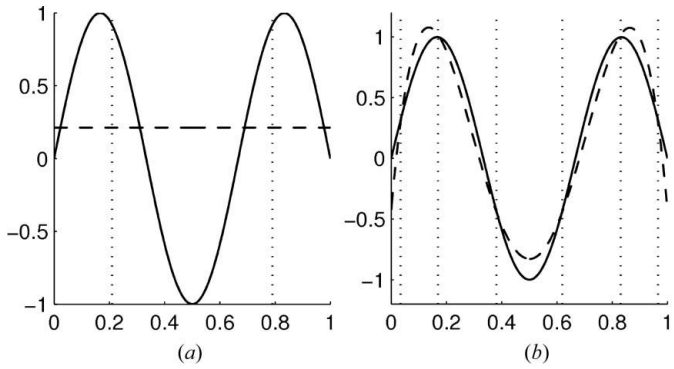

Figure 2
(*a*) The approximation error at $x \simeq 0.21$ and $x \simeq 0.79$ may be large when the graph is far from a parabola. The solid line shows the function $\sin(3\pi x/2)$ and the dashed line is the linear least-squares fit. (*b*) A fit with a fifth-order polynomial (dashed) gives rise to six favourable points (indicated by dotted lines), which are again close to the points of minimum approximation error.

$$\langle f, g\rangle := (1/N)\sum_{i=1}^{N} f(x_i)g(x_i), \quad f, g \in \mathbb{L}^2(X). \tag{12}$$

A particular basis of the space $\mathbb{L}^2(X)$ is obtained by orthonormalizing the monomials $1, x, \ldots, x^{N-1}$ with respect to the scalar product. This leads to orthonormal polynomials $p_0, p_1, \ldots, p_{N-1}$, where p_k has degree k.

Using this orthonormal basis, the least-squares approximation of Y with a polynomial $P^{(q)}$ of degree q is nothing but the projection of Y onto the space Π_q spanned by $p_0, \ldots, p_q$:

$$P^{(q)} = \langle Y, p_0\rangle p_0 + \langle Y, p_1\rangle p_1 + \cdots + \langle Y, p_q\rangle p_q. \tag{13}$$

In order to choose favourable points for the evaluation of $P^{(q)}$, we compare it with the higher-degree approximation $P^{(q+1)}$. In view of equation (13), the difference is

$$P^{(q+1)}(x) - P^{(q)}(x) = \langle Y, p_{q+1}\rangle p_{q+1}(x), \tag{14}$$

which obviously vanishes at the roots of p_{q+1}, independent of the specific measurement Y. Hence, at these favourable points, the least-squares fit of degree q has the same values as the least-squares approximation of degree $q + 1$. Since the maximum degree that can reasonably be used in the fitting procedure is restricted because of the limited supply of data and the corresponding danger of overfitting, this observation can be used to increase the degree of the approximation without actually increasing the degree of the fitting procedure.

The following theorem summarizes our considerations and shows that the favourable points defined in the more general scenario generalize those described in Theorem 1. The detailed proof can be found in Appendix *A*. Again, it uses the notation of minimum distance $d_r(f)$ between f and polynomials of degree r, *i.e.*

$$d_r(f) = \min_{Q \in \Pi_r} \|f - Q\|_\infty. \tag{15}$$

Theorem 2. Let $q \in \mathbb{N}_0$, $X = \{x_1, \ldots, x_N\}$ with $N > q + 1$, $f \in \mathbb{L}^2(X)$ and $Y \in \mathbb{L}^2(X)$ defined by equation (11) with least-squares approximation $P^{(q)}$ of degree q. Further, let $\overline{x}$ be a root of p_{q+1}. Then, the expected squared approximation error at $\overline{x}$ is estimated by

$$E\big(|f(\overline{x}) - P^{(q)}(\overline{x})|^2\big) \leq C_q\big[\sigma^2/N + (2q + 4)d_{q+1}(f)^2\big], \tag{16}$$

where $C_q = \sum_{k=0}^{q} p_k(\overline{x})^2$.

In other words, the degree-q fit of Y at a favourable point $\overline{x}$ is, up to the noise and a multiplicative constant, as accurate as the best approximation of f with polynomials of degree $q + 1$.

Our first remark concerns the noise-related part σ^2/N of the total error. In general, it is advisable to increase the number of measurements N to reduce this part of the error estimate. However, in macromolecular X-ray experiments, the situation is somewhat different, because the number of obtainable intensity measurements per unique reflection is limited owing to the deterioration of the crystal in the course of the experiment.

In order to increase the number of measurements by some factor, the intensity I of the beam, and therefore the strength of f, has to be reduced by the same factor to ensure that the total dose applied to the crystal does not change. In other words, NI and f/I are intensity-independent. In contrast with this, the main part of the noise level σ^2 reduces only in proportion with the intensity, as a consequence of the Poisson statistics underlying the photon count. Assuming an additional basic noise level σ_0^2 and a contribution that is proportional to the square of the intensity, we have $\sigma^2 = \sigma_0^2 + c_1 I + c_2 I^2$ with proportionality constants c_1 and c_2.

To see how the choice of the total intensity influences the measurement error, we apply Theorem 2 to the scaled data:

$$Y_*(x_i) = \frac{f(x_i)}{I} + \frac{\sigma}{I}\varepsilon_i. \tag{17}$$

Since $f_* = f/I$ is independent of the total intensity I, an intensity dependence can only be found in the error component related to the random fluctuations with variance σ^2/I^2:

$$\frac{\sigma^2}{I^2 N} = \frac{\sigma_0^2}{NI}\frac{1}{I} + \frac{c_1}{NI} + \frac{c_2}{NI} I. \tag{18}$$

Since NI is constant, we see that the fluctuation effect inherent in the photon counting is not influenced at all by a change in intensity. As a consequence, the decision for low intensity/high multiplicity *versus* high intensity/low multiplicity should be based on an estimate of the additional error contributions. If the intensity-independent level dominates that proportional to I, a high intensity should be used, and, conversely, a low intensity should be employed if the error contribution proportional to the intensity dominates.

In a macromolecular data collection, the term σ_0^2 is given mainly by background and detector dark noise, and c_2 is strongly influenced by the nonlinearity and counting properties of the detector. However, other factors originating from the characteristics of the X-ray beam and experimental setup (shutters, slits, spindle) also come into play. If all these factors are known, which would be highly desirable, a program like *BEST* (Popov & Bourenkov, 2003) may be used to find the optimum strategy for data collection.

The second remark concerns the decision for the order of the fitting polynomial which is the focus of this article. In particular, we wish to investigate under which conditions the linear fit ($q = 1$) outperforms the constant fit ($q = 0$). Comparing the error bounds from Theorem 2 for the scaled data $f_* = f/I$, this may be the case when

$$C_1\left[\frac{\sigma^2}{I^2 N} + 6d_2(f_*)^2\right] \le C_0\left[\frac{\sigma^2}{I^2 N} + 4d_1(f_*)^2\right], \tag{19}$$

or, after solving for $d_2(f_*)$, when the condition

$$d_2(f_*)^2 \le \frac{2C_0}{3C_1} d_1(f_*)^2 - \left(1 - \frac{C_0}{C_1}\right)\frac{\sigma^2}{I^2 N} \tag{20}$$

is satisfied. From this condition, one can read off that the linear fit is likely to win over the constant one, provided the multiplicity N and the signal-to-noise ratio I/σ are high and the measured profile f_* is sufficiently better approximated by quadratic polynomials than by linear ones.

For a given distribution of nodes x_i, the constants C_0 and C_1 can be computed as shown below. However, the condition of equation (20) is not really a comparison of errors but a comparison of worst-case error bounds. A more realistic answer to the question of whether a linear fit is better than a constant one is given in §3, where the expected approximation error is evaluated for synthetic data which mimic the intensity measurements in a macromolecular X-ray experiment.

2.3. Favourable point locations

For a given distribution of nodes x_i, the computation of the favourable points $\overline{x}$ used in Theorem 2 is straightforward. First, let us consider the construction in the case $q = 0$ of fitting with a constant. In this case, we require the first polynomials p_0 and p_1, which are obtained by orthonormalization of the monomials $r_k(x) = x^k$. As a result of the definition of the scalar product [equation (12)], the constant monomial r_0 is already normalized, so that $p_0(x) = r_0(x) = 1$, implying also $C_0 = 1$. The next polynomial p_1 is given by

$$p_1 = \alpha_1(r_1 - \langle r_1, p_0\rangle p_0) \tag{21}$$

with a suitable normalization constant α_1. Introducing the abbreviation

$$\overline{X} = \langle r_1, p_0\rangle = (1/N)\sum_{i=1}^{N} x_i \tag{22}$$

for the barycentre of X, we thus have

$$p_1(x) = \alpha_1(x - \overline{X}), \quad \alpha_1 = 1/\|r_1 - \overline{X}\|. \tag{23}$$

Based on p_0, the least-squares fit $P^{(0)}$ is

$$P^{(0)}(x) = \langle Y, p_0\rangle p_0(x) = (1/N)\sum_{i=1}^{N} Y(x_i), \tag{24}$$

which is simply the average of the given measurements. According to Theorem 2, the average should be considered as an approximation of the true function f at the favourable point $\overline{x} = \overline{X}$, which is the root of the polynomial p_1.

Proceeding to the case $q = 1$, the favourable points for the linear fit $P^{(1)}$ are defined as roots of

$$p_2 = \alpha_2(r_2 - \langle r_2, p_1\rangle p_1 - \langle r_2, p_0\rangle p_0). \tag{25}$$

Using the centred power moments

$$m_k = (1/N)\sum_{i=1}^{N}(x_i - \overline{X})^k, \tag{26}$$

the polynomial p_2 can be written as

$$p_2 = \alpha_2\left[(x - \overline{X})^2 - \frac{m_3}{m_2}(x - \overline{X}) - m_2\right] \tag{27}$$

with roots

$$\overline{x} \in \left\{\overline{X} + \frac{m_3}{2m_2} + s\left(\frac{m_3^2}{4m_2^2} + m_2\right)^{1/2} \;\middle|\; s \in \{-1, 1\}\right\}. \tag{28}$$

Table 1
Left and right favourable points for a regular distribution of the nodes $x_1, \ldots, x_N$.

In the limit $N \to \infty$, we recover the two points of the continuous case discussed in Theorem 1.

N	4	5	6	8	10	20	∞
Left	0.221	0.217	0.215	0.214	0.213	0.212	0.211
Right	0.780	0.783	0.785	0.786	0.787	0.788	0.789

In particular, if the point distribution is symmetric around the barycentre, the odd moments vanish ($m_3 = 0$), resulting in the symmetric favourable points $\overline{X} + s(m_2)^{1/2}$ with $s \in \{-1, +1\}$. In this case, the formula for C_1 is also particularly simple. Noting that $\alpha_1 = 1/(m_2)^{1/2}$, we find $p_1(\overline{x})^2 = 1$, so that $C_1 = p_0(\overline{x})^2 + p_1(\overline{x})^2 = 2$.

As specific example, let us consider the case of regularly distributed nodes

$$x_i = (1/N)\,(i - 1/2) \in [0, 1], \quad i = 1, \ldots, N. \tag{29}$$

In this case, $\overline{X} = \frac{1}{2}$, $m_3 = 0$ and $m_2 = (1 - 1/N^2)/12$, giving rise to the favourable points

$$\overline{x} \in \left\{ 1/2 + (s/6)\left(3 - 3/N^2\right)^{1/2} \;\middle|\; s \in \{-1, 1\} \right\}. \tag{30}$$

For various values of N, the resulting left and right favourable points are listed in Table 1. It should be noted that the locations change only slightly with N. This is important in cases such as crystallographic data collection, where the point distribution X varies from one experiment to another so that the favourable points differ. Since the variation is not very strong as long as the points are distributed sufficiently regularly, it still makes sense to use a common evaluation point. This aspect is analysed carefully in the next section.

We close this section by reporting the favourable points in the case $q = 2$ for node distributions that are symmetric about their barycentre. Computing the roots of

$$p_3(x) = \alpha_3 \left[(x - \overline{X})^3 - \frac{m_4}{m_2}(x - \overline{X}) \right], \tag{31}$$

we find three favourable points for $q = 2$:

$$\overline{x} \in \left\{ \overline{X} - \left(\frac{m_4}{m_2}\right)^{1/2},\ \overline{X},\ \overline{X} + \left(\frac{m_4}{m_2}\right)^{1/2} \right\}. \tag{32}$$

3. Testing the favourable points

Theorem 2 estimates the averaged squared approximation error

$$e_q = E(|f(\overline{x}) - P^{(q)}(\overline{x})|^2) \tag{33}$$

when the fit polynomial $P^{(q)}$ of degree q is used at the favourable point $\overline{x}$ to approximate the true value $f(\overline{x})$ from noisy data $Y(x_i) = f(x_i) + \sigma\varepsilon_i$.

In the practically relevant case of intensity measurements in a macromolecular X-ray experiment, the values $f(x_i)$ would be intensities corresponding to certain reciprocal lattice points of the Fourier transform of the unit cell, where x_i is a rotation angle of the crystal, or equivalently, a dose value of radiation. The number n of dose values at which the intensity is available is called its multiplicity.

As far as the x_i are concerned, we are faced with the complication that they are not fixed in advance but depend on the experimental setup and vary from one reciprocal lattice point to another. As a consequence, the favourable points are also different for different reflections, which leads to a certain quality reduction when we interpolate several intensities at a common dose value. Instead of the situation studied in Theorem 2, it would therefore be more realistic to allow for a reasonable variability of x_i while using some effective favourable point $\overline{x}_q$.

Secondly, the dose-dependent intensity functions f in X-ray experiments possess certain structural features which suggest the consideration of a restricted range of possible target functions f and the estimation of the average interpolation error for this whole class.

In the following, we will therefore include a variation of the nodes x_i as well as the target functions f in the averaging process (described in §§3.1 and 3.2). The resulting average squared approximation error e_q then characterizes the behaviour of the fitting procedure for a whole range of scenarios which are reasonably related to those found in macromolecular X-ray experiments. Clearly, the advantage of using synthetic values for f and x_i over real ones is that the errors can be computed exactly.

3.1. Effective favourable points

As already indicated in Table 1, the locations of the favourable points vary only slightly with multiplicity for regular node distributions. To check the behaviour in cases that are irregular, we compare various dose value distributions in the unit interval. Our goal is to find out whether a definition of common favourable points is still feasible.

In the m-box random case, we split the unit interval into $\lceil n/m \rceil$ equal-sized boxes and choose in each box m random points, *i.e.*

$$x_i = \frac{1}{\lceil n/m \rceil}(\lceil i/m \rceil - u_i), \quad i = 1, \ldots, n, \tag{34}$$

with u_i being independent and uniformly distributed in [0, 1]. This case (often with $m = 4$ when the detector is not swung out) appears appropriate for an experiment in crystallography in which the total rotation range is several times larger than the asymmetric unit of reciprocal space given by the crystal symmetry and alignment (Dauter, 1999).

In the random case, the points are chosen fully random, *i.e.*

$$x_i = u_i, \quad i = 1, \ldots, n, \tag{35}$$

where again u_i are independent and uniformly distributed in [0, 1].

In order to assess the distribution of favourable points, we conduct an experiment with $M = 10\,000$ node sets and multiplicity $n = 10$. The favourable points are computed in the case

of linear fitting and the results are collected in histograms, shown in Fig. 3.

Mean values for the favourable points are summarized in Table 2. The general observation is that the variation of the favourable points is strongly related to the regularity of the points x_i. In addition, the variation reduces if the multiplicity is increased. Finally, the average over the favourable points is only weakly dependent on the structure of the experiment and is quite close to the corresponding values listed in Table 1 for the case of regular node distribution.

In the following, we will work with the 4-box random node distribution because it reflects to some extent the structure of real macromolecular X-ray data. As an effective favourable point we take $\overline{x}_1 = 0.21$.

With a similar approach, we find in the case of a constant least-squares fit ($q = 0$) that $\overline{x}_0 = 0.5$ is very close to the average favourable points for all types of experiments. In the case of quadratic fits, $\overline{x}_2 = 0.12$ is a reasonable compromise, although for fittings of higher degree, the dependence of the favourable points on the multiplicity is generally stronger.

3.2. Model for intensity functions

Next, we briefly describe the structure of a generic synthetic data set Y used in our numerical tests (further details can be found in Appendix B). In accordance with the model described in the previous section, we assume

$$Y(x_i) = f(x_i) + \sigma\varepsilon_i, \tag{36}$$

where f models the true dose-dependent intensity and ε_i are independent standard normal-distributed random numbers which simulate measurement errors. The variance parameter σ is chosen in such a way that a certain signal-to-noise ratio r is obtained. More specifically, we set

$$\sigma = s/r, \quad s = (1/n)\sum_{i=1}^{n} f(x_i) \tag{37}$$

so that $r = s/\sigma$ is the ratio of average intensity to noise level σ.

To obtain a suitable structure for the intensity values, we assume that the dose-dependent contribution $F(x)$ of the Fourier transform has the form

$$F(x) = A\exp(i\varphi) + B(x)\exp(i\psi), \tag{38}$$

where $A\exp(i\varphi)$ is the transform of the undisturbed electron density at the reciprocal lattice point under consideration,

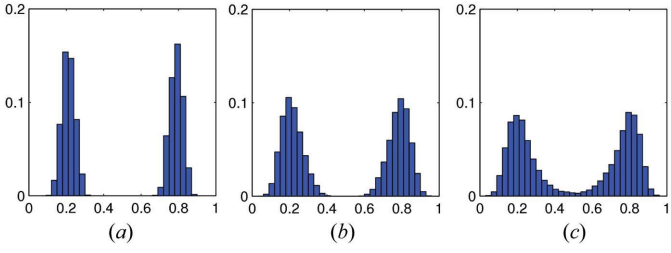

Figure 3
Distributions of favourable points for $M = 10\,000$ node sets with multiplicity $n = 10$. The node distributions are (*a*) 2-box random, (*b*) 4-box-random and (*c*) random.

Table 2
Mean values of favourable points.

n	1-Box random	2-Box random	4-Box random	Random
4	0.214 0.787	0.224 0.773	0.254 0.747	0.254 0.748
8	0.211 0.789	0.212 0.788	0.217 0.782	0.226 0.773
16	0.211 0.789	0.211 0.789	0.212 0.788	0.219 0.782

while $B(x)\exp(i\psi)$ represents the dose-dependent transform of the damaged density. Our structural assumptions on B are $B(0) = 0$ and monotonicity, reflecting the growing influence of radiation damage. Restricting to the case $A > 0$, the scaled intensity $f(x) = |F(x)|^2/A^2$ has the form

$$f(x) = |a + b(x)|^2, \tag{39}$$

with $a = \exp[i(\varphi - \psi)]$ and $b = B/A$ also being monotonic. Since b is unknown and definitely dependent on many independent factors, we randomly choose it from a reasonably large set of functions which share the monotonicity property and the zero initial value. Similarly, a is chosen randomly on the unit circle in the complex plane. For further details on the construction of f, we refer to Appendix B.

3.3. Relevance of effective favourable points

In the following, we investigate the usefulness of effective favourable points for the evaluation of the least-squares approximations. To this end, we compute the averaged squared approximation error

$$E[|f(x) - P^{(q)}(x)|^2] \tag{40}$$

between the true value $f(x)$ and the degree-q polynomial fit $P^{(q)}(x)$ of the noisy data at 100 equidistant points $x \in [0, 1]$. The average is based on $M = 10\,000$ sets of 4-box random distributed nodes in connection with M random functions constructed as described above.

In Fig. 4, results are shown for various multiplicities and signal-to-noise ratios in the case of linear fits. Obviously, the effective favourable point $\overline{x}_1 = 0.21$ offers no advantage in the case of low multiplicities and small signal-to-noise ratios. Conversely, for high average multiplicities and low measurement errors, the point is almost optimum.

In order to assess the influence of the degree of the fitting polynomial, we repeat the simulation for constant ($q = 0$) and quadratic ($q = 2$) fitting.

In the case of low multiplicity and high noise level, the linear fit has larger average approximation errors than the constant fit at all points (see Fig. 5). Thus, it is not advisable to use linear fits in this situation. The error of the quadratic fit is so large that the curve is not visible on the chosen scale. We note in passing that the error of the constant fit is minimum close to the effective favourable point $\overline{x}_0 = 0.5$. We can also see that the constant approximation is poor if used as an approximation for the zero-dose value, because the average approximation error is comparably large at $x = 0$.

At a medium multiplicity $n = 10$, but still at a low signal-to-noise ratio $r = 8$, the linear approximation is better than either

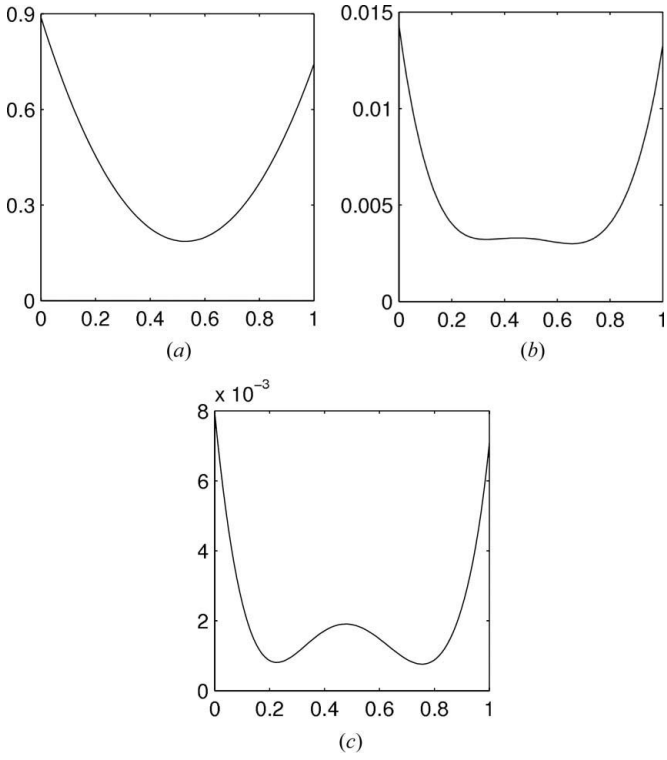

Figure 4
Averaged squared approximation error [equation (40)] of linear fits (q = 1) for $x \in [0, 1]$.The multiplicities n and signal-to-noise ratios r are (*a*) n = 4, r = 2, (*b*) n = 10, r = 8, and (*c*) n = 16, r = 14.

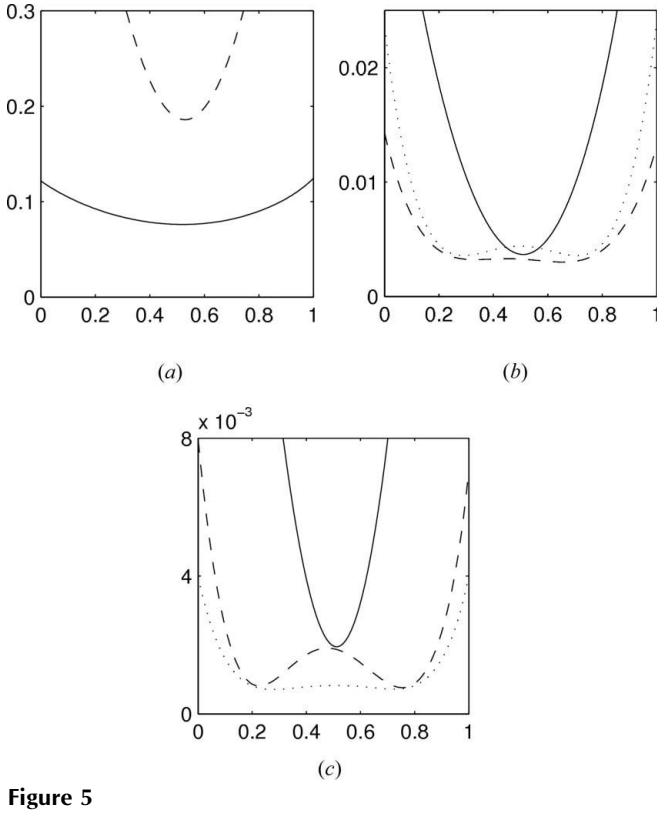

Figure 5
Averaged squared approximation errors of constant (solid), linear (dashed) and quadratic (dotted) fits. The multiplicities n and signal-to-noise ratios r are (*a*) n = 4, r = 2, (*b*) n = 10, r = 8, and (*c*) n = 16, r = 14.

the constant or the quadratic fit over the complete interval. Although at the favourable point $\overline{x}_1 = 0.21$, the approximation error of the linear fit shows no local minimum and the accuracy is quite close to that of the constant fit at $\overline{x}_0 = 0.5$. Thus, the linear fit may be used instead of the constant approximation if an interpolation point closer to the origin is desirable. A glance at the error of the quadratic fit shows that it is now better than the constant approximation at the boundaries of the interval but less accurate at the centre.

This situation changes for high multiplicities and better signal-to-noise ratios, as indicated in Fig. 5(*c*). Now the quadratic fit generally has the lowest average approximation error, although a local minimum at the effective favourable point $\overline{x}_2 = 0.12$ has not yet developed. In contrast, the approximation error of the linear fit possesses a clear local minimum at $\overline{x}_1 = 0.21$ and the error value at this point is smaller than the minimum error of the constant approximation. Even if the error values were almost equal, it would still pay off to use the linear fit, because it delivers a good approximation at a smaller dose value. We note that, also in this case of accurate measurements, the zero-dose extrapolation based on the evaluation of the polynomial approximation at $x = 0$ is accompanied by a relatively large approximation error.

3.4. Systematic assessment of errors at effective favourable points

In order to summarize the observations of the previous section, we compute the errors at favourable points $\overline{x}_q$ for a range of characteristics

$$n \in \{4, 5, 6, \ldots, 20\}, \quad r \in [1, 21]. \tag{41}$$

By e_q we denote the averaged squared approximation error of the polynomial approximation of degree q at the corresponding effective favourable point derived in §3.1. The ratios e_0/e_1 and e_0/e_2 are shown in Fig. 6. As expected, it is preferable to use the constant approximation in the case of low multiplicity and poor signal-to-noise ratio. However, if the signal-to-noise ratio exceeds ten and the multiplicity is at least

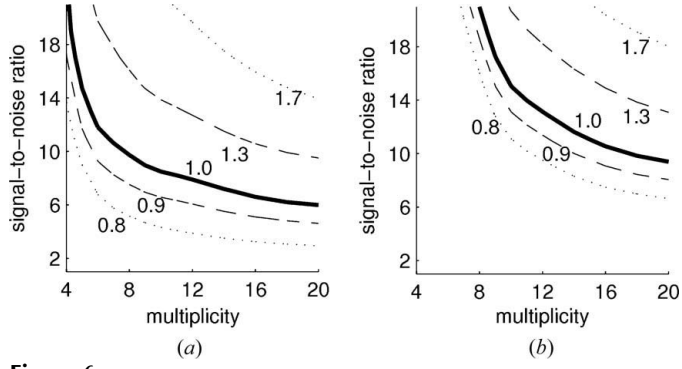

Figure 6
(*a*) Level lines of the ratio e_0/e_1. Above the solid line, the error of the constant fit at $\overline{x}_0$ is larger than that of the linear approximation at $\overline{x}_1 = 0.21$. (*b*) Level lines of the ratio e_0/e_2 with e_2 computed at the favourable point $\overline{x}_2 = 0.12$.

eight, the error e_1 of the linear approximation is generally better than e_0 and it is obtained at a lower dose value. For multiplicities less than eight, increasingly large signal-to-noise ratios are necessary to obtain a linear fit that outperforms the constant one.

Comparing the error e_2 of the quadratic approximation with e_0, we observe a similar behaviour, with the difference that the break-even line is reached only at higher multiplicities and larger signal-to-noise ratios. However, a direct comparison between e_1 and e_2 shows that, for the synthetic experiment considered here, the quadratic fitting never leads to considerably smaller errors than the linear one.

4. Results and discussion

The results presented above support a number of direct conclusions and offer explanations for previous findings. Most importantly from a practical viewpoint, this paper states the conditions under which linear extrapolation can be expected to yield better reduced data than the (usually employed) estimate using a constant, where the latter disregards the dose information attached to each observation of a unique reflection. It is found that, if either a sufficiently large multiplicity or a high enough signal-to-noise ratio is present, the linear fit is superior to the constant fit when calculated at the favourable value of interpolation. It is shown that this favourable point corresponds to 21% of the full dose, similar to the value suggested by Diederichs (2006) without proper theoretical justification.

A most welcome aspect of this finding is that the biological interpretation of electron densities and atomic positions may be performed based on estimates of intensities obtained at a lower level of radiation damage than the fit of a constant to the data (which always results in intensities corresponding to an average level of radiation damage). Thus, the models of biological structures obtained by linear interpolation are less altered by radiation damage, and certain types of misinterpretation (like giving significance to radiation-induced changes of side-chain conformations) may be less likely.

Within a given macromolecular data set, the signal-to-noise ratio is usually highest at low resolution, and low at high resolution. Thus, the low-resolution structure factors benefit most from the linear interpolation, as noted by Diederichs *et al.* (2003), whereas the effect at high resolution (with low signal-to-noise ratio) may be an increase in averaged approximation error. The method thus seems to be best applicable to substructure detection, which requires a high signal-to-noise ratio, but is less suitable for producing data for high-resolution refinement, unless specific procedures are implemented to guard against an increase in noise in the interpolated data or the multiplicity is unusually high (Weiss *et al.*, 2004).

Our results also show that extrapolation to zero dose, using the fitting function directly, leads to an increase in errors. This is most relevant to MAD and SAD phasing and explains the finding of Ravelli *et al.* (2005), who observed increased noise in anomalous difference Pattersons obtained from zero-dose extrapolation. Furthermore, it rationalizes the finding that zero-dose extrapolation does not appear to be as helpful in MAD phasing as initially expected by Diederichs *et al.* (2003). Rather than interpolating towards the middle of the total dose range, as suggested by Diederichs *et al.* (2003), our results indicate that a better strategy would be to interpolate the intensities of each of the MAD data sets (or rather, each wavelength) at its favourable dose value (at either 21 or 79% of the dose range covered by this wavelength).

A final comment concerns snapshots of structure factors corresponding to different states of radiation damage. It may be suggested to interpolate between about 15 and 85% of the dose interval of the measurement; values near 0 and 100% have significantly more error (see Figs. 4 and 5).

Future work should investigate whether an even better model for radiation damage and its correction could be obtained by generating more realistic data from synthetic molecular models with simulated (known) radiation damage.

APPENDIX *A*
Proofs of the theorems

A1. Proof of Theorem 1

Defining the $\mathbb{L}^2([0,1])$-orthonormal polynomials

$$p_0(x) = 1, \quad p_1(x) = (12)^{1/2}(x - 1/2), \tag{42}$$

the linear least-squares fit of a function f is given by

$$P_f = \langle f, p_0\rangle p_0 + \langle f, p_1\rangle p_1. \tag{43}$$

Further, let Q be any quadratic polynomial and define the distance to f as

$$d = \max_{x\in[0,1]} |f(x) - Q(x)|. \tag{44}$$

Using the fact that the linear least-squares fit P_Q coincides with Q at the favourable points $\overline{x}$, we have

$$|P_f(\overline{x}) - f(\overline{x})| \le |P_f(\overline{x}) - P_Q(\overline{x})| + |Q(\overline{x}) - f(\overline{x})|. \tag{45}$$

To estimate the first term, we use the representation of the linear fit

$$|P_f(\overline{x}) - P_Q(\overline{x})| \le |\langle f - Q, p_0\rangle|\,|p_0(\overline{x})| + |\langle f - Q, p_1\rangle|\,|p_1(\overline{x})|. \tag{46}$$

Since $|p_0(\overline{x})| = |p_1(\overline{x})| = 1$, we conclude

$$|P_f(\overline{x}) - P_Q(\overline{x})| \le d + d\int_0^1 |p_1(x)|\,\mathrm{d}x = d + (3^{1/2}/2)d \le 2d. \tag{47}$$

Inserted into equation (45), we obtain the required estimate $|P_f(\overline{x}) - f(\overline{x})| \le 3d$.

A2. Proof of Theorem 2

The proof is similar to that of Theorem 1. We first observe that, for some arbitrary polynomial $Q \in \Pi_{q+1}$, the best approximation $P_Q^{(q+1)}$ in Π_{q+1} is Q itself. With the definition of the favourable points $\overline{x}$ as roots of p_{q+1}, we have

$$P_Q^{(q)}(\overline{x}) = P_Q^{(q)}(\overline{x}) + \langle Q, p_{q+1}\rangle p_{q+1}(\overline{x}) = P_Q^{(q+1)}(\overline{x}) = Q(\overline{x}), \tag{48}$$

which means that the best fit of Q in Π_q equals Q at the favourable points. Using this property, we can split the difference

$$P_Y^{(q)}(\overline{x}) - f(\overline{x}) = P_Y^{(q)}(\overline{x}) - P_Q^{(q)}(\overline{x}) + Q(\overline{x}) - f(\overline{x}). \tag{49}$$

Further, in view of the linearity of the least-squares procedure, we have

$$P_Y^{(q)}(\overline{x}) - P_Q^{(q)}(\overline{x}) = P_{f-Q}^{(q)}(\overline{x}) + P_\varepsilon^{(q)}(\overline{x}), \tag{50}$$

where $\varepsilon = Y - f \in \mathbb{L}^2(X)$ is the function $\varepsilon(x_i) := \varepsilon_i$ which assigns to each x_i the corresponding measurement error. Altogether, we have

$$P_Y^{(q)}(\overline{x}) - f(\overline{x}) = P_\varepsilon^{(q)}(\overline{x}) + R(\overline{x}), \quad R = P_{f-Q}^{(q)} + Q - f. \tag{51}$$

For the squared error, we conclude

$$|P_Y^{(q)}(\overline{x}) - f(\overline{x})|^2 = P_\varepsilon^{(q)}(\overline{x})^2 + R(\overline{x})^2 + 2P_\varepsilon^{(q)}(\overline{x})R(\overline{x}). \tag{52}$$

Since ε has a mean of zero at every x_i, the linearity of both the expected value and the least-squares projection yields

$$E[P_\varepsilon^{(q)}(\overline{x})R(\overline{x})] = P_{E(\varepsilon)}^{(q)}(\overline{x})R(\overline{x}) = 0. \tag{53}$$

We therefore have

$$E[|P_Y^{(q)}(\overline{x}) - f(\overline{x})|^2] = E[P_\varepsilon^{(q)}(\overline{x})^2] + R(\overline{x})^2. \tag{54}$$

Expanding the least-squares approximation of ε, we find

$$P_\varepsilon^{(q)}(\overline{x})^2 = \sum_{k,l}\langle\varepsilon, p_k\rangle\langle\varepsilon, p_l\rangle p_k(\overline{x})p_l(\overline{x}). \tag{55}$$

For the expected value of the product of the scalar products, we obtain

$$E(\langle\varepsilon, p_k\rangle\langle\varepsilon, p_l\rangle) = (1/N^2)\sum_{i,j}E(\varepsilon_i\varepsilon_j)p_k(x_i)p_l(x_j) \tag{56}$$

and, in view of the independence and the common variance,

$$E(\langle\varepsilon, p_k\rangle\langle\varepsilon, p_l\rangle) = (\sigma^2/N^2)\sum_i p_k(x_i)p_l(x_i) = (\sigma^2/N)\langle p_k, p_l\rangle. \tag{57}$$

Inserting this relation into the previous sum and using $\langle p_k, p_l\rangle = \delta_{kl}$, we finally obtain

$$E[P_\varepsilon^{(q)}(\overline{x})^2] = (\sigma^2/N)\sum_k p_k(\overline{x})^2. \tag{58}$$

It remains to estimate the deterministic term R^2 in terms of the distance

$$d = \max_{x\in[0,1]}|f(x) - Q(x)|. \tag{59}$$

First, we rewrite the squared sum as a sum of squares

$$R(\overline{x})^2 \le 2\left\{P_{f-Q}^{(q)}(\overline{x})^2 + [Q(\overline{x}) - f(\overline{x})]^2\right\} \le 2\left[P_{f-Q}^{(q)}(\overline{x})^2 + d^2\right]. \tag{60}$$

Noting that

$$P_{f-Q}^{(q)}(\overline{x})^2 = \left[\sum_{k=0}^{q}\langle f - Q, p_k\rangle p_k(\overline{x})\right]^2 \le (q+1)\sum_{k=0}^{q}\langle f - Q, p_k\rangle^2 p_k(\overline{x})^2 \tag{61}$$

and applying the Schwarz inequality

$$\langle f - Q, p_k\rangle^2 \le \|f - Q\|^2\|p_k\|^2 \le d^2, \tag{62}$$

we obtain the estimate

$$R(\overline{x})^2 \le 2\left[(q+1)d^2\sum_{k=0}^{q}p_k(\overline{x})^2 + d^2\right] \le 2(q+2)\sum_{k=0}^{q}p_k(\overline{x})^2 d^2, \tag{63}$$

where we have used $p_0 = 1$ so that $1 \le \sum_{k=0}^{q}p_k(\overline{x})^2$. Combining the two estimates in equation (54), the result of the theorem follows.

APPENDIX *B*
Construction of the intensity functions

In this section, we give additional information on the construction of the dose-dependent intensity function

$$f(x) = |a + b(x)|^2, \quad x \in [0, 1]. \tag{64}$$

From the form of f it is clear that we can restrict our considerations to monotonically growing functions b because the decreasing case differs from the increasing one only by a sign change of a, where a is in any case chosen to be uniformly distributed on the unit circle in $\mathbb{C}$.

To ensure that $b(x)$ is a reasonably general monotonic function, we write

$$b(x) = \beta[x + \mu Q(x)], \tag{65}$$

where β is any real number, $\mu \in (0, 1)$ and Q is a polynomial which satisfies $Q(0) = 0$ and $|Q'(x)| \le 1$ (we took $\mu = 0.8$ in our test case). This guarantees that $b(0) = 0$ and that the slope of $x + \mu Q(x)$ is always positive, implying monotonicity of b.

The derivative $R(x) = Q'(x)$ is constructed as a convex combination,

$$R(x) = \lambda_0 T_0(-x) + \lambda_1 T_1(-x) + \cdots + \lambda_k T_k(-x), \tag{66}$$

of the Chebychev polynomials T_i which are known to obey $|T_i(x)| \le 1$ (we took $k = 5$ in our experiment). In order to obtain a wide variety of such polynomials, we choose the coefficients $(\lambda_0, \ldots, \lambda_k)$ uniformly from the standard $(k + 1)$ simplex according to Devroy (1984).

While the parameter a in equation (64) is chosen to be uniformly distributed on the unit circle, the value of β is chosen such that the total intensity variation amounts to a fraction λ of the initial intensity a^2. The factor λ is a random variable with average $\overline{\lambda} = 0.4$ which is uniformly distributed in the two intervals $[\lambda_-, \overline{\lambda}]$ and $[\overline{\lambda}, \lambda_+]$ to the left and the right of the average, where $\lambda_- = 0.2$ and $\lambda_+ = 1.1$.

Altogether, the algorithm for the construction of f is as follows. First, the complex number a is chosen randomly with $|a| = 1$ and R is constructed using a random sample from the $(k + 1)$ simplex, followed by integration to obtain Q. Then, the

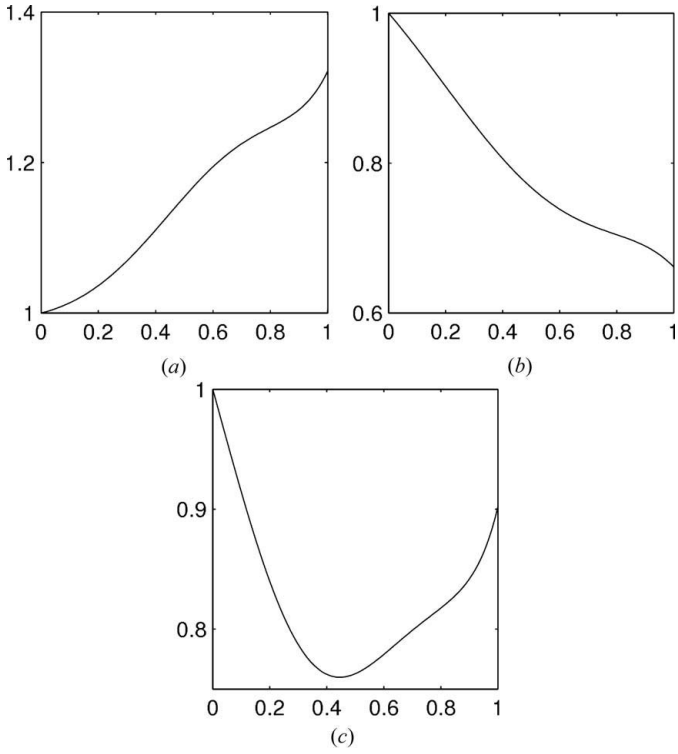

Figure 7
Some typical normalized synthetic intensity functions with an intensity variation of around 40%. (*a*) Monotonic growth corresponds to phase factors *a* with a positive real part. (*b*) A monotonic decrease typically appears for a negative real part of *a* and a small imaginary part. (*c*) A minimum generally occurs in the case when *a* has a negative real part and an imaginary part close to 1.

sign of $\beta > 0$ is selected based on the random fraction λ, specifying the allowed total variation of the intensity f on $[0, 1]$.

Some typical examples of the resulting functions f are shown in Fig. 7.

We conclude with a description of the overall characteristics of our synthetic data. Choosing 4-boxed random points, multiplicity $n = 12$ and signal-to-noise ratio $r = 8$, we construct $M = 100\,000$ intensity functions $Y_1, \ldots, Y_M$ which are set up with the parameters given above. Computing the difference between the average intensity of the first $Y_k(x_{k1})$ and last $Y_k(x_{kn_k})$ measurements, we find

$$\left|(1/M)\sum_{k=1}^{M} Y_k(x_{k1}) - (1/M)\sum_{k=1}^{M} Y_k(x_{kn_k})\right| \simeq |1.027 - 1.026|. \tag{67}$$

Relative to the average initial intensity, this amounts to only 0.1%, which is a consequence of the symmetry in our construction of the intensity functions where an increase in the intensity is as probable as a decrease. This corresponds to the fact that, during processing of experimental data, the overall trend of weakening diffraction is removed by the scaling procedure.

A second characterizing quantity is the average absolute intensity change relative to the absolute initial intensity

$$\frac{\sum_{k=1}^{M} |Y_k(x_{k1}) - Y_k(x_{kn_k})|}{\sum_{k=1}^{M} Y_k(x_{k1})} \simeq 20\%, \tag{68}$$

which describes the typical variation in each of the 100 000 intensity functions. Again, this value appears realistic for an experiment with moderate radiation damage.

References

Blake, C. C. F. & Phillips, D. C. (1962). *Effects of X-irradiation on Single Crystals of Myglobin*, in *Biological Effects of Ionizing Radiation at the Molecular Level*. Brno: International Atomic Energy Agency, Vienna.

Burmeister, W. P. (2000). *Acta Cryst.* D**56**, 328–341.

Dauter, Z. (1999). *Acta Cryst.* D**55**, 1703–1717.

Devroy, Luc. (1984). *Generation of Random Numbers.* Heidelberg: Springer.

Diederichs, K. (2006). *Acta Cryst.* D**62**, 96–101.

Diederichs, K., McSweeney, S. & Ravelli, R. B. G. (2003). *Acta Cryst.* D**59**, 903–909.

Popov, A. N. & Bourenkov, G. P. (2003). *Acta Cryst.* D**59**, 1145–1153.

Ravelli, R. B. G. & McSweeney, S. (2000). *Structure*, **8**, 315–328.

Ravelli, R. B. G., Nanao, M. H., Lovering, A., White, S. & McSweeney, S. (2005). *J. Synchrotron Rad.* **12**, 276–284.

Weik, M., Ravelli, R. B. G., Kryger, G., McSweeney, S., Raves, M. L., Harel, M., Gros, P., Silman, I., Kroon, J. & Sussman, J. L. (2000). *Proc. Natl Acad. Sci. USA*, **97**, 623–628.

Weiss, M. S., Mander, G., Hedderich, R., Diederichs, K., Ermler, U. & Warkentin, E. (2004). *Acta Cryst.* D**60**, 686–695.