Use of a Maximum Entropy Method as a Regularization Technique during the Retrieval of Trace Gas Profiles from Limb Sounding Measurements

J. Steinwagner

Max Planck Institute for Nuclear Physics, Heidelberg, Germany

G. Schwarz and S. Hilgers

German Aerospace Center (DLR), Oberpfaffenhofen, Germany

(Manuscript received 7 October 2005, in final form 19 April 2006)

ABSTRACT

The retrieval of trace gas profiles from radiance measurements of limb sounding instruments represents an inverse problem: vertical profiles of mixing ratios have to be extracted from sequences of horizontally measured radiances recorded by a spectrometer. Typically, these retrievals are plagued by random noise and systematic errors, necessitating the use of regularization techniques during inversion calculations. In the following, the use of selected maximum entropy operators as a regularization tool is discussed and their performance with conventional optimal estimation and Tikhonov-type regularization techniques is compared. The main gain with the proposed maximum entropy operators is that no a priori knowledge is needed; a reasonable initial guess profile is fully sufficient. The approach is verified by using simulated data of the Michelson Interferometer for Passive Atmospheric Sounding (MIPAS) instrument, an infrared Fourier transform spectrometer flown on the European Envisat mission.

1. Introduction

Typical limb sounding spectrometers being flown on earth observation or planetary missions measure spectral radiances when scanning the limb above a selected location. A primary goal of these instruments is to provide vertical profiles of pressure, temperature, and/or trace gas profiles.

In general, the retrieval of vertical profiles represents an inverse problem, as one observes spectral radiances emitted and/or absorbed by the atmosphere along horizontal lines of sight of several thousand kilometers, whereas the required results are vertical profiles of atmospheric conditions or mixing ratios of species above the selected locations. Thus, the issue is to infer the requisite quantities from the given measurements where, on the one hand, we can often use a large number of spectral emission or absorption lines identifying a selected species, but on the other hand, the data and results are affected by measurement noise and systematic errors of our retrieval approach, for instance by imperfect knowledge of instrument characteristics and/or insufficient accuracy of spectroscopic parameters.

Thus, one has to cope with the potential ill-posedness of the retrieval problem, that is, the potential lack of a unique solution, and the ill-conditioning of the computations with nonnegligible impact of the measurement noise affecting the shape of the profiles to be retrieved. Hence, dedicated processing techniques have to be applied leading to stable solutions despite a potentially degraded quality of input data and processing parameters.

2. Retrieval and regularization approaches

The retrieval problem can be formulated, with vectors and matrices shown in boldface, in general as

\[ y^d = F(x), \]

where an atmospheric profile (i.e., the “state parameters”) \( x \in \mathbb{R}^n \) is sought, where \( y \in \mathbb{R}^m \) represents the mostly overdetermined—measurement vector including noise \( d \), and \( F \) denotes the relation between \( y^d \) and...
The problem in Eq. (1) can be solved by a least squares approach but, due to the potential difficulties mentioned above, the solution may not be stable, physically meaningful, and/or unique. However, the inclusion of a suitable constraint allows us to solve the constrained inverse problem

$$\min_{x}(\rho(y, F(x)) + c(x))$$  \hspace{1cm} (2)$$

Usually, $\rho$ is a mathematical norm like the two-norm ||·||. The goal of expression (2) is to find a profile vector $x$ for which the difference between the real measurement $y$ and a simulated measurement $F(x)$ becomes minimal when taking into account the constraint $c(x)$. Common examples for constraints are smoothness of profiles or a priori assumptions on natural variability.

A common solution is to apply iterative techniques minimizing the nonlinear least squares differences between measured and simulated (i.e., forward modeled) spectra together with additional regularization terms derived from the constraint $c(x)$. The iterations proceed until sufficient convergence has been reached. For each iteration step, the parameters of the simulated measurements $F$ will be updated, typically driven by linearized derivatives of the measured radiances. In most cases, the regularization terms will be derived from smoothness criteria. A general example of an approach to handle whole vertical profiles simultaneously is given below (cf. von Clarra mm et al. 2003; Carlotti 1998).

Here, the regularization term is based on a dedicated regularization matrix $R$

$$x_{i+1} = x_i + (K^T S^{-1} K + R + \lambda I)^{-1} \times (K^T S^{-1} (y - F(x_i)) - R(x_i - x_0)),$$  \hspace{1cm} (3)$$

where $x_i$ is iteration $i$ of the profile $x$ to be retrieved, $i$ the iteration index, $S_y$ a covariance matrix of the measurement, $R$ the regularization matrix (scheme-dependent), $\lambda$ a damping factor, $I$ the identity matrix, $y$ the sequence of measured data, $F(x_i)$ the simulated (i.e., forward modeled) measurements, $K$ a matrix of partial derivatives $\partial F(x)/\partial x$, and $x_0$ an a priori profile derived from climatological data.

This general equation may be shaped to specific application requirements by selecting appropriate regularization matrices, damping factors, and a priori profiles. This leads to an adaptation of Eq. (3) to optimal estimation, Tikhonov, or Levenberg–Marquardt schemes (Rodgers 2000).

In the case of optimal estimation, the regularization matrix is the inverse a priori covariance matrix; common Tikhonov techniques often use synthetic regularization matrices derived from difference operators and do not need an a priori profile nor a $\lambda$ damping term; in contrast, Levenberg–Marquardt techniques rely on a nonzero $\lambda$ damping term with $R$ being set to zero. For more details on how to derive a regularization matrix $R$ from a given constraint $c(x)$, see appendix A.

If there is nothing or little known about a reasonable a priori profile, one may rather use a regularization functional that does not depend on a priori knowledge. In addition, there may be features that are likely to be dampened by a traditional regularization scheme that, typically, aims at smoothing result profiles. In these cases, it is more suitable to choose a regularization scheme that accounts for the “importance” of potential small-scale profile features. This leads to the use of maximum entropy methods, as shown below.

3. Second differences maximum entropy

The constraint used for maximum entropy methods is based on Shannon’s information content (Shannon 1948): the maximum entropy method (MEM) finds the maximum gain in information content. In principle, one evaluates the information content of a given quantity prior to and after an improvement (i.e., an iteration step). As the information content can be interpreted as generalized entropy, iterations leading to the highest gain in information content are linked to maximum entropy. When, instead of single element scalars, multielement vectors like a height-resolved profile are being retrieved, the regularization derived from the constraint $c(x)$ searches for solutions with total gain in information content of the entire vector.

In the literature, one can find several alternative MEM implementations. The method used here is a generalization of the standard MEM approach: it seeks to maximize the entropy of the vector of second differences of the unknown profile. Hence, it is called MEM2. The approach tries to find maximum entropy values of the discrete formulation of a second derivative of the profile. This means that the retrieved profile will tend to follow its true variability (for details, see section 5d).

To this end, we have to express the constraint $c(x)$ in the form of Shannon’s missing information measure (for further details, see section 4):

$$c(x) = \gamma_{\alpha} S_{\alpha}(x) = -\gamma \sum_{i=1}^{N} q_i^{\alpha} \log q_i^{\alpha},$$  \hspace{1cm} (4)$$

where $\gamma_{\alpha} < 0$ is the scalar regularization parameter, $\alpha$ is defined in section 4, $q_i^{\alpha}$ is defined in Eq. (11), and $S_{\alpha}(x)$ is a generalized expression for Shannon’s missing information measure (see section 4).

This constraint has to be translated into a regulariza-
Table 1. Comparison of the MEM, optimal estimation, and Tikhonov schemes.

<table>
<thead>
<tr>
<th>Principle</th>
<th>MEM2</th>
<th>Optimal estimation</th>
<th>Tikhonov</th>
</tr>
</thead>
<tbody>
<tr>
<td>A priori knowledge</td>
<td>Aims at a uniform distribution of the second derivatives of a profile vector. This leads to a height-resolved profile when the best possible solution is determined.</td>
<td>Weights the importance of a priori knowledge about the atmospheric profile against the measurement information.</td>
<td>Aims at smoothing the atmospheric profile or the difference between an atmospheric profile and a priori knowledge by applying the discrete form of a 0th-, 1st-, 2nd-, or nth-order derivative.</td>
</tr>
<tr>
<td>Kind of a priori knowledge</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

| Constraint functional | \( c(x) = -y \sum_{q} \log q' \) cf. Eq. (4) | \( c(x) = \| x - x_{\text{obs}} \|^{2} \) | \( c(x) = \alpha L \| x - x_{u} \|^{2} \) ** |

| Linear constraint functional | No | Yes | Yes |

Finding a solution for Eq. (2) leads to

| Shape of functional when applied to Eqs. (3) or (5) | \( x_{k+1} = x_{k} + [2K^T S_{x}^{-1} K_{k} + \nabla (\nabla c(x_{k}))]^{-1} [2K^T S_{x}^{-1} y - \nabla F(x_{k}) - \nabla (\nabla c(x_{k}))] \) | \( x_{k+1} = x_{k} + [2K^T S_{x}^{-1} K_{k} + \nabla (\nabla c(x_{k}))]^{-1} [2K^T S_{x}^{-1} y - \nabla F(x_{k}) - \nabla (\nabla c(x_{k}))] \) | \( x_{k+1} = x_{k} + [2K^T S_{x}^{-1} K_{k} + \nabla (\nabla c(x_{k}))]^{-1} [2K^T S_{x}^{-1} y - \nabla F(x_{k}) - \nabla (\nabla c(x_{k}))] \) |

| (See also appendix B) | | | |

* A common method for determining the matrix \( S_{x} \) is to define noise-based values for the diagonal elements and to use exponentially decaying off-diagonal elements. Details can be found in Steck and von Clarmann (2001).

** The \( \alpha \) is the regularization parameter that controls the strength of the regularization. A method for choosing an optimal \( \alpha \) was given by Hansen (1992) using the L-curve criterion.

4. Derivation of MEM2 from MEM

For a better understanding of the MEM2 details it is useful to start with the general MEM expression and natural logarithms. A similar derivation can also be found in Ramos et al. (1999a). We start with the expression for Shannon’s missing information measure,

\[
S(x) = - \sum_{i=1}^{N} q_i \ln(q_i). \tag{8}
\]

Each \( q_i \) is defined as

\[
q_i = x_i / \sum_{i=1}^{N} x_i, \tag{9}
\]

where \( x_i \) is the \( i \)th profile element.

It can easily be shown that

\[
0 \leq S(x) \leq \ln(N), \tag{10}
\]

where \( N \) is the maximum number of profile elements. Equation (10) illustrates that \( S \) has both a well-defined minimum and a maximum.

Equation (8) can be generalized even further when \( q_i \) becomes a function of the profile elements. To distinguish these \( q_i \) from the ones in Eq. (9) we write

\[
q_i^* = p_i / \sum_{i=1}^{N} p_i. \tag{11}
\]

The \( p_i \) are the components of a discrete representation of the profile vector \( x \)

\[
p = \Delta^n x. \tag{12}
\]
with $\alpha = 0, 1, 2, \ldots$, and $\Delta^\alpha$ is a discrete difference operator. Setting $\alpha = 0$ yields the standard MEM formulation ($p_i = x_i$); $\alpha = 1$ corresponds to the case where $p_i = |x_{i+1} - x_i|$. For $\alpha = 2$ (which corresponds to MEM2) the relation between $p$ and $x$ is defined as

$$p = \Delta^2 x$$

(13)

with

$$i = 1, \ldots, N$$

and $x_0, x_{N+1} = 0$;

$x_{\text{min}}, x_{\text{max}}$ are used for proper scaling, and $\zeta \approx 10^{-15}$ is an auxiliary computational variable. Then we obtain a compact expression for the denominator in Eq. (11):

$$p = \sum_{i=1}^{N} p_i = -(x_1 + x_N) + N(2(x_{\text{max}} - x_{\text{min}}) + \zeta).$$

(14)

Using Eqs. (11)–(14) leads to the formulation of $c(x) = \gamma \sigma^2 S_a(x)$ as in Eq. (4), which is formally identical to Eq. (8) but where all quantities have a generalized meaning and a magnitude depending on the value of $\alpha$. A detailed calculation of the derivatives in Eqs. (6) and (7) for nonperiodic boundary conditions ($x_j$ not equal to $x_N$) under consideration of the generalizations described above is given in appendix B.

An analytical calculation of the important regularization parameter $\gamma$ is difficult. Hence, we determined $\gamma$ empirically. In cases without excessive systematic errors, the admissible $\gamma$ can be found by comparing the spectral residual after the final iteration with the assumed noise level of the measured spectra. (The same holds for the regularization parameter of the Tikhonov method as a prerequisite for a fair comparison of methods.)

An algorithm for the numerical determination of the $\gamma$ parameter is contained in Ramos et al. (1999b).

5. Comparative tests

a. The MIPAS instrument

The data used for our comparisons were simulated Michelson Interferometer for Passive Atmospheric Sounding (MIPAS: ESA 2000) measurements. MIPAS is a Fourier transform interferometer having a spectral resolution of 0.035 cm$^{-1}$ (unapodized) and designed to study the chemistry of the middle atmosphere by detecting trace gases in the midinfrared. It is flown on Envisat, a European environmental mission, in a sun-synchronous orbit (98° inclination, 101-min orbit period, 800-km orbit height). The MIPAS instrument is a backward-looking limb sounder. The tangent point of the line of sight is about 3000 km away from the satellite. This calls for precise pointing and accurate ray-path modeling.

A vertical scan from the top to the bottom of the atmosphere above a selected tangent point consists of up to 17 spectral measurements. The vertical step width between the measurements is 3 km at lower heights and increases within the stratosphere. Technically, the measurements can start from 6 km and reach up to 120 km. Each measurement comprises about 60 000 sample points. Thus, we have to limit the amount of spectral data used for each species to be retrieved.

Further details on MIPAS can be found in several upcoming publications about the Envisat mission and the MIPAS instrument.

b. Simulated data

As we had to compare the MEM2 algorithm in detail with conventional schemes, simulated data of various species with precisely known reference profiles were used to assess the performance of diverse retrieval algorithms. In particular, we used a MIPAS level 2 processing system described by Hilgers et al. (2001). The internal forward model used for all our simulations is described by Stiller (2000). This model, called the Karlsruhe Optimized and Precise Radiative Transfer Algorithm (KOPRA), performs precise line-by-line calculations with variable parameter settings (e.g., for ray tracing, geometric and spectral grids, absorption coefficients, line shapes, cross sections, continuum parameterizations, field-of-view parameters, etc.).

Our reference profiles were compiled with regular 1-km spacing. Then simulated measurements were generated for the height range from 8 up to 53 km with 1-km spacing (46 height steps). Since we knew the “real” solution in these simulations, we could determine a relative and absolute error for each subsequent retrieval performed on the simulated height grid. Because, apart from noise effects, the deviation from the ideal solution is influenced by the regularization, the absolute and relative errors are indicative of the goodness of the regularization.

All a priori profiles (required for optimal estimation retrievals) and initial guess profiles were taken from tables of standard atmospheres; identical a priori and initial guess profiles were selected.

The use of real MIPAS data for comparative retrieval tests is described by Steinwagner and Schwarz (2006). For the sake of simplicity, we skipped the use of these data in the case of maximum entropy methods.
As can be seen below, the characteristic features of MEM2 retrievals can clearly be derived from simulated measurements.

c. Selected retrieval criteria

For an analysis of the MEM2 performance, we selected the following quantitative criteria (Rodgers 2000; Steinwagner and Schwarz 2006):

- the relative root-mean-square (rel. rms):
  \[
  \text{rel. rms} = \sqrt{\sum_i \left( (x_i^k - x_{i,\text{rel}}) / x_{i,\text{ref}} \right)^2} / n
  \]  

- the absolute root-mean-square (abs. rms):
  \[
  \text{abs. rms} = \sqrt{\sum_i (x_i^k - x_{i,\text{rel}})^2} / n
  \]  

- the relative difference per height step (rel. diff):
  \[
  \text{rel. diff}_i = (x_i^k - x_{i,\text{rel}}) / x_{i,\text{ref}}
  \]  

- the height resolution per height step (height res.):
  \[
  \text{height res.}_i = \text{FWHM of normalized matrix } A \text{ rows}
  \]  

\[
A = [K_x^c S_y^{-1} K_x + \nabla(\nabla c(x_k))]^{-1} K_y^c S_y^{-1} K_x
\]  

- the degrees of freedom (dfs) describing the regularization strength:
  \[
  \text{dfs} = \text{tr}(A), \quad \text{and}
  \]  

- the height-resolved estimated standard deviation of the retrieval (esd):
  \[
  \text{esd} = \left[ \text{diag}\left( [K_x^c S_y^{-1} K_x + \nabla(\nabla c(x_k))]^{-1} \right) \right]^{1/2}
  \]  

where \(k\) is the number of iterations, \(i\) the height step index, \(n\) the number of height steps (where the retrieved profile value differs from the reference profile value), FWHM is the full width half maximum, and \(\text{tr}(A)\) is the trace of matrix \(A\).

The reader should be aware that the last three metrics depend on the selected constraining term \(c(x)\) and therefore on the regularization.

d. Typical retrieval results

We selected ozone as our preferred species to be shown here as it is well known for its global characteristics and its potential fine structures. Typical ozone profile retrieval results are depicted in Figs. 1–3. Table 2 summarizes the numerical results.

In Fig. 1 we show simulated and retrieved ozone profiles together with their initial guess profile. The zigzag patterns contained in the reference profiles symbolize artificial fine structures that should be preserved within the retrieved profiles. Figure 2 illustrates the height-resolved relative deviation of the retrieved profiles from the reference profile (for a general discussion of characteristic maximum entropy effects, see below).
Typical additive noise was added to the simulated spectra; both the initial guess profiles as well as the a priori profiles were shifted vertically by 5 km with respect to the reference profiles. This prevents the calculations from starting too close to the true result. The regularization parameter for MEM2 was set to \(6.5 \times 10^{-4}\); the values for the diagonal elements of the optimal estimation weighting matrix resulted from the assumed noise level and were set to 75. The method for how entries in the covariance matrix are correlated to explicit variations in ppmv is given in Steck (2002). The off-diagonal elements of the matrix were set to decrease exponentially with their distance from the main diagonal, for details about the correlation length and the roll-off definition, see Steck and von Clarmann (2001). The regularization parameter in the Tikhonov case is also noise driven and was set to 65 (we selected a first-order Tikhonov scheme).

Table 2. Overview of results for simulated ozone retrievals; terms first column as in section 5c.

<table>
<thead>
<tr>
<th></th>
<th>MEM2</th>
<th>Optimal estimation</th>
<th>Tikhonov</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean height res. (km)</td>
<td>1.0–2.5</td>
<td>3.0</td>
<td>3.0</td>
</tr>
<tr>
<td>dfs</td>
<td>24.3</td>
<td>19.5</td>
<td>19.9</td>
</tr>
<tr>
<td>rel. rms</td>
<td>55.0%</td>
<td>51.9%</td>
<td>50.8%</td>
</tr>
<tr>
<td>abs. rms (ppmv)</td>
<td>0.28</td>
<td>0.27</td>
<td>0.29</td>
</tr>
<tr>
<td>Max value of esd (ppmv)</td>
<td>−0.1</td>
<td>~0.06</td>
<td>~0.06</td>
</tr>
</tbody>
</table>

It is obvious from Figs. 1 and 2 that all methods basically preserve our ozone profile and its artificial zigzag patterns (all results were obtained with a maximum number of three iterations). When we analyze the local deviations, the differences between MEM2 and the classical methods (optimal estimation and Tikhonov) become more obvious. It seems that MEM2 yields better results within the region of high ozone mixing ratios between 20 and 45 km but exceeds the 10% limit we set (dashed vertical line in Fig. 2) at heights above 45 km, where we have lower ozone mixing ratios.

The classical methods yield averaged absolute deviations ranging from 0.27 ppmv (optimal estimation) to 0.29 (Tikhonov) appearing to be insignificant. When expressed as averaged relative deviations, more pronounced differences appear: 55.0% (MEM2) versus 51.9% (optimal estimation) and 50.8% (Tikhonov).

The number of degrees of freedom also shows characteristic MEM2 effects: 24.3 (MEM2), versus 19.5 (optimal estimation), and 19.9 (Tikhonov). These regularization strengths are still considered to be of the same order of magnitude, because—in a strict mathematical sense—the numbers from differently regularized retrieval results cannot be compared.

However, while the height resolution is of the same order for the classical approaches ranging from 2.5 to 3.5 km, MEM2 yields considerably better values from 1.0 to 2.5 km (cf. Table 2). This is one of the fundamental MEM2 results and one cannot surmise that a classical maximum entropy solution would lead to rather smooth profiles with inferior height resolution (note that we use a second differences approach).

Finally, Fig. 3 shows the height-resolved estimated standard deviations (esds). These again show obvious differences between the classical approaches and MEM2. The esds belonging to optimal estimation and Tikhonov are more or less constant over the whole height range with values around 0.06 ppmv. In contrast, the MEM2 esd figures show a much higher dynamic range and lie in the range between 0.02 and 0.1 ppmv. This lower quality compensates for the advantages of MEM2’s improved height resolution.

The general behavior of the retrieved profiles described above demonstrates the typical characteristics and expected performance of a maximum entropy approach. In principle, MEM searches for a solution taken from the maximum allowable volume of state space within the limits imposed by the assumed noise level of the measurements and the entropy-specific solution constraints. When we maximize the entropy of second (i.e., double) differences of a profile, we obtain a MEM2 solution within the allowed state space vol-
ume with maximum vertical variability of the profile (Urban 1996). This maximum vertical variability solution is equivalent to a profile with no externally driven vertical smoothing or smearing constraints (i.e., with full vertical resolution and, as a consequence, conspicuous preservation of fine details). Of course, the spanned state space volume depends on the noise level specified by the covariance matrix of the measurement [i.e., $S_y$ within Eq. (3)], but does not depend on any given climatological profile covariances. Thus, typical MEM2 profiles are noise dependent with an upper limit of fine structures on the verge of annoying overshoots. If minimum noise were added to the simulated spectra, one would obtain near-perfectly retrieved profiles; the basic convergence behavior of MEM2 is comparable to the other methods being described here and mainly depends on the initial guess profiles $x_0$ of Eq. (3).

This general behavior can also be studied when we retrieve profiles of other species. A typical example of an abundant species with strong emission lines is carbon dioxide being used for temperature profiling. In this case, the signal-to-noise ratio is rather high. As shown in Fig. 4, the retrieved temperature profile after the first iteration is still affected by the (excessively displaced) initial guess profile; however, there are virtually no signs of noise-induced artifacts and no smoothing effects due to overregularization. A counterexample is nitric acid, a species with much lower mixing ratio and weak noise-affected emission lines. Here we can see that the lack of any a priori information leads to imperfect profile shape retrievals below and above the trace gas layer when the useful signal approaches zero (see Fig. 5). In addition, the retrieved profile after the first iteration still represents a coarse fit between the initial guess and the true profile with noticeable deviations from the ideal result.

6. Conclusions

The proposed MEM2 technique is a promising retrieval alternative if no reliable a priori information is available that could be used in an optimal estimation approach and/or if unexpected profile features would limit the usefulness of synthetic $n$th order Tikhonov derivatives. Then MEM2 provides a robust tool for profile retrievals with high nominal height resolution compensated by somewhat lower profile accuracy. The initial guess profiles may deviate from the final results; however, we expect a faster convergence when appropriate initial guess profiles are used.

Finally, results yielded by MEM2 preretrievals could be used by subsequent optimal estimation or Tikhonov postretrievals using the MEM2 results as a priori profiles or for the selection of $n$th order Tikhonov derivatives.

Acknowledgments. J. S. acknowledges the financial support provided by DLR, Remote Sensing Technology Institute, Oberpfaffenhofen, Germany, during his time as a Ph.D. student. We thank three anonymous
convert for the noise (diagonal elements) and its correlation between samples (nondiagonal elements), as noisy measurements may not be independent of each other (e.g., after apodization).

To obtain the minimum of $M(x)$, we have to compute the first derivative $M_1(x) = \nabla(M(x))$ and set it to zero:

$$M_1(x) = 0. \quad (A2)$$

Using (A1) and (A2) and setting $K$ as the Jacobian matrix $\nabla_x F(x)$ we obtain

$$\nabla_1[(y' - F(x))^T S_y^{-1}(y' - F(x)) + c(x)] = (-2\nabla_x F(x))^T S_y^{-1}(y' - F(x)) + \nabla c(x)$$

$$= - 2K^T S_y^{-1}(y' - F(x)) + \nabla_x c(x)$$

$$= 0 = M_1(x). \quad (A3)$$

This implicit nonlinear problem can be solved by applying Newton’s method (Press et al. 2002). Let $k$ be the current iteration. Then we can write

$$x_{k+1} = x_k - [\nabla M_1(x_k)]^{-1} M_1(x_k). \quad (A4)$$

Here we need $\nabla M_1(x)$, that is, the second derivative of $M(x)$:

$$\nabla M_1(x) = -2K^T S_y^{-1}(y' - F(x)) - 2K^T S_y^{-1}(y') - 2K^T S_y^{-1} \nabla F(x) - \nabla c(x)$$

$$\approx -K^T S_y^{-1}(-K) + \nabla c(x) = K^T S_y^{-1}K + \nabla c(x). \quad (A5)$$

**APPENDIX A**

Conversion of Constraints

As mentioned in section 2, we need a more general formulation of Eq. (3) based on Eq. (2) showing explicitly how the constraint $c(x)$ is converted into the terms $\nabla(c(x_k))$ and $\nabla c(x_k)$ (Steinwagner and Schwarz 2006). When applying a scalar product rule, we obtain from Eq. (2)

$$M(x) = (y' - F(x))^T S_y^{-1}(y' - F(x)) + c(x), \quad (A1)$$

where $M(x)$ is a scalar expression to be minimized.

The noise contained in the measurements will be modeled by a weighting matrix $S_y$ accounting for the “importance” of the noise (diagonal elements) and its correlation between samples (nondiagonal elements), as noisy measurements may not be independent of each other (e.g., after apodization).

Fig. 5. Nitric acid profile retrieved from simulated data. Solid: retrieved profile (after the first iteration), dashed: reference profile, dotted: initial guess profile.
The first term of the second derivative has been omitted because its entries are supposed to be small and due to its nature as a matrix of matrices the gain in accuracy compared to the computing effort is considered negligible. Inserting Eqs. (A5) and (A3) into Eq. (A4) leads to

\[
x_{k+1} = x_k + [2K_x^{-1}K_y + \nabla c(x_k)]^{-1} \times [2K_x^{-1}(y^d - F(x_k)) - \nabla c(x_k)].
\]

(A6)

This is equivalent to Eq. (3) for \( \lambda = 0 \).

**APPENDIX B**

**First and Second Derivatives for MEM2**

In the following, the calculations of the first and second derivatives needed in Eqs. (6) and (7) are given in detail (vectors are boldface). We use the definitions given in section 4. Additionally, we define

\[
\xi_i = \frac{p_i}{p} \Rightarrow S(x) = -\sum_{i=1}^{N} \xi_i \log \xi_i.
\]

(BD)

a. First derivative

Using Eq. (BD) we can rewrite Eq. (6)

\[
\frac{dS}{dx_k} = \frac{dS}{\xi_1} \frac{\partial \xi_1}{\partial x_k} + \frac{dS}{\xi_2} \frac{\partial \xi_2}{\partial x_k} + \cdots + \frac{dS}{\xi_n} \frac{\partial \xi_n}{\partial x_k} = \sum_{i=1}^{N} \frac{\partial S}{\partial \xi_i} \frac{\partial \xi_i}{\partial x_k}.
\]

(BG1)

Hence, we must calculate the following expressions

\[
\frac{dS}{\xi_i},
\]

(B1)

\[
\frac{\partial \xi_i}{\partial x_k}.
\]

(B2)

For Eq. (B1) we obtain

\[
\frac{dS}{\xi_i} = -\log \xi_i - \xi_i \frac{1}{\xi_i} = -(1 + \log \xi_i),
\]

(B1a)

and for (B2)

\[
\frac{\partial \xi_i}{\partial x_k} = \frac{\partial}{\partial x_k} \left( \frac{p_i}{p} \right) = \frac{1}{p} \frac{\partial p_i}{\partial x_k} - \frac{p_i}{p^2} \frac{\partial p}{\partial x_k}.
\]

(B2a)

To calculate (B2a) we need

\[
\frac{\partial p_i}{\partial x_k},
\]

(B3)

\[
\frac{\partial p}{\partial x_k}.
\]

(B4)

Then we obtain for (B3)

\[
\frac{\partial p_i}{\partial x_k} = A_{ik}.
\]

(B3a)

This relation can be derived by using Eqs. (13) and (14). This leads to

\[
p_i = \sum_{k=1}^{N} A_{ik} x_k + C,
\]

where \( \frac{\partial p_i}{\partial x_k} = A_{ik} \), \( C = 2(x_{\text{max}} - x_{\text{min}}) + \epsilon \) is constant with respect to \( x_k \). From this we get

\[
A_{ik} = \begin{cases} 
-2 & \text{for } l = k \\
1 & \text{for } l + 1 = k \\
1 & \text{for } l - 1 = k \\
0 & \text{otherwise}.
\end{cases}
\]

(B3b)

For (B4) we find

\[
\frac{\partial p}{\partial x_k} = \sum_{i=1}^{N} \frac{\partial p_i}{\partial x_k} = \sum_{i=1}^{N} A_{ik} = p^*.
\]

(B4a)

The entries of \( p^* \) can then be determined by

\[
p^*_k = \begin{cases} 
-1, & k = 1, N \\
0, & k = 2, n - 1.
\end{cases}
\]

(B4b)

Using (B3a) and (B4a), we get for (B2a)

\[
\frac{\partial S}{\partial x_k} = \frac{\partial}{\partial x_k} \left[ \frac{1}{p} \sum_{i=1}^{N} A_{ik} - \frac{p_i}{p} \sum_{i=1}^{N} A_{ik} \right] = \frac{1}{p} A_{ik} - \frac{p_i}{p} \sum_{i=1}^{N} A_{ik} - A_{ik} = \frac{1}{p} A_{ik}.
\]

(B2b)

With (B1a) and (B2b) one obtains for \( \partial S/\partial x_k \)

\[
\frac{\partial S}{\partial x_k} = \sum_{i=1}^{N} \left( \frac{1}{p} \left[ 1 + \log \left( \frac{p_i}{p} \right) \right] \right) \left( A_{ik} - \frac{p_i}{p} p^*_k \right).
\]

(B5)

Or, more generally,

\[
\frac{\partial S}{\partial x_k} = \sum_{i=1}^{N} \frac{\partial S}{\partial \xi_i} \frac{\partial \xi_i}{\partial x_k} = \text{matrix} \times \text{vector} = \text{vector},
\]

(B5a)

that is, the explicit form of Eq. (6).
b. Second derivative

We also need the second derivative

$$\frac{\partial}{\partial x_n} \left( \frac{\partial S}{\partial x_k} \right) = \sum_{i=1}^{N} \left[ \frac{\partial}{\partial x_n} \left( \frac{\partial S}{\partial \xi_i} \right) \frac{\partial \xi_i}{\partial x_k} + \frac{\partial S}{\partial \xi_i} \frac{\partial}{\partial x_n} \left( \frac{\partial \xi_i}{\partial x_k} \right) \right]. \tag{BG2}$$

$$\delta \xi_i/\delta x_k$$ and $$\delta S/\delta \xi_i$$ were already calculated in (B2b) and (B1a) for the first derivative.

Therefore, we still need to determine the following expressions:

$$\frac{\partial}{\partial x_n} \left( \frac{\partial S}{\partial \xi_i} \right), \tag{B6}$$
$$\frac{\partial}{\partial x_n} \left( \frac{\partial \xi_i}{\partial x_k} \right). \tag{B7}$$

From (B6) together with Eq. (BG1) it follows that

$$\frac{\partial}{\partial x_n} \left( \frac{\partial S}{\partial \xi_i} \right) = \frac{\partial}{\partial x_n} \left[ \frac{1}{p} \frac{\partial p_i}{\partial x_n} \right] = \frac{1}{p} \frac{\partial p_i}{\partial x_k} \frac{\partial}{\partial x_k} \frac{1}{p} \frac{\partial p_i}{\partial x_k} + \frac{1}{\partial x_n} \frac{\partial}{\partial x_k} \frac{\partial^2 p_i}{\partial x_n \partial x_k} - \frac{1}{\partial x_n} \frac{\partial p_i}{\partial x_k} \frac{\partial}{\partial x_k} \frac{\partial p_i}{\partial x_k} + \frac{2}{\partial x_n} \frac{\partial}{\partial x_k} \frac{\partial^2 p_i}{\partial x_n \partial x_k}. \tag{B7a}$$

Some of the mixed terms cancel out:

$$p_i = \sum_{k=1}^{N} A_{ik} x_k + C \Rightarrow \frac{\partial p_i}{\partial x_k} = A_{ik} \Rightarrow \frac{\partial^2 p_i}{\partial x_n \partial x_k} = 0, \tag{B7b}$$
$$\frac{\partial}{\partial x_n} \left( \frac{\partial p}{\partial x_k} \right) = \frac{\partial}{\partial x_n} \left( \frac{\partial}{\partial x_k} \left( \sum_{i=1}^{N} p_i / \xi_i \right) \right) = \frac{\partial}{\partial x_n} \left( \sum_{i=1}^{N} p_i / \xi_i \right) = \frac{\partial}{\partial x_n} \left( \sum_{i=1}^{N} A_{ik} \right) = 0. \tag{B7c}$$

Here $$\mathbf{A}$$ is a matrix of constants.

For the second derivative we obtain the following expression, which is derived from (B6c)–(B2b) together with (B1a)–(B7a) and using (B7b) as well as (B7c):

$$\frac{\partial}{\partial x_n} \left( \frac{\partial S}{\partial x_k} \right) = \sum_{i=1}^{N} \left\{ \sum_{m=1}^{N} \left[ -\frac{p_i}{p_m} \log \frac{1}{p} A_{mn} - \frac{p_m}{p^2} \sum_{i=1}^{N} A_{im} \right] \left[ \frac{1}{p} A_{ik} - \frac{1}{p^2} \sum_{i=1}^{N} A_{ik} \right] \right\} \right\} \right\}, \tag{B8}$$

As in the case of the first derivative, this expression can also be simplified by canceling terms and applying the Kronecker notation, Eq. (B8) reduces to a sum and we finally obtain
This is the explicit form of Eq. (7).

\[
\frac{\partial}{\partial x_n} \left( \frac{\partial S}{\partial x_k} \right) = - \sum_{l=1}^{N} \left[ \frac{1}{p_l} \left( A_{ln} - \frac{p_l}{p_n} p_n^k \right) \right] \left( A_{lk} - \frac{p_l}{p_k} p_k^n \right) + \left( 1 + \log \frac{p_n}{p} \right) \frac{1}{p} \left( A_{lk} p_n^* - A_{ln} p_k^* \right) - \frac{2p_l}{p} p_n^k p_k^*.
\]

(B9)

REFERENCES


