Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computer Physics Communications 153 (2003) 59–65
Computer Physics Communications
www.elsevier.com/locate/cpc

Iteratively regularized Gauss–Newton method for bound-constraint problems in atmospheric remote sensing

Adrian Doicu *, Franz Schreier, Michael Hess

DLR–Remote Sensing Technology Institute, Oberpfaffenhofen, Wessling, Germany

Received 10 October 2002

Abstract

In this paper two algorithms for the solution of nonlinear ill-posed problems with simple bounds on the variables are presented. The proposed algorithms are bound-constraint versions of the iteratively regularized Gauss–Newton method. The numerical performances of the algorithms are studied by means of simulations concerning the retrieval of molecular concentrations from limb sounding observations. For these examples, the unconstrained algorithm leads to unreasonable solutions.

1. Introduction

One of the most efficient regularization methods for nonlinear ill-posed problems is the iteratively regularized Gauss–Newton method. This method can be regarded as a Tikhonov regularization with a variable regularization parameter. The iteratively regularized Gauss–Newton method was first studied by Bakushinskii [1]. Convergence results for solving nonlinear ill-posed problems were given by Blaschke et al. [2], Hohage [3] and Deuflhard et al. [4]. The performances of the algorithm on the choice of the regularization matrices and sequences of regularization parameters for atmospheric remote sensing were discussed by Doicu et al. [5].

However, in some applications the iteratively regularized Gauss–Newton method may lead to unrealistic solutions. This is the case in atmospheric remote sensing by means of high resolution spectroscopy, when the iterative process leads to gas concentrations with negative values. In fact, if a negative solution occurs, the forward model used to compute the new iterate fails. In this context, the solution of a bound-constraint inversion problem is justified. In the present paper, we discuss two extensions of the iteratively regularized Gauss–Newton algorithm for the solution of nonlinear ill-posed problems with simple bounds on the variables. The benefits of using a bound-constraint algorithm stem from two observations about most practical problems. Firstly, the restriction on the expected size of the variables is frequently encountered in atmospheric remote sensing. Secondly, even if no bounds are active at the solution, their presence can prevent the function from being evaluated at unreasonable or non-physical points during the iterations.

In Section 2 we describe the basic unconstrained algorithm for atmospheric remote sensing and outline

* Corresponding author.
E-mail address: adrian.doicu@dlr.de (A. Doicu).

0010-4655/03/$ – see front matter
doi:10.1016/S0010-4655(03)00138-3

some of the concepts that will be used. The bound-constraint algorithms are discussed in Section 3. In Section 4 we illustrate these approaches by calculating the vertical profiles of molecular concentrations from limb sounding observations.

2. Basic unconstrained algorithm for the retrieval problem

The retrieval problem in atmospheric remote sensing is to extract vertical profiles of atmospheric state parameters from spectral radiance measurements. The relationship between the atmospheric state parameters and the spectral radiance is given by the radiative transfer equation. The discrete form of the radiative transfer equation can be written as

$$y = F(x), \tag{1}$$

where the mapping $F:\mathbb{R}^n \to \mathbb{R}^m$ is the radiative transfer model, $x \in \mathbb{R}^n$ is the state vector containing the atmospheric parameters (temperature or molecular density profiles) to be retrieved and $y \in \mathbb{R}^m$ is the exact data vector containing the spectral radiances (at a finite number of typically equidistant wavenumbers) measured by a "perfect" instrument. Here $\mathbb{R}^n$ stands for the n-dimensional real Euclidean space with the usual inner product $\langle x, y\rangle = x^{\mathrm{T}}y$, while $\|\cdot\|$ denotes the l_2 vector norm and the subordinated l_2 matrix norm. The width of the grid on which the atmospheric state parameters are represented depends on the altitude resolution and representational errors, while the spectral interval is given by the characteristics of the spectroscopic instrument. In our analysis we assume that the exact data are attainable, i.e. that there exists the exact solution $\hat{x}$ such that $y = F(\hat{x})$. Measurements are made to a finite accuracy and in practice only the noise data vector y^δ,

$$y^\delta = y + \delta, \tag{2}$$

is available. In the present analysis we consider a semi-stochastic data model in the sense that the exact solution $\hat{x}$ is deterministic but the measurement error δ is stochastic with zero mean and the covariance matrix S_δ, $S_\delta = \sigma^2 I$, where I is the identity matrix.

While the forward model maps the state space into the measurement space, we are interested in the inverse mapping, or equivalently, in the determination of an appropriate estimate of the exact state vector $\hat{x}$. An estimate of the exact solution can be found by minimizing the output least squares function

$$\mathcal{F}(x) = \tfrac{1}{2}\left\|F(x) - y^\delta\right\|^2 \tag{3}$$

possibly by an iterative method.

In the framework of the iteratively regularized Gauss–Newton method one considers the augmented objective function

$$\begin{aligned}\mathcal{F}_k(x) &= \tfrac{1}{2}\left\|r(x)\right\|^2 \\ &= \tfrac{1}{2}\left[\left\|F(x) - y^\delta\right\|^2 + \alpha_k\left\|L(x - x_a)\right\|^2\right],\end{aligned} \tag{4}$$

where L is some regularization matrix, (α_k) is a monotonically decreasing sequence and x_a is the a priori state vector, the best beforehand estimate of $\hat{x}$. The generalized residual vector

$$r(x) = \begin{bmatrix} F(x) - y^\delta \\ \sqrt{\alpha_k}\,L(x - x_a) \end{bmatrix}$$

is a mapping $r:\mathbb{R}^n \to \mathbb{R}^{m+p}$ for $L \in \mathbb{R}^{p\times n}$. The new iterate is given by

$$x_{k+1}^\delta = x_k^\delta + p_k, \tag{5}$$

where p_k is the solution of the unconstrained subproblem:

$$\min_{p\in\mathbb{R}^n}\left(r^{\mathrm{T}}\left(x_k^\delta\right)J_r\left(x_k^\delta\right)p + \tfrac{1}{2}p^{\mathrm{T}}J_r\left(x_k^\delta\right)^{\mathrm{T}}J_r\left(x_k^\delta\right)p\right), \tag{6}$$

and $J_r(x) = r'(x)$ denotes the Jacobian matrix of r evaluated at x. The iteration is stopped accordingly to the discrepancy principle, that is, at the first index $k_* = k_*(\Delta)$ for which

$$\begin{aligned}&\left\|F\left(x_{k_*}^\delta\right) - y^\delta\right\| \leqslant \tau\Delta < \left\|F\left(x_k^\delta\right) - y^\delta\right\|, \\ &\quad 0 \leqslant k < k_*,\end{aligned} \tag{7}$$

where $\tau > 1$ and Δ is an upper bound for the error, $\|\delta\| \leqslant \Delta$. In practice, this bound can be chosen as the expected value of $\|\delta\|$, i.e., $\Delta = \sqrt{\mathcal{E}\{\|\delta\|^2\}} = \sigma\sqrt{m}$, where $\mathcal{E}$ is the expected value operator.

The regularization matrix L is typically either the identity matrix ($L = L_0 = I$), a discrete approximation to the first ($L = L_1$) or second ($L = L_2$) derivative operator or some approximation to the a priori covariance matrix [5]. Information about the magnitude and smoothness of the state vector can be simultaneously taken into account by combining several derivative orders. The sequence of regularization parameters (α_k) can be constructed by using the noise-level

criterion or the L-curve criterion for each linear subproblem. The second parameter choice method allows enough regularization to be applied at the beginning of iterations and then to be gradually decreased. For more details we refer to [5].

The process of finding the new iterate can be summarized as follows:

Compute the Gauss–Newton direction p_k by solving the unconstrained subproblem (6) and take $x_{k+1}^{\delta} = x_k^{\delta} + p_k$.

Since the iteratively regularized Gauss–Newton method is a version of the Tikhonov regularization with variable α, Eriksson [6] computes the new iterate by using a step-length algorithm, i.e.

Compute the Gauss–Newton direction p_k by solving the unconstrained subproblem (6) and determine a positive step-length a_k for which it holds that $\mathcal{F}_k(x_k^{\delta} + a_k p_k) < \mathcal{F}_k(x_k^{\delta})$. Take $x_{k+1}^{\delta} = x_k^{\delta} + a_k p_k$.

3. Modified bound-constraint algorithms

When the constraints are simple bounds on the variables, so that the minimization problem is

$$\begin{cases} \min_{x\in\mathbb{R}^n} \ \mathcal{F}(x) = \frac{1}{2}\|F(x) - y^{\delta}\|^2, \\ \text{subject to the simple bounds } l \leqslant x \leqslant u, \end{cases} \tag{8}$$

it is reasonable to preserve feasibility, i.e. to consider only iterates x_k^{δ} that satisfy $l \leqslant x_k^{\delta} \leqslant u$.

The first algorithm we propose is a simplified version of an active-set algorithm for well-posed problems [7]. In order to retain feasibility of the next iterate we impose that the step length does not violate any bound. The process of computing the new iterate then takes the form:

Compute the Gauss–Newton direction p_k by solving the unconstrained subproblem (6). Compute $\bar{a}$, the maximum non-negative feasible step along p_k, i.e. $l \leqslant x_k^{\delta} + a p_k \leqslant u$ for all a with $0 < a \leqslant \bar{a}$. Determine a positive step-length a_k for which it holds that $\mathcal{F}_k(x_k^{\delta} + a_k p_k) < \mathcal{F}_k(x_k^{\delta})$ and $a_k \leqslant \xi\bar{a}$, where $\xi < 1$. Take $x_{k+1}^{\delta} = x_k^{\delta} + a_k p_k$.

Certainly it may happen that no such step-length a_k exists; in this case a_k is taken as $\xi\bar{a}$, even though $\xi\bar{a}$ does not satisfy the criteria usually required of a step length for unconstrained optimization. At the next iteration, the point $x_k^{\delta} + \xi\bar{a}\,p_k$ can be regarded as initial guess for the new objective function (4). The choice $\xi < 1$ guarantees that the new iterate is in the interior of the feasible region. In fact the variables are not fixed on their bounds, that is, the variables remain free during the iterative process. This algorithm will be referred to as the bound-constraint algorithm A.

The second algorithm relies on the use of a collection of optimization routines contained in the public part of the PORT library. For our purpose we use the routine DRN2GB which is the double-precision version of an algorithm for solving nonlinear least-squares problems with simple bounds [8]. The algorithm is a trust-region method with a local active-set strategy to select the step, while the active set is made afresh every iteration. The routine works by "reverse communication". With reverse communication the caller invokes the minimizing routine with an approximate minimization point and the values of the residual function and of the Jacobian matrix at that point. The minimizing routine returns a new approximate solution at which the user should evaluate the residual function and the Jacobian matrix and call the minimizing routine again. In essence, the method provides x_{k+1} as $x_k + p_k$, where p_k is the solution of the trust region subproblem:

$$\begin{cases} \min_{p\in\mathbb{R}^n} \ r^{\mathrm{T}}(x_k) J_r(x_k)\, p + \frac{1}{2} p^{\mathrm{T}} G(x_k)\, p \\ \text{subject to the simple bounds } l \leqslant x \leqslant u \\ \text{and the trust region constraint } \|p\| \leqslant \rho_k. \end{cases} \tag{9}$$

Here ρ_k is the radius of the trust region and $G(x_k)$ is some approximation to the Hessian of $\mathcal{F}$ computed at x_k. G can be computed in the framework of a Gauss–Newton model as $G(x_k) = J_r(x_k)^{\mathrm{T}} J_r(x_k)$, or in the framework of a Quasi-Newton model as $G(x_k) = J_r(x_k)^{\mathrm{T}} J_r(x_k) + S_k$, where S_k is a secant approximation to the second-order part of the least-squares Hessian [9]. The algorithm starts with $S_0 = 0$. With this choice, the first iteration is equivalent to an iteration of the Gauss–Newton method.

In this context the step of computing the new iterate is:

Compute the new iterate x_{k+1}^{δ} by a single call of the subroutine DRN2GB.

Since the DRN2GB is invoked a single time, $G(x_k^{\delta}) = J_r(x_k^{\delta})^{\mathrm{T}} J_r(x_k^{\delta})$ and the radius of the trust region is the Gauss–Newton step. If the Gauss–Newton step is too long the trust region is shrunk repeatedly to obtain an acceptable feasible x_{k+1}^{δ}. An acceptable x_{k+1}^{δ} means at least a feasible solution for which the objective function $\mathcal{F}_k$ decreases. In contrast to the previous algorithm the variables can be fixed on their bounds during the iterative process. This algorithm will be referred to as the bound-constraint algorithm B.

The new algorithms maintain the peculiarities of the unconstrained version of the iteratively regularized Gauss–Newton method: a descending sequence of regularization parameters and the use of the discrepancy principle as an a posteriori stopping rule.

4. Numerical simulations

In our numerical simulations we consider the Michelson Interferometer for Passive Atmospheric Sounding (MIPAS) that has been designed to measure the Earth's atmospheric composition with respect to a large number of species. MIPAS is one of the instruments on the ESA's Environmental Satellite (Envisat) which was launched successfully into its orbit on 1 March 2002. The MIPAS instrument provides information about temperature, ozone (O_3), nitrogen family (NO_2, HNO_3, N_2O, etc.), dynamic tracers (H_2O, CH_4) and other species.

In limb sounding the line of sight of a space-borne instrument is oriented at the Earth's limb. A sequence of observations with different angles corresponding to different tangent altitudes can be used for the retrieval of gas concentration as a function of altitude. The geometry of the atmospheric remote sensing by a limb sounding instrument is shown in Fig. 1.

Although the retrieval grid can be chosen independently of the tangent grid a close relation between the resolution of the retrieved profile and the tangent altitude spacing is expected. Scanning the atmosphere with smaller altitude steps might improve the vertical resolution, but due to the finite field of view of the instrument the contributions of adjacent atmospheric layers are more correlated and consequently the Jacobian matrix becomes more ill-conditioned. An equidistant retrieval grid with a spacing of 3 km is considered between 12 km and 42 km, while a grid spacing of 5 km is considered between 42 km and 52 km. The tangent grid is assumed to be identical to the retrieval grid. The field of view of the instrument is a symmetric trapezium with 1.4 km and 2 km base lengths.

In our first example we consider the retrieval of NO_2 in the spectral interval between 1600.50 and 1601.40 cm^{-1}. This corresponds to Channel C of the MIPAS instrument. A number of 37 equidistant data points were chosen in this spectral interval. Because H_2O, O_2 and CH_4 are dominant in the observed spectral region, no other gases were considered. The exact gas profile $\hat{x}$ was taken from the U.S. standard atmosphere. The a priori and initial gas profiles x_a and x_0^{δ}, respectively, were assumed to be identical and were chosen as a constant profile, i.e. $x_a = x_0^{\delta} = 0.006$ ppmv. For the exact gas profile

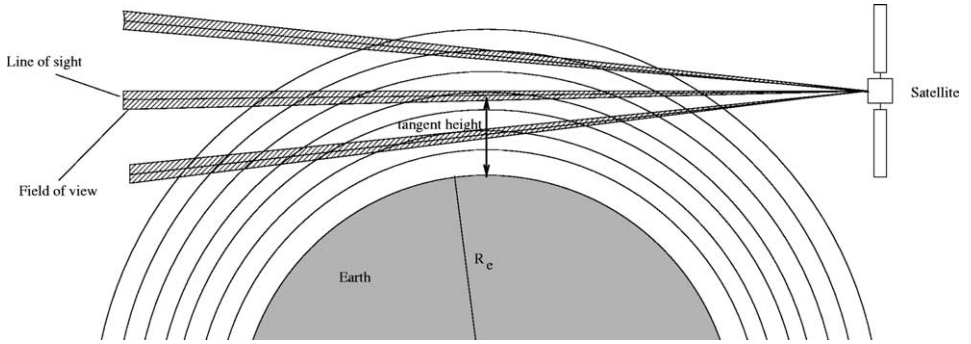

Fig. 1. Geometry of atmospheric remote sensing by a limb sounding instrument.

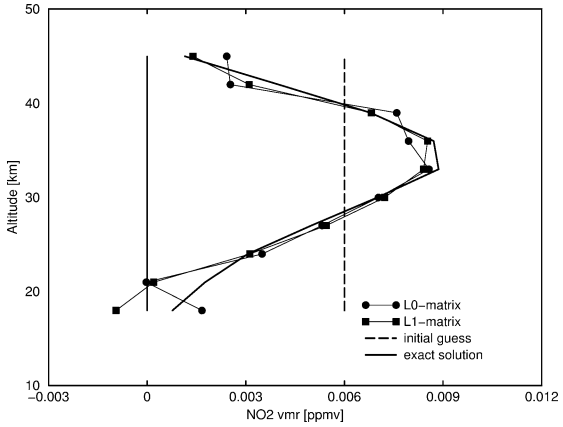

Fig. 2. Result of NO_2 retrieval using the unconstrained algorithm.

a contaminated spectrum was generated. The noise δ was described by a Gaussian probability distribution with zero mean and covariance matrix $S_\delta = \sigma^2 I$. We choose $\sigma = 4$ nW/(cm^2 sr cm^{-1}). In Fig. 2 the inversion results of the unconstrained algorithm are plotted. The profiles correspond to the L_0 and L_1 regularization matrices. In both situations the gas concentration has negative values. Note that the unconstrained algorithm is stopped as soon as negative solutions occur. The results in Fig. 3 are obtained by using the bound-constraint algorithms. For the bound-constraint algorithm A we imposed $x \geqslant l$, while for the bound-constraint algorithm B we imposed $l \leqslant x \leqslant u$, where $l = 0.0001 x_a$ and $u = 100 x_a$. Both algorithms lead to realistic solutions with comparable accuracy. The relative errors $\varepsilon = \|\hat{x} - x_{k_*}^\delta\|/\|\hat{x}\|$ decrease from 57 to 10% for the L_0 regularization matrix and from 57 to 6% for the L_1 regularization matrix.

Next we consider the retrieval of N_2O. For this simulation the intensity spectrum between 1270.825 and 1272.025 cm^{-1} was analyzed. This spectral interval corresponds to Channel B of the MIPAS instrument. The number of equidistant data points for each spectrum was 49. H_2O, CH_4 and HOCl were considered as active gases. The exact gas profile $\hat{x}$ corresponds to the U.S. standard atmosphere, while the a priori and initial gas profiles were chosen as $x_a = x_0^\delta = 0.15$ ppmv. The standard deviation of the added noise was $\sigma = 12$ nW/(cm^2 sr cm^{-1}). In Figs. 4 and 5 we plot the inversion results obtained by using the unconstrained and the bound-constraint algorithms. If no constraints on the variables are imposed,

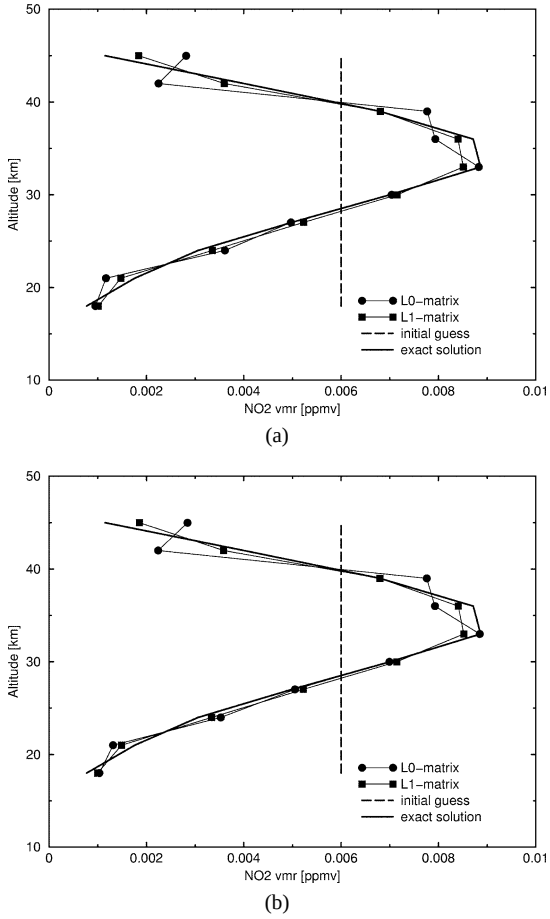

Fig. 3. Result of NO_2 retrieval using: (a) the bound-constraint algorithm A and (b) the bound-constraint algorithm B.

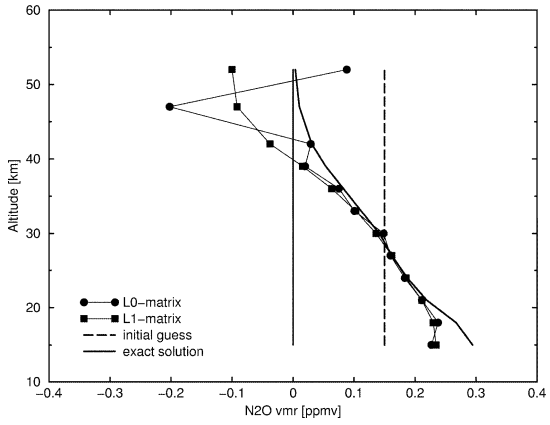

Fig. 4. Result of N_2O retrieval using the unconstrained algorithm.

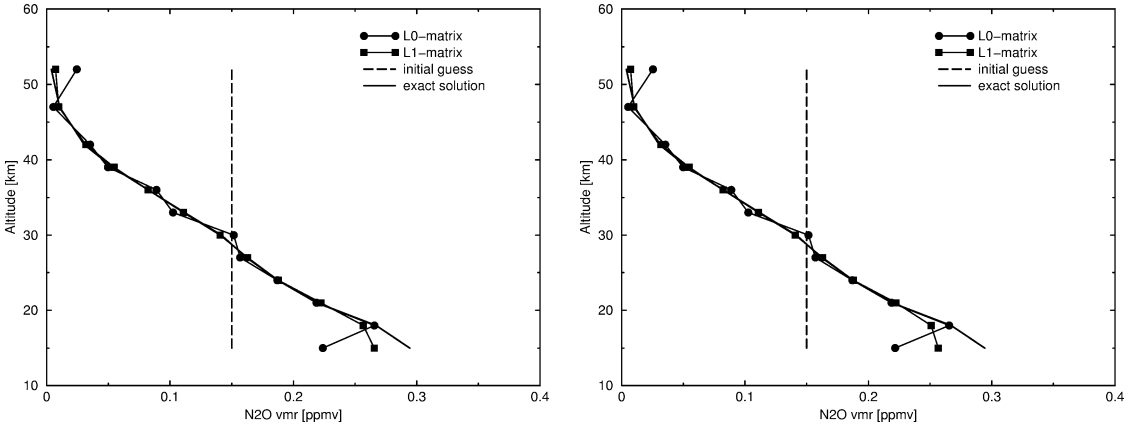

Fig. 5. Result of N_2O retrieval using: (a) the bound-constraint algorithm A and (b) the bound-constraint algorithm B.

Table 1
History of regularization parameters α_k, relative residuals $\mathcal{F}(x_k^\delta)/(m\sigma^2)$ and relative errors ε_k for the retrieval of N_2O using the bound-constraint algorithm B with the L_1 regularization matrix

k	α_k	$\mathcal{F}(x_k^\delta)/(m\sigma^2)$	ϵ_k
0	7.34	59.5752	59.90
1	7.09	16.1443	17.19
2	6.64	3.7312	14.46
3	6.27	0.6813	8.99
4	5.98	0.5524	7.10
5	5.56	0.5130	6.30
6	5.07	0.5042	5.806
7	5.02	0.5037	5.803
8	–	0.5035	5.802

negative solutions occur. The bound-constraint algorithms have similar inversion performances and lead to solutions with acceptable accuracy. The relative errors for the bound-constraint algorithms decrease from 60 to 11% for the L_0 regularization matrix and from 60 to 6% for the L_1 regularization matrix. Table 1 shows the history of regularization parameters α_k, relative residuals $\mathcal{F}(x_k^\delta)/(m\sigma^2)$ and relative errors ε_k for the bound-constraint algorithm B with the L_1 regularization matrix. These results serve as an evidence for the convergence of the sequence of regularization parameters and the convergence of the relative residual to 0.5 (cf. Eqs. (3) and (7)).

We want to point out that we can also prevent the occurrence of negative solutions by an adequate choice of the initial guess. If the initial guess is "sufficiently" close to the exact solution, the unconstrained algorithm may lead to correct solutions. However, this strategy is time consuming and restricts the capabilities of the unconstrained algorithm to handle unexpected situations. This aspect is especially important in case of automated operational data processing of large amounts of satellite data.

Our numerical analysis indicates that the bound-constraint algorithm B is more time consuming than the bound-constraint algorithm A. The explanation lies in the fact that the model and the trust radius selection requires more function evaluations than the step-length procedure.

5. Conclusions

In the present paper two implementations of the iteratively regularized Gauss–Newton algorithm for the solution of bound-constraint problems arising in remote sensing are discussed. The first algorithm uses a "constrained" step-length procedure to compute the new iterate. The second algorithm uses the optimization routines DRN2GB from the PORT library to minimize the quadratic function subject to simple bounds on the variables. In order to cope with the ill-posedness of the problem, a decreasing sequence of regularization parameters and the discrepancy principle as an a posteriori stopping rule are used. Whether the proposed algorithms are regularization methods in the sense of [10] remains an open question. However, the inversion performances of the algorithms are accept-

able, at least for the examples of gas concentration retrieval considered in Section 4. The retrieval of N_2O and NO_2 from limb sounding observations are sufficiently accurate for an initial guess lying far away from the exact solution. These encouraging results suggest that the present approach is suitable for inversion of the radiative transfer equation to analyze limb sounding measurements.

References

[1] A.B. Bakushinskii, The problem of the convergence of the iteratively regularized Gauss–Newton method, Comput. Math. Math. Phys. 32 (1992) 1353–1359.
[2] B. Blaschke, A. Neubauer, O. Scherzer, On convergence rates for the iteratively regularized Gauss–Newton method, IMA J. Numer. Anal. 17 (1997) 421–436.
[3] T. Hohage, Logarithmic convergence rates of the iteratively regularized Gauss–Newton method for an inverse potential and an inverse scattering problem, Inverse Problems 13 (1997) 1279–1300.
[4] P. Deuflhard, H.W. Engl, O. Scherzer, A convergence analysis of iterative methods for the solution of nonlinear ill-posed problems under affinely invariant conditions, Inverse Problems 14 (1998) 1081–1106.
[5] A. Doicu, F. Schreier, M. Hess, Iteratively regularized Gauss–Newton method for atmospheric remote sensing, Comput. Phys. Commun. 148 (2002) 214–226.
[6] J. Eriksson, Optimization and regularization of nonlinear least squares problems. Ph.D. Thesis, Department of Computing Science, Umea University, Sweden, 1996.
[7] P.E. Gill, W. Murray, M.H. Wright, Practical Optimization, Academic Press, London, 1981.
[8] D.M. Gay, A trust-region approach to linearly constrained optimization, in: Numerical Analysis, Proceedings, Dundee 1983, in: D.F. Griffiths (Ed.), Lecture Notes in Math., Springer, Berlin, 1984.
[9] J.E. Dennis, D.M. Gay, R.E. Welsch, An adaptive nonlinear least-squares algorithm, ACM Trans. Math. Software 7 (1981) 348–368.
[10] H.W. Engl, M. Hanke, A. Neubauer, Regularization of Inverse Problems, Dordrecht, Kluver, 1996.