Inverse Problems **17** (2001) 571–590 www.iop.org/Journals/ip PII: S0266-5611(01)14776-3

Grid refinement and scaling for distributed parameter estimation problems

U M Ascher[1,3] **and E Haber**[2]

[1] Department of Computer Science, University of British Columbia, Vancouver, BC, Canada V6T 1Z4
[2] Departments of Computer Science and of Earth and Oceanography Sciences, University of British Columbia, Vancouver, BC, Canada V6T 1Z4

E-mail: ascher@cs.ubc.ca and haber@cs.ubc.ca

Received 16 June 2000, in final form 22 November 2000

Abstract
This paper considers problems of distributed parameter estimation from data measurements on solutions of differential equations. A nonlinear least squares functional is minimized to approximately recover the sought parameter function (i.e. the model). This functional consists of a data fitting term, involving the solution of a finite volume or finite element discretization of the forward differential equation, and a Tikhonov-type regularization term, involving the discretization of a mix of model derivatives.

The resulting nonlinear optimization problems can be very large and costly to solve. Thus, we seek ways to solve as much of the problem as possible on coarse grids. We propose to search for the regularization parameter first on a coarse grid. Then, a gradual refinement technique to find both the forward and inverse solutions on finer grids is developed.

The grid spacing of the model discretization, as well as the relative weight of the entire regularization term, affect the sort of regularization achieved and the algorithm for gradual grid refinement. We thus investigate a number of questions which arise regarding their relationship, including the correct scaling of the regularization matrix. For nonuniform grids we rigorously associate the practice of using unscaled regularization matrices with approximations of a weighted regularization functional. We also discuss interpolation for grid refinement.

Our results are demonstrated numerically using synthetic examples in one and three dimensions.

1. Introduction

Many inverse problems in applications involve the recovery of a coefficient function of a system of differential equations. Instances include geophysics and medical imaging such as

[3] This work was completed while the author was visiting at IMPA, Rio de Janeiro.

DC resistivity [48], magnetotelluric inversion [42], diffraction tomography [15], impedance tomography [6]; other applications include oil reservoir and aquifer simulation [16, 19]. The forward problem is (a discretization of) a differential system, usually assumed linear:

$$\mathcal{A}(m)\, u \;=\; q, \tag{1}$$

where $\mathcal{A}$ refers to a differential operator defined on an appropriate domain Ω and equipped with suitable boundary conditions. This operator depends on a model, m, which is to be approximately recovered based on measurement data b on the solution u of (1).

Let the operator Q indicate the projection onto the locations in Ω to which the data are associated. Thus, the data are viewed as a nonlinear function of the model:

$$b = Q\mathcal{A}(m)^{-1}q + \epsilon$$

where ϵ is the measurement noise. Since the data are noisy, and the inverse problem of recovering m from it is often ill-posed even without noise [17, 49], there is no unique model which generates the data. Therefore, a process of regularization is used to recover a relatively smooth (or piecewise smooth) solution to a nearby problem which is unique, at least locally.

To solve the forward problem in practice it must be discretized, and we assume that this is done using a stable, accurate finite volume or finite element method. Moreover, the model m must be discretized as well. Thus, our forward model is written as

$$Au = q \tag{2}$$

where the $N_u \times N_u$ matrix A is assumed nonsingular for simplicity. This matrix depends on a parameter vector m of length N_m, $A = A(m)$. In addition, Q becomes a matrix with N_u columns. We have thus implicitly introduced two grids, one with which u is associated and one for m. We assume an appropriate interpolation scheme to connect between the two grids, which is embedded in $A(m)$. Often in practice, these two grids are intimately and simply related, e.g. we envision a grid of rectangular cells for u (in case that Ω is in 2 dimensions), and m is then assumed to consist of the values of the model at cell centres. But here we keep these two grids separate, as is common, for example, in control applications [46].

A regularization method often utilized in practice minimizes the Tikhonov functional [17, 49][4]

$$\min_m \frac{1}{2}\|QA(m)^{-1}q - b\|^2 + \frac{\beta}{2}\|W(m - m_{\mathrm{ref}})\|^2, \tag{3}$$

where W is typically a weighting matrix which does not depend on m, m_{ref} is a reference model and $\beta \geqslant 0$ is the regularization parameter. Note that the choices of W and m_{ref} involve *a priori* information available on the model. Smoothness knowledge is incorporated into W, which typically consists of a mix of discretizations of the zeroth, first and second derivatives of m [42, 50].

The optimization problem (3) can be very large indeed, involving up to a million variables and more in three-dimensional applications. So, advantage must be taken of sparsity of A, W and Q, and special-purpose software is called for [13, 22, 23]. Each typical iteration for solving (3) may involve several solutions of the forward problem (2), hence the latter must be performed very efficiently.

In this paper we consider a complementary approach for further significant reduction in computational cost, namely, that of solving as much of the problem as possible on coarse grids in u and m. Note that for problems where $\Omega \subset \mathbb{R}^3$ a solution of the forward problem on a grid with 16^3 cells, say, typically costs less than 1% of the cost of solving the same problem on 64^3 cells. Thus, we consider the following algorithmic framework:

[4] Throughout this paper the least squares norm is assumed unless specifically indicated otherwise.

(1) Solve problems (3) on as coarse a grid as possible in both u and m to determine an appropriate value for the parameter β.

(2) Once β has been determined, refine the grid gradually, i.e. solve a sequence of problems (3) with the same value of β on finer and finer grids, using the solution of the most recently solved problem, appropriately interpolated, as the first iterate for the solution process of (3) on the next finer level.

(3) On the finer grid, adjust the value of β if necessary.

The idea of grid continuation in m or in u is hardly new (e.g. [2, 20, 35, 52]), but the combination of finding β on a coarse grid and subsequently refining in both m and u is, and it has a great potential for efficiency gain. Note that, as long as the u-grid is fine enough, the grid refinement in u is not expected to make a marked difference in terms of regularization; yet, the refinement in both m and u is crucial for achieving a significant reduction in computational effort. A number of questions arise when attempting to carry out this approach in practice, including the scaling of W and β on different grids and the interpolation between grids. We examine these questions in what follows.

A significant volume of literature has been generated regarding the regularization process (3), including the practical estimation of the regularization parameter β [17, 23, 24, 27, 32, 37, 42, 45, 49, 50]. It is also well known that this regularization functional tends to smear discontinuities in the model, so if such discontinuities are known to be present and the quality and abundance of the data is sufficient to recover them, then one may resort to other regularizations such as total variation (TV) or the Huber norm [8, 18, 31]. However, these latter regularization functionals tend to yield optimization problems which in practice are much harder to solve [9], and the determination of β may be less clear as well. We consider (3) for the discussion following.

In this paper we use the Morozov discrepancy principle [17, 37, 42] in order to determine a value for β on a specific grid. Although this principle is not always the most practical, it is the simplest for the purpose of our illustrations. For other methods to evaluate β, see [17, 24, 28, 42, 50]. Any of these methods involves solving (3) for a few β-values, and a standard continuation in the parameter β, while keeping the grids fixed, is applied. In order to be able to do this successfully, as well as use results from different grids, we assume that the model m depends continuously on β. This assumption does not always hold for nonlinear problems, especially if the TV norm is used and m may have jump discontinuities. However, there are many cases of practical interest where this assumption appears to hold; see, e.g. [12, 17, 37, 47, 50] and references therein. Indeed, if continuity does not appear to hold then, rather than further pursuing the search for a value for β, we must conclude that the modelling process is inadequate and replace the discrepancy principle.

In fact, the discretization in m is also a form of regularization (e.g. [17, 28] and references therein). Assuming a uniform grid with spacing h for m, the regularization effect is stronger the larger h is. However, the dependence of the model's regularity on h is often more erratic, and h itself, unlike β, cannot be varied at will during an efficient multilevel computation.

The plan of the paper is as follows. We gather preliminaries in section 2. These include a sketch of the continuation process in β for a fixed grid. We also indicate why, although parameter continuation as outlined above appears natural, for small values of β which occur in practice care must be taken during the continuation process, both in β and in the grids. Thus, it is important in practice to obtain the best initial iterate possible, especially when the continuation process involves grid refinement, where the cost of each iteration increases significantly. This, in turn, necessitates paying special attention to the question of scaling in terms of β and the m-grid.

In section 3 we discuss the ingredients necessary for a successful grid continuation process. At first we discuss the scaling of the discrete regularization matrix W. Practitioners in data inversion often tend to employ unscaled differences in W, even if the grid in m is nonuniform [4, 18, 23, 42, 45]. We recall in section 3.1 the correct scaling of W so that β remain constant as the grid is refined, both for uniform and for nonuniform grids (cf [17, 41]). For several uniform grids, the regularization parameters of an unscaled W must be modified so that the scaled β remains constant for an effective grid-continuation process—see example 3.1. Even for one uniform grid it can be argued that a correct scaling can be important to observe, because this allows obtaining a good guess for the value of β for one problem based on a prior investigation of another, similar problem, independent of the grid involved and the combination of different derivatives used in W.

Nonuniform grids are often used in remote sensing applications [23, 39, 42], frequently with unscaled differences. This practice does not appear to have been rigorously explained before; rather, heuristic justifications have been given. We consider in section 3.3 a grid transformation which allows a rigorous explanation for employing an unscaled W, even for a nonuniform grid. We do not attempt, though, to propose or completely justify the actual choice of such a grid.

In section 3.4 we further consider continuation approaches in the grid. When refining the grid and using the solution on a coarser grid to define a starting iterate for the finer grid, the simplest approach is to use a piecewise constant interpolation. This is natural if m is interpreted as a piecewise constant grid function, which makes sense when the regularization functional contains no derivatives of m. However, in other cases piecewise constant interpolation can be improved upon, as we demonstrate in example 3.3.

Examples in one dimension are used in section 3 to demonstrate various aspects of the proposed techniques and analysis in a tractable way. However, grid continuation becomes much more practically important in more space dimensions. In section 4 we experiment with a three-dimensional example, where grid continuation both in u and in m is used to advantage, reducing the total computational cost by a factor of more than 4.

Conclusions and further discussion are offered in section 5.

2. Preliminaries

In this section we recall some well known facts and procedures regarding the determination of a value for β along with a solution of (3). We then make an observation regarding a simple continuation approach, in either β or the grid, for small β values.

2.1. Reducing β gradually to fit the data

Let

$$J = -QA^{-1}G \qquad \text{where} \quad G = \frac{\partial(A(m)u)}{\partial m}. \tag{4}$$

The sensitivity matrix J describes the change of the predicted data $Qu(m)$ with respect to changes in the model m. Assume that a tolerance Tol is known and that the inverse problem is solved using the discrepancy principle, that is, the inverse problem is posed as that of finding m such that (2) holds and

$$\|Qu - b\| \leqslant \text{Tol}. \tag{5}$$

If the data vector has length N_b and contains noise with known standard deviation σ then set

$$\text{Tol} = N_b\sigma \tag{6}$$

and require (approximate) equality in (5).

Next, consider the constrained optimization problem

$$\min \tfrac{1}{2}\|W(m - m_{\text{ref}})\|^2 \tag{7a}$$

$$\text{s.t. } \|QA^{-1}q - b\|^2 \leqslant \text{Tol}^2. \tag{7b}$$

Forming the Lagrangian of (7) and differentiating to obtain the necessary conditions for an optimum, and then comparing to similar necessary conditions of (3), it becomes clear that β should in fact approximate the inverse of the Lagrange multiplier of the data fitting constraint. Since this multiplier is typically increased from zero until equality is reached in (7*b*), it is natural to consider a practical procedure in which β is started at a large positive value and is then subsequently decreased until (5) is satisfied. The value of β is thus related to the noise level in the data. It should not be 'too small', to avoid fitting the noise rather than the data.

Moreover, for small values of β the nonlinear problem of minimizing (3) can be difficult to solve, because J is typically rank-deficient and/or ill-conditioned, while for large values of β this problem becomes easier, because the simple quadratic term $\frac{\beta}{2}\|W(m - m_{\text{ref}})\|^2$ dominates. Thus, the continuation procedure of gradually reducing β also has the effect of gradually increasing the nonlinearity of the problem, allowing the option of considering only local convergence properties of optimization strategies.

2.2. An observation on simple continuation with small β

In practice, when using a Newton or a Gauss–Newton method for solving the nonlinear least squares problem (3), it has often been our experience that continuation in either the grid or in β is straightforward and effective when β is so large that the regularization term $\beta W^T W$ dominates the Jacobian matrix. For instance, the Jacobian matrix for the Gauss–Newton case is

$$C = J^T J + \beta W^T W. \tag{8}$$

However, for small β, we have observed that more Newton-like iterations are needed for each continuation step.

In fact, often the Newton-like correction appears to be not particularly small in norm for the first iterate (i.e. starting with the iterate provided by the continuation process), even if the objective function value appears to be close to minimum. This may occur because C typically has a small singular value when β is small. Correspondingly, the residual norm of the gradient of (3) may rise at first before starting to descend (for the latter to occur the variation in C must not be large), and the whole process requires a few iterations per grid or the β-value to converge.

Thus, for small values of β, we must attempt to find a first iterate at each continuation step which is particularly close to the minimizer if we want the process to terminate after only one or two iterations per continuation step.

3. The scaling of W and β, and grid continuation

Throughout this section we consider inverse problems in one space dimension, i.e. the differential equation (1) is defined on an interval $[0, x_f]$. The results generalize directly to multi-dimensional problems.

3.1. Discretizing a continuous regularization term

Here we introduce notation and quickly derive some known results which prove useful in what follows. Following [17, 41] and many others, let us consider the Tikhonov term as a discretization of a continuous regularization term, which in one dimension on an interval $[0, x_f]$ may be written as

$$T = \tfrac{1}{2}\int_0^{x_f} \{\beta_0(m - m_{\rm ref})^2 + \beta_1((m - m_{\rm ref})')^2 + \beta_2((m - m_{\rm ref})'')^2\}\,{\rm d}x.$$

The weights β_i all relate to one parameter β by assuming that there are fixed values $k_0, k_1 = 1$ and k_2 such that $\beta_i = k_i\beta$, $i = 0, 1, 2$. Assuming for the time being that $\beta_2 = 0$, and setting $m_{\rm ref} = 0$ to save on notation, we will concentrate below on the functional

$$T = \tfrac{1}{2}\int_0^{x_f} \{\beta_0(m)^2 + \beta_1(m')^2\}\,{\rm d}x. \tag{9}$$

Next we discretize (9) on a generally nonuniform grid:

$$0 = x_0 < x_1 < \cdots < x_{N_m-1} < x_{N_m} = x_f. \tag{10}$$

Note that in principle this grid has nothing to do with the grid on which u is discretized. (In practice, however, the u-grid and the m-grid often coincide.) Let

$$\begin{aligned}
&h_{i-1/2} = x_i - x_{i-1}, \qquad 1 \leqslant i \leqslant N_m,\\
&h_i = (x_{i+1} - x_{i-1})/2 = (h_{i+1/2} + h_{i-1/2})/2, \qquad 0 \leqslant i \leqslant N_m,\\
&\qquad (x_{-1} = -x_1, \quad x_{N_m+1} = 2x_f - x_{N_m-1})\\
&h = \max_{1\leqslant i\leqslant N_m} h_{i-1/2},\\
&\bar{h} = x_f/N_m = \text{mean}\,(h).
\end{aligned} \tag{11}$$

The vector of unknowns m has elements which are considered as approximations to m at midpoints,

$$m_{i-1/2} \approx m(x_i - h_{i-1/2}/2), \qquad 1 \leqslant i \leqslant N_m. \tag{12}$$

We obtain

$$T_h = \frac{\beta_0}{2}\sum_{i=1}^{N_m} h_{i-1/2} m_{i-1/2}^2 + \frac{\beta_1}{2}\sum_{i=0}^{N_m} c_i h_i \left(\frac{m_{i+1/2} - m_{i-1/2}}{h_i}\right)^2, \tag{13}$$

where $c_i = 1$, $1 \leqslant i \leqslant N_m - 1$, $c_0 = c_{N_m} = 1/2$. We set either Dirichlet or Neumann homogeneous boundary conditions for $m_{-1/2}$ and $m_{N_m+1/2}$: in the absence of more information we prefer Neumann:

$$m_{-1/2} = m_{1/2}, \qquad m_{N_m+1/2} = m_{N_m-1/2}. \tag{14}$$

When we write (13), (14) in matrix-vector form, the matrix W, which appears in the term $\|W(m - m_{\rm ref})\|^2$ in (3) and (7a), is given by

$$W = \begin{pmatrix} \sqrt{k_0}W_0 \\ W_1 \end{pmatrix} \tag{15a}$$

$$W_0 = \text{diag}\{\sqrt{h_{i-1/2}}\} \tag{15b}$$

$$W_1 = \text{diag}\{\sqrt{h_i}\}\begin{pmatrix} -1/h_1 & 1/h_1 & & & \\ & -1/h_2 & 1/h_2 & & \\ & & \ddots & \ddots & \\ & & & -1/h_{N_m-1} & 1/h_{N_m-1} \end{pmatrix}. \tag{15c}$$

Note that W_0 is $N_m \times N_m$, whereas W_1 is $(N_m - 1) \times N_m$.

Consider a uniform grid, where $h_i = h$, $h_{i-1/2} = h$, for all relevant i. Then W_1 is in a divided difference form, and both W_0 and W_1 are scaled by $\sqrt{h}$. Similarly, if $\beta_2 \neq 0$ we keep $k_2 = \beta_2/\beta_1$ independent of the grid and obtain a matrix W_2 of size $(N_m - 2) \times N_m$ consisting of discretizations of the second derivative in divided form and scaled by $\sqrt{h}$.

Remark 3.1. In d space dimensions the scaling factor $\sqrt{h} = h^{1/2}$ becomes $h^{d/2}$, as can be easily verified; see also section 4.

With this scaling of W it is expected that β be independent of the grid spacing, provided that m is smooth and the grid is sufficiently fine, because T_h in (13) clearly tends to a well-defined limit. Indeed, the two terms of the integrand in (9) are approximated in (13) using midpoint and trapezoidal quadrature, respectively:

Theorem 3.1. *If $m \in C^1[0, x_f]$ then, with β_0 and β_1 held fixed in (9) and (13),*

$$T - T_h = \mathrm{O}(h^2).$$

If m contains a jump discontinuity, however, then T of (9) is unbounded. If T_h were to remain faithful then, as the grid is refined, there would be an index $i = j$ such that $\frac{m_{j+1/2} - m_{j-1/2}}{h_j} = \mathrm{O}(h_j^{-1})$, so T_h would blow up too as $h \to 0$. To keep T and T_h both finite and close, we must change the regularization functional, e.g., switch to the l_1 (TV) norm in (9).

Thus, keeping β fixed with the above scaling of W should generate similarly looking solution profiles on different grids, provided the coarser grid is already sufficiently fine for the model.

For a uniform grid it may be argued that the general regularization matrix

$$W = \begin{pmatrix} \sqrt{\beta_0} W_0 \\ \sqrt{\beta_1} W_1 \\ \sqrt{\beta_2} W_2 \end{pmatrix}$$

may be written in an unscaled form, i.e. with $W_0 = I$, and W_1 and W_2 written using unscaled differences. The dependence on h is then relegated to the constants, namely, $\beta_0 \propto h,\ \beta_1 \propto h^{-1},\ \beta_2 \propto h^{-3}$. However, in practice the determination of these coefficients is often a hard, open problem, and their dependence on the grid (i.e. the scaling involved) must therefore be kept in mind.

3.2. A numerical example

For the numerical examples in this paper we have implemented versions of two Newton-like iterations, see [23], for the nonlinear optimization problems obtained upon setting the grids in u and m and determining a value for β. The first, denoted GN, is the Gauss–Newton method for the nonlinear least squares problem (3).

The second method, denoted NP, uses a two-phase iteration: (i) apply a full Newton step to the nonlinear equations

$$\mathbf{grad}\, \mathcal{L} = 0, \tag{16a}$$

where

$$\mathcal{L} = \mathcal{L}(u, m, \lambda) = \frac{1}{2}\|Qu - b\|^2 + \frac{\beta}{2}\|W(m - m_{\mathrm{ref}})\|^2 + \lambda^T\,(A(m)u - q) \tag{16b}$$

is the Lagrangian of the constrained formulation for the same optimization problem,

$$\begin{aligned} &\min_{m,u} \frac{1}{2}\|Qu - b\|^2 + \frac{\beta}{2}\|W(m - m_{\mathrm{ref}})\|^2 \\ &\text{s.t. } A(m)u = q. \end{aligned} \tag{17}$$

Table 1. Numerical experiment 1: iteration counts and misfit levels. Uniform grid spacings $h^u = 2^{-7}$ for u and h for m are used.

Noise	β	h	Method	Iterations	$\|Qu - b\|$
5%	3.0e−6	2^{-4}	GN	7	2.80e−2
		2^{-5}		7	2.81e−2
		2^{-6}		7	2.80e−2
		2^{-7}		7	2.80e−2
5%	3.0e−6	2^{-4}	NP	5	2.80e−2
		2^{-5}		5	2.81e−2
		2^{-6}		5	2.80e−2
		2^{-7}		5	2.80e−2
5%	$2^{-3} \times$ 3.0e−6	2^{-7}	NP	6	2.61e−2
	$2^{3} \times$ 3.0e−6	2^{-7}		5	3.43e−2
10%	8.0e−6	2^{-4}	GN	7	5.59e−2
		2^{-7}		7	5.60e−2
10%	8.0e−6	2^{-4}	NP	5	5.59e−2
		2^{-7}		5	5.60e−2

(ii) Set

$$u = A(m)^{-1}q, \tag{18}$$

once m has been updated.

See [23] for additional details and discussion. The objective function (3) is used for both GN and NP as a merit function.

Example 3.1. We generate data for the model

$$-(m(x)u')' = 1,$$
$$u(0) = u(1) = 0,$$
$$m(x) = 1 + 3\mathrm{e}^{-(x-0.18)^2/0.01} + 1.5\mathrm{e}^{-(x-0.6)^2/0.02}.$$

This is a smooth model. The data are generated at 127 equidistant points, and 5 and 10% random noise levels, using a normal distribution with uniform standard deviation, are added. This generates data fitting tolerances Tol $= 2.83 \times 10^{-2}$ and Tol $= 5.66 \times 10^{-2}$, respectively. The forward model uses a straightforward centred discretization on a uniform grid with step size $h^{(u)} = 2^{-7} = 1/128$. We set $m_{\text{ref}} = 0$, $\beta_0 = 0$ and $\beta_2 = 0$, and concentrate on $\beta = \beta_1$ in (9), (13). Some experimentation yields that the values $\beta = 3.0 \times 10^{-6}$ and $\beta = 8.0 \times 10^{-6}$ result in data misfits which are close to the tolerances for the 5 and 10% noise levels, respectively.

Table 1 displays iteration counts and misfit levels, starting in all cases from the 'cold start' iterate $m \equiv 1$, u by (18) and, for NP, $\lambda \equiv 0$. The stopping criterion for the nonlinear iteration is when $\|\delta m\| < 10^{-4}\|m\|$ or when the norm of the gradient of (3) is below 10^{-8}. No line search is needed in any of the results reported here.

In figure 1 we display the curves obtained for m on the coarse grid $h = 2^{-4}$ as well as on 3 fine grids $h = 2^{-7}$ obtained by (i) keeping β fixed; (ii) keeping $h^{-1}\beta$ fixed; and (iii) keeping $h\beta$ fixed. When β is kept fixed the results on the fine and the coarse grids look rather close. This agrees with theorem 3.1. Keeping $h\beta$ fixed results in an undue flattening of the model, and the data fitting error is unacceptably large. Keeping $h^{-1}\beta$ fixed results in a wiggly curve which fits too much of the noise.

Next we solve for $h = 2^{-7}$, $\beta = 3.0 \times 10^{-6}$, starting from the simply interpolated solution of the coarser problem $h = 2^{-4}$ and the same value for β. Whereas starting from the 'cold start' requires five iterations according to table 1, here it takes NP four iterations to converge,

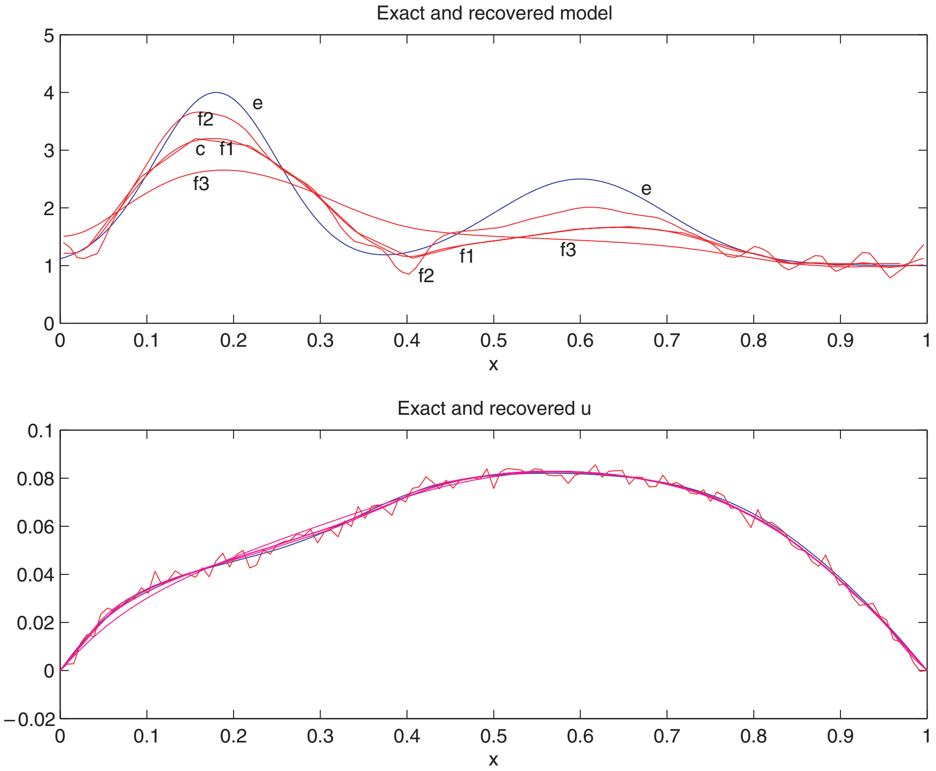

Figure 1. Recovered m for a smooth model with 5% noise using (1) the coarser grid spacing $h = 2^{-4}$ for m with $\beta = 3.\mathrm{e}{-6}$, denoted 'c'; (2) the finer grid spacing $h = 2^{-7}$ for m with $\beta = 3.\mathrm{e}{-6}$, denoted 'f1'; (3) the finer grid spacing $h = 2^{-7}$ for m with $\beta = (3.\mathrm{e}{-6})/8$ (keeping $h^{-1}\beta$ fixed) denoted 'f2'; and (4) the finer grid spacing $h = 2^{-7}$ for m with $\beta = 8 * (3.\mathrm{e}{-6})$ (keeping $h\beta$ fixed) denoted 'f3'. The case where β is kept fixed as the grid is refined yields the results closest to those on the coarser grid.

(This figure is in colour only in the electronic version, see `www.iop.org`)

displaying the phenomenon discussed in section 2.2 where the norm of the gradient increases at the first iteration before starting to descend. A more gradual grid refinement does better. A similar phenomenon occurs when continuing in β for a fixed grid, when the value of β is small while the relative change in β is not.

Next we repeat the same experiment with noise level 10%. Some results are displayed in table 1. The behaviour is similar to that observed for the lower noise level. The recovered profile contains slightly less detail than before, but the scaling and convergence features captured in figure 1 and theorem 3.1 remain unchanged.

Note that for β at the correct level for the given noise, the number of Newton or Gauss–Newton iterations does not change as the grid in m becomes finer, in agreement with [29]. We return to this point in section 5.

We have also experimented with $\beta_2 \neq 0$, i.e., using a mix of first and second derivatives in the regularization. When employing a divided difference scaling with $\sqrt{h}$ in W and keeping β fixed, the resulting solution profiles for m are very close on different grids, while they vary more significantly when β_i values are changed.

3.3. Weighted scaling for a nonuniform grid

The general nonuniform grid (10) is discussed in section 3.1 without regard to where it came from. The regularization functional (9) is uniformly weighted over the interval of integration, which reflects a tacit assumption that information is equally available and equally reliable everywhere in x.

This is not the case in a typical geophysical application, where the measurements are normally taken at the surface of the earth, and where information quality deteriorates as depth (which we call here x) increases. Increasingly large step sizes in the grid are then taken, not because m is expected to be smoother, but because less is possible to recover as depth increases.

We therefore pursue the assumption that similar importance should be placed on each grid element, regardless of its size, in the discretized regularization functional. This leads to a weighted Tikhonov functional and its discretization. Similar assumptions are used in [36, 42] and elsewhere. In this section we do not attempt to justify this choice of *a priori* information. Rather, we show how to interpret this information when working with different grids. This is important for the performance of grid continuation algorithms.

We think of the grid (10), (11), as a sufficiently smooth transformation of a uniform grid with spacing $\bar{h}$. Thus, assume there is a *differentiable*, monotonically increasing function $\rho : [0, x_f] \to [0, x_f]$ such that

$$\rho(\mathrm{i}\bar{h}) = x_i, \qquad i = 0, 1, \ldots, N_m. \tag{19}$$

Clearly this function is invertible, so there is a corresponding inverse transformation ψ such that $\psi(\rho(\xi)) = \xi, \quad 0 \leqslant \xi \leqslant x_f$, and we may consider in place of (9) the regularization functional

$$T_\psi = \frac{1}{2}\int_0^{x_f} \left\{\beta_0 m^2 + \beta_1 \left(\frac{\mathrm{d}m}{\mathrm{d}\psi}\right)^2\right\} \mathrm{d}\psi. \tag{20}$$

Letting

$$\omega(x) = \frac{\mathrm{d}\rho}{\mathrm{d}\xi}(\psi(x)) = \left.\frac{\mathrm{d}\rho}{\mathrm{d}\xi}\right|_{\xi=\psi(x)}$$

we have $\mathrm{d}\psi = \omega^{-1}\,\mathrm{d}x$, $\frac{\mathrm{d}m}{\mathrm{d}\psi} = \omega m'$. Thus,

$$T_\psi = \tfrac{1}{2}\int_0^{x_f} \{\beta_0 m^2 + \beta_1(\omega m')^2\}\omega^{-1}(x)\,\mathrm{d}x.$$

Example 3.2. A possible choice is an exponential grid

$$\rho(\xi) = \frac{c^\xi - 1}{d}, \qquad \text{whence} \tag{21a}$$

$$\psi(x) = \log_c[dx + 1], \tag{21b}$$

where $c > 1$ is a fixed parameter and $d = c^{x_f} - 1$. This is indeed an exponential grid:

$$x_i = \frac{(c^{\bar{h}})^i - 1}{d}. \tag{21c}$$

The weight function is then

$$\omega(x) = \frac{(dx + 1)\ln c}{d}. \tag{21d}$$

The function $\omega(x)$ increases as x increases.

In the discretization we now obtain

$$T_h^{\psi} = \frac{\beta_0}{2}\sum_{i=1}^{N_m} h_{i-1/2}\omega^{-1}(x_{i-1/2})m_{i-1/2}^2 + \frac{\beta_1}{2}\sum_{i=0}^{N_m} c_i h_i \omega(x_i)\left(\frac{m_{i+1/2}-m_{i-1/2}}{h_i}\right)^2. \tag{22}$$

Theorem 3.1 still applies to $T_\psi - T_h^\psi$, with the obvious modifications in notation.

Note that with an exponential grid (21), or any other grid such that $h_{i+1/2} > h_{i-1/2}$, the weights in (22) for the discretization of m' increase in size as x (or i) increases, while those for m decrease: as we get further from the surface boundary, the influence of the data deteriorates, hence the regularization term should induce more flatness (i.e. closeness to the *a priori* m_{ref}) and be less mindful of the actual values of m needed for the data fitting.

Often, for a given grid (10) we do not really have a grid function ρ defined everywhere, nor do we know $\rho(\xi)$ explicitly as we do in (21). Thus, it makes sense to approximate

$$\begin{aligned}\omega(x_{i-1/2}) &= \frac{\rho(\mathrm{i}\bar{h}) - \rho((\mathrm{i}-1)\bar{h})}{\bar{h}}\\ &= \frac{x_i - x_{i-1}}{\bar{h}}\\ &= h_{i-1/2}/\bar{h}.\end{aligned}$$

Likewise, $\omega(x_i) = h_i/\bar{h}$. Thus, the weights in (22), not surprisingly, turn out to be precisely those for a uniform grid, with the average step size $\bar{h}$ for spacing:

$$T_h^{\psi} = \frac{\bar{h}\beta_0}{2}\sum_{i=1}^{N_m} m_{i-1/2}^2 + \frac{\beta_1}{2\bar{h}}\sum_{i=0}^{N_m} c_i(m_{i+1/2}-m_{i-1/2})^2. \tag{23}$$

Since the approximations for ω are also second-order accurate we have obtained the following theorem.

Theorem 3.2. *If $m \in C^1[0, x_f]$ and the grid is defined as in (19) by a monotonically increasing function $\rho \in C^1[0, x_f]$ then, with β_0 and β_1 held fixed in (20) and (23),*

$$T_\psi - T_h^\psi = \mathrm{O}(\bar{h}^2).$$

Remark 3.2.

- The expression (23) depends on the grid only through its size N_m and the presumed location of the $m_{i-1/2}$, which in turn appear in the discretization $A(m)$. (The values of m which are needed in the midpoints of the u-grid for $A(m)$ are obtained by a piecewise constant interpolation, and this is where the abscissae of the m-grid, (10), come in.)
- The rationale for (23) is similar to that for error equidistribution (e.g. [2], chapter 9), in that the underlying assumption is that each grid element has equal importance, which its relative size should reflect.

The above remarks on (23) are in contrast to the unweighted expression (13). For each of these two expressions, if we keep β constant and refine the grid in m then (for a sufficiently fine coarsest grid) this process is expected to yield a similarly looking profile for m. This profile would generally be different, though, for the two regularization functionals. We have experimented with nonuniform grids on the setup described in example 3.1. The essential observations are indeed that using the unweighted expression (13) or the weighted expression (23) may produce a significantly different profile for m, while changing the resolution $\bar{h}$ for the same grid function $\rho(\xi)$ has a much smaller effect in this sense.

The experiments in [22, 23] were performed using the expression (23) in the Tikhonov functional. This is a usual practice which theorem 3.2 justifies. A question which remains unanswered, however, is how the grid (10) is actually determined. Indeed, in instances such as the magnetotelluric problem [23] the solution u of the forward problem, not $m - m_{\mathrm{ref}}$, has an evanescing behaviour to which an exponential grid such as (21) can be tailored. It is of course convenient to use the same grid for m as well, but such a choice appears to be more specific than has actually been rigorously justified in the literature.

3.4. More on grid continuation

The simplest grid continuation procedure for m would use a piecewise constant interpolation as the grid is refined in order to define the next first iterate. This is commensurate with the interpretation of the model as piecewise constant. However, if the regularization functional involves derivatives then the term $W^T W m$ effectively takes discrete derivatives of m, and a simple piecewise constant interpolation increases the magnitude of these discrete derivative approximations. Generally, if T involves n derivatives then m should be best interpreted as representing a piecewise polynomial function in C^{n-1}. This leads to a conforming finite element interpretation for the discretization of T.

In case of T of (9), if $\beta_1 = 0$ then a piecewise constant interpolation suffices, but for the more interesting case where $\beta_1 \neq 0$ a local piecewise linear interpolation is preferred: for each pair of consecutive coarse grid points in turn, apply linear interpolation for all fine grid points which fall within that coarse grid subinterval. The process is symmetrized by repeating from right to left and averaging.

The field u for both GN and NP is evaluated on the finer grid by solving the forward problem (2), once an iterate for m has been established. However, for large problems the forward solution process also involves an iterative method and thus requires a starting iterate. The discretization scheme of the forward problem often provides a natural interpolation scheme for u.

Example 3.3. We consider the same model as in example 3.1. Again a 5% noise is added to the synthetic data. But now we assume that the data are known only at 15 equidistant points. Moreover, as we refine the grid we do so simultaneously for the grids in m and in u, $N_m = N_u + 1 = h^{-1}$. The work per iteration on coarser grids is therefore considerably smaller than on finer grids, so we are particularly interested in reducing the number of iterations required on the finer grids to a minimum.

We set $\beta_0 = \beta_2 = 0$, and concentrate on $\beta = \beta_1$. Our initial guess is $m = m_{\mathrm{ref}} \equiv 1$, u by (18) and $\lambda \equiv 0$. We solve the problem on a uniform grid with $h = 2^{-4}$, for $\beta = 10^{-3}$, using either NP or GN (see section 3.2).

Table 2 lists iteration counts for different combinations of method and continuation technique in the grid. We recorded the following continuation combinations:

- The continuation sequence in the grid involves four grids with step sizes $h = 2^{-4}, 2^{-5}, 2^{-6}, 2^{-7}$. The continuation methods are:

  (i) cns—the simplest piecewise constant interpolation.

  (ii) lin—local linear interpolation.

  These interpolations are used for m, and then (18) is used for u and a piecewise linear interpolation for λ, if needed, to begin the iteration on each grid.
- The continuation sequence in β involves four values (three steps), $\beta = 1.0 \times 10^{-3}, 5.0 \times 10^{-5}, 1.0 \times 10^{-5}, 3.0 \times 10^{-6}$. The nonlinear problem gets harder as β gets smaller. A simple continuation method in β is used.

Table 2. Iteration counts for various methods of determining first iterates by continuation in m- and u-grids or in β.

Method	β	h^{-1}	Cont	Itns	Method	β	h^{-1}	Cont	Itns
NP	1.0e−3	16	no	4	GN	1.0e−3	16	no	6
		32	cns	2			32	cns	2
		64	cns	2			64	cns	2
		128	cns	2			128	cns	2
NP	1.0e−3	16	no	4	GN	1.0e−3	16	no	6
		32	lin	2			32	lin	2
		64	lin	1			64	lin	1
		128	lin	1			128	lin	1
NP	1.0e−3	16	no	4	GN	1.0e−3	16	no	6
	5.0e−5			4		5.0e−5			4
	1.0e−5			3		1.0e−5			4
	3.0e−6			3		3.0e−6			6
NP	3.0e−6	16	no	5	GN	3.0e−6	16	no	8
		32	cns	3			32	cns	5
		64	cns	3			64	cns	4
		128	cns	2			128	cns	3
NP	3.0e−6	16	no	5	GN	3.0e−6	16	no	8
		32	lin	3			32	lin	3
		64	lin	3			64	lin	3
		128	lin	2			128	lin	2

From table 2 we can see that the more sophisticated interpolation technique for grid refinement does improve the process occasionally by one iteration per grid continuation, which can be important.

We have also experimented with a cubic spline interpolation for the grid refinement, but this yielded no further improvement over the local linear one.

Note, finally, that the last steps of continuation in the grid are much more expensive than the first ones. In these, only one or two iterations, rather than four or five from a 'cold start', are needed for the best variants, even for the smallest value of the regularization parameter!

4. A three-dimensional example

In this section we conduct some numerical experiments with synthetic data using the forward model arising in DC resistivity or impedance tomography [6, 48] in three dimensions. Thus, the following differential equations are defined in a domain $\Omega \subset \mathbb{R}^3$:

$$\operatorname{div} \boldsymbol{J} = q, \tag{24a}$$

$$\boldsymbol{J} + \mathrm{e}^m \,\mathbf{grad}\, u = \mathbf{0}. \tag{24b}$$

This system is equipped with the following boundary conditions: on the boundary $\partial\Omega$ the flux in the normal direction vanishes,

$$\boldsymbol{J} \cdot \boldsymbol{n}\big|_{\partial\Omega} = 0, \tag{24c}$$

which specifies u up to a constant. To fix that constant we also require

$$\int_\Omega u \,\mathrm{d}V = 0. \tag{24d}$$

Note that the flux $\boldsymbol{J}$ can be eliminated from (24). The differential equation can then be seen as the three-dimensional extension of the differential equation used in examples 3.1 and 3.3, except that the model appears in the exponent in $\sigma = \mathrm{e}^m$. This handles the practical difficulty often arising when the conductivity σ varies widely over a few orders of magnitude, and it automatically respects the physical requirement that $\sigma > 0$. The price is a loss of resolution. The resulting forward problem is in the form (1), and it is well posed for a given model $m(x, y, z)$.

Below we assume that the domain Ω is the unit cube,

$$\Omega = [-1, 1]^3,$$

and consider discretization on a uniform, staggered tensor grid. Thus, with N a positive integer and $h = 2/N$, we consider Ω as the union of N^3 cubic cells of side h each. Corresponding to the notation of section 3 we approximate m in the (i, j, k)th cell by a constant $m_{i-1/2,j-1/2,k-1/2}$, $1 \leqslant i, j, k \leqslant N$. Thus, $N_m = N^3$.

We use the same cell grid to construct a finite volume discretization for the differential system (24). For details see, e.g., [21, 52]. Only short, centred differences are used. The components of $\boldsymbol{J} = (J^x, J^y, J^z)^T$ are placed at the face centres of the cells. Integrating (24*a*) over cell (i, j, k) (whose centre coordinates are denoted $(i-1/2, j-1/2, k-1/2)$) then gives

$$J^x_{i,j-1/2,k-1/2} - J^x_{i-1,j-1/2,k-1/2} + J^y_{i-1/2,j,k-1/2} - J^y_{i-1/2,j-1,k-1/2} + J^z_{i-1/2,j-1/2,k} \\ - J^z_{i-1/2,j-1/2,k-1} = hq_{i-1/2,j-1/2,k-1/2}, \qquad 1 \leqslant i, j, k \leqslant N. \tag{25a}$$

The boundary conditions (24*c*) are naturally used to eliminate values of J^x, J^y, or J^z at cells which are next to a boundary in (25*a*).

The components of u are placed, like m, at cell centres. The x-component, say, of (24*b*) is then discretized centred at the x-face of the cell, yielding

$$h^{-1}(u_{i+1/2,j-1/2,k-1/2} - u_{i-1/2,j-1/2,k-1/2}) = \sigma^{-1}_{i,j-1/2,k-1/2} J^x_{i,j-1/2,k-1/2}, \tag{25b}$$

where

$$\sigma_{i,j-1/2,k-1/2} = 1/(\mathrm{e}^{-m_{i+1/2,j-1/2,k-1/2}} + \mathrm{e}^{-m_{i-1/2,j-1/2,k-1/2}}) \tag{25c}$$

is a harmonic average of two neighbouring cell properties. This can be written in matrix notation as

$$\hat{G}u = S(m)^{-1}J \tag{26}$$

where the matrix $\hat{G}$ discretizes the gradient operator and depends only on the grid and the ordering of the variables. The matrix $S(m)$ is a diagonal matrix which depends on m as per (25*c*). Note that it is easy to differentiate the product $S(m)v$ with respect to m; such a directional differentiation yields a bidiagonal matrix.

Using expressions similar to (25*b*) also in the y-and z-directions, the components of $\boldsymbol{J}$ in (25*a*) which are inside Ω can be eliminated. This yields a system of N^3 linear algebraic equations for u based on a seven-point discretization stencil. This system of equations has a constant null-space, but it becomes nonsingular upon imposing a discrete version of (24*d*).

We now turn to the discretization of the regularization operator which is written in three dimensions as

$$T = \tfrac{1}{2}\int_\Omega \left(\beta_0 m^2 + \beta_1 |\mathbf{grad}\, m|^2\right) \mathrm{d}V. \tag{27}$$

The gradient operator in (27) is discretized using short differences, precisely as for u in (26). This yields

$$T_h = \frac{\beta_0}{2} m^T V_c m + \frac{\beta_1}{2} m^T \hat{G}^T V_f \hat{G} m \tag{28}$$

Table 3. Numerical experiment 3: iteration counts and misfit levels for different grids and β values.

β	h	Iterations	$\|Qu - b\|/\|b\|$
1.0e−4	2^{-2}	5	3.1e−3
1.0e−3	2^{-2}	5	1.5e−2
1.0e−2	2^{-2}	6	7.9e−2
1.0e−4	2^{-3}	5	2.9e−3
1.0e−3	2^{-3}	5	1.4e−2
1.0e−2	2^{-3}	5	7.2e−2
1.0e−4	2^{-4}	5	2.8e−3
1.0e−3	2^{-4}	5	1.3e−2
1.0e−2	2^{-4}	5	7.1e−2

where V_c and V_f are diagonal matrices with the integration cell volumes as their elements.

The matrix W, which appears in the term $\|W(m - m_{\text{ref}})\|^2$ in (3) and (7*a*), is given by (15*a*) with

$$W_0 = V_c^{\frac{1}{2}} \qquad \text{and} \qquad W_1 = V_f^{\frac{1}{2}} \hat{G}.$$

If we use uniform grid spacing h then the diagonal elements of V_c and V_f are all equal to h^3, cf remark 3.1.

Example 4.1. Using the forward problem discretization described above we generate data for (24) using the smooth model

$$m(x, y, z) = \exp\big(-(x - 0.3)^2 - (y - 0.3)^2 - (z - 0.3)^2\big)$$
$$-3\,\exp\big(-(x + 0.3)^2 - (y + 0.3)^2 - (z + 0.3)^2\big).$$

We also choose

$$q(x, y, z) = \delta(x - 0.3)\delta(y + 0.3)\delta(z + 0.3) - \delta(x + 0.3)\delta(y - 0.3)\delta(z - 0.3).$$

This yields a unique u by (24).

The data is sampled at 4^3 uniformly spaced grid points in the interval $[-0.6, 0.6]$. The forward problem (2) is solved to a tolerance of 10^{-6}, using BiCG-stab with SSOR preconditioners [3]. We set $\beta_0 = 0$. The optimization for (3) is performed using an inexact GN method [33, 34, 40, 43], where the linear system

$$(J^T J + \beta W^T W)\delta m = r$$

is solved by its equivalent least-squares system to accuracy of 0.1 using the CGLS method. The nonlinear iteration is terminated when the relative gradient norm is below 10^{-3}.

We conduct two experiments.

(1) First we add a 1% noise to the data and solve (3) for three values of β, iterating for each from a 'cold start' on three different grids.

Note of the results gathered in table 3 that the relative misfit varies little with the grid. This demonstrates that indeed it is possible here to zoom in on an adequate β value using calculations only on the coarse grid with $N_m = 8^3$, in preparation for recovering the model with one β value only on the fine grid with $N_m = 32^3$. Furthermore, using a smaller β or h does not increase the cost in iterations for this example.

Table 4. Numerical experiment 4: iteration counts and misfit levels for grid continuation.

Noise	β	h	Iterations	$\|Qu - b\|/\|b\|$
1%	8.0e−4	2^{-2}	5	1.2e−2
	8.0e−4	2^{-3}	2	1.0e−2
	8.0e−4	2^{-4}	1	1.0e−2
5%	7.0e−3	2^{-2}	4	5.8e−2
	7.0e−3	2^{-3}	1	5.4e−2
	7.0e−3	2^{-4}	1	5.2e−2
10%	2.0e−1	2^{-2}	4	1.1e−1
	2.0e−1	2^{-3}	1	1.0e−1
	2.0e−1	2^{-4}	1	1.0e−1

(2) Next, having selected β based on coarse grid information we use this value and calculate solutions on the gradually finer grids with $N_m = 16^3$ and $N_m = 32^3$. We use the same model and data but with 1, 5 and 10% noise levels. In order to interpolate m and also u from one grid to the next we use continuous, piecewise linear interpolation. This is a local operation.

The results are recorded in table 4. Note the improved iteration counts for the finer grids. Since runs on the finest grid are so much more expensive, an improvement factor of about 4 in CPU time is obtained by the gradual refinement for a fixed value of β.

Remarks.

- The results demonstrate clearly the advantage of our proposed method. Note that each iteration on the coarsest grid is about 300 times cheaper than the same iteration on the finest grid; thus our continuation method enjoys substantial improvement over existing methods where β is obtained on the fine grid.
- The consistency in the results of table 3 over different grids can be obtained only if the coarse grid is not too coarse; that is, if we can find u_H such that $\|Qu_H - b\|$ is small enough. If the noise level is lower than the discretization error, then the coarse grid may not be able to produce the data accurately enough. This has happened in the above example for a noise level of roughly 0.1%. In this case we could not deduce about the regularization parameter based on the coarsest grid calculations.
- This raises the question, what is the coarsest grid that can be used for the indicated purpose? We do not have a rigorous answer. A similar basic lack of rigour occurs also in other multilevel and multigrid methods for practical problems: the coarsest grid must be ‘fine enough’ to set the multilevel machinery in motion.
- In table 4 we see that, as expected, the misfit is slightly higher on the coarse grid then on the fine grid. This suggests that when the β search is conducted on the coarsest grid it need not be very accurate.

5. Conclusions and further discussion

In this paper we have investigated a number of issues related to a multilevel continuation approach for the rapid solution of the nonlinear inverse problem (3). These include the use of grids for the model m and the forward solution $u = A(m)^{-1}q$, as well as the regularization matrix W and the parameter β.

- We have repeatedly observed that, for small values of β which occur in practice, Newton-like iterations may stall at first, even when starting from a close initial iterate. Thus:
  - (i) Although it is possible[5] to additionally regularize C, care must be taken when designing algorithms which attempt to change β at each iteration.
  - (ii) When contemplating grid continuation, care must be taken to have as close a starting iterate as possible, and this highlights the importance of the correct scaling of W.
- We have investigated the scaling of W for both uniform and nonuniform grids and interpreted rigorously the frequent use in practice of unscaled W. In particular, for nonuniform grids, theorem 3.2 indicates how a uniform scaling may be interpreted as corresponding to a weighted regularization functional. The principle is that each grid subinterval is equally important, regardless of its length. The question how the m-grid should be rigorously chosen with this principle in mind seems nontrivial and is beyond the scope of this paper.
- Even though in practice m is often interpreted as representing a piecewise constant function, for the purpose of grid continuation in the presence of a regularization functional which involves derivatives of m, it is preferable to use a smoother than piecewise constant interpolation. For a regularization functional involving up to first-order derivatives a continuous piecewise linear interpolation is particularly suitable.
- The above findings were combined into a continuation algorithm which first uses a coarse grid, where calculations are cheap, to find an appropriate β commensurate with the quality of the data; and then proceeds with gradual grid refinement in both m and u, using one β value with properly scaled W and a sufficiently smooth interpolation, until a satisfactory resolution is achieved. Substantial efficiency gains are obtained in each of these two stages.

We emphasize again that the choice of the coarsest u-grid must be dictated by the discretization error. If this error is smaller than the measurement error, then the value of β used on this grid is a good guess for finer grids. However, if the discretization error is larger than the measurement error, then one cannot infer about β on the fine grids based on β obtained at the coarse grid.

Further remarks.

- As we show in the numerical examples, keeping β fixed with W properly scaled allowed for a good continuation process from one grid to the next, provided that the coarser grid is already sufficiently fine for the model and that a sufficiently smooth interpolation procedure is used for m. But for the examples in one dimension it did not seem to buy much otherwise, because the solution profile is already captured on the coarser grid. This raises the question why refine the model at all? There are several reasons for this. One is that perhaps new features can be found (which we do not know in advance) with the increased degrees of freedom. If no features are found then this gives a measure of security that none can be found with the present information. Another reason for refinement in the model is for purposes of model presentation and evaluation. Finally, recall that in order to achieve a major efficiency gain we refine the grids in both m and u. The refinement then ensures that the predicted data—which are matched against the given data b—are accurate enough.
- The continuation technique used in β was the simplest: take the previously obtained solution as the first iterate for the next problem (3) or (17) in the continuation chain.

[5] This can be done, say, by using a trust region method, or by increasing β in (8) without changing (3); iterative methods for solving large-scale problems may achieve such an effect on the fly as well.

There are well known, more involved techniques for constructing the first iterate for each continuation step using more information, see e.g. [40,51]. For instance, we can consider the gradient of (3) for variable β on fixed grids as a family of problems depending on a parameter, which we can *choose* to be $\tau = -\ln\beta$. It is then possible to use the derivative with respect to the parameter in order to construct a hopefully better initial iterate, at the cost of essentially one iteration. However, this approach does not always yield results which justify the additional cost. Specifically, we have not found it useful in the case of example 3.3.

- The case $\beta = 0$, $Q = I$, has attracted the attention of a number of numerical analysts, although it is rather rare in practice. For a sequence of finer and finer grids the problem (3) (with $\beta = 0$) has been noticed to become harder to solve (requiring more Newton-type iterations) as the grid is refined [19,35]. Some supporting theory is supplied in [10,11]. In contrast, it is well known that if we fix m and consider a sequence of finer and finer grids for u in one dimension, the corresponding matrices $A(m)$ would have uniformly bounded inverses (approximating the Green function of (1)). This forms the basis of a grid independence principle [1,14] which states that the number of Newton iterations required for (local) convergence for a nonlinear forward problem of this type (i.e., if instead of (2) we had a discretization of a nonlinear differential system in u, say $f(u; m) = 0$, with $A = \frac{\partial f}{\partial u}$) remains independent of the u-grid, even though the eigenvalue spread of A increases as the u-grid is refined. For the inverse problem (3), a grid-independence principle can also be formulated for a fixed $\beta > 0$ [29]. This apparent contradiction with [10,11,19,35] is resolved if we pay attention to the distinction of which parameter, β or the discretization parameter h, is being kept fixed, and which is allowed to approach the limit of 0. The setup is similar to that arising in discretization methods for singular perturbation problems (e.g. [2] and references therein). Moreover, in practice we do expect both $h > 0$ and $\beta > 0$ to have fixed values, and the relationship between their relative sizes may vary. Our experiments with realistic values of β and h agreed with [29], namely no significant increase in difficulty of solving the nonlinear problem was observed as the grid regularization parameter h was decreased.
- Unlike, for example, citenash,hlmr,borcea, we are not aiming here at using one multigrid approach in the construction of a whole solution method. Thus, each problem in a given chain of problems typified by a particular grid and a value of β is solved separately, once an appropriate initial iterate is established. Of course, it may well be natural to solve the forward model using a multigrid technique [7,25], what with some multilevel data structures already at hand, but such is not the focus of this paper.
- Finally, we make a further comment about *inexact methods*. In large-scale problems, which typically arise in distributed parameter estimation in more than one dimension, iterative linear algebra methods are routinely employed in the solution of both (2) and each iteration of a method for (3) (or (17)). In such techniques the residual is reduced repeatedly until its norm falls below a prescribed tolerance. But practical necessity, especially in three dimensions, often indicates that such tolerances should not be very small, giving rise to inexact methods (see, e.g., [33,34] and references therein). In this paper we have considered a number of issues which arise independently of the use of inexact methods. Thus, we have tried to isolate their effect by also calculating solutions using different tolerances when solving large, sparse linear systems, especially in example 4.1. Such additional calculations have yielded similar results to those reported here. Generally speaking, most of our conclusions hold in practice, even when inexact methods are used. See also [26,44] for a different approach.

Acknowledgments

We wish to thank Doug Oldenburg, Volker Schulz and Dhavide Aruliah for valuable comments on our exposition.

References

[1] Allgower E, Bohmer K, Potra F and Rheinboldt W 1986 A mesh independence principle for operator equations and their discretizations *SIAM J. Numer. Anal.* **23** 160–9

[2] Ascher U, Mattheij R and Russell R 1995 *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations* (Philadelphia: SIAM)

[3] Barrett R, Berry M, Chan T F, Demmeland J, Donato J, Dongarra J, Eijkhout V, Pozo R, Romine C and Van der Vorst H 1994 *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods* (Philadelphia: SIAM)

[4] Blake A and Zisserman A 1987 *Visual Reconstruction* (Cambridge, MA: MIT Press)

[5] Borcea L 2001 A nonlinear multigrid for imaging electrical conductivity and permittivity at low frequency *Inverse Problems* **17** 329–59

[6] Borcea L, Berryman J G and Papanicolaou G C 1996 High-contrast impedance tomography *Inverse Problems* **12** 835–58

[7] Brandt A 1984 *Multigrid Techniques: 1984 Guide with Applications to Fluid Dynamics* (Rehovot, Israel: The Weizmann Institute of Science)

[8] Chan T and Tai X 1997 Augmented Lagrangian and total variation methods for recovering discontinuous coefficients from elliptic equations *UCLA Technical Report* TR 97-2

[9] Chan T F, Golub G H and Mulet P 1999 A nonlinear primal-dual method for total variation-based image restoration *SIAM J. Sci. Comput.* **20** 1964–77

[10] Chavent G 1989 A non-linear least-square theory for inverse problems *Inverse Methods in Action* ed P Sabatier (Berlin: Springer)

[11] Chavent G and Liu J 1989 Multiscale parametrization for the estimation of a diffusion coefficient in elliptic and parabolic problems *Proc. IFAC Symp. on Control of Distributed Parameter Systems*

[12] Claerbout J 1985 *Imaging the Earth's Interior* (Oxford: Blackwell)

[13] Dennis J E, Heinkenschloss M and Vicente L N 1998 Trust-region interior-point SQP algorithms for a class of nonlinear programming problems *SIAM J. Control Optim.* **36** 1750–94

[14] Deuflhard P and Potra F 1992 Asymptotic mesh independence of Newton–Galerkin methods via a refined Mysovskii theorem *SIAM J. Numer. Anal.* **29** 1395–412

[15] Devaney A J 1989 The limited-view problem in diffraction tomography *Inverse Problems* **5** 510–23

[16] Ewing R (ed) 1983 *The Mathematics of Reservoir Simulation* (Philadelphia: SIAM)

[17] Engl H W, Hanke M and Neubauer A 1996 *Regularization of Inverse Problems* (Dordrecht: Kluwer)

[18] Farquharson C and Oldenburg D 1998 Non-linear inversion using general measures of data misfit and model structure *Geophys. J.* **134** 213–27

[19] Gomez S, Perez A and Alvarez R 1998 Multiscale optimization for aquifer parameter identification with noisy data *Computational Methods in Water Resources XII* vol 2

[20] Haber E 1997 Numerical strategies for the solution of inverse problems *PhD Thesis* University of British Columbia

[21] Haber E, Ascher U, Aruliah D and Oldenburg D 2000 Fast simulation of 3D electromagnetic using potentials *J. Comput. Phys.* **163** 150–71

[22] Haber E, Ascher U and Oldenburg D 1999 Solution of the 3D electromagntic inverse problem *3rd Int. Symp. on 3D Electromagnetics (Salt Lake City, Oct. 1999)*

[23] Haber E, Ascher U and Oldenburg D 2000 On optimization techniques for solving nonlinear inverse problems *Inverse Problems* **16** 1263–80

[24] Haber E and Oldenburg D 2000 A GCV based method for nonlinear inverse problem *Comput. Geosci.* **4**

[25] Hackbusch W 1985 *Multi-Grid Methods and Applications* (Berlin: Springer)

[26] Hanke M 1997 Regularizing properties of a truncated Newton-cg algorithm for nonlinear inverse problems *Numer. Funct. Anal. Optim.* **18** 971–93

[27] Hanke M and Hansen P C 1993 Regularization methods for large scale problems *Surv. Math. Industry* **3** 253–315

[28] Hansen P C 1998 *Rank Deficient and Ill-Posed Problems* (Philadelphia: SIAM)

[29] Heinkenschloss M 1993 Mesh independence for nonlinear least squares problems with norm constraints *SIAM J. Optim.* **3** 81–117

[30] Henson V, Limber M, McCormick S and Robinson B 1996 Multilevel image reconstruction with natural pixels *SIAM J. Sci. Comput.* **17** 193–216

[31] Huber P J 1964 Robust estimation of a location parameter *Ann. Math. Stat.* **35** 73–101

[32] Ito K and Kunisch K 1990 The augmented lagrangian method for parameter estimation in elliptic systems *SIAM J. Control Optim.* **28** 113–36

[33] Kelley C T 1995 *Iterative Methods for Linear and Nonlinear Equations* (Philadelphia: SIAM)

[34] Kelley C T 1999 *Iterative Methods for Optimization* (Philadelphia: SIAM)

[35] Liu J 1993 A multiresolution method for distributed parameter estimation *SIAM J. Sci. Comput.* **14** 389–405

[36] Menke W 1989 *Geophysical Data Analysis: Discrete Inverse Theory* (San Diego: Academic)

[37] Morozov V A 1984 *Methods for Solving Incorrectly Posed Problems* (Berlin: Springer)

[38] Nash S G 1999 A multigrid approach to discretized optimization problems *Manuscript*

[39] Newman G A and Alumbaugh D L 2000 Three-dimensional magnetotelluric inversion using nonlinear conjugate gradients *Geophys. J. Int.* **140** 410–8

[40] Nocedal J and Wright S 1999 *Numerical Optimization* (New York: Springer)

[41] Oldenburg D W and Li Y 1994 Inversion of induced polarization data *Geophysics* **59** 1327–41

[42] Parker R L 1994 *Geophysical Inverse Theory* (Princeton, NJ: Princeton University Press)

[43] Pernice M and Walker H 1998 Nitsol: a Newton iterative solver for nonlinear systems *SISC* **19** 302–18

[44] Rieder A 1999 On the regularization of nonlinear ill-posed problems via inexact Newton iterations *Inverse Problems* **15** 309–27

[45] Scales J 1999 Introduction to inverse problems webpage http://landau.Mines.EDU/ jscales/gp605/

[46] Schulz V 1996 Reduced SQP methods for large scale optimal control problems in DAEs with application to path planning problems for satellite mounted robots *PhD Thesis* University of Heidelberg

[47] Schulz V, Bardossy A and Helmig R 1999 Conditional statistical inverse modeling in groundwater flow by multigrid methods *Comput. Geosci.* **3** 49–68

[48] Smith N C and Vozoff K 1984 Two-dimensional DC resistivity inversion for dipole dipole data *IEEE Trans. Geosci. Remote Sens.* **22** 21–8 (special issue on electromagnetic methods in applied geophysics)

[49] Tikhonov A N and Arsenin V Ya 1977 *Methods for Solving Ill-posed Problems* (New York: Wiley)

[50] Wahba G 1990 *Spline Models for Observational Data* (Philadelphia: SIAM)

[51] Watson L 1986 Numerical linear algebra aspects of globally convergent homotopy methods *SIAM Rev.* **28** 529–45

[52] Wesseling P 1992 *An Introduction to Multigrid Methods* (Chichester: Wiley)