Proceedings of the Inverse Problems from Theory to Applications Conference (IPTA 2014)

26–28 August 2014, @Bristol, UK

Alfred K Louis, Simon Arridge and Bill Rundell
Introduction

IPTA 2014 was held in Bristol at the science museum, At-Bristol on 26–28th August 2014. The conference represents a new forum to disseminate interdisciplinary inverse problems research and to facilitate increased engagement and collaboration within the international community.

The conference was organised and hosted by Inverse Problems, Institute of Physics Publishing – iopscience.org/ip.

The scientific committee organised a diverse scientific programme combining mathematical and experimental work with theoretical, numerical and practical approaches to solving inverse problems.

The proceedings collection includes proceedings papers from the conference plenary lectures and from mini symposia presentations. We hope that you enjoy reading the proceedings.

Scientific committee

Alfred K Louis, Saarland University, Germany
Simon Arridge, University College London, UK
Bill Rundell, Texas A and M University, United States

Chair of industrial liaison

Bill Lionheart, University of Manchester, UK

Organising committee

Leanne Mullen, IOP Publishing, UK
Claire Garland, Institute of Physics, UK
Christopher Wileman, IOP Publishing, UK
Edward Jost, IOP Publishing, UK
# Contents

## Introduction

## I Plenary Talks

1 Seismic inverse problems: recent developments in theory and practice

2 An active set algorithm for \( \ell^0 \) optimization with decaying solutions

3 Emergence of ideas in inverse theory

4 Regularization methods

## II Minisymposia

M2 Asymptotic Expansions

5 Reconstruction of inclusions in photothermal imaging

M3 Inverse Source Problems

6 Reconstruction of refractive index from blind experimental data using approximate globally convergent and adaptive finite element methods

7 An inverse problem for the heat equation in view of practical application

M4 Applied Inverse Problems in Industry

8 Large scale adaptive optics

M5 Tomography

9 Compensating local affine deformations in 2D computerized tomography

10 X-rays fluorescence computed tomography: analytic inversion and iterative methods

M6 Inverse Spectral Problems

11 Inverse scattering for energy-dependent Schrödinger equations

M7 Regularisation Methods – Theory

12 On stable parameter estimation in epidemiology using abstract discrepancy principle

M8 Inverse Problems in Cell Biology

13 3D migration of cells solving an inverse problem

14 An optimal control approach to cell tracking
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>TGV-based flow estimation for 4D leukocyte transmigration</td>
<td>79</td>
</tr>
<tr>
<td>M9</td>
<td><strong>Physical Imaging</strong></td>
<td>84</td>
</tr>
<tr>
<td>16</td>
<td>A CS-based approach for physical imaging</td>
<td>85</td>
</tr>
<tr>
<td>17</td>
<td>Differential equation based inversion method for solving inverse scattering problems</td>
<td>90</td>
</tr>
<tr>
<td>M10</td>
<td><strong>Inverse Scattering</strong></td>
<td>95</td>
</tr>
<tr>
<td>18</td>
<td>Imaging of anomalous components in unknown background</td>
<td>96</td>
</tr>
<tr>
<td>M11</td>
<td><strong>Inverse Problems for Wave Phenomena</strong></td>
<td>101</td>
</tr>
<tr>
<td>19</td>
<td>Optimization approach to combined inverse tsunami problem</td>
<td>102</td>
</tr>
<tr>
<td>20</td>
<td>Inverse force problems for the wave equation</td>
<td>108</td>
</tr>
<tr>
<td>21</td>
<td>Gelfand–Levitan, Marchenko and Krein equations. Theory, numerics and applications</td>
<td>113</td>
</tr>
<tr>
<td>22</td>
<td>The methods of the field continuation from the part of the boundary</td>
<td>118</td>
</tr>
<tr>
<td>M13</td>
<td><strong>Regularisation Methods – Algorithms</strong></td>
<td>123</td>
</tr>
<tr>
<td>23</td>
<td>ARTs penalized by non-smooth convex functions</td>
<td>124</td>
</tr>
<tr>
<td>M15</td>
<td><strong>Compressive Sensing</strong></td>
<td>130</td>
</tr>
<tr>
<td>24</td>
<td>Non-linear compressed sensing and its application to beam hardening correction in x-ray tomography</td>
<td>131</td>
</tr>
<tr>
<td>25</td>
<td>Inverse problems regularized by sparsity</td>
<td>136</td>
</tr>
<tr>
<td>M16</td>
<td><strong>Seismic Imaging</strong></td>
<td>141</td>
</tr>
<tr>
<td>26</td>
<td>A Lie group associated to seismic velocity estimation</td>
<td>142</td>
</tr>
<tr>
<td>M17</td>
<td><strong>Inverse Problems in Astronomy</strong></td>
<td>147</td>
</tr>
<tr>
<td>27</td>
<td>Alternating minimization for Poisson blind deconvolution in astronomy</td>
<td>148</td>
</tr>
<tr>
<td>28</td>
<td>Restoration of astrophysical images, statistical influence of the camera</td>
<td>153</td>
</tr>
<tr>
<td>29</td>
<td>Sparse point source removal for full sky CMB missions</td>
<td>158</td>
</tr>
<tr>
<td>M18</td>
<td><strong>Inverse Statistical Methods</strong></td>
<td>163</td>
</tr>
<tr>
<td>30</td>
<td>An ensemble variational filter for sequential inverse problems</td>
<td>164</td>
</tr>
</tbody>
</table>
Part I

Plenary Talks
Seismic inverse problems: recent developments in theory and practice

W Symes
Rice University, USA
E-mail: symes@caam.rice.edu

Abstract: The widely used least-squares approach to seismic inversion may fail due to insufficient data spectrum. The model extension modification to least-squares inversion can eliminate this full bandwidth requirement. For linearized extended modeling about smooth background parameters, a geometric-optics based analysis explains the increased robustness of extended inversion, and relates its characteristics explicitly to those of travel-time inversion.

1 Introduction

Inference of earth structure from seismic data may be formulated as an inverse problem, via a choice of earth model domain \( M \), data range \( D \), and forward or modeling map \( F : M \rightarrow D \). \( M \) and \( D \) may be (subsets of) Hilbert spaces, and the inverse problem set as a nonlinear least squares problem: given \( d \), choose \( m \) to minimize

\[
J_{\text{OLS}}[m; d] = \frac{1}{2} \| F[m] - d \|^2.
\]

Specific choices of \( M \), \( D \), and \( F \) appropriate for seismology are suggested below. For now, note that seismic measurements are time series, so that \( D \) consists of functions of time, amongst other experimental variables, with definite and limited bandwidth. This formulation, along with many variations and refinements, has been studied since the late '70s. In the last ten years, 3D computational realizations have become feasible, and successful least squares inversions have revealed subsurface maps of unprecedented clarity - see for example [Virieux and Operto, 2009, Plessix et al., 2010, Vigh et al., 2010]. However, the oscillatory nature of seismic data combines with the non-linearity of \( F[m] \) to create numerous stationary points of \( J_{\text{OLS}} \), most of which are uninformative about earth structure. Because of problem size, only rapidly convergent iterations related to Newton’s method are computationally feasible. These are local optimization methods, finding stationary points, whereas only the global minimizer or a close approximation contains useful information. Low frequency data helps by providing a larger region of attraction for the global minimizer, but it is not always possible to acquire data with sufficiently high signal/noise ratio at sufficiently low frequencies to take advantage of this observation [Plessix et al., 2010, ].

This paper focuses on two ideas for modification of \( J_{\text{OLS}} \) to produce optimization tasks more amenable to calculus-based methods: (i) relaxation via extension, that is, enlargement of the domain of \( F \), and (ii) linearization and scale separation, which permit analysis of the extended inverse problem via high frequency asymptotics. Both ideas originate in industrial seismic data processing; the specific replacement for (1) presented here is a variant of Wave Equation Migration Velocity Analysis, or WEMVA. Our principal result is this: the objective Hessian at a consistent model-data pair constrains perturbations in travel time slopes. That is, this variant of WEMVA is related to stereotomography [Billette and Lambaré, 1998], at least in quadratic approximation. Similar assertions seem to be widely believed of WEMVA in general, but so far as I know this paper sketches the first explicit demonstration.

The extension described below is only one of many possible. See [Symes, 2009], [Symes, 2008] for other possibilities and extensive overview.
The abbreviated discussion to follow glosses over many details of analysis and physics, not all of which are trivial or acknowledged explicitly.

## 2 Extension and Linearization

The simplest useful physical description of seismic wave propagation is provided by \textit{constant density acoustics}, in which earth structure is characterized by a \textit{sound velocity} $v(x)$, a positive function of position $x \in \mathbb{R}^3$, or equivalently by $m = v^{-2}$. The extension discussed here replaces $m$ with a bounded coercive self-adjoint operator $\tilde{m}$ on $L^2(\mathbb{R}^3)$. In this extended acoustics, the pressure field $p$ resulting from a point isotropic radiator with time dependence $w(t)$ located at position $x_s$ is the solution of the initial value problem

$$
\begin{align*}
\left(\tilde{m} \frac{\partial^2 p}{\partial t^2} - \nabla^2 p \right)(x,t;\mathbf{x}_s) &= w(t)\delta(x - x_s), \\
p(x,t;\mathbf{x}_s) &= 0, \quad t << 0.
\end{align*}
$$

(2)

The physical significance of this extension is the relaxation of the “no action at a distance” axiom of continuum mechanics. Abusing notation by writing $\tilde{m}$ also for the distribution kernel of $\tilde{m}$, if $(x,y) \in \text{supp} \, \tilde{m}$, then a change in volume at $x$ may result in a change in pressure at $y$. [Stolk, 2000] and [Blazek et al., 2013] show that problems like (2) have sensible solutions, depending smoothly on the coefficient $\tilde{m}$ in an appropriate sense. Define the extension operator $E$ by $Em(x,y) = m(x)\delta(x,y)$: that is, $E$ maps a function of $x$ into the corresponding multiplication operator. Then with $m = v^{-2} \in L^\infty(\mathbb{R}^3)$ and $\tilde{m} = Em$ (the “physical” case), (2) becomes the ordinary acoustic wave equation.

A simplified model of seismic data is the trace of $p$ on a horizontal surface $\{x : x_1 = z_r\}$ (that is, $x_1$ is the depth coordinate, traditionally denoted $z$ - thus $z_r$ for the constant depth of recording). I will ignore all issues arising from finite sampling, and regard the receivers as occupying a continuum of dimension $n - 1$. Likewise, regard the positions of the sources as occupying a continuum, for convenience a subset of another horizontal surface $\{x : x_1 = z_s\}$. Denote the horizontal coordinates by $(x_2, ..., x_n) = \mathbf{x}'$, and the source and receiver coordinate vectors by $x_r = (z_r, \mathbf{x}'_r), x_s = (z_s, \mathbf{x}'_s)$.

With these conventions, the data recorded in a seismic survey define a function of $\tilde{m}$, the the \textit{forward map} or \textit{modeling operator} $\mathcal{F}$:

$$
\mathcal{F}[\tilde{m}](\mathbf{x}', t; \mathbf{x}'_s) = \chi_{r,s}(\mathbf{x}', \mathbf{x}'_s)\chi_t(t)p((z_r, \mathbf{x}'_r), t; (z_s, \mathbf{x}'_s))
$$

(3)

in which $\chi_{r,s} \in C^\infty_0(\mathbb{R}^{2(n-1)})$ and $\chi_t \in C^\infty(\mathbb{R})$ account for the finite extent of source and receiver positions and duration of recording. Write $\mathcal{F}_\delta[\tilde{m}]$ for the special case defined by $w = \delta$ in (2). Then $\mathcal{F}_\delta[\tilde{m}]$ is the distribution kernel of an operator closely related to the Dirchlet-to-Neumann map which plays a role in many inverse problems.

Introduce the linearization about a physical coefficient $m$:

$$
F_\delta[m] = D\mathcal{F}_\delta[E[m]], \quad f[m] = w * F_\delta[m]
$$

(4)

and the “horizontal offset” constraint: apply $F$ only to operator perturbations of $m$ satisfying $\tilde{m}(x,y) = \tilde{m}'(x_1, \mathbf{x}', y')\delta(x_1 - y_1)$. That is, a volume perturbation can only result in a pressure perturbation at the same depth. [This constraint can be weakened]. We assume from now on that $m$ is smooth, so that solutions of (2) can be approximated by geometric optics. Then under certain conditions on the geometry of bicharacteristics,

1. $F_\delta[m]$ is a bounded operator on $L^2(\mathbb{R}^3)$;

2. $F_\delta[m]^*F_\delta[m]$ is a member of the class $OPS^0$ of \textit{pseudodifferential operators} of order 0, elliptic in the conic set of “illuminated reflectors”. In fact, for any pseudodifferential operator $P$ of order 0, $F_\delta[m]^*PF_\delta[m]$ is pseudodifferential. Similarly, for any pseudodifferential $P$ of order 0, $F_\delta[m]^*PF_\delta[m]^*$ is also a pseudodifferential operator of order 0. The symbols of both operators, hence the operators themselves, depend smoothly on $m$.

3. For each smooth $m, \delta m$, there exists a pseudodifferential operator $Q[m, \delta m]$ of order 1, for which

$$
D\mathcal{F}_\delta[m, \delta m] = D^2\mathcal{F}_\delta[E[m]](E[\delta m], \cdot) = F_\delta[m]Q[m, \delta m].
$$


\( Q \) is essentially skew-symmetric:
\[
Q[m, \delta m] + Q[m, \delta m]^* = R[m, \delta m] \in OPS^0(\mathbb{R}^5).
\]

\( Q \) is smooth in \( m \), linear in \( \delta m \).

The necessary geometric conditions, and the meaning of “illuminated reflector”, are explained for example by [Stolk et al., 2009] and [ten Kroode, 2012]. The first result in item 2 is well-known, the second is peculiar to extended modeling. The factorization of \( DF \) does not seem to have been described in the prior literature.

3 A Well-Posed Inverse Problem

The linearized version of the inverse problem introduced earlier is: given \( \delta d = d - F[m] \), find \( m, \delta m \) to minimize \( ||F[m]E[\delta m] - \delta d|| \). Introduce a linear map \( A \) on the space of distribution kernels whose kernel equals the range of \( E \). For example, under the “horizontal offset” condition, a natural choice would be \( A\tilde{m} = [x', \tilde{m}'] \).

In terms of distribution kernels, \( A\tilde{m} = (x' - y')\tilde{m}'(x_1, x', y') \). This choice of \( A \) acts on the kernel of \( \tilde{m} \) as a pseudodifferential operator of order 0. For consistent data, \( \delta d = F[m]\delta \tilde{m}, A\delta \tilde{m} = 0 \), the solution \( m, \delta \tilde{m} \) minimizes
\[
J_{DS}[m, \delta \tilde{m}] = \frac{1}{2} (||F[m]\delta \tilde{m} - \delta d||^2 + \alpha^2 ||A\delta \tilde{m}||^2),
\]
for any positive \( \alpha \).

\( J_{DS} \) is just as likely to exhibit severe nonconvexity as \( J_{OLS} \) - indeed the former includes the latter as a section. [Kern and Symes, 1994] observed that for this type of problem, the reduced objective
\[
\tilde{J}_{DS}[m] = \min_{\delta \tilde{m}} J_{DS}[m, \delta \tilde{m}]
\]
is smooth, independently of the spectrum of \( w \) - in fact even for limit case \( w = \delta \). This conclusion follows immediately from the facts cited above. The normal operator
\[
N[m] = F_\delta[m]^*F_\delta[m] + \alpha^2 A^*A \in OPS^0
\]
is microlocally elliptic, and can be made elliptic and indeed invertible with suitable regularization of \( F_\delta \) - assume that this has been done. Then
\[
\tilde{J}_{DS}[m] = J_{DS}[m, \delta \tilde{m}[m]], \quad \delta \tilde{m}[m] = N[m]^{-1}F_\delta[m]^*\delta d.
\]

Thus \( \tilde{J}_{DS}[m] \) can be rewritten as a sum of terms of the form \( \langle d, F_\delta[m](...)F_\delta[m]^*d \rangle \): the elided factors are products of pseudodifferential operators depending smoothly on \( m \), whence the second fact cited in item 2 implies the claimed smoothness.

Remark: The analogue of \( F_\delta[m](...)F_\delta[m]^* \) for ordinary, rather than extended, modeling is not pseudodifferential, which implies that the linearized problem stated at the beginning of this section has non-smooth objective for \( w = \delta \), leading to the observed proliferation of stationary points for band-limited \( w \).

As \( J_{DS} \) is smooth, it is well-approximated near a global minimizer \( m_* \), by its Hessain quadratic form. This form is easiest to analyze in case the data is consistent with \( m_* \), that is, \( F[m_*]\delta \tilde{m}[m_*] = \delta d, A\delta \tilde{m}[m_*] = 0 \). A page or so of algebra, in which one systematically makes use of consistency to drop various terms, results in
\[
D^2\tilde{J}_{DS}[m_*]|(\delta m_1, \delta m_2) = \langle [A, Q[m_*], \delta m_1]^*[\delta \tilde{m}[m_*]], [A, Q[m_*], \delta m_2]^*[\delta \tilde{m}[m_*]] \rangle.
\]

To understand the significance of this expression, it is necessary to compute the principal symbol \( q_1 \) of \( Q \), a smooth function on the punctured cotangent bundle \( T^*(\mathbb{R}^5) \setminus \{0\} \), homogeneous in the fiber variables \( \zeta, \xi, \eta \). In the simplest case, in which each source or receiver point is connected by a unique ray to any point in the domain of \( \delta \tilde{m} \), a computation of the type pioneered by [Beylkin, 1985], [Rakesh, 1988] yields
\[
q_1[m, \delta m](x_1, x', y', \zeta, \xi, \eta') = i\zeta(DT[m, \delta m](x, x', (x_1, x', y', \zeta, \xi, \eta')))
\]
\[
+ DT[m, \delta m](y, x'(x_1, x', y', \zeta, \xi, \eta')).
\]

(5)
In (5), $DT$ is the perturbation of geometrical-optics travelt ime along the ray (assumed unique) between $x$ and $(z_r, x'_r)$ or $(z_s, x'_s)$, with $x'_r$ and $x'_s$ defined as follows (see [ten Kroode, 2012] for more on this construction).

Denote by $\Theta'_r(x, \xi)$ the horizontal coordinates of the intersection, if any, with $x_1 = z_r$ of the ray of geometric optics with initial data $(x, \xi)$. Define $\Theta'_s(x, \xi)$ similarly. The notation reminds the reader that the ray, hence the location and time of surface arrival, depends only on the unit vector $\xi$ of $\xi$. For an open conic set of $(x, x', y', \zeta, \xi, \eta')$, there exist unique data $(x, \xi(x_1, x', y', \zeta, \xi, \eta'))$ and $(y, \eta(x_1, x', y', \zeta, \xi, \eta'))$ so that

1. $\zeta = \xi_1 + \eta_1$
2. $m(x)|\xi|^2 = m(y)|\eta|^2$
3. $\xi_1, \eta_1 > 0$
4. $\xi_i = \xi'_i, \eta_i = \eta'_i$ for $i > 1$

Then

$$
\Theta'_r(x_1, x', y', \zeta, \xi, \eta') = \Theta_r(x, \xi(x_1, x', y', \zeta, \xi, \eta'))
$$

$$
\Theta'_s(x_1, x', y', \zeta, \xi, \eta') = \Theta_s(y, \eta(x_1, x', y', \zeta, \xi, \eta'))
$$

(6)

The symbol of the “annihilator” $A$ is simply $a_0(x_1, x', y', \zeta, \xi, \eta') = x' - y'$. According to the calculus of pseudodifferential operators (for instance [Taylor, 1981]), the principal symbol of $[A, Q]$ is

$$
-i\{a_0, q_1\} = -i(\nabla \zeta - \nabla \eta')q_1
$$

$$
= \nabla_{x'} DT_r \cdot \nabla_{r} \Theta_r \cdot \zeta(\nabla \zeta' - \nabla \eta') \hat{\xi}
$$

$$
+ \nabla_{x'} DT_s \cdot \nabla_{s} \Theta_s \cdot \zeta(\nabla \zeta' - \nabla \eta') \hat{\eta}
$$

(7)

in which the subscripts on $DT$ signify the collections of arguments at which it is to be evaluated, per (5).

It follows from the calculus, in particular from Gårding’s inequality, that for sufficiently oscillatory $\delta \tilde{m}$ - that is, for a sufficiently dramatic separation of scales - the Hessian form is positive definite on subspaces of $\delta \tilde{m}$ for which $\{a_0, q_1\}$ is nonvanishing in the support of $\delta \tilde{m}$. The expression (7) suggests the meaning of this condition. The first factor in each of the two summands can be interpreted as the slope of arrival time curves - all of them, not just a few that are picked for analysis. The second factor in each case is proportional to a geometric amplitude (see for example [Zhang et al., 2003]). This factor must be present: if the amplitudes are zero, then the presence or absence of reflectors (that is, support of $\delta \tilde{m}$) cannot furnish the data with information about kinematics. The third factor is a real symbol of order zero, as the derivatives of the unit vector are homogeneous of order $-1$.

4 Conclusion

Apparently, under certain circumstances, the WEMVA objective function described in this paper is sensitive to the same aspects of the model $m$ as is slope tomography [Billette and Lambaré, 1998, Chauris and Noble, 2001], but without requiring the picking of travel times. Picking is implicitly accomplished by data weighting, via the presence of the perturbational model $\delta \tilde{m}$ and geometric amplitude in the Hessian kernel. The explicit symbol computation (5), (7) suggest that a detailed analysis of the Hessian singular spectrum may be feasible.

The computations presented here relied on strict ray geometric assumptions, which may to some extent be relaxed - see [ten Kroode, 2012] for a closely related discussion.

Acknowledgements

The author thanks Fons ten Kroode, Christiaan Stolk, and Jie Hou for many insights and helpful discussions. The author acknowledges the generous long-term support provided by the sponsors of The Rice Inversion Project.
References


An active set algorithm for $\ell^0$ optimization with decaying solutions

Yuling Jiao$^1$, Bangti Jin$^2$ and Xiliang Lu$^1$

$^1$School of Mathematics and Statistics, Wuhan University, Wuhan, P.R. China
$^2$Department of Mathematics, University of California, Riverside, 900 University Ave., Riverside CA 92521, USA

E-mail: bangti.jin@gmail.com

Abstract: One of the main tasks in compressed sensing is to develop and analyze efficient algorithms for recovering sparse solutions of underdetermined linear systems. In [1], we proposed a primal dual active set with continuation algorithm (PDASC) for $\ell^0$-regularized minimization problems. The convergence of the algorithm was shown under the restricted isometry property $\delta_{T+1} \leq \frac{1}{\sqrt{T}+1}$ [1], which depends on the sparsity level $T$. In this work, for a class of signals with a strong decay property, we relax the condition on the restricted isometry constant to be independent of the sparsity level $T$. Further, three choice rules are proposed to choose a proper regularization parameter. Several numerical examples are presented to verify its efficiency and accuracy.

Keywords: primal dual active set method; global convergence; $\ell^0$ optimization

1 Introduction

We consider the following $\ell^0$-regularized minimization problem arising in compressive sensing

$$\min_{x \in \mathbb{R}^p} J_\lambda(x) = \frac{1}{2}\|\Psi x - y\| + \lambda\|x\|_0,$$

where the sensing matrix $\Psi \in \mathbb{R}^{n \times p}$ ($p \gg n$) has normalized column vectors, i.e., $\|\psi_i\| = 1$, $\|x\|_0$ denotes the number of nonzero components in the vector $x \in \mathbb{R}^p$, and the data $y \in \mathbb{R}^n$ may be noisy. Since many natural signals have an approximately sparse representation, compressive sensing has been widely applied in signal processing, imaging and machine learning [2, 3, 4]. One of the main tasks in compressive sensing is to design efficient algorithms for recovering sparse solutions of an underdetermined linear system $\Psi x = y$. One popular approach is convex relaxation, i.e., basis pursuit denoising [5] or Lasso [6]. Its popularity can largely be attributed to the fact that it admits efficient numerical solution. Recently, nonconvex models were shown to be theoretically advantageous over convex ones [7]. However, this calls for efficient algorithms for solving problem (1) and effective choice rules for choosing the parameter $\lambda$.

Since the objective function $J_\lambda(x)$ in (1) is nonconvex and discontinuous, it is highly nontrivial to find its global minimizer(s), which represents one of the main obstacles in the wide-spread applications of the model (1). Most existing algorithms for problem (1), including iterative hard thresholding [8] and penalty decomposition [9] etc., can at most converge to a local minimizer or a stationary point. Recently, based on a novel necessary optimality condition for the global minimizer (also a sufficient condition for the coordinatewise minimizer), a PDAS algorithm was developed to solve (1) in [10, 11]. It enjoys a local superlinear convergence, but requires a good initial guess; and its convergence analysis is challenging. Very recently, the convergence of the PDASC algorithm, when coupled with a continuation strategy, was shown under the restricted isometry property with $\delta_{T+1} \leq 1/(2\sqrt{T}+1)$ on the sensing matrix $\Psi$ [1]. In this work, we focus on the analysis of the PDASC algorithm. By assuming a decay property on the true signal, the condition in [1] is weakened in that the restricted isometry property (RIP) constant $\delta_{T+1}$ is independent of the sparsity level $T$. Further, three a posterior rules are proposed to choose a proper regularization parameter. These represent our essential contributions, and will be described in Section 2 and numerically verified in Section 3.
2 PDASC algorithm

First we describe the problem setting. Let the true signal \( x^* \) have \( T \) nonzero components, and \( y = \sum_{i \in A^*} x_i^* \psi_i + \eta \), where \( A^* \) is the exact active set (indices of nonzero components) with size \( T = |A^*| \). Further, we may assume \( A^* = \{1, 2, ..., T\} \) and \( |x_1^*| \geq |x_2^*| \geq \cdots \geq |x_T^*| > 0 \). The noise vector \( \eta \) satisfies \( \|\eta\| \leq \epsilon \) where \( \epsilon \) is the noise level. We denote \( S = \{1, 2, ..., p\} \) and \( T^* = S \setminus A^* \). For any index set \( A \subset S \), we denote by \( x_A \in \mathbb{R}^{|A|} \) (or \( \Psi_A \in \mathbb{R}^{n \times |A|} \)) the subvector of \( x \) (or the submatrix of \( \Psi \)) whose (column) indices appear in \( A \). The RIP constant \( \delta_s \in (0, 1) \) is the smallest positive constant satisfying \((1 - \delta_s)\|x\|^2 \leq \|\Psi x\|^2 \leq (1 + \delta_s)\|x\|^2\), for any \( x \) with \( \|x\| \leq s \).

The PDASC algorithm in [1] combines the strengths of the PDAS algorithm [10] and the continuation strategy, which respectively enjoys fast local convergence and controlling the size of the active set during the iteration. The complete procedure is described in Algorithm 1.

### Algorithm 1 PDASC algorithm

1: Set \( \lambda_0 = \frac{1}{2\beta} \|\Psi y\|^2_2, A(\lambda_0) = \emptyset, x(\lambda_0) = 0 \) and \( d(\lambda_0) = \Psi^T y, \beta \in (0, 1) \).
2: for \( k = 1, 2, \ldots \) do
3: Let \( \lambda_k = \beta \lambda_{k-1}, A_0 = A(\lambda_{k-1}) \), \((x_0, d_0) = (x(\lambda_{k-1}), d(\lambda_{k-1})) \).
4: for \( j = 1, 2, \ldots \) do
5: Define the sets \( A_j \) and \( I_j \) by \( A_j = \{i : |x_i^j| + d_i^{j-1}| > \sqrt{2\lambda_k}\} \) and \( I_j = A_j^c \).
6: Check \( A_j = A_{j-1} \). If fulfilled, let \( A(\lambda_k) = A_j \) and \((x(\lambda_k), d(\lambda_k)) = (x^{j-1}, d^{j-1}) \).
7: Update the primal and dual variables \( x^j \) and \( d^j \) by
   \[
   x^j_{I_j} = 0, \quad \Psi^T_{A_j} \Psi_{A_j} x^j_{A_j} = \Psi^T_{A_j} y, \quad d^j = \Psi^T (\Psi x^j - y). 
   \]
8: end for
9: Check the stopping criterion.
10: end for

To analyze the algorithm, we make the following assumptions.

\textbf{(H1)} The true signal satisfies a decay property \( \max_{1 \leq i \leq T-1} |x^*_{i+1}|/|x^*_{i}| = \rho < 1 \).

\textbf{(H2)} The sensing matrix \( \Psi \) satisfies the RIP condition \( \delta \triangleq \delta_T \leq (1 - \rho) \sqrt{1 - \rho^2}/6 \).

\textbf{(H3)} The noise level is small: \( \epsilon \leq |x^*_{T}|(1 - \rho)/30 \).

If the noise level \( \epsilon \) is known and we choose the stopping criterion at line 9 of Algorithm 1 by

\[
\|\Psi x(\lambda_k) - y\| \leq \epsilon. 
\]

Then we have the following global convergence result. The proof of the theorem is technical and deferred to Appendix A. We note that the RIP condition is Assumption (H2) is independent of the sparsity level \( T \) of the true signal \( x^* \), which improves a known result in [1].

\textbf{Theorem 2.1.} Let Assumptions (H1)-(H3) hold, \( \beta = 0.25 \). Then Algorithm 1 stops in finite steps \( k \), and \( A(\lambda_k) = A^*, x(\lambda_k) = \Psi^T_{A^*} y \).

Now we briefly discuss the choice of the parameter \( \lambda \). First, if the noise level \( \epsilon \) is known, we can apply the discrepancy principle [12], i.e., select the \( \lambda \) value as the first \( \lambda_k \) such that (2) holds. If this knowledge is unavailable, we employ either Bayesian information criterion (BIC) or the size of the active set to determine \( \lambda \). Specifically, we first run Algorithm 1 until \( \|x(\lambda_k)\|_0 \geq \theta n, \theta \in [0.5, 1] \) and can be fixed at 0.9 in practice. Then we select \( \lambda \) and \( x(\lambda) \) by following rules. (a) BIC: The BIC functional is defined by \( \text{BIC}(\lambda_k) := \frac{1}{2}\|\Psi x(\lambda_k) - y\|^2_2 + \ln p d\lambda_k \), where \( d\lambda_k = \|x(\lambda_k)\|_0 \) is the degree of freedom of \( x(\lambda_k) \) [13]. Then we take \( \lambda \in \arg \min_{\lambda_k} \text{BIC}(\lambda_k) \). (b) The size of the active set: For any \( d = 1, \ldots, p \), denote by \( S_d = \{\lambda_k : |A_k| = d\} \) and \( d = \arg \max_d \|S_d\| \). Then we choose \( \lambda = \max_k \lambda_k, \) s.t. \( \lambda_k \in S_d \).

3 Numerical examples

Now we illustrate the performance of Algorithm 1. All the experiments are run on a dual-core desktop with 3.16 GHz and 4 GB RAM. First we describe data generation and input parameters in Algorithm 1. The
random $T$-sparse true signals $x^*$ with a dynamic range $\text{DR} := \max\{|x^*_i| : x^*_i \neq 0\}/\min\{|x^*_i| : x^*_i \neq 0\}$ are generated as follows. Due to the randomness, a large DR likely implies a faster decay of magnitude of $x^*$. The sensing matrix $\Psi \in \mathbb{R}^{n \times p}$ is taken to be random Gaussian. The observation vector $y$ is given by $y = \Psi x^* + \eta$, where $\eta$ is the measurement noise, with the entries $\eta_i$ following an i.i.d. normal distribution $N(0, \sigma^2)$ and noise level $\epsilon = \|\eta\|$. We use $(T, n, p, \text{DR}, \sigma)$ to denote the data. The decreasing factor $\beta$ is fixed at $\beta = 0.7$.

Figure 1. (a) The support recovery probability on data set $(10 : 20 : 400, 500, 1000, 10^{0.1\cdot 5}, 0.01)$; and (b) the CPU time (b) and (c) relative $\ell^2$ error for $(500 : 100 : 800, 2 \times 10^3, 10^2, 0.01)$

First we show the dependence of Algorithm 1 on the DR of the signal $x^*$ (i.e., the decay ratio of $x^*$). To this end, for any $\text{DR} \in \{1, 10, \ldots, 10^5\}$ we generate $(T, n, p, \text{DR}, \sigma) = (10 : 20 : 390, 200, 1000, \text{DR}, 0.01)$. Then we run Algorithm 1 on 100 independent realizations of the setup $(T, n, p, \text{DR}, \sigma)$ using (2) to select $\lambda$, and the results are shown in Fig. 1(a). The probability of Algorithm 1 to exactly recover the true support increases with the increase of the DR of the signal. Surprisingly, Algorithm 1 can exactly recover a target support of size larger than $n/2$, if $\text{DR} > 10^3$.

Next we show the efficiency and accuracy of Algorithm 1 with the choice rules: discrepancy principle [12], BIC and the size of the active set (AS), which are denoted by PDASC-DP, PDASC-BIC, and PDASC-AS, respectively. We compare them with two state-of-the-art methods, i.e., GreedyGp [14] and AIHT [15]. The problem setup is $(T, n, p, \text{DR}, \sigma) = (500 : 100 : 800, 2 \times 10^3, 10^2, 0.01)$, and the results are based on 100 independent realizations. The CPU time (in seconds), and the relative $\ell^2$ error are shown in Fig. 1(b)-(c). It is observed that all three methods PDASC-DP, PDASC-BIC, and PDASC-AS are competitive with and outperform existing methods in terms of the CPU time and the error. Although not presented, we would like to note that the regularization parameters selected by the three rules are close to each other, which is also confirmed by Fig. 1(c).

A Proof of Theorem 2.1

First we recall properties of the RIP constant [16]. Let $A, B$ be disjoint subsets of $S$. Then

$$
\|\Psi_A^t \Psi_A x_A\| \overset{\varepsilon}{\leq} (1 + \delta_{|A|}) \|x_A\|, \quad \|\Psi_A^t \Psi_A\|^{-1} \|x_A\| \overset{\varepsilon}{\geq} (1 + \delta_{|A|})^{-1} \|x_A\|,
$$

$$
\|\Psi_B^t \Psi_B\| \overset{\varepsilon}{\leq} \delta_{|A|+|B|}, \quad \|\Psi_B^t y\| \overset{\varepsilon}{\leq} (1 - \delta_{|A|})^{-1/2} \|y\|,
$$

$\delta_s$ is monotone increasing with respect to $s$.

Let Assumptions (H1) - (H3) hold. Given any $A \subset A^s$, let $B = A^s \setminus A$ and $I = S \setminus A$. Consider one step iteration: $x_A = \Psi_A^t y$, $x_I = 0$, $d = \Psi_I^t (y - \Psi x)$. Let $\tilde{x}_A = x_A - x_A^*$. Then one can verify direct computation that

$$
\|\tilde{x}_A\| \leq \frac{\delta_{|A|+|B|}}{1 - \delta_{|A|}} \|x_B^*\| + \frac{1}{\sqrt{1 - \delta_{|A|}}} \|x_B^*\| \leq \frac{\epsilon y}{1 - \delta_{|A|}} \|x_B^*\| + \frac{1}{\sqrt{1 - \delta_{|A|}}} \epsilon. \tag{3}
$$
Now we consider the two cases separately.

**Case 1.** If $\mathcal{B} = \emptyset$, i.e., $\mathcal{A} = \mathcal{A}^*$, then $\|\hat{x}_\mathcal{A}\| \leq \frac{1}{\sqrt{1-\rho}} \epsilon$, and

$$|d_j| \leq \epsilon + \delta |\mathcal{I}+1|\|\hat{x}_\mathcal{A}\| \leq \epsilon + \frac{\delta \rho + 1}{\sqrt{1-\rho}} \epsilon, \quad \forall j \in \mathcal{I},$$

$$|x_i| \geq |x_i^*| - \|\hat{x}_\mathcal{A}\| \geq |x_i^*| - \frac{1}{\sqrt{1-\rho}} \epsilon, \quad \forall i \in \mathcal{A}.$$

**Case 2.** If $\mathcal{B}$ is nonempty, let $i_\mathcal{A}$ be the smallest index in $\mathcal{B}$ and $c = |x_i^*|$. Consequently,

$$|d_{i_\mathcal{A}}| - c | \leq \delta |\mathcal{B}| \|x_{\mathcal{B}\setminus\{i_\mathcal{A}\}}^*\| + \epsilon + \delta |\mathcal{A}| + 1 \|\hat{x}_\mathcal{A}\|,$$

and

$$|d_j| = |\psi_{\mathcal{B}}(\Psi_{\mathcal{B}}x_{\mathcal{B}} + \eta - \Psi_{\mathcal{A}}\hat{x}_\mathcal{A})| \leq \left\{ \begin{array}{ll} \delta |\mathcal{B}| + 1 \|x_{\mathcal{B}}^*\| + \epsilon + \delta |\mathcal{A}| + 1 \|\hat{x}_\mathcal{A}\| & \forall j \in \mathcal{I}^* \cap \mathcal{I}, \\ \delta |\mathcal{B}| + 1 \|x_{\mathcal{B}}^*\| + \epsilon + \delta |\mathcal{A}| + 1 \|\hat{x}_\mathcal{A}\| & \forall j \in \mathcal{B}, j \neq i_\mathcal{A}. \end{array} \right.$$

Now the decay property of the signal from Assumption (H1) directly implies

$$\|x_{\mathcal{B}}^*\| \leq \frac{c}{\sqrt{1-c^2}} \leq \frac{c}{\sqrt{1-c^2}} \text{ and } \|x_{\mathcal{B}\setminus\{i_\mathcal{A}\}}^*\| \leq \frac{\epsilon}{\sqrt{1-c^2}}.$$ (6)

Now appealing to Assumption (H1)-(H3) and estimates (3)-(5), we deduce

$$\|\hat{x}_\mathcal{A}\| \leq \frac{\delta}{1-c} \frac{c}{\sqrt{1-c^2}} + (1 + \delta) \epsilon \leq \frac{1}{2}(1 + \rho) c,$$ (7)

$$|d_{i_\mathcal{A}}| - c | \leq \frac{\delta \rho c}{\sqrt{1-c^2}} + \epsilon + \delta \|\hat{x}_\mathcal{A}\| \leq \frac{1}{2}(1 + \rho) c,$$ (8)

$$|d_j| \leq \left\{ \begin{array}{ll} \frac{1}{2}(1 - \rho) c & \forall j \in \mathcal{I}^* \cap \mathcal{I}, \\ c - \frac{\epsilon(1-c)}{c} & \forall j \in \mathcal{B}, j \neq i_\mathcal{A}. \end{array} \right.$$ (9)

As a direct consequence of these estimates, we have the following useful corollary.

**Corollary 3.1.** Let Assumptions (H1)-(H3) hold, and $l \leq \mathcal{I}^* \cup \mathcal{I}^*$, $\mathcal{A} = \{1, \ldots, l\}$, $x_{\mathcal{A}} = \Psi_{\mathcal{A}}^* y$, $d = \Psi_{\mathcal{A}}^* (y - \Psi_{\mathcal{A}} x_{\mathcal{A}})$. Then there holds $\min_{i \in \mathcal{A}} |x_i| \geq \|d\|_{\mathcal{E}^*}.$

**Lemma 3.1.** Let Assumption (H2) hold. Let $l \leq \mathcal{I}^* \cup \mathcal{I}^*$, $\mathcal{A} = \{1, \ldots, l\}$, $\mathcal{A}_1 = \{1, \ldots, l - 1\}$, $\mathcal{A}_2 = \{1, \ldots, l\}$, $x_{\mathcal{A}_2} = \Psi_{\mathcal{A}_2}^* y$, $d^i = \Psi_{\mathcal{A}_1}^* (y - \Psi_{\mathcal{A}_1} \Psi_{\mathcal{A}_2}^* y)$. Then $|x_{\mathcal{A}_1}| \geq |d^i|$.

**Proof.** Let $D = \Psi_{\mathcal{A}_1}^* \Psi_{\mathcal{A}_1}$, and $b = (b_1, \ldots, b_{l-1})^T = \Psi_{\mathcal{A}_1}^* \eta$. Then for any $z \in \mathbb{R}^{l-1}$, there holds

$$z^T(D - bb^T) z = \|\Psi_{\mathcal{A}_1} z\|^2 - \|\Psi_{\mathcal{A}_1} z\| \geq (1 - \delta)\|z\|^2 - \delta^2 \|z\|^2 = (1 - \delta - \delta^2)\|z\|^2,$$

$$\geq 1 - \delta^2 - b^T D^{-1} b \geq 1 - \|b\|_2 \|D^{-1} b\| \geq 1 - \frac{1}{1 - \delta} \|\psi_{\mathcal{B}}^* \Psi_{\mathcal{A}_1}\| \geq 1 - \frac{\delta^2}{1 - \delta} > 0,$$

i.e., $D - bb^T$ is symmetric and positive definite. Further, $|b| \leq \delta$ yields $0 < b^T D^{-1} b < 1$. Hence we deduce

$$\left(\psi_{\mathcal{A}_1}^* \Psi_{\mathcal{A}_1} \right)^{-1} = \begin{bmatrix} D & b \\ b^T & 1 \end{bmatrix}^{-1} = \begin{bmatrix} (D - bb^T)^{-1} \\ -b^T (D - bb^T)^{-1} \end{bmatrix}.$$ Clearly, $-b^T (D - bb^T)^{-1} = \frac{b^T D^{-1} - b b^T D^{-1} b}{1 - b^T D^{-1} b}$. Upon letting $\mathcal{B} = \mathcal{A}_2 \setminus \mathcal{A}_1$, $h = \psi_{\mathcal{A}_1}^* (\Psi_{\mathcal{B}} x_{\mathcal{B}} + \eta) \in \mathbb{R}^{l-1}$ and $c = \psi_{\mathcal{B}}^* (\Psi_{\mathcal{A}_1} x_{\mathcal{B}} + \eta)$, we deduce

$$d^i = \psi_{\mathcal{B}}^* (\psi_{\mathcal{I}} x_{\mathcal{B}} + \eta - \Psi_{\mathcal{A}_2} D^{-1} \psi_{\mathcal{A}_1}^* (\psi_{\mathcal{A}_1} x_{\mathcal{B}} + \eta)),$$

$$= x_{\mathcal{B}}^* + (c - b^T D^{-1} b) x_{\mathcal{B}}^* + \eta = (1 - b^T D^{-1} b) x_{\mathcal{B}}^* + (c - b^T D^{-1} b) h,$$

where the second line follows from $b = \Psi_{\mathcal{A}_1} \psi_{\mathcal{I}}^*$. Similarly

$$x_{\mathcal{B}}^* = x_{\mathcal{B}}^* + (\psi_{\mathcal{A}_2}^* \Psi_{\mathcal{A}_2}^{-1} \psi_{\mathcal{A}_1}^* (\Psi_{\mathcal{B}} x_{\mathcal{B}} + \eta))_i = x_{\mathcal{B}}^* + \left(\psi_{\mathcal{A}_2}^* \Psi_{\mathcal{A}_2}^{-1} \left( \begin{array}{c} h \\ c \end{array} \right) \right)_i = \frac{d^i}{1 - b^T D^{-1} b}.$$ Therefore $|x_{\mathcal{B}}^*| \geq |d^i|$, and this completes the proof.

---

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, book title and publisher. Published under licence by IOP Publishing Ltd.
The next lemma shows the behavior of the inner iterations.

**Lemma 3.2.** Let Assumptions (H1)-(H3) hold, and \( A_0 = \{1, \ldots, s\}, s < T, I_0 = A_0^0, x_0 = \Psi_{A_0} y, x_0^0 = 0, d^0 = \Psi^l (y - \Psi) \). Then lines 3-8 in Algorithm 1 terminate in finite steps and the stopping criterion \( A_j = A_{j-1} \) is reached.

**Proof.** By Corollary 3.1, \( ||d^0||_{\ell^\infty} \leq \min_{i \in A_0} |x_i^0| \). Then let \( \sqrt{2\lambda} \in [||d^0||_{\ell^\infty}, \min_{i \in A_0} |x_i^0|] \), and \( \lambda = 0.25 \hat{\lambda} \). Now suppose \( \sqrt{2\lambda} \in [||x_{q+1}^0||, ||x_q^0||] \) for some \( q < T \) (we discuss the case \( q = T \) later), and let \( i_k = \min \{ i : i \in I_k \} \), \( a = |x_{q+1}^0|, d = |x_q^0|, b = a + \frac{d-a}{4}, c = d - \frac{d-a}{4} \). Next we discuss separately the following three scenarios.

(i) \( \sqrt{2\lambda} \in (b, c) \). By (7)-(9), one can verify: \( i_k \leq q \Rightarrow i_{k+1} > i_k, i_k \geq q + 1 \Rightarrow A_{k+1} = \{1, 2, \ldots, q\} \), and \( A_k \subset A^*, \) moreover \( \{1, 2, \ldots, q\} \) is stable.

(ii) \( \sqrt{2\lambda} \in [a, b] \). One can check: \( i_k \leq q \Rightarrow i_{k+1} > i_k, i_k \geq q + 1 \Rightarrow A_{k+1} = \{1, 2, \ldots, q\} \) or \( A_{k+1} = \{1, 2, \ldots, q+1\} \). After a finite number of steps, \( A_k \) reaches either \( \{1, 2, \ldots, q\} \) or \( \{1, 2, \ldots, q+1\} \). By Lemma 3.1, if \( \sqrt{2\lambda} < x_{q+1} \) (definition for \( x_{q+1} \) is similar as Lemma 3.1), then \( \{1, 2, \ldots, q+1\} \) is stable. If \( \sqrt{2\lambda} \geq d_{q+1} \) then \( \{1, 2, \ldots, q\} \) is stable. This shows the convergence.

(iii) \( \sqrt{2\lambda} \in [c, d] \). A similar argument implies that either \( \{1, 2, \ldots, q\} \) or \( \{1, 2, \ldots, q-1\} \) is stable. Further, one can prove \( q > s \).

In case \( q = T \), it suffices to show \( i_k \leq T - 1 \Rightarrow i_{k+1} > i_k \) and \( A_k \subset A^* \). Then it reaches either \( \{1, \ldots, T - 1\} \) or \( \{1, 2, \ldots, T\} \). A similar argument implies that the algorithm stops. \( \square \)

Now we can state the proof of Theorem 2.1.

**Proof.** The proof follows from Lemma 3.2. The active set \( A_k \) is monotonically increasing but always contained in \( A^* \). Hence, when \( \sqrt{2\lambda} \leq \frac{2}{3} |x_T^*| \), the active set \( A \) must be identical with \( A^* \), and the algorithm stops when the active set \( A_k \) equals to \( A^* \). \( \square \)

### References

Emergence of ideas in inverse theory

Pierre C Sabatier

Laboratoire Univers et Particules de Montpellier, Université de Montpellier; 34095 MONTPELLIER Cedex 05,
France

E-mail: pierre.sabatier@univ-montp2.fr

Abstract: This paper builds on several reviews in which the interdisciplinary research on inverse problems was highlighted by the author, particularly a review published in 1971, which showed their common evolution in all fields, another one in 2000, almost comprehensive, except for statistical methods, and the 25th anniversary review (2009), where the consequences of the ‘natural’ origin of inverse problems was analysed. Their references already give a good indirect bibliography for the present paper (a longer version, with more direct references, can be obtained from the author). One focuses here on the emergence of ideas making what already was called in the seventies Inverse Theory. The emphasis is on ‘Seminal’ ideas, S, i.e. those which appear when a set of immature ideas for scientific working ‘crystallizes’ into a well defined program. Defended first by only one or a few people, later they guide works, scientific results, and more ‘technical’ ideas, of several others. The most important technical ideas, can themselves be called ‘seminal techniques’, S.T., considering the results or generalisations they create! In fact, it may be good to recall a maxim from C Lanczos: “a lack of information cannot be remedied by any mathematical trickery”. Another one, given by G B Shaw, ends the paper.

1 Before Inverse Theory...

EVE’s diary: Whenever a new creature comes along I name it... The minute I set eyes on an animal I know what it is. I don’t have to reflect a moment: the right name comes out instantly, just as if it were an inspiration, as no doubt it is, for I am sure it wasn’t in me half a minute before. I seem to know just by the SHAPE of the creature and the WAY IT ACTS what ANIMAL it is [1].

We call ‘data’ what is observable or measurable in the image, and call ‘parameters’ values describing the objects to be identified. Make dictionaries of both, and write down all correspondences between them. Such a management of the IP Image ➔ Object is called now

S Shape (or Behaviour) Recognition.

Example: friend or foe interrogations during World Wars 1 and 2.

Bad or noisy data necessitated further:

a Extending the definition of solutions.

b Getting algorithms to calculate them.

c Seeking ways of describing (coding?) image and/or object.

Hereafter, we will follow the evolution of the theory, but please allow the author to begin with his own suggestion, concluding the interdisciplinary NASA meeting of 1971 [2]:

Never ask for ‘solutions’ but:

SQ ‘Here are the experimental results. What information do they contain?’

This seminal question was a cornerstone for the meetings we organised between 1972 and 2000 with logo RCP 264, and for the journal Inverse Problems we founded in 1985 with Roy Pike and IOP: IOP, journal staff, authors, boards, editors, particularly A Louis for the present paper, deserve all our gratitude. All this led to the analysis of how inverse problems [3, 4] usually arise:

1) we get ‘data’ from observations and/or measurements;

2) we know, or we presume, that a set of physical quantities, called ‘parameters’, or of any other elements of interest in natural sciences, and collectively called a ‘physical state’, or an ‘object’, is related to the information presumably contained in these data;
3) we try to identify precisely from the data the ‘state’; or the ‘object’. This is a ‘natural inverse problem’. The mathematical one begins with SETTINGS:

One first gathers and defines ‘parameters’ in such a way that they describe the ‘state’ or ‘object’. Then, one defines a mathematical ‘space of parameters’, \( C \), including all the physical elements and others, fulfilling the two requirements that its definition as a mathematical space is ‘simple’, and that one can express the difference between two ‘states’, i.e. with a topology, the simplest and most usual choice being that of a metric space. The natural phenomenon that relates ‘our parameters to our data’ can now be represented by a mapping \( M \) from the space of parameters \( C \) to a space of data \( E \) constructed with the same requirements and that contains all possible data, (usually also a metric space). The triple \((C, M, E)\) represents a ‘mathematical model’ for our problem, defining a mathematical direct problem. A method constructing for any element \( e \) of \( E \) an element \( c \) of \( C \) such that \( e = M(c) \), or, respectively, \( e \) close to \( M(c) \), is called a solution, or, respectively, a generalised solution of the mathematical inverse problem. As we noticed in a previous paper, relations between the ‘natural inverse problem’ and the ‘mathematical inverse problem’ are always present, even if those do not appear in many published papers.

2 History of Seminal Ideas

S1 (Prehistory): Memorise two dictionaries of elements of \( C \) and of \( E \), with all relations between them (convenient for finite sets only).

S2 (1826, Abel): Try to derive formulas of profile inversion. Abel’s well-known one is:

\[
g(y) = \int_0^y f(z)(y-z)^{-\frac{1}{2}}dz \quad (0 \leq y \leq a)
\]

\[
\int_0^x f(z)dz = \pi^{-1}\left[ \int_0^x g(y)(x-y)^{-\frac{1}{2}}dy \right] \quad (0 \leq x \leq a)
\]

(1900…Hadamard … Holmgren, Petrovski) A well-posed inverse problem is such that for any result, there exists a unique, and stable, solution. “On a pu trouver des cas très étendus dans lesquels l’un ou l’autre de ces problèmes se présentait comme parfaitement bien posé, je veux dire comme possible et déterminé… il est remarquable… que deux problèmes tout à fait analogues en apparence puissent être l’un possible et l’autre impossible, suivant qu’ils correspondent ou non à une donnée physique”. (1902, J. Hadamard, comparing Laplace and Helmholtz equations with Dirichlet and Cauchy conditions).

(1940) Identification or construction? Physicists, or practitioneers, only seek solutions, or approximate solutions, that can be derived by constructive methods.

S3. (1890-1940) Weak nonuniqueness of inverse problem solution can be represented by a bound of the measurements resolving power. (Lord Rayleigh.)

S4. Find a probable value of parameters. (Statistical approaches. Bayes, Laplace,…)

1850-1940. Routes in the spaces \( C \) and \( E \) (Darboux, Crum, Krein, Bargmann…)

1940-1970 Inverse Theory is BORN!!

3 Inverse Theory with perfect data.

NONUNIQUENESS (nonlinear and spectral problems: famous analyses of Krein, Gelfand, Levitan, Marchenko, Kay, Levinson, Moses, Regge, Newton, Faddeyev. [5])

1) Approximate mappings leading to several turning points and Abel’s integral have been used to describe nonuniqueness, (e.g. in geophysics and in neutron scattering at fixed energy), and they explain the ambiguities, (numerically verified
later) by physical reasons: parts of the profile escaping paths of propagation, guided waves, shell of minimum velocity, etc. [3, 4, 5] The nullspace of local linearizations of \( M \) may also show nonuniqueness. \( ST1. \) approximate mappings from \( C \) to \( E, \) can show the emergence of ambiguities in the range they are valid. But exercise extreme caution...!

2) Recall: \textit{Signal Theory and linear self adjoint operator}: Let \( D = -\frac{\partial^2}{\partial x^2} + V, \) with \( V \) real and ‘regular’, e.g. inside the usual Fourier transform

\[
L_1^1(\mathbb{R}) = \left\{ V \mid \int_{-\infty}^{+\infty} (1 + |x|)|V(x)| \, dx < \infty \right\}
\]

The spectrum of \( D \) ((i.e. the set of numbers \( \lambda \) for which \( (D - \lambda I) \) is not invertible) is made of the positive half axis (values \( \lambda = k^2 \)) and a finite number of negative eigenvalues \( \lambda_n = -\kappa_n^2 \), for which the operator \( D \) has an eigenfunction \( \psi_n \) in \( L^2(\mathbb{R}) \): \( D\psi_n(x) = -\kappa_n^2\psi_n(x) \) The \( \psi_n \) definition is achieved by normalising it in \( L^2(\mathbb{R}) \). For real \( k \), (so-called continuous spectrum), the eigenfunctions are replaced by the ingoing or outgoing Jost solutions of \( (D - k^2)f = 0 \), respectively:

\[
f^\pm(k,x) = \exp[\pm ikx] + o(1) \quad x \to \pm \infty
\]

Now the spectral signal analysis uses the fact that any element \( h \) in \( L^2(\mathbb{R}) \) is equal to its projection on these eigenfunctions, i.e. the sum of an integral involving \( f^\pm \) and an expansion on the \( \psi_n \), e.g., for \( V = 0 \), with \( f^+ = e^{ikx} \), we get so the usual Fourier transform \( h(x) = \int_{-\infty}^{+\infty} e^{-ikx} H(k)dk \)

3) \textit{Wave problems}: Schrödinger’s wave equation \( D - k^2 u = 0 \), where \( k^2 \) is the wave number \( (k^2 = k,k) \) is related to Helmholtz or Impedance Equations. In 3d problems, \( D \) is equal to \(-\Delta + V \). [3, 4, 5, 6] From now on, assume \( V \) decreases fast enough at \( \infty \).

1d scattering: \( k \) real (positive energy) ‘outgoing’ and ‘ingoing’ Jost solutions define the scattering data (reflection and transmission coefficient): \( T(k)f^- (k,x) = f^+(-k,x) + R(k)f^+(k,x) \), \( R(k) \) is supposed ‘measurable’. A discrete spectrum may exist (‘bound states’). For \( V \) ‘small’ linearized formulas suggest: \( V(x) \simeq \frac{1}{2\pi} \int_{-\infty}^{+\infty} k R(k) e^{2ikx} \, dk \) but it is WRONG: \textit{Information on discrete spectrum must be used!} The famous example is the ‘transparent’ \( V \) for which \( R = 0 \), but one discrete eigenvalue \(-\kappa^2 \): \( V_0(x) = -\frac{2\kappa^2}{2\kappa^2 - (x-x_0)^2} \)

3d scattering

\[
u(x) = \exp[ikx] + x^{-1} \exp[ikx] F \left( \hat{x}, \hat{k}, k \right) + o(x^{-1})
\]

The inverse problem \( F \) to \( V \) is:

(a) overdetermined if \( F \) given for all \( \hat{x}, \hat{k}, k \): obvious on linearized behavior (large \( k \)).
(b) not determined uniquely at fixed \( k \), unless strong restrictions, e.g. \( |V(x)| \leq \alpha \exp(-\beta x) \)

If not, ‘transparent potentials’ (Newton-Sabatier, or singular)
(c) So as to understand physical reasons enabling transparent ‘long tail’ potentials, see the spherically symmetric case, linearized for small interactions

\[
F(\theta, k) = \left[ 2k \sin \frac{\theta}{2} \right]^{-1} \int_0^{\infty} \sin \left[ 2k \sin \frac{\theta}{2} \right] V(r) \, dr
\]

IP reduces to a well known Fourier inversion, with a band \((0,2k)\) filtering: uniqueness only if the a priori decreasing of \( V \) guarantees a unique continuation.

Hence, a very precious property of an inverse problem is strong \textbf{NONUNIQUENESS}: it suggests what is lacking in our information.

4 Inverse Theory with real data, but controlled nonuniqueness.

Algorithms. Going back to modelling

For details on regularisations or algorithms see our review [3] and its classical references, (e.g. by Louis or Natterer), and keep in mind the rules:

R1 There are always measurement errors.
R2 If \( E \) linear space, define a norm, describing a “size”.
R3 A priori knowledge only defines “admissible” parameters.
R4 First choose a branch of \( M \), and study IP “posedness”.
R5 A generalised solution must be admissible and stable.
R6 Linearising \( M \) may give informations.

As Gauss and Legendre had done before, one first minimized the misfit \(|\mu(x)|^2 = \sum_{p=1}^{N} [e_p - (Mx)_p]^2 \) But for the first well known “least-squares” algorithms, Google citations credit Levenberg(1944, 6000 cit), Marquardt (1963, 20000 cit), Tikhonov (1963, 11000 cit). Compromise came gradually. Mathematical results in the sixties, mainly due to Tikhonov’s
School, led to

**S1: A regularised solution** is a generalised one achieving a ‘compromise’ between misfit, physical relevance, and stability

Example: Minimise the “cost function” $C(x) = ||x - x_0||^2 + \lambda^2 ||y - Mx||^2$

**Regularisation with guides appraising nonuniqueness (linear $M$):**

There are two methods [3]: a) the Singular value decomposition, which enables pointing out irrelevant, impossible, and important parameters, defining with Bertero de Mol and Pike the degrees of freedom in our measurement device, and regularising by shunting too small singular values; b) the Backus and Gilbert method, which appraises the resolving power of measurements.

**Algorithms** Their development relies on various strategies. **Search of fixed points:** but nonunique in most inverse problems. **Search of asymptotic expansions towards a solution:** which often needs justifications. **Algorithms coupled with regularisations:** as optimisations, mollifiers, etc **Algorithms based on the analysis of the problem,** as NS matrix methods, Herglotz functions,... **Algorithms suggested by statistics:** Monte Carlo, hedgehog, etc. **Algorithms suggested by other sciences:** as simulated annealing, genetic algorithms... **Algorithms imposed by very special a priori information. See below.**

**Descriptive Modelling** is not the only possible one in $C^0$, the admissible elements of $C$ : Examples being : **Mathematical morphology.** (1964 Matheron, Serra [3]); **Special physical Requirements. Distinguished features:** Positivity, convex sets of solutions, represented by their extremals. (1977 Sabatier [3, 7]). Related seminal ideas are:

**S1: Reduce the numbers of degrees of freedom in $C^0$:** Natural reduction: Geologic examples; Imaging problems: Coding, Sparsity (2004 Daubechies, Defrise and De Mol [8])

**S2. Decisive modelling.** (1983 Sabatier [3, 9]): For some illposed problems, there is a set of wellposed questions whose answer leads to a decision. For instance, it determines a bound (e.g., on density, or temperature), which is common to all the admissible solutions, and whose value justifies a decision. Remodelling by selecting such questions forgets any descriptive image.

5 Inverse Problems and experiments

**S Time Reversal: Theory or Experiments enable time reversing propagation of acoustic, elastodynamic, or e.m. waves governed by a linear pde, solving the source IP.** **Acoustical Example:** (Fink and Praha [10])

**Experiment:** TRC is a 2d transducer array that samples the wavefield; receiving amplifiers, a storage memory and programmable transmitters are able to synthetize a time reversed version of the stored signal. On the surface $S$ surrounding a source, TRC recombine into it the signal (which is a Green function in the case of a point source at $r_0$). **Theory:** Assume we measure the (pressure) field and its normal derivative at any point of $S$ during the interval $[0,T]$. Then, removing the initial (point) source at $r_0$, “create” on the cavity surface monopole and dipole sources that correspond to time reversal of these same measured components: one can show that a time reversed pressure field $\delta^{tr}(r_1,t_1)$ propagates inside the cavity. Spatial reciprocity and time reversal invariance of the (linear) wave equation yield $\delta^{tr}(r_1,t_1) = G(r_1,T-t_1 | r_0,t_0) - G(r_1,t_1 | r_0, T-t_0)$ which can be interpreted as the superposition of incoming and outgoing waves, centered on the initial source position: thus the time reversed field shows two wavefronts: the second one is a replica of the first one, multiplied by $-1$. It follows that source can be reconstructed from recorded values on the surface of the cavity, provided long enough time of recording before time reversing. Furthermore, elastic scattering on inhomogeneities do not alter efficiency. Method works in complex media, with scattering by many targets, which can increase the apparent mirror size. Hence it enables in special cases superresolution or detecting an object behind an opaque obstacle.

**S Invisibility:** Absorbers in scattering (1963, Weston...), strong nonuniqueness, and scattering ‘transparent’ parameters were the results on invisibility before 2000. They suggested similar studies (Uhlmann ... ) on Calderon Problem. A new idea of research on invisibility appeared with Invisibility cloaking. The theory suggests how electromagnetic (or other) currents can be managed (Pendry...) to surround a cavity by a ‘cloak’ of invisibility, using ‘metamaterials’ to match continuously the waves propagation.
6 Inverse Methods: (since 1967), and Final Remarks

One seeks a solution of a nlpde, e.g. \( u(x, t) \) of KdV, by using direct and inverse scattering (or spectral) problems of a linear operator as one solves linear p.d.e. by using Fourier and Inverse Fourier transforms on the scattering (or spectral) data \( s(t) \):

\[
\begin{align*}
  u(., 0) &\Rightarrow d.s.p. \Rightarrow s(0) \Rightarrow s(t) \Rightarrow i.s.p. \Rightarrow u(., t)
\end{align*}
\]

Write down the nlpde as consistency condition ("Lax equation") for spectral-scattering problems of linear operators depending on a common parameter, the nlpde's solution.

I reviewed KdV case and several generalisations for the 25th anniversary of Inverse Problems: solving the Schrödinger 1d scattering IP gives a solution of KdV, the transparent potential giving a soliton. A similar sequence of similar formulas can be managed with the Zacharov-Shabat spectral Problem and is associated to the Nonlinear Schrödinger Equation. Many other generalisations showed nonlinear coherent structures, e.g. boomerons, trappons, kinks, and hierarchies of "integrable" nlpde [4]. As a matter of fact, developments are now concerned not only with integrability and nonlinear evolution equations but also with many other points of dynamical systems. However, and even in methods dealing with 'inverse of inverse problems', Inverse Methods remain a huge branch of Inverse Theory: The author worked mainly on them after 1985, together with many people coming from our field (e.g. Ablowitz, Boiti, Calogero, Degasperis, Fokas, Kaup, Kruskal, Lax, Léon, Segur, Shabat, Zacharov).

**Final Remarks** Ideas and techniques naturally appeared in applications of inverse problems to various imagings (Soundings of Solid Earth, sonar, radar, lidar, Xraytomography, gammagraphy, nmr, magnetoecephalography, thermography, muons or neutrinos tomography of volcanos, optical diffuse imaging [3, 11, 12]), and in those of inverse methods, e.g. nonlinear signal analysis (Osborne) of waves governed by integrable equations (Calogero-Eckhaus class) [3, 4]. The remarkable 'interdisciplinarity' of this research continuously made us problems with specialists after 1990 [13], illustrating the GB Shaw maxim: "no man can be a pure specialist without being in the strict sense an idiot".

References

Regularization methods

A K Louis

Saarland University, Germany

E-mail: louis@num.uni-sb.de

Abstract: Characteristic for ill-posed problems is the lack of information for determining an acceptable solution in the equation under consideration. The unavoidable data noise necessitates an inexact, of course controlled inexact, solution with a trade-off between damping of noise and accuracy in the solution. Principle strategies are the reduction of the information content in the solution or the addition of hopefully correct information to the problem. We collect results on linear problems, where a detailed analysis is available, and then we present actual developments for nonlinear problems.

Keywords: Regularization methods, Tikhonov-Phillips, approximate inverse, optimization.

1 Introduction

The definition of ill-posed problems dates back to the work of Hadamard in the beginning of the 20th century. He first calls a problem well or correctly posed if the problem has a unique solution for all admissible right-hand sides. He then added the demand for continuous dependency of the solution on the data for suitable topologies. In the case of linear problems in Hilbert spaces the concept of pseudo solutions resolves the problems of existence and uniqueness of the solution. The choice of a smaller space for the data or a larger space for the solutions is often proposed as a remedy for the instability. But for the modeling of the problem the image space has to be large enough to contain the typical data errors. Here continuity assumptions for example would exclude the typical data noise. Consequently the dependence of the solution on the data is not continuous. Continuity is established by the concept of stabilization or regularization, where a solution is defined for all possible data and dependent on a regularization parameter. The definition of regularization can be described by the condition that the regularized solution depending on a regularization parameter tends to the pseudo solution if the data error tends to zero. For theoretical analysis this is very helpful, for practical problems with unavoidable non-zero noise level the selection of the suitable parameter is a crucial task.

For the sake of simplicity we present the main ideas first for linear problems in Hilbert spaces. Recent advances in Banach spaces are extremely important but a detailed presentation would exceed as well as a complete list of references the page limit for this short presentation. Similarly we present roughly ideas for nonlinear problems.

2 Linear problems

Linear inverse problems have in this fast developing field a rather long history of around 60 years. Actually what we now call Tikhonov-Phillips regularization was already used for solving nonlinear problems in each step for the linearized problems by Levenberg-Marquardt. The basic idea is to introduce additional information independently and on complete different level of abstraction achieved in the early 60ies by Tikhonov in the former Soviet Union and by Phillips in the U.S.A. The history was repeated at the end of the 60ies reducing, as we nowadays call it, the resolution content of the solution by Likht and Backus-Gilbert.

Recent advances comprise additional information like sparsity or compressed sensing.
We start off from the definition of regularization methods for linear problems in Banach spaces to solve for \( f \) the equation

\[
Af = g .
\]

**Definition 2.1** Let \( X, Y \) be Banach spaces and let \( A : X \to Y \) be linear, continuous and injective. Let \( S_\gamma : Y \to X \) be continuous and

\[
\lim_{\gamma \to 0} S_\gamma Af = f
\]

for all \( f \in X \). Then we call \( S_\gamma \) a regularization method.

If we restrict our considerations to Hilbert spaces the injectivity is not imposed. This definition can be extended to include the reconstruction not only of the function \( f \) itself, even the calculation of features, where for a given observation operator \( L \) directly \( Lf \) is computed, avoiding that way first calculating \( f \) and then \( Lf \). For edge detectors \( L \) typically is a differential operator, \([4]\).

### Addition of information

The classical way of stabilizing a problem is by the introduction of additional information. Very often this is achieved by variational methods where besides the data confidential term a side condition with wanted properties of the approximate solution is minimized leading to Tikhonov-Phillips type methods. Already simple examples however show that the introduction of correct information may lead in the case of erroneous data to worse results. When introducing in a Hilbert space setting a weighted residual for the data error as

\[
\|Af - g\|^2 = (Af - g)^* C (Af - g)
\]

with a positive definite, selfadjoint operator \( C : Y \to Y \), then the variational problem can be formulated as

\[
\min_{f \in X} \|Af - g\|^2 + \gamma \Omega(f)
\]

where in \( \Omega \) the additional information is encoded. In the simplest case \( \Omega \) is represented as

\[
\Omega(f) = \|f\|^2 = f^* B^* B f
\]

with \( B^* B : X \to X \) positive definite operator, then the optimization problem reduces to the solution of a linear problem

\[
\left( A^* C A + \gamma B^* B \right) f_\gamma = A^* C g .
\]

The selection of the operators \( B \) and \( C \) depends on the user’s preferences. If the operators are considered as weighting of the solution and of the residuals respectively the approach is considered as deterministic, if they are constructed as covariance operators then one faces a Bayesian approach for Gaussian noise, hence a stochastic approach.

In the case of Poisson noise one is bound to a Bayesian approach, here the problem becomes nonlinear and methods like expectation maximization or multiplicative ART methods are powerful tools. We come back to this in the course of iterative methods.

Other realizations of Tikhonov Phillips type methods comprise

<table>
<thead>
<tr>
<th>( \Omega(f) )</th>
<th>name</th>
</tr>
</thead>
<tbody>
<tr>
<td>( - \int f(x) \log(f(x)) dx )</td>
<td>maximum entropy</td>
</tr>
<tr>
<td>( \int</td>
<td>\nabla f(x)</td>
</tr>
<tr>
<td>( | f |<em>{L_0} ) or ( | f |</em>{L_1} )</td>
<td>sparsity</td>
</tr>
</tbody>
</table>

In each of these cases the variational problem becomes nonlinear, even if the original problem was linear. The same statement is true if problems in Banach spaces are considered.

If the problem is formulated in Hilbert spaces and the operator is compact, the singular value decomposition allows for simple convergence analysis. The speed of convergence can be analyzed if additional smoothness of the solution \( f \) is considered. This is not surprising when one compares with classical numerical approaches: the difference quotient converges to the derivative for differentiable functions. If the function fulfills further smoothness conditions the speed of convergence can be specified. The role of the smoothness is played by source
conditions, often expressed by norms generated by powers of $A^*A$ or by the singular value decomposition.

In many situations the regularization method can be represented as a numerical filter on the expansion with respect to the singular value decomposition \{v_n, u_n; \sigma_n\} where $Av_n = \sigma_n u_n$ as

$$T_\gamma g = \sum_{n=0}^{\infty} F_\gamma(\sigma_n) \frac{1}{\sigma_n} (g, u_n) v_n$$

(7)

The case $F = 1$ represents the minimum norm solution. Conditions for being regularization methods can be expressed by boundedness of $F_\gamma$ and its pointwise convergence to 1. For the above mentioned Tikhonov-Phillips method with $B$ and $C$ being the identity the filter $F_\gamma$ is given as

$$F_\gamma(\sigma_n) = \frac{\sigma_n^2}{\sigma_n^2 + \gamma}.$$

Another type of regularization methods is represented by iterative methods. Here one observes in the case of noisy data a so-called semiconvergence: the error starts decreasing up to the point where the influence of the data error becomes sufficiently strong, then the error increases. Consequently the regularization parameter is the stopping of the iteration. Again the solution is calculated inexactly. Prominent examples are Landweber iteration or the method of conjugate gradients applied to the normal equation. Both methods can be expresses in the form of filtered versions of the singular value decomposition. For the Landweber iteration

$$f^{m+1} = (I - \beta A^*A) f^m + \beta A^* g$$

it takes for initial guess $f^0 = 0$ on the form

$$F_m(\sigma) = 1 - (1 - \beta \sigma^2)^m$$

with $m$ being the iteration number. For the conjugate gradient method the filter comprises discrete orthogonal polynomials, and especially the stability analysis, which goes back to Nemirovski, is very tedious. If the iteration operator is not a function of $A^*A$ the analysis becomes much more difficult. For the well-known Kaczmarz method, often used in computerized tomography a semiconvergence result was proven only recently. In tomography the method is called ART as abbreviation of algebraic reconstruction method. The itates are projected on the hyperplanes representing the linear equations. Multiplicative ART applied for expectation maximization has the form

$$f^{m+1} = \frac{1}{A^* b} f^m A^* - \frac{b}{A^* f^m}$$

where the quotient is to be understood pointwise. Here positivity is preserved and the convergence analysis is completely different from the above mentioned methods.

Critical in all these methods is the selection of the regularization parameter $\gamma$ or the stopping index in the iterations $m$ where $\gamma$ can be interpreted as $\gamma = 1/m$. If a priori information about data errors and bounds on the solution are known, then a priori rules can be formulated. This mostly is not the case. Then criteria like choosing the parameter such that the defect is of the same size as the noise level are known as Morozov type rules. For many application Hansen’s L curve criterion is a useful technique, although this approach is occasionally criticized as not leading to an order-optimal method. But if only measured data with unknown noise level are available it is often the method of choice in the applications.

**Reducing the information content in the solution**

An obvious way to reduce the information content is truncated singular value decomposition. This can be interpreted as a filter method where the filter $F_\gamma(\sigma_n) = 1$ for singular values larger than a certain threshold, and 0 otherwise. In contrast to the above mentioned methods the expansion coefficients are not changed to something that does prevent the noise to prevail in the solution, they are simply cut off. But the application necessitates the knowledge of the singular value decomposition. For analyzing the solution process this is a helpful, but expansive tool.

The projection of the solution to a finite dimensional subspace is also known to be a regularization method.
This result goes back to Natterer. The regularization parameter is the stepsize in the discretization, but changing the regularization parameter means to completely restart the calculation.

A completely different approach was presented by Likht and Backus-Gilbert. Likht showed theoretically that calculating linear functionals on the solution can stabilize the problem. At a first glance that looks like changing the topology in the space \( X \) and so as changing the problem. Backus and Gilbert numerically calculated such functionals, although they considered their approach completely different. Especially their method is extremely time consuming so that its perception was a method to evaluate the solution only at a few points.

The method of approximate inverse starts from a different approach and was motivated by applications in computerized tomography. The aim was to precompute a solution operator starting from an approximation of the delta distribution. The regularization parameter here is the ‘difference’ to the delta distribution. In Backus-Gilbert they consider the ‘deltaness’ of their reconstruction kernel, which they could not influence.

If \( \delta_x^x \) is the approximation of the delta distribution for the point \( x \) the following auxiliary problem has to be solved

\[
A^* \psi_x^x = \delta_x^x
\]  

which can be replaced by the normal equation \( A A^* \psi_x^x = A \delta_x^x \) if \( \delta_x^x \) is not in the range of \( A^* \), leading to a stabilized minimum norm solution of the original problem. The reconstruction kernel \( \psi_x^x \) is precomputed independent of the data. The regularized solution is then determined as

\[
S^x g(x) = \psi_x^x g .
\]  

If the same problem frequently has to be solved for different measured data this approach is very efficient, especially if invariances of the problem can be applied. For details see [2].

The above mentioned filter methods can be interpreted as special cases when the approximation of the delta distribution is considered as \( \delta_x^x = \sum F_n(\sigma_n) v_n v_n(x) \). Also feature reconstructions are possible, [4].

Here mostly the ideas are presented in Hilbert spaces. Generalizations to Banach spaces are in the centre of present theoretical research, but also for numerical evaluations. As monography [5] is recommended.

### 3 Nonlinear problems

In contrast to the well-developed theory for linear problems a similar statement for nonlinear problem does not hold. The seminal work by Engl, Kunisch, Neubauer [6] opened the door for the development of the regularization theory. As many practical problems are nonlinear numerical approaches were developed. In the well-known Levenberg-Marquardt method from the 50ies the nonlinear problem was linearized and in each iteration step the linear problems were stabilized in a way later known as Tikhonov-Phillips method.

The first problem in nonlinear equations is that the solution is typically non unique. Hence in the definition of the minimum norm solution one looks for a minimum norm solution near a prior \( f^* \). In [6] the minimization of the sum of defect and penalty term is then considered. As numerical methods then inexact Newton methods were studied; inexact means that the linear problems in each iteration step are not exactly solved. As the whole problem is practically never exactly solvable, the experience from the linear regularization theory tells to stop the iteration dependent on the noise in the data. For a monograph for iterative solvers see [7].

In recent times also sparsity are considered as additional information. The sparsity is controlled by the \( L_0 \) norm which leads to NC complete problems. As compromise this norm is replaced by the \( L_1 \) norm. The minimization is then formulated as optimization problem and the powerful techniques developed there are applied. For details see the Topical Review [8]. With compressed sensing the story continues.

### Acknowledgments

The author thanks the Deutsche Forschungsgemeinschaft for supporting the work under grant Lo310/11-2 and Lo310/13-1.
References

Part II
Minisymposia
Minisymposium M2
Asymptotic Expansions
Reconstruction of inclusions in photothermal imaging

A Carpio\(^1\) and M L Rapun\(^2\)

\(^1\)Universidad Complutense, Spain
\(^2\)Universidad Politécnica de Madrid, Spain

E-mail: carpio@mat.ucm.es

Abstract: Photothermal imaging allows to inspect the structure of composite materials by means of non-destructive tests. The surface of a medium is heated at a number of locations. The resulting temperature field is recorded on the same surface. Thermal waves are strongly damped. Robust schemes are needed to reconstruct the structure of the medium from the decaying time dependent temperature field. The inverse problem is formulated as a weighted optimization problem with a time dependent constraint. The inclusions buried in the medium and their material constants are the design variables. We propose an approximation scheme in two steps. First, Laplace transforms are used to generate an approximate optimization problem with a small number of stationary constraints. Then, we implement a descent strategy alternating topological derivative techniques to reconstruct the geometry of inclusions with gradient methods to identify their material parameters. Numerical simulations assess the effectivity of the technique.

Keywords: photothermal imaging, topological derivatives, gradient methods, constrained optimization

1 Introduction

Photothermal imaging has arisen as an effective tool for nondestructive testing of composite materials. The surface of a semi-infinite medium is heated with a laser beam. The temperature is then measured at a number of receptors placed on the surface during a time interval, see Fig. 1. This technique has been employed in recent experiments by [1, 2]. Here, we develop a theoretical framework to process the measured data and reconstruct the structure of inclusions buried in the medium.

The imaging set-up is depicted in Fig. 1. Let \(\kappa_e\) be the thermal conductivity of the exterior medium and \(\rho_e\) the density multiplied by its specific heat. Inclusions with thermal parameters \(\kappa_i\) and \(\rho_i\) are buried in it. The inverse problem consists in finding the inclusions \(\Omega\) and their material parameters \(\kappa_i, \rho_i\) such that
the temperature field measured at the detector locations $x_1, \ldots, x_M \in \Pi$ at times $t_1, \ldots, t_N$ agrees with the solution of the corresponding forward problem. This is a transmission problem for the heat equation governing the temperature field

$$U(x, t) := \begin{cases} U_e(x, t), & \text{in } \Omega_e \times (0, \infty) := (\mathbb{R}^2 \setminus \Omega) \times (0, \infty), \\ U_i(x, t), & \text{in } \Omega \times (0, \infty), \end{cases}$$

given by

$$\begin{cases} \rho_e \partial_t U_e - \kappa_e \Delta U_e = 0, & \text{in } \Omega_e \times (0, \infty), \\ \rho_i \partial_t U_i - \kappa_i \Delta U_i = 0, & \text{in } \Omega \times (0, \infty), \\ U_e - U_i = U_{\text{inc}}, & \text{on } \partial \Omega \times (0, \infty), \\ \kappa_i \partial_n U_i - \kappa_e \partial_n U_e = \kappa_e \partial_n U_{\text{inc}}, & \text{on } \partial \Omega \times (0, \infty), \\ U_e(x, 0) = U_i(x, 0) = 0, & \forall x \in \mathbb{R}^2, \end{cases} \tag{1}$$

where $\mathbb{R}^2 := \{(x, y) \in \mathbb{R}^2, \ y < 0\}$ and $\Pi := \{(x, 0), \ x \in \mathbb{R}\}$. The surface of the sample $\Pi$ is thermically excited at a source point $x_0 \in \Pi$ with a delta-pulse source, producing a thermal wave

$$U_{\text{inc}}(x, t) = \frac{1}{t} \exp\left(-\frac{\rho_e |x - x_0|^2}{4 \kappa_e t}\right), \quad x \in \mathbb{R}^2, \ t > 0. \tag{2}$$

Adiabatic boundary conditions are imposed on the upper boundary $\Pi$.

The inverse problem is regularized using a constrained variational reformulation: Determine regions $\Omega$ and parameters $\kappa_i, \rho_i$ minimizing the functional

$$J(\mathbb{R}_-^2 \setminus \Omega, \kappa_i, \rho_i) = \frac{1}{2} \sum_{k=1}^M \sum_{j=1}^N f(t_j) (U_{\text{total}}(x_k, t_j) - U_{\text{meas}}(x_k, t_j))^2, \tag{3}$$

where $U_{\text{total}}$ is the solution of the time dependent forward problem (1) for an inclusion $\Omega$ with thermal parameters $\kappa_i$ and $\rho_i$. The forward problem acts as a constraint. The weight function $f(t)$ normalizes the time decay of solutions of the heat equation:

$$f(t) = \max_{x \in \{x_1, \ldots, x_M\}} |U_{\text{meas}}(x, t)|^{-1}.$$

Thermal waves are strongly damped. Time dependent weights prevent losing information as time grows.

2 Approximation using Laplace transforms

The time dependent heat problem can be efficiently solved combining Laplace transforms and stationary boundary element formulations. This suggests an alternative approximate variational formulation for the inverse problem, involving a small number of stationary constraints.

For each value of $s$, the Laplace transform of $U$

$$u_s(x) = \int_0^\infty e^{-st}U(x, t)dt,$$

is a solution of a Helmholtz transmission problem with complex wave numbers depending on the parameter $s$ [3]:

$$\begin{cases} \kappa_e \Delta u_s - s \rho_e u_s = 0, & \text{in } \Omega_e, \\ \kappa_i \Delta u_s - s \rho_i u_s = 0, & \text{in } \Omega, \\ u_s^+ - u_s^- = u_{\text{inc},s}, & \text{on } \partial \Omega, \\ \kappa_i \partial_n u_s^+ - \kappa_e \partial_n u_s^- = \kappa_e \partial_n u_{\text{inc},s}, & \text{on } \partial \Omega, \\ \partial_n u_s = 0, & \text{on } \Pi, \end{cases} \tag{4}$$

where $u_{\text{inc},s} = \int_0^\infty e^{-st}U_{\text{inc}}(x, t)dt$. The Laplace transform is then inverted choosing hyperbolic paths [4]:

$$\gamma(\theta) := \mu(1 - \sin(\pi/4 + i \theta)), \quad \theta \in \mathbb{R}, \text{ where } \mu > 0.$$ 

The solution of (1) takes the form

$$U(x, t) = \frac{1}{2\pi i} \int_{-\infty}^{\infty} e^{\gamma(\theta)} u_{\gamma(\theta)}(x) \gamma'(\theta)d\theta.$$
A truncated trapezoidal rule yields an approximation of $U$:

$$
U(x, t) \approx \sum_{\ell=-L}^{L} c_{\ell} e^{i\ell s} u_{s\ell}(x)
$$

with nodes and weights:

$$
s_{\ell} = \gamma \left( \frac{\log(L)}{L} \ell \right), \quad c_{\ell} = \frac{\log(L)}{2\pi L} \gamma' \left( \frac{\log(L)}{L} \ell \right).
$$

As in [5, 6], we replace the original cost functional (3) by the approximate functional

$$
J(\mathbb{R}^2 \setminus \Omega, \kappa_i, \rho_i) = \frac{1}{2} \sum_{k=1}^{M} \sum_{j=1}^{N} f(t_j) \left( \sum_{\ell=-L}^{L} c_{\ell} e^{i\ell s} u_{s\ell}(x_k) - U_{\text{meas}}(x_k, t_j) \right)^2,
$$

(5)

involving now $2L + 1$ stationary constraints: the Helmholtz transmission problems for the coefficient functions $u_{s\ell}(x)$, $\ell = -L, \ldots, L$.

3 Descent strategy with respect to inclusions and parameters

To approximate solutions of the optimization problem we resort to a descent technique. Sequences of approximate inclusions and parameters are generated along which the cost functional (5) decreases. Descent with respect of the domains is implemented using topological derivatives. Descent directions for the parameters are determined using gradient methods.

Let us fix an initial guess of the parameters $\kappa_i = \kappa_i^0$ and $\rho_i = \rho_i^0$. A first guess of the inclusions is obtained calculating the topological derivative of the resulting shape functional.

The topological derivative $D_T(x, \mathcal{R})$ of a shape functional $J(\mathcal{R})$ is a scalar function of $x \in \mathcal{R}$ that provides asymptotic expansions of the form [7]:

$$
J(\mathbb{R}^2 \setminus B_\varepsilon(x)) = J(\mathcal{R}) + D_T(x, \mathcal{R})\pi \varepsilon^2 + o(\varepsilon^2), \quad \text{as} \quad \varepsilon \to 0.
$$

Placing small inclusions $B_\varepsilon(x) = B(x, \varepsilon)$ at the points $x \in \mathcal{R}$ where the topological derivative is negative the value of the functional decreases. Points where the topological derivative attains the largest negative values are likely to belong to inclusions of different materials. A first approximation $\Omega_1$ of $\Omega$ is defined as the set of all the points where the topological derivative of $J(\mathbb{R}^2 \setminus \Omega_1, \kappa_1, \rho_1)$ falls below a negative threshold (see [5, 8] for guidelines of the selection of such constant).

The evaluation of the topological derivative is effectively performed exploring explicit expressions in terms of forward and adjoint fields. Similarly to Theorem 3.2 in [5] we can prove

**Theorem.** Fixing $\kappa_i, \rho_i$, the topological derivative of the functional $J(\mathbb{R}^2 \setminus \Omega, \kappa_i, \rho_i)$ defined in (5) is

$$
D_T(x) = \text{Re} \left( \sum_{\ell=-L}^{L} \frac{2\kappa_i \Delta p_{s\ell} - s_{\ell} \rho_s p_{s\ell}}{\kappa_e + \kappa_i} \nabla u_{\text{total}, s\ell}(x) \nabla p_{s\ell}(x) + (\rho_e - \rho_i) s_{\ell} u_{\text{total}, s\ell}(x)p_{s\ell}(x) \right) \quad (6)
$$

when $x \in \mathbb{R}^2 \setminus \Omega$, where $u_{\text{total}, s\ell} = u_{\text{inc}, s\ell} + u_{s\ell}$, and $u_{s\ell}$ is a solution of (4) for $s = s_{\ell}$. The fields $p_{s\ell}$ are solutions of adjoint problems:

$$
\begin{align*}
\begin{cases}
\kappa_e \Delta p_{s\ell} - s_{\ell} \rho_s p_{s\ell} = g_{s\ell}, & \text{in } \Omega_e, \\
\rho_e p_{s\ell}^+ - \kappa_e \Delta p_{s\ell}^+ - s_{\ell} \rho_s p_{s\ell}^+ = 0, & \text{in } \Omega, \\
p_{s\ell} - p_{s\ell}^+ = 0, & \text{on } \partial\Omega, \\
\partial_n p_{s\ell} = 0, & \text{on } \partial\Omega,
\end{cases}
\end{align*}
$$

(7)

where $g_{s\ell}(x) := \sum_{i=1}^{M} \sum_{j=1}^{N} f(t_j) c_{s\ell} e^{i\ell s} \left( U_{\text{meas}}(x_i, t_j) - \sum_{k=-L}^{L} c_k e^{i\ell s} u_{s\ell}(x_i) \right) \delta(x_i)$.

To determine $\Omega_1$, we set $\Omega = \emptyset$, $\kappa_i = \kappa_i^0$, $\rho_i = \rho_i^0$. Given an approximation $\Omega_d$ for fixed $\kappa_i, \rho_i$, a better one can be determined computing the topological derivative of $J(\mathbb{R}^2 \setminus \Omega_d, \kappa_i, \rho_i)$ and adding to $\Omega_d$ points where it falls below a negative threshold.
Once a first approximation of the inclusion \( \Omega_1 \) is selected, the parameters are updated using correctors provided by a gradient method. Given approximations \( \kappa_{i}^{q-1}, \rho_{i}^{q-1} \) of the material parameters, we define \( \kappa_{i}^{q} = \kappa_{i}^{q-1} + \eta \phi_{i}^{q} \) and \( \rho_{i}^{q} = \rho_{i}^{q-1} + \eta \psi_{i}^{q} \), for small \( \eta > 0 \). The numbers \( \phi_{i}^{q}, \psi_{i}^{q} \) are selected differentiating

\[
J(\eta) := J(\Omega_{d}^{0}, \kappa_{i}^{q-1} + \eta \phi_{i}^{q}, \rho_{i}^{q-1} + \eta \psi_{i}^{q}),
\]

with respect to \( \eta \) and imposing \( J'(0) < 0 \). Following \([5, 8]\) we prove that the choice

\[
\phi_{i}^{q} = \text{Re} \left( \frac{1}{\text{meas}(\Omega_{d})} \int_{\Omega_{d}} \sum_{\ell=-L}^{L} \nabla u_{s_{\ell}} \nabla p_{s_{\ell}} \right), \quad \psi_{i}^{q} = \text{Re} \left( \frac{1}{\text{meas}(\Omega_{d})} \int_{\Omega_{d}} \sum_{\ell=-L}^{L} s_{\ell} u_{s_{\ell}} p_{s_{\ell}} \right),
\]

(8) ensures \( J'(0) < 0 \). The fields \( u_{s_{\ell}}, p_{s_{\ell}} \) are solutions of (4) and (7), respectively, taking \( \Omega = \Omega_{d}, \kappa_{i} = \kappa_{i}^{q-1} \) and \( \rho_{i} = \rho_{i}^{q-1} \).

Our algorithm alternates a few gradient iterations to correct the parameters with a topological derivative evaluation to update the domains. Once the parameters are corrected, we compute again the topological derivative to update the domains, and update the current approximation by adding to the previous domain the points where the topological derivative attains now the largest negative values. Once the approximation of the domains is improved we perform further gradient iterations to update the parameters and so on. The algorithm stops when either \( \text{meas}(\Omega_{d} \setminus \Omega_{d-1}) \), \( |\kappa_{i}^{q} - \kappa_{i,1}^{q-1}| + |\rho_{i}^{q} - \rho_{i,1}^{q-1}| \) and \( \|U_{\text{meas}} - U_{\text{total}}\| \) are small, or \( J(\mathbb{R}^{2} \setminus \Omega_{d}, \kappa_{i}^{q}, \rho_{i}^{q}) \) is small.

When \( \Omega = \bigcup_{j=1}^{D} \Omega_{d,j} \) and the parameters \( \kappa_{i}, \rho_{i} \) are piecewise constant with values \( \kappa_{i,j}, \rho_{i,j} \) inside \( \Omega_{d,j} \), the method can be generalized as follows. Given an approximation \( \Omega_{d} := \bigcup_{j=1}^{D'} \Omega_{d,j} \) \( (D' \) is not necessarily equal to the true number of defects \( D) \), the values of the parameters are updated by \( \kappa_{i,j}^{q} = \kappa_{i,j}^{q-1} + \eta \phi_{i,j}^{q} \) and \( \rho_{i,j}^{q} = \rho_{i,j}^{q-1} + \eta \psi_{i,j}^{q} \) with \( \phi_{i,j}^{q} = \text{Re} \left( \frac{1}{\text{meas}(\Omega_{d,j})} \int_{\Omega_{d,j}} \sum_{\ell=-L}^{L} \nabla u_{s_{\ell}} \nabla p_{s_{\ell}} \right) \) and \( \psi_{i,j}^{q} = \text{Re} \left( \frac{1}{\text{meas}(\Omega_{d,j})} \int_{\Omega_{d,j}} \sum_{\ell=-L}^{L} s_{\ell} u_{s_{\ell}} p_{s_{\ell}} \right) \).

Now \( u_{s_{\ell}} \) and \( p_{s_{\ell}} \) solve (4) and (7) with \( \Omega = \bigcup_{j=1}^{D'} \Omega_{d,j} \), and thermal parameters \( \kappa_{i} = \kappa_{i,j}^{q-1} \) and \( \rho_{i} = \rho_{i,j}^{q-1} \) in \( \Omega_{d,j} \). The topological derivative is computed using (6) with \( u_{s_{\ell}} \) and \( p_{s_{\ell}} \) defined as above (considering the current piecewise constant values of the parameters). The values of \( \kappa_{i}, \rho_{i} \) in \( \mathbb{R}^{2} \setminus \Omega_{d} \) are taken as the initial guesses \( \kappa_{i}^{0}, \rho_{i}^{0} \).

4 Numerical example

We consider the reconstruction of two inclusions of different sizes located at different depths. To simplify computations, we assume that the conductivities of the inclusions are known and equal to the exterior one: \( \kappa_{e} = \kappa_{i,1} = \kappa_{i,2} = 1 \). In the exterior media \( \rho_{e} = 0.2 \). The unknowns of the inverse problem (represented in Fig. 2(a)) are the two defects \( \Omega_{1} \) and \( \Omega_{2} \) and the values of their densities \( \rho_{i,1} = 1 \) and \( \rho_{i,2} = 2 \). We generate 7 incident excitations at the sources represented by ‘*’ marks in all the plots in Fig. 2. The temperature is then measured at the 8 observation points represented by ‘x’ marks at 10 uniformly distributed times in the time interval \([0.05, 0.5]\). Data were generated solving the corresponding direct problem using the numerical method detailed in \([3, 8]\) and adding a 1% gaussian relative error at each observation point.

We start the algorithm setting \( \rho_{i}^{0} = 0.5 \) everywhere. The topological derivative (TD) of the functional \( J(\mathbb{R}^{2}, \rho_{i}^{0}) \) at the sampling region \([-3.5, 3.5] \times [-2.5, 0]\) is represented in Fig. 2(b). The regions where the TD attains large negative values (dark blue colors on the plot) characterize the expected location of the defects. We observe that all these points are concentrated in the same region, suggesting that only one defect is buried in the medium. Notice that these points belong to the true defect \( \Omega_{2} \), which is the biggest and is closest to the boundary II. Furthermore, the true value \( \rho_{i,2} \) provides a higher contrast with \( \rho_{e} \) than \( \rho_{i,1} \). The initial guess for \( \Omega \) is represented in Fig. 2(c). Now we update the initial value \( \rho_{i,2}^{0} = 0.5 \) by the gradient technique. After ten iterations we obtained the value \( \rho_{i,2}^{10} = 0.9644 \). In the next step a new TD computation is performed, yielding the domain represented in Fig. 2(d). The algorithm continues by alternating 10 gradient iterations with a TD computation. After three TD computations the smaller defect is detected. We take then as initial guess \( \rho_{i,1}^{30} = 0.5 \) (the superscript 30 means that globally we have already performed 30 gradient iterations, although they only affect the approximation of \( \rho_{i,2} \), which is \( \rho_{i,2}^{30} = 1.3313 \) at this stage). The current domains are represented in Fig. 2(e). After 9 iterations the algorithm stopped. The final reconstructed objects are given in Fig. 2(f). The estimated values of the parameters are \( \rho_{i,1}^{30} = 0.7292 \) and \( \rho_{i,2}^{30} = 1.8227 \). The overall
reconstruction is quite satisfactory taking into account that few data were available, distorted by noise. It is also remarkable that we were able to characterize not only the big object, but also the smaller one, which has less contrast with the exterior and is located further from the observation/source points.

Acknowledgments

The authors are partially supported by the Spanish Government research project TRA2010–18054 and the Spanish Ministerio de Economía y Competitividad Grant No. FIS2011–222888–C02–02.

References

Minisymposium M3

Inverse Source Problems
Reconstruction of refractive index from blind experimental data using approximate globally convergent and adaptive finite element methods

M V Klibanov¹, L Beilina², N T Thân¹ and J B Malmberg²
¹Department of Mathematics and Statistics University of North Carolina at Charlotte, USA
²Department of Mathematical Sciences, Chalmers University of Technology and Gothenburg University, Sweden
E-mail: mklibanv@uncc.edu

Abstract: We consider a coefficient inverse problem for a wave-like equation in which the unknown coefficient represents the refractive index of the medium of interest. A two-stage inverse method, which combines a globally convergent method and an adaptive finite element method, is proposed for reconstructing the coefficient. The key advantage of this two-stage method is that it can provide accurate reconstruction results without requiring a good initial guess of the coefficient of interest as opposed to conventional optimization approaches. In particular, we present reconstruction results for blind real experimental data. These results confirm the feasibility of the proposed algorithm in practical situations.

Keywords: Coefficient inverse problem, globally convergent method, adaptive finite element method, wave-like equation, experimental data.

1 Introduction

Our aim is the reconstruction of dielectrics from blind backscattered experimental data. The reconstruction is done from time domain data, as opposed to a more conventional case of frequency domain data. Experimental data were collected using a microwave scattering facility which was built at the University of North Carolina at Charlotte. This system sends electromagnetic pulses into the medium and collects the time resolved backscattered data on a part of a plane. The spatially distributed dielectric permittivity function \( \varepsilon_r(x) \), \( x \in \mathbb{R}^3 \) (the square of the refractive index) is the unknown coefficient in a Maxwell’s system. This coefficient is reconstructed from backscattered measured data in blind cases. To do this, an approximate globally convergent numerical method in combination with an adaptive finite element method is used in a two-stage procedure. The globally convergent method provides an approximation of the coefficient of interest without requiring a good first guess. This approximation is then improved by the adaptive finite element method. For more detail, we refer to the references cited herein.

In our model problem, the electromagnetic wave is approximated by a Cauchy problem for a wave-like equation [3, 5]

\[
\varepsilon_r(x) D_t^2 E_2(x, t) - \Delta E_2(x, t) = 0, \ x \in \mathbb{R}^3, \ t \in (0, \infty), \quad (1)
\]

\[
E_2(x, 0) = 0, \ D_t E_2(x, 0) = \delta(x - x_0). \quad (2)
\]

where \( E_2(x, t) \) is the initialized component of the electric field [3, 5]. Here \( D_t \) is the derivative operator with respect to \( t \). Let \( \Omega \subset \mathbb{R}^3 \) be a convex bounded domain with a piecewise smooth boundary and assume that
\( \varepsilon_r(x) \) is unknown in \( \Omega \) and \( \varepsilon_r(x) \equiv 1, \ x \in \mathbb{R}^3 \setminus \Omega \). Furthermore, we assume that \( x^0 \notin \bar{\Omega} \). Denote by \( \Gamma \) the part of \( \partial \Omega \) where the backscattered data is measured. The coefficient inverse problem is aimed at reconstructing \( \varepsilon_r(x), x \in \Omega \), given the following boundary data

\[
E_2(x, t) = g(x, t), x \in \Gamma, t \in (0, \infty).
\] (3)

2 The approximately globally convergent method

Using the Laplace transform \( w(x, s) = \int_0^\infty E_2(x, t)e^{-st}dt \), for \( s \geq \delta > 0 \), and defining the function \( v(x, s) = \ln(w(x, s))/s^2 \), we obtain the following equation:

\[
\varepsilon_r(x) = \Delta v(x, s) + 2s^2(\nabla v(x, s))^2, x \in \Omega, s \geq \delta.
\] (4)

To eliminate the unknown coefficient \( \varepsilon_r \), we differentiate (4) with respect to \( s \). Define \( q = D_s v \) and write \( v = V - \int_0^s q(\cdot, \tau)d\tau \), where \( V = v(\cdot, \bar{s}) \) is the so-called tail function. Functions \( q \) and \( V \) satisfy the following integral differential equation

\[
\Delta q - 2s^2\nabla q \cdot \int_s^\bar{s} \nabla q(x, \tau)d\tau + 2s^2\nabla V \cdot \nabla q + 2s \int_s^\bar{s} \nabla q(x, \tau)d\tau \left(2s^2(\nabla V)^2 + 2s\nabla |V|^2\right) = 0, x \in \Omega.
\] (5)

Moreover, \( q \) satisfies the boundary condition \( q(x, s) = \varphi(x, s), x \in \Gamma \) with \( \varphi(x, s) \) derived from the measured data \( g(x, t) \). The missing boundary data of \( q \) on \( \partial \Omega \setminus \Gamma \) is completed using simulated data for homogeneous medium. Therefore, in the following, we assume that \( \varphi \) is given on the whole boundary \( \partial \Omega \).

To solve (5), we consider a partition \( s = s_N < s_{N-1} < \cdots < s_0 = \bar{s} \) for a chosen interval \( [s, \bar{s}] \), with step size \( s_n - s_{n+1} = h, n = 0, 1, \ldots, N - 1 \). We approximate \( q \) by piecewise constant function \( q(\cdot, s) = q_n \) on \( [s_{n+1}, s_n) \), for \( n = 0, 1, \ldots, N - 1 \). We obtain a boundary value problem for a nonlinear system on each interval \( [s_{n+1}, s_n) \):

\[
\Delta q_n + A_n \nabla W_n \cdot \nabla q_n = B_n(\nabla q_n)^2 + C_n(\nabla W_n)^2,
\] (6)

\[
q_n = \varphi_n, x \in \partial \Omega,
\] (7)

where \( W_n = V - \sum_{k=0}^{n-1} h q_k \) and \( \varphi_n \) is computed from \( \varphi \).

The functions \( q, v, V \) and \( \varepsilon_r \) are iteratively computed from an initial approximation for \( V \), derived from an approximate asymptotic expansion of \( V \) as \( s \to \infty \) [2, 3].

3 The adaptive finite element method

We derive the stabilized Maxwell’s system in a nonmagnetic medium with permeability \( \mu = 1 \) in the spatial domain \( \Omega \):

\[
\mathcal{M}_s \varepsilon \varepsilon \mathcal{E} = \varepsilon_r D_s^2 \mathcal{E} + \nabla \times (\nabla \times \mathcal{E}) - s \nabla (\nabla \cdot (\varepsilon_r \mathcal{E})) = 0,
\]

where \( \mathcal{E} = (E_1, E_2, E_3) = \mathcal{E}(x, t) \) is the electric field and \( s \geq 1 \) is a suitably chosen parameter. Boundary conditions for \( \mathcal{E} \) are obtained from experimental data and simulations in the same way as in the globally convergent method. Appropriate initial conditions are supplied.

In this stage, we improve the approximation \( \varepsilon_{r, \text{glob}} \) obtained in stage 1 by minimizing the Tikhonov functional:

\[
\Phi(\varepsilon_r) = \frac{1}{2} ||\mathcal{E} - \mathcal{E}_{\text{obs}}||^2_{L_2(S_T)} + \frac{\alpha}{2} ||\varepsilon_r - \varepsilon_{r, \text{glob}}||^2_{L_2(\Omega)}.
\]
Inverse Problems – from Theory to Applications (IPTA2014)

Here, $E_{\text{obs}}$ is the backscattered data on $S_T = \Gamma \times (0, T)$. $\alpha > 0$ is a regularization parameter. The minimization is performed via the Lagrangian

$$L(\varepsilon_r, E, \lambda) = \Phi(\varepsilon_r, E) + (M_{\varepsilon_r} E, \lambda).$$

Setting the Fréchet derivative $L'(\varepsilon_r, E, \lambda) = 0$ implies that $E$ solves the forward problem weakly and $\lambda$ solves an adjoint problem weakly. The gradient $\nabla L$ with respect to $\varepsilon_r$ is given by [4]:

$$\nabla L(\varepsilon_r, E, \lambda) = \alpha(\varepsilon_r - \varepsilon_r, \text{glob}) + \int_0^T (s(\nabla \cdot E)(\nabla \cdot \lambda) - D_1 E \cdot D_1 \lambda) dt.$$

(8)

In numerical computation, $E$, $\lambda$, $\varepsilon_r$ are approximated by $E_h$, $\lambda_h$, $\varepsilon_{r,h}$ from finite-dimensional spaces over triangulations $K_h$ of the domain $\Omega$ and $I_T$ of the time interval $(0, T)$.

A posteriori error estimates of [1] allow for adaptive mesh refinements. These refinements combined with gradient based minimization iterations gives the adaptive algorithm [4].

The flowchart of the two-stage algorithm is summarized in Figure 1.

---

**Figure 1.** Flowcharts for the two stages: (a) the approximately globally convergent method and, (b) the adaptive finite element method.

---

### 4 Numerical examples

Figures 2–4 show the results of the proposed algorithm for three different objects. The first object (Figure 2) was a wooden block. The second one (Figure 3) was a wooden doll which was empty inside, i.e., it was an example of heterogeneous objects. These first two objects were placed in air. The third object (Figure 4) was an empty ceramic mug buried in a sand box. This case modeled buried objects. We emphasize that since the refractive index of the ceramic mug was smaller than that of the sand, its signal was quite weak which
Figure 2. Reconstruction of a wooden block placed in air. Reconstruction in stage 1: (a) perspective view, (b) front view. The final mesh (c). Reconstruction in stage 2: (d) perspective view, (e) front view, and (f) zoomed in. Refractive index: Reconstructed 2.10, measured 2.14. The thin line indicate the true shape.

Figure 3. Reconstruction of a wooden doll placed in air. Reconstruction in stage 1: (a) perspective view, (b) front view. The final mesh (c). Reconstruction in stage 2: (d) perspective view, (e) front view, and (f) zoomed in. Refractive index: Reconstructed 1.89, measured 1.85. The thin line indicate the true shape.
made the reconstruction more difficult than the first two objects. In spite of that, we still could accurately
reconstruct it. Note that all these objects were blind to the authors until the reconstruction was completed.

Several numerical results presented in [3, 5, 6] have indicated that the proposed algorithm can work well
even for the very challenging case of blind real experimental data.

![Figure 4. Reconstruction of a ceramic mug buried in a sand box: (a) reconstruction in stage 1, (b) reconstruction in stage 2. Refractive index: Reconstructed 1.23, measured 1.39. The thin line indicate the true shape.](image)

**Acknowledgements**

This research was supported by US Army Research Laboratory and US Army Research Office grants W911NF-
11-1-0335 and W911NF-11-1-0399, the Swedish Research Council, the Swedish Foundation for Strategic Re-
search (SSF) through the Gothenburg Mathematical Modelling Centre (GMMC) and by the Swedish Institute,
Visby Program.

**References**

York: Springer)
[4] Beilina, L, Thành, NT, Klibanov, MV, Malmberg, JB 2014 Reconstruction of shapes and refractive indices from blind backscat-
An inverse problem for the heat equation in view of practical application

Takashi Takiguchi
National Defense Academy of Japan, Japan
E-mail: takashi@nda.ac.jp

Abstract: We discuss an inverse problem to detect an inclusion in a homogeneous medium applying the heat conduction, on which there are several prior researches. We discuss the points to be modified in the prior researches on this problem in order that they are applied for practice. We also try to give some modifications.

Keywords: inverse problem for the heat equation, generalized functions.

1 Introduction

We discuss an inverse problem to detect an inclusion in a homogeneous medium. As an approach to this problem, the application of the X-ray tomography is being studied. With application of the X-ray tomography, the following problems are under investigation; non-destructive testing for mixed materials of the two kinds, non-destructive testing for the fuel tank of the rockets, non-destructive testing for die casting of the aluminium and so on. The problem of non-destructive testing for mixed materials of the two kinds arose from the development of the three dimensional CAD system which would enable us to describe the inner structure of the pillars and the walls in the buildings. In this problem, it is necessary to investigate the internal structure of the pillars consisted of the steel and the aluminium, which is not clear from their production process. It seems that it is not difficult to understand the motivation to study the other two examples, which are typical problems in non-destructive testing. For the time being, the same algorithm as the computerized tomography (CT) is applied to all of the above examples. Since the objects in these problems are much simpler than the interior structure of the human body, it is expected to reduce the X-ray data for the reconstruction of the object. This problem is closely related to the geometric tomography and there are many studies on it both in the viewpoint of theory and in the viewpoint of application. For example, confer [1, 6, 7] for the results in the viewpoint of theory and [2, 8, 9, 10] for the studies in the viewpoint of practical application. Unfortunately, the results mentioned above are not still satisfactory for practical application in view of the following points.

- In the case where we project parallel beams of the X-rays from two directions, we can classify the shapes of the inclusions into the three classes, one is the uniquely determined ones by these data, another is the non-uniquely determined ones and the other is the class of the sets which allows no solution with such data ([6, 7, 8]). In practice, the study the last class has no meaning, since the existence of the solution is proved by the phenomenon itself. For the class of the uniquely determined sets, reconstruction formulas ([6, 8]) were given and further studies were developed on the treatment of the errors, the construction of a reconstruction algorithm and its implementation by computers and so on, which are satisfactory for practical application ([2, 8]). It was, however, proved that there are very few sets reconstructed by their two projections ([9]) and it is not known how to find the exact two directions for the reconstruction for the uniquely reconstructible sets, even if they exist.
In most industrial CT devices, cone beams of the X-rays are applied, in view of which we have to develop
the counterpart of the above theory for the cone beams.

For general inclusions, the exact data of the beams of the X-rays for the reconstruction are not known.
Needless to say their reconstruction methods, treatment of the errors, construction of an approximate
reconstruction algorithm, its implementation by computers and so on.

There are other problems of the use of the X-ray tomography.

(i) The cost of the testing is very expensive if we apply the X-ray tomography.

(ii) We cannot ignore harmful influence of the X-rays on the human body.

In order to solve the problems (i) and (ii), we try another approach. We study to detect an inclusion in a
homogeneous medium applying the heat conduction.

2 Prior researches

In this section, we review the known results on the inverse problems to detect an inclusion in a homoge-

In this case, is it possible to reconstruct $D$ by the boundary data $u|_{\partial \Omega}$ if we suitably control the heat flux

This is an inverse problem to apply “Neumann to Dirichlet” boundary data. In [3], M.Ikehata first studied
one-spatial case, where he investigated the essence of Problem 1. This result was extended by M.Ikehata and
M.Kawashita [4, 5] to the higher spatial dimensional cases. They proved that the convex hull, as well as some
other information, of the inclusion $D$ is reconstructed with the choice of a suitable adjoint solution of the heat
equation. Their theory being very excellent and beautiful as mathematical one, it seems that there are several
points to be modified in view of practical application.

In practical application, it is not easy give the heat flux as the boundary data. In addition to it, its
observation is not easy, either.

Though Ikehata-Kawashita controlled the input of the heat (flux) $f(x, t)$ on the whole boundary points
$x \in \partial \Omega$, in view of practical application, it is much easier to give only one point source $\delta(x_0)f(t)$ on a
fixed boundary point $x_0 \in \partial \Omega$.

3 Main problem

For the solution to the problems mentioned at the end of the last section, we study the following problem.
Problem 2. Let Ω be a bounded domain in \( \mathbb{R}^n \) \( n = 1, 2, 3 \) with smooth boundary. Let \( D \) be an open subset of \( \Omega \) with smooth boundary and satisfy that \( \overline{D} \subset \Omega \) and \( \Omega \setminus D \) is connected. We denote the unit outward normal vectors to \( \partial \Omega \) and \( \partial D \) by the same symbol \( \nu \). Let \( T > 0 \) be an arbitrary constant. Given \( f = f(t), \ t \in (0,T) \) and \( x_0 \in \partial \Omega \), let \( u = u(x,t) \) be the solution of the initial boundary value problem for the heat equation

\[
\begin{align*}
\partial_t u - \Delta u &= \delta(x_0)f(t) \quad \text{in } (\Omega \setminus D) \times (0,T), \\
\partial_n u &= 0 \quad \text{on } \partial D \times (0,T), \\
\partial_n u &= 0 \quad \text{on } \partial \Omega \times (0,T), \\
u(x,0) &= 0 \quad \text{in } \Omega \setminus D.
\end{align*}
\]

In this case, is it possible to reconstruct \( D \) by the boundary data \( u|_{\partial \Omega} \) if we suitably control the point heat source \( f(t) \) at \( x_0 \in \partial \Omega ?\)

It is our main purpose to study Problem 2, the solution of which solves the two problems mentioned at the end of the last section. As M. Ikehata [3] did, we first study the one-spatial dimensional case, the solution to which would tell us the essence of the problem. Specifically, we study the following problem.

Problem 3. Let \( \Omega = (0,X), D = (a,b), \ 0 < a < b < X \) and Let \( u(x,t) \) be the solution of the following initial and boundary value problem for the heat equation.

\[
\begin{align*}
\partial_t u - \partial_x^2 u &= \delta(x)f(t) \quad \text{in } (0,a) \times (0,T), \\
\partial_x u &= 0 \quad \text{on } \{0\} \times (0,T), \\
\partial_x u &= 0 \quad \text{on } \{a\} \times (0,T), \\
u(x,0) &= 0 \quad \text{for } x \in (0,a).
\end{align*}
\]

In this case, is it possible to recover \( a \) by the boundary data \( u(0,t) =: g(t) \) if we suitably control the heat source \( f(t) \) ?

In the initial and the boundary value problem (3), an inclusion \( D = (a,b) \) is included in a homogeneous medium \( \Omega = (0,X) \). We pose an inverse problem to reconstruct the inclusion by observing the boundary data at \( x = 0, X \) with controlling the input heat source at the boundary \( x = 0, X \). This case, it is impossible to reconstruct some information about the point \( x = b \) from the observation at \( x = 0 \) and vice versa, that is, it is also impossible to reconstruct some information about the point \( x = a \) from the boundary value at \( x = X \). Therefore, we posed the Problem 3.

In order to solve Problems 2 and 3, we apply the idea of hyperfunctions to treat the Delta functions on the boundary \( \partial \Omega \). It is a pity that we have to control the heat source \( f(t) \) for our solutions to Problems 2 and 3 in such high temperature that test object would be melt down, which is exactly the same problem as the solution to Problem 1 by Ikehata-Kawashita. It is the fatal problem left to be solved for the study of this problem to be applied for practice.

4 Open problems for further development

As was mentioned at the end of the last section, even if the reconstruction formulas for the inclusions are obtained, the known results on Problems 1, 2 and 3 are far from being applied for practice. For the conclusion, we mention open problems left to be solved for this problems to be applied for practice.

Problem 4.

(i) Let us first remark the most important open problem. In the solutions to Problems 1, 2 and 3, the heat source \( f \) to be controlled on the boundary is required to be very high. It must tend to infinity to obtain the information of the inclusion, which is impossible in practice for the following two reasons. One is very simple: we cannot give infinitely high heat sources or heat flux. The other is that the object will be melt down at high temperature. Therefore, we have to develop a method which enables us to detect
the inclusion without tending the temperature to infinity. For this purpose, we propose two ways to
generalize our main theorems.

- One approach to this problem is to develop another method to extract some information of the
  inclusion in the low temperature state.

- The other way is to give suitable error estimates for the limiting processes in the solution to Problems
  1, 2 and 3, which may be helpful to establish a theory of the approximation of the solution at the
  reasonable temperature state.

We claim that both approaches are interesting and are to be studied.

(ii) Even if the above problem is solved, there still are a number of problems left to be solved for prac-
tice; the treatment of the errors, the construction of an approximation algorithm for the inclusions, its
implementation by computers and so on.

(iii) There are many generalization of our approach, the study of which is interesting and important.

Acknowledgements

The author is supported in part by JSPS Grant-in-Aid for Scientific Research (C) 26400184.

References

[2] Huang L and Takiguchi T 1998 The reconstruction of uniquely determined plane sets from two projections J. Inv. Ill-Posed
  Problems, 6 pp 217–242
[3] Ikehata M 2007 Extracting discontinuity in a heat conductive body. One-space dimensional case Applicable Analysis 86 pp
  963–1005
[5] Ikehata M and Kawashita M 2010 On the reconstruction of inclusions in a heat conductive body from dynamical boundary
  data over a finite time interval Inverse Problems 26 095004
[6] Kuba A. and Volčič A. 1988 Characterization of measurable plane sets which are reconstructable from their two projections,
  Inverse Problems 4 pp 513–527
[8] Takiguchi T 2004 Reconstruction of measurable plane sets from their orthogonal projections Contemporary Mathematics 348
  pp 199–208
[9] Takiguchi T 2007 Reconstruction of the measurable sets in the two dimensional plane by two projections Journal of Physics,
  Conference Series 73 012022
[10] Takiguchi T 2012 Non-uniqueness of the reconstruction for connected and simply connected sets in the plane by their fixed
    finite projections Acta Mathematica Scientia 32B pp 1637–1646
Minisymposium M4
Applied Inverse Problems in Industry
Large scale adaptive optics

A Binder\textsuperscript{1} and M Yudytskiy\textsuperscript{2}

\textsuperscript{1}MathConsult GmbH, Linz
\textsuperscript{2}Johann Radon Institute for Computational and Applied Mathematics (RICAM), Linz

E-mail: andreas.binder@mathconsult.co.at, mykhaylo.yudytskiy@ricam.oeaw.ac.at

Abstract: In ground-based telescopes, the light coming from the astronomical objects is being distorted by the turbulent nature of the atmosphere caused by, e.g., changing temperature and wind conditions. Adaptive Optics (AO) is a technology which allows to compensate for these distortions and to enhance the resolution of the images. Equipped with the AO correction, large ground-based telescopes are able to compete with space telescopes with respect to the quality of scientific observations. The AO mechanism is based on the measurements of the wavefront aberration and produces - via solving the inverse problem - commands which are applied to deformable mirrors that correct the aberration. In online control systems, this inverse problem has to be solved up to 3000 times per second. We present results that were obtained in the Austrian Adaptive Optics project carried out for the European Southern Observatory between 2009 and 2013.

Keywords: adaptive optics, wavefront reconstruction, atmospheric tomography, fast iterative methods

1 Introduction

Images from ground-based telescopes suffer from turbulence in the atmosphere, which leads to serious image degradation. Powerful adaptive optics (AO) techniques have been developed in the last few decades to remedy this problem. AO is a hardware-based technique for the correction of the phase of the incoming light which aims to compensate for rapidly changing optical distortions in the atmosphere by deforming a mirror in real time.

The correction is based on the reconstruction of the turbulence in the atmosphere from measurements in the direction of one or several guide stars. A guide star is a bright star in the sky that can be used as a light source (natural guide star, NGS) or an artificially generated light source using a laser (laser guide star, LGS). The detector that measures the phase aberrations is called a wavefront sensor (WFS). Due to the rapid changes in the atmosphere wavefront compensation must be performed in a millisecond time frame.

In the classical AO, also known as single conjugate adaptive optics (SCAO), a single guide star, i.e., a single light source is observed. The AO system corrects the cumulative effect of the turbulence towards a direction close to the guide star using a deformable mirror (DM).

In more complex AO modalities, such as the multi object adaptive optics (MOAO) or the multi conjugate adaptive optics (MCAO), this idea is extended to make it possible for several celestial objects to be observed simultaneously. These AO capabilities are enabled by utilizing several DMs and several WFSs. The key ingredient here is the solution of an atmospheric tomography problem, in which the refractive index of the atmosphere is reconstructed over the region observed by the telescope, i.e., the field of view.

In an atmospheric tomography problem, wavefront sensor measurements are obtained in several guide star directions, see Figure 1. From these measurements, turbulence layers of the atmosphere are reconstructed. This problem resembles a limited angle tomography problem. Hence, due to the presence of measurement noise, regularization methods are vital to obtain a stable solution. The classical approach is to formulate the problem in the Bayesian setting and to postulate the solution as the maximum a-posteriori (MAP) estimate. The approach is motivated from the availability of turbulence layer models, such as Kolmogorov or von Karman,
and the statistical information on the distribution of the measurement noise at the sensor. These and other mathematical challenges in adaptive optics have been comprehensively reviewed by Ellerbroek and Vogel in [1].

Current numerical methods for implementing the MAP estimate are based on direct matrix inversion. These algorithms, such as the matrix-vector multiplication (MVM) method, scale as $O(n^2)$, where $n$ is the dimension of the data. Due to the steady growth in telescope size and the amount of processed data, there is a strong increase in the computational load on the system that performs the computation. In combination with the millisecond time-constraint for the reconstruction, this poses a significant hurdle for AO systems using such algorithms.

Within the last decade, the research tended more towards iterative methods. In particular, the introduction of matrix-free approaches helped reduce the asymptotic computational cost to as low as linear complexity, i.e., $O(n)$. 

In this paper we give a brief overview of several recently developed methods for various AO systems by the Austrian Adaptive Optics team: CuReD, P-CuReD, Kaczmarz reconstructor and the FEWHA. These algorithms are very fast, of excellent quality and can provide enormous savings in the computing power required to reconstruct the wavefront on future AO systems for the European Extremely Large Telescope (E-ELT) of the European Southern Observatory (ESO).

The paper is organized as follows. In Sections 2 and Section 3, we discuss the methods for wavefront reconstruction and for atmospheric tomography, respectively. In Section 4, we state a short conclusion.

2 Wavefront reconstruction

In SCAO, a single WFS, which indirectly measures the wavefront aberrations, is used to derive the DM shape for correction. The process of inferring the wavefront from the WFS measurements is called wavefront reconstruction.

Various sensors are used in AO, each with their own functionality, advantages and a mathematical model that relates the wavefront to the measurements. Two commonly used sensors in AO are the Shack-Hartmann wavefront sensor (SH-WFS) and the pyramid WFS (P-WFS). Below we present two wavefront reconstruction algorithms for these sensors.
Cumulative reconstructor for Shack-Hartmann WFS

The Shack-Hartmann wavefront sensor relates the wavefront to the measurements via the average gradients over the sub-domains, called subapertures,

\[ s_{ij} \propto \int_{\Omega_{ij}} \nabla \varphi(x,y) d(x,y), \]  

Here, \( s_{ij} \in \mathbb{R}^2 \) are measurements at subaperture \( \Omega_{ij} \subset \mathbb{R}^2 \) at indices \( (i,j) \) and \( \varphi(x,y) \) is the wavefront aberration at the telescope pupil.

A method, called the cumulative reconstructor (CuRe) has been developed by Zhariy et al. in [2] to recover \( \varphi \) from a vector of measurements \( s = (s_{ij}) \) over all subapertures directly. The method is based on the fundamental relation between a differentiable function and its average gradient.

Later, the method has been successfully adapted to fit more realistic telescope-related problems by Rosensteiner in [3] and CuRe with domain decomposition (CuReD) has been introduced in [4]. The CuReD method is especially effective for large apertures, where noise propagation plays a crucial role.

The method has been demonstrated to produce similar quality results as the MVM in simulated environments and optical bench tests and has been successfully tested on a telescope. From the computational standpoint, the CuReD algorithm scales as \( O(n) \), is highly parallelizable and is roughly 140 times faster than the MVM on a SCAO system in the E-ELT configuration.

Preprocessed cumulative reconstructor for pyramid WFS

A variation of the SCAO system called extreme adaptive optics (XAO) is planned to deliver planet-imaging performance. Pyramid wavefront sensors allow for a higher contrast of the corrected images as well as a higher resolution of the wavefronts and are therefore used for XAO systems.

Although the P-WFS model is significantly more complex than that of the SH-WFS, a simple relation between the two types of measurements has been established, which has led to the development of a preprocessed CuReD (P-CuReD) method by Shatokhina et al. in [5].

The method is implemented in two steps, where in the first step, the P-WFS measurements are transformed into SH-WFS-like measurements by means of a convolution with a sparse kernel. In the second step of the wavefront reconstruction, the CuReD algorithm is applied.

In combination, the two step method delivers the required quality, but what is more important is the computational advantage of the method over the MVM. As the method scales linearly with the data, the speed-up factor of P-CuReD over the MVM is 1100 for the XAO system in E-ELT configuration. With this results, the P-CuReD method brings a complex XAO system into the domain of feasible implementations with the hardware available today.

3 Atmospheric tomography

The aim of the atmospheric tomography problem is the reconstruction of \( L \) turbulence layers \( \phi_1, \ldots, \phi_L \) located at altitudes \( 0 \leq h_1 < \ldots < h_L \), see Figure 1, which we write in a vector form \( \phi = (\phi_1, \ldots, \phi_L) \). The input data for the tomographic problem are the WFS measurements from \( G \) guide star directions, which we denote by \( s_1, \ldots, s_G \). Thus, the tomographic problem can be formulated as the reconstruction of \( \phi \) in

\[ s_g = \Gamma_g P_g \phi \]  

for \( g = 1, \ldots, G \), where \( \Gamma_g \) is the WFS operator, e.g., given by (1), and \( P_g \) is a geometric propagation operator towards a guide star \( g \), which maps layers to wavefronts via

\[ (P_g \phi)(r) = \sum_{\ell=1}^{L} \phi_{\ell} (\gamma_{g\ell} r + \theta_{g\ell} h_{\ell}). \]  

Here, \( r = (x,y) \) is the position at the aperture, \( \theta_g = (\theta_x, \theta_y) \) is a direction vector towards the guide star \( g \) and \( \gamma_{g\ell} \) is a positive scaling constant related to the guide star type (NGS or LGS) and the altitude of the layer.
After the layers have been obtained, the DMs are fit onto the reconstructed layers, by solving a minimization equation,

$$\argmin_{\phi_{DM}} \left( \int_F \| \hat{P}_\theta \phi_{DM} - P_\theta \phi \|^2 d\theta \right),$$

where the unknown $\phi_{DM}$ is a vector of DM shapes at various altitudes and $F$ is set of directions of interest or the field of view. Here, $\hat{P}_\theta$ is a propagation operator similar to (3) with respect to the mirror shapes. This step is referred to as the fitting step.

Kaczmarz reconstructor

The Kaczmarz approach to the atmospheric tomography problem has been introduced by Ramlau and Rosensteiner in [6, 7]. The problem (2) is solved in two subsequent steps.

First, the wavefronts are reconstructed in all guide star directions $g = 1, ..., G$ from the measurements. Depending on the WFS type, this can be done by, e.g., the CuReD or the P-CuReD algorithms for Shack-Hartmann or pyramid WFS, respectively.

In the second step, the tomographic problem is solved with respect to the reconstructed wavefronts. The authors use the Kaczmarz algorithm for this step, which is an efficient method in the context of tomographic problems, and shows a very fast convergence in this particular setting.

The advantage of the sequential splitting of the problem is that the computational effort of each of the sub-steps is very low. The algorithm can be formulated in a matrix-free way, scales linearly with the data and is, to a certain degree, parallelizable. Due to the fast convergence of the Kaczmarz algorithm, the reconstructor meets the computational requirements of several AO systems in the E-ELT configuration and shows a speed-up over the MVM of up to a factor of 200.

The quality of the algorithm also meets the requirements, however only if the AO systems uses a constellation of natural guide stars. Complexities brought out by noise associated with laser guide stars diminish the qualitative performance of this method.

Finite element-wavelet hybrid algorithm

An alternative method is the finite element-wavelet hybrid algorithm (FEWHA), which has been introduced in [8, 9]. The FEWHA is a conjugate gradient based approach in which the Bayesian MAP estimate of the turbulence layers of the atmosphere is discretized using a finite element and a wavelet basis simultaneously. This dual-domain strategy induces a very efficient matrix-free representation of the underlying operators. The method utilizes the locality properties of compactly supported orthonormal wavelets, both in spatial and frequency domains.

The convergence of this iterative scheme is accelerated by designing an efficient preconditioner and by utilizing multi-scale techniques. Altogether, the computational complexity of the FEWHA scales linearly with the dimension of the problem; the method converges in a few iterations, is highly parallelizable and has a small memory footprint.

While the tomographic problem in the Kaczmarz method is split into sub-problems, the FEWHA approaches (2) as a single problem. One of the key advantages of the coupled tomographic approach of the FEWHA is that it allows to incorporate the statistical information on the measurement noise and the turbulence directly. This makes the algorithm versatile for different AO systems as it can be applied to configurations utilizing NGS as well as LGS. The drawback of the coupled approach of the FEWHA is that it is somewhat slower than the Kaczmarz method. Nevertheless, the method is up to a factor of 50 times faster than the MVM on a conventional computing system.

4 Conclusion

In this paper we gave a brief overview of four methods developed by the Austrian Adaptive Optics team for the European Extremely Large Telescope of the ESO. The methods are in line with the quality of the benchmark methods, and perform significantly faster than the MVM in terms of reconstruction speed. A table summarizing the computational performance of the methods is given in Table 1.
Table 1. Algorithms, AO systems, and speed-up compared to the reference method, the MVM, in the E-ELT configuration.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Applicable system</th>
<th>Speed-up</th>
</tr>
</thead>
<tbody>
<tr>
<td>CuReD</td>
<td>SCAO</td>
<td>140</td>
</tr>
<tr>
<td>P-CuReD</td>
<td>XAO</td>
<td>1100</td>
</tr>
<tr>
<td>Kaczmarz</td>
<td>MCAO, MOAO</td>
<td>200</td>
</tr>
<tr>
<td>FEWHA</td>
<td>MCAO, MOAO</td>
<td>50</td>
</tr>
</tbody>
</table>

Acknowledgments

The presented work was done in the framework of the project “Mathematical Algorithms and Software for E-ELT Adaptive Optics” in cooperation with the ESO. The project was funded by the Austrian Federal Ministry of Science and Research and the work was partially funded by the Austrian Hochschulraum-Strukturfonds. The authors express their gratitude to the members of the Austrian Adaptive Optics team.

References

Minisymposium M5

Tomography
Compensating local affine deformations in 2D computerized tomography

B N Hahn

Department of Applied Mathematics, Saarland University, D-66041 Saarbrücken, Germany

E-mail: hahn@num.uni-sb.de

Abstract: We consider an object which changes locally during the scanning in computerized tomography. Deformations of the specimen lead to inconsistent data. To reduce motion artefacts in the images, the dynamic behaviour has to be included in the reconstruction method. In 2D, the inversion formula is global, i.e. the reconstruction at a specific point calls for all data. In the case of local deformations, this especially means that data from stationary parts of the object are used for the reconstruction of dynamic areas, and vice versa. This mixture can be avoided by using local algorithms. Thus, within this article, we propose a local analytic reconstruction method, which is illustrated with numerical examples.

Keywords: Dynamic tomography, local tomography, motion compensation, moving objects, local deformations.

1 Introduction

The data acquisition in computerized tomography is time consuming since the x-ray source has to rotate around the specimen. Within this article, we consider specimens that change locally during the data acquisition, i.e. the support $\Omega \subset \mathbb{R}^2$ of the object can be decomposed in a stationary part $\Omega^{\text{stat}}$ and a dynamic area $\Omega^{\text{dyn}}$ with $\Omega^{\text{stat}} \cap \Omega^{\text{dyn}} = \emptyset$.

The mathematical model for stationary 2D computerized tomography is given by the Radon transform $R$. To determine the attenuation coefficient $f$ of the object from measured data $g$, the equation $Rf = g$ has to be solved. The inversion formula for the Radon transform

$$f(x) = \frac{1}{4\pi} \int_{S^1} (\mathcal{H}g')((x^T \theta) \theta, \theta) \, d\theta,$$

is global due to the presence of the Hilbert transform $\mathcal{H}$ [1]. Thus, to reconstruct $f$ at a specific point, all data are required. For a locally changing object, this means that the reconstruction of the dynamic area $\Omega^{\text{dyn}}$ calls even for data of the stationary part $\Omega^{\text{stat}}$, and vice versa.

In order to get a local inversion formula in $\mathbb{R}^2$, one reconstructs $\Lambda f$ instead of $f$, leading to the so-called Lambda-tomography [1]. Formally, the operator $\Lambda$ is defined as the square root of the Laplacian. The two functions $\Lambda f$ and $f$ share the same singular support, i.e. they have the same singularities. Thus, $\Lambda f$ yields the edges of the investigated object.

Within the following sections, we derive an analytic reconstruction algorithm to compensate for local deformations. Therefore, we use the framework of the approximate inverse [2], which has already been successfully applied in the case of globally deforming objects [3, 4].
2 The model of dynamic CT

The Radon transform maps a function \( f \in L^2(\Omega) \) with \( \Omega \) being the unit circle to its line integrals

\[
Rf(\theta,s) = \int_{\mathbb{R}^2} f(x) \delta(s-x^T \theta) \, dx, \quad \theta \in S^1, \ s \in [-1,1],
\]

where \( \delta \) denotes the delta distribution. Let \( g_\theta \) denote the measured data \( g \) for fixed \( \theta \in S^1 \). The time-dependent step of the data acquisition in computerized tomography is the rotation of the x-ray source around the specimen. Thus, the unit vector \( \theta \) acts as temporal component, \([4, 5]\). Within this article, we want to consider objects which change during this time period.

According to \([3]\), additional information about the object’s deformation are required. Therefore, we consider a dynamic behaviour where the state of the object at the moment of measuring \( g_\theta \) is given by \( f \circ \Gamma_\theta \) with diffeomorphic motion functions \( \Gamma_\theta \), see also \([4]\).

Then, the mathematical model of dynamic computerized tomography is given by the integral operator

\[
R_\Gamma f(\theta,s) = \int f(\Gamma_\theta x) \delta(s-x^T \theta) \, dx,
\]
i.e. we have to solve the inverse problem

\[
R_\Gamma f = g
\]
instead of \( Rf = g \).

3 The method of the approximate inverse

The method of the approximate inverse is a regularization scheme to determine a feature function \( Lf \) with linear operator \( L \) from an ill-posed inverse problem

\[
Af = g, \quad (1)
\]
where \( A : L^2(\Omega) \to L^2(\Omega) \) is a linear integral operator \([6]\). If \( L \) equals the identity, we obtain an algorithm for the classical density reconstruction. Due to the ill-posedness of \((1)\), a smoothed version

\[
(Lf)^\gamma(x) := \langle Lf, \delta_x^\gamma \rangle = \langle f, L^* \delta_x^\gamma \rangle
\]
with a prescribed mollifier \( \delta_x^\gamma \) is calculated instead of \( Lf \). Using the precomputed solution \( \psi_x^\gamma \) of the auxiliary problem

\[
A^* \psi_x^\gamma = L^* \delta_x^\gamma,
\]
we obtain

\[
(Lf)^\gamma(x) = \langle g, \psi_x^\gamma \rangle.
\]
The approximate inverse is now defined as follows, \([6]\).

**Definition 3.1.** The operator \( S^\gamma : L^2(\Omega) \to L^2(\Omega), \)

\[
S^\gamma g(x) = \langle g, \psi_x^\gamma \rangle
\]
is called the approximate inverse of \( A \) to compute an approximation of \( Lf \), and \( \psi_x^\gamma \) is called the reconstruction kernel.

If \( \psi_x^\gamma \) is compactly supported, only data in a small neighbourhood of \( x \) are required for the reconstruction at point \( x \). Especially, different reconstruction kernels \( \psi_x^\gamma \) could be applied, depending on \( x \in \Omega^{\text{dyn}} \) or \( x \in \Omega^{\text{stat}} \). Thus, we can use the following reconstruction scheme for local deformations,

\[
S^\gamma g(x) := \begin{cases} 
(g, \psi_x^{\gamma,\text{dyn}}) & x \in \Omega^{\text{dyn}} \\
(g, \psi_x^{\gamma,\text{stat}}) & x \in \Omega^{\text{stat}}
\end{cases}
\]
with
\[ R^*_1 \psi^\gamma_{\text{dyn}} = \mathcal{L}^* \delta^\gamma_x \] (2)
and
\[ R \psi^\gamma_{\text{stat}} = \mathcal{L}^* \delta^\gamma_x. \] (3)

In [3, 4], we derived two methods to solve the auxiliary problem (2) in the dynamic case depending on the type of deformation. If the motion functions \( \Gamma_\theta, \theta \in S^1 \) correspond to affine deformations, the solution of (2) can be computed analytically based on an exact inversion formula for \( R \) [3]. In the case of non-affine deformations, \( R \) integrates a function along arbitrary curves instead of straight lines, and therefore, no inversion formula is known so far. Thus, we derived a reconstruction procedure which uses the kernel
\[ \psi^\gamma_{\text{dyn}}(\theta, s) = (2 \sqrt{1 - s^2} + \alpha)^{-1} \left( R_1 \delta^\gamma_x(\theta, s) + \alpha \psi^\gamma_{\text{stat}}(\theta, s) \right) \]
with parameter \( \alpha > 0 \), \( \psi^\gamma_{\text{stat}} \) solving (3) and
\[ \delta^\gamma_x(z) := \left( \int_{S^1} | \det \Gamma^{-1}_\theta(z)| d\theta \right)^{-1} \delta^\gamma_x(z). \]
The resulting method then compensates even for non-affine deformations [4].

However, the derived kernels \( \psi^\gamma_{\text{dyn}} \) are in general not compactly supported. Thus, we compute in the following a reconstruction kernel which leads to a local algorithm.

### 4 Computing suitable reconstruction kernels

The aim of this section is to determine a compactly supported reconstruction kernel for the dynamic problem with affine deformations. The following lemma states a link between the operators \( R \) and \( R_1 \), see also [7, 8] for a proof.

**Lemma 4.1.** Let the object’s deformation be described by affine motion functions \( \Gamma_\theta x := A_\theta x + b_\theta \) with \( A_\theta \in \mathbb{R}^{2 \times 2} \) and \( b_\theta \in \mathbb{R}^2 \) for all \( \theta \in S^1 \).

Then, the dynamic operator \( R_1 : L_2(\Omega) \to L_2(S^1 \times \mathbb{R}) \) is related to the static Radon transform \( R \) via
\[ R_1 = \mathcal{V} R \]
with
\[ \mathcal{V} g(\theta, s) = | \det A_\theta |^{-1} ||A_\theta^{-1} \theta||^{-1} g \left( \frac{A_\theta^{-1} \theta}{||A_\theta^{-1} \theta||} s + b_\theta^T A_\theta^{-1} \theta \right). \]

This link can be used to solve the auxiliary problem (2).

**Theorem 4.1.** Let \( \delta^\gamma_x \) be a mollifier, and \( \psi^\gamma_{\text{stat}} \) the corresponding reconstruction kernel for the static problem, i.e.
\[ R^* \psi^\gamma_{\text{stat}} = \mathcal{L}^* \delta^\gamma_x. \]

Then, the reconstruction kernel for the dynamic setting is given by
\[ R^* \psi^\gamma_{\text{dyn}}(\theta, s) = \mathcal{V}^* \psi^\gamma_{\text{stat}}(\theta, s) \]
\[ = | \det A_\theta | | h(\theta) ||A_\theta^{-1} \theta||^{-2} \psi^\gamma_{\text{stat}} \left( \frac{A_\theta^{-1} \theta}{||A_\theta^{-1} \theta||} s + b_\theta^T A_\theta^{-1} \theta - x T A_\theta^{-1} \theta \right) \]
\[ \text{with } h(\theta) := (A_\theta^{-1} \theta)_1 \cdot \frac{\partial}{\partial \varphi} (A_\theta^{-1} \theta)_2 - (A_\theta^{-1} \theta)_2 \cdot \frac{\partial}{\partial \varphi} (A_\theta^{-1} \theta)_1 \text{ and } \varphi \text{ being the phase angle of the unit vector } \theta. \]

**Proof** For \( \psi^\gamma_{\text{dyn}} := \mathcal{V}^* \psi^\gamma_{\text{stat}}, \) it holds with lemma 4.1
\[ R^*_1 \psi^\gamma_{\text{dyn}} = R^*_1 \mathcal{V}^* \psi^\gamma_{\text{stat}}(\theta, s) = (\mathcal{V}^{-1} R)_1^* \psi^\gamma_{\text{stat}}(\theta, s) = R^* \psi^\gamma_{\text{stat}} = \mathcal{L}^* \delta^\gamma_x. \]
The representation (4) of the operator $V^{-s}$ can be simply verified by computing the adjoint of $V^s$. □

According to theorem 4.1, a suitable reconstruction kernel for the dynamic problem can be computed based on a known kernel for the stationary setting. In [9], the authors derived a reconstruction kernel $\psi_{x,\text{stat}}^\gamma$ for stationary Lambda-tomography

$$\psi_{\Lambda,x,\text{stat}}^\gamma(\theta, s) = v\left(\frac{s - x^T \theta}{\gamma}\right), \quad v(s) = -\frac{1166167275}{1048576} (s^2 - 1/23)(1 - s^2)^{10}.$$  (5)

Altogether, using the kernel $\psi_{x,\text{dyn}} = V^{-s}\psi_{\Lambda,x,\text{stat}}^\gamma$ leads to a local reconstruction method which compensates for affine deformations.

5 Numerical results

The method is tested for a numerical phantom whose objects within the circle $V_{\Lambda,0}(0)$ of radius 0.4 around the origin are rotating, as illustrated in Figure 1. Precisely, the motion model is given by $\Gamma_{\delta}^x := A_{\delta}x$ with

$$A_{\theta(\varphi)} := \begin{pmatrix} \cos \frac{\varphi}{3} & -\sin \frac{\varphi}{3} \\ \sin \frac{\varphi}{3} & \cos \frac{\varphi}{3} \end{pmatrix} \quad \text{for } \varphi \in [0, \frac{\pi}{2}]$$

and

$$A_{\theta(\varphi)} := A_{\theta(\pi - \varphi)} \quad \text{for } \varphi \in \left[\frac{\pi}{2}, \pi\right],$$

where $\varphi$ is the phase angle of the unit vector $\theta(\varphi)$. The data are computed analytically for 300 different source positions and 451 detector points. Then, a uniformly distributed noise on $[-0.007, 0.007]$ has been added to these data. The result of the algorithm using the kernel (5) for stationary Lambda tomography is shown in Figure 2. The edges of the moving objects are seriously distorted and their true shapes cannot be recognized from this reconstruction. Especially, this illustrates that the dynamic behaviour has to be taken into account by the reconstruction method.

Next, we apply the proposed local algorithm to compute an image of the phantom at the initial time. The result is displayed in Figure 3. Comparing it with the original phantom in Figure 1 shows that the edges of all objects are reconstructed very well, no matter whether they belong to the stationary or the dynamic area. Thus, the example proves the good local motion compensation properties of our algorithm.

Acknowledgements

This work was supported by grant LO 310/11-2 of the Deutsche Forschungsgemeinschaft.
Figure 2. Static reconstruction of the phantom from noisy data

Figure 3. Dynamic reconstruction at the initial time from noisy data

References

X-rays fluorescence computed tomography: analytic inversion and iterative methods

A R De Pierro\textsuperscript{1} and E X Miqueles\textsuperscript{2}

\textsuperscript{1}University of S\~{a}o Paulo, S\~{a}o Carlos, Brazil
\textsuperscript{2}Brazilian Synchrotron Light Source, CNPEM, Campinas, Brazil

E-mail: depierro.alvaro@gmail.com

Abstract: X-Rays Fluorescence Computed Tomography (XFCT) is a relatively new technique aiming at reconstructing the distribution of an element, usually a metal, inside a body. This is achieved by irradiating the object with high intensity X-rays at a given energy level. This induces emission radiation by the element, that is counted by an outside detector. The model for the problem is the Generalized Attenuated Radon Transform (GART), the operator to be inverted. We present a brief overview on recent developments for the inversion of GART: an analytic formula [11], iterative methods based on Radon inversion [12] as well as others that consider the Poisson nature of the noise [16].

Keywords: computed tomography, fluorescence, attenuated Radon transform

1 Introduction

This paper presents a brief overview of recent results in X-Rays Fluorescence Computed Tomography (XFCT). This is a relatively new tomographic technique aiming at determining the concentration distribution of an element (Copper, Iron, Zinc, Iodine and others) inside a body. This concentration distribution could be used to detect malignancy in a tissue, determine 3D rock structure in mineralogy as well as many other applications [19]. When the object is irradiated by high intensity monochromatic synchrotron X rays at a specific energy of a given element, this stimulates fluorescence emission, that is detected by an external detector, as shown in Figure 1. The X-rays flux goes through the object and non attenuated photons are detected as in X-rays computed tomography (CT). These X-rays excitate the element that emits radiation with an energy, in principle, different from the one of the incoming X-rays. Therefore, the mathematical problem is the inversion of the Generalized Attenuated Radon Transform (GART) given by

\[ \mathcal{R}_W: f = f(x) \in U \mapsto d = d(t, \theta) \in V, \quad d(t, \theta) = \mathcal{R}_W f(t, \theta) = \int_{x \xi = t} f(x)W(x, \theta)dx, \quad (1) \]

where \( f(x) \) is the (fluorescence) emission density at \( x \), and \( W(x, \theta) \) is given by,

\[ W(x, \theta) = \omega_\lambda(x, \theta)\omega_\mu(x, \theta), \quad (2) \]

where \( \mu \) is the fluorescence attenuation and \( \lambda \) is the attenuation of the X-rays. \( \omega_\lambda \) represents the survival probability of reaching \( x \), \( \omega_\mu \) the probability of the radiation from \( x \) to reach the detector, and they are defined respectively by

\[ \omega_\lambda(x, \theta) = e^{-\mathcal{D}_\lambda(x, \theta+\pi)}, \quad \omega_\mu(x, \theta) = \int_\Gamma e^{-\mathcal{D}_\mu(x, \theta+\gamma)}d\gamma \quad \text{and} \quad \mathcal{D}_h(x, \theta) = h(x + q\xi^+)dq \quad (3) \]
It is worth noting that $\lambda$ is obtained from the X-rays data by a simple Radon inversion, but $\mu$ is unknown, although, for low energy values, it can be approximated by $\lambda$. $U$ and $V$ defined in (1) can be regarded as $L^2$ normed spaces, of compactly supported functions. We use the notation $R_{\text{xfct}}$ or $R_{\text{spect}}$ to distinguish between the scanning modalities.

Until very recently, most of the results had been obtained just by Radon inversion, without considering any kind of attenuation, or by weighted Radon inversion \cite{3}. What follows describes two better alternatives: Radon iteration and analytic inversion through a new formula that generalizes work by A. Fokas \cite{4} and Novikov \cite{5}.

2 Iterative Inversion

As noted in the Introduction, the emission attenuation $\mu$ is unknown, but for low energies, it is approximately equal to $\lambda$. The new problem to be solved is: Given the data $d$, find $\{f, \mu\}$ such that $R_W(\mu)f - d = 0$. The following alternate iteration procedure was proposed in \cite{12}:

$$f^{(k+1)} = L\left(d, f^{(k)}, \mu^{(k)}\right), \quad \mu^{(k+1)} = N\left(d, f^{(k+1)}, \mu^{(k)}\right),$$

where $L$ stands for an approximate inversion of $R_W$ given $\mu^{(k)}$ - i.e. the linear part - and $N$ stands for the non-linear part, e.g., application of (say) Newton’s method to (1) for $f^{(k+1)}$ given. Retrieving $\mu$ through the algorithm $N$ is the most difficult part of the inverse problem. For the algorithm $L$, we use two iterative strategies, described below.

Using Radon inverse

An alternative for the inversion is to use the inverse of the Radon transform and iterate using the residual of the attenuated one, $R_W$, a common technique in numerical analysis, first suggested for SPECT by Kunyansky \cite{9}. In our case this gives for $k \in \mathbb{N}$ (starting with the image obtained by Radon inverse applied to the data)

$$f^{(k+1)} = f^{(k)} + e^{(k)}, \quad e^{(k)} = \frac{R_W^{-1}(d - R_W f^{(k)})}{a}, \quad a(x) = \frac{1}{2\pi} \int_0^{2\pi} W(x, \theta) d\theta. \quad (5)$$

where $a$ is a weight that can be defined as in \cite{3} for the case of the attenuated transform for SPECT. In \cite{14} it was proven that the contraction constant for the iteration operator is:

$$K = \left(\mathcal{I} - \frac{1}{a} R_W^{-1} R_W\right) f \Rightarrow \|K\| \leq c = \sup_{u \in \mathbb{R}^2} \sup_{\theta \in [0,2\pi]} \left|1 - \frac{1}{2\|a\|_{\infty}} [W(u, \theta) + W(u, \theta + \pi)]\right| \quad (6)$$

For an exponential weight, such as SPECT or XFCT, the contraction bound satisfies $c < 1$, which guarantees convergence of (5).
Continuous Expectation Maximization (EM)

For the sake of comparison we have tried the EM [16] algorithm in its continuous version, given by the following iteration:

$$f^{(k+1)}(x) = f^{(k)}(x) \frac{B_W d^{(k)}(x)}{B_W e(x)}, \quad d^{(k)}(t, \theta) = \frac{d(t, \theta)}{R_W f^{(k)}(t, \theta)}$$

where $B_W$ is the attenuated backprojection (adjoint operator of $R_W$), and $e = 1$.

$$B_W d(x) = \int_0^{2\pi} W(x, \theta) d(x \cdot \xi, \theta) d\theta$$

Figure 2 shows a 32 x 32 representation of functions $\{f, \mu, \lambda\}$ and the simulated attenuated Radon transform with 80 projections views and 60 rays per view. The reconstructions are shown in the same Figure. AKT stands for the iteration using the weighted Radon transform.

3 Analytic Inversion

It follows, a brief description of the rationale leading to an analytic formula for the inversion of $R_W$. Details could be found in [11]. The tomographic modalities that we are considering can be modeled from a photon transport equation [8]

$$\eta \cdot \nabla u(x) + \lambda u(x) = f(x)$$

with a scattering term neglected (because we are assuming that the source is monochromatic), $u$ standing for the energy transport, $\eta$ the transport direction, $\lambda$ the linear attenuation coefficient and $f$ the unknown radioactive source. Except for the case of XFCT (which is another partial differential equation), the theory of Fokas [4] applied to (9) gives the analytic inverse of $R_\omega$. In our notation, $\cdot$ denotes the real dot product.

Starting with the simplest case of CT, where $\lambda = 0$ and therefore $\eta \cdot \nabla u(x) = f(x)$, the claimed theory, shows by a spectral analysis of the above PDE, that the solution $u$ is written as $u(x) = I_\eta R f(x)$. Hence, the unknown source $f$ is determined by $f(x) = [\eta \cdot \nabla I_\eta] R f(x)$. Using a radiation condition, it can be shown that $\eta \cdot \nabla I_\eta \rightarrow I$ as $\|\eta\| \rightarrow \infty$, meaning that $I$ is a left inverse operator for $R$. The details for the theory can be found in [4, 7, 11]. Since many calculations are involved, we show the main steps in the fluxogram of Figure 3, summarized below:
Using this same approach for a non-null attenuation map λ ∈ U - this is the case of SPECT, the partial differential equation (9) also gives us a fluxogram quite similar to the one in Figure 3, although the part (b) gives rise to a different d-bar equation, which is \( \partial_z u + \lambda(x)u/j(\lambda) = f(x)/j(\lambda) \). At this point, an Euler factor can be introduced, decoupling this d-bar equation on two others

\[
\partial_z (ue^{\lambda}) = \frac{fe^{\lambda}}{j(\lambda)}, \quad \partial_z q = \frac{\lambda}{j(\lambda)}
\]

Now, the fluxogram-algorithm can be applied simultaneously at both d-bar equations. Since the computations involved are very cumbersome, we refer to [11, 4]. It appears to be surprising that now, as for the Radon transform with \( \lambda = 0 \), the solution \( u \) is represented in terms of the attenuated Radon transform of \( u \) with a spect weight, that is \( u(x) = I_\eta \rho \text{spect} f(x) \). Here, the operator \( I_\eta \) is not the same as before, although quite similar. Hence, assuming the radiation condition \( u = O(1/\lambda) \) with \( \lambda \to \infty \), we go back to (9) and find that the unknown source \( f \) is given by \( f = I \rho \text{spect} f \), that is \( I = \rho \text{spect}^{-1} \), which is Novikov’s inversion.

In the case of XfCT, the case is very similar, but with a different partial differential equation. Indeed, instead of (9) we consider \( \eta \cdot \nabla u(x) + a(x, \eta)u(x) = f(x) \). We decouple the PDE on two d-bar equations like (12), with \( \lambda \) replaced by \( a \). Using some properties of the vector \( \eta = \eta(\lambda) \) with \( \lambda \) approaching \( S \) from inside and outside, we can solve the Riemann-Hilbert problem (see Figure 3). Finally, requiring \( a \) to satisfy a given property, we obtain the following algorithm to compute the left inverse operator \( \rho \text{fct}^{-1} \).

Figure 3. Fluxogram to construct a left inverse operator for \( \rho_W \) in XfCT.

(a)-(b) A change of variables \( z = x \cdot v \) (and therefore \( x = z \cdot v \)) transforms the PDE \( \eta \cdot \nabla u(x) = f(x) \) to \( (\eta \cdot v)\partial_z u(z, \bar{v}) + (\eta \cdot \bar{v})\partial_{\bar{v}} u(z, \bar{v}) = f(z, \bar{v}) \).

(b)-(c) Assuming the condition \( \eta \cdot v = 0 \) and \( (\eta \cdot \bar{v}) = j(\lambda) \) with \( j \) being the Jacobian of the change of variables \( x \to (z, \bar{v}) \) we arrive at a d-bar equation that measures the departure from analyticity of \( u \). Set \( S = \{ \lambda : j(\lambda) = 0 \} \subseteq \mathbb{C} \) stands for the singularity set. We assume that \( S \) determines a closed curve, dividing the complex plane in two regions \( S^+ \) and \( S^- \).

(c)-(d) Using the generalized Cauchy integral Theorem [1] and a radiation condition \( u = O(1/\lambda) \) it can be shown [4, 7, 11] that the analytic solution for \( \lambda \notin S \) is given by

\[
u(x, \lambda) = \frac{\text{sign}(j(\lambda)/2\pi)}{2\pi i} \int_{S^\pm} \frac{f(y)dy}{v(\lambda) \cdot (y - x)} \]

(10)

Using this same approach for a non-null attenuation map \( \lambda \in U \) - this is the case of SPECT, the partial differential equation (9) also gives us a fluxogram quite similar to the one in Figure 3, although the part (b) gives rise to a different d-bar equation, which is \( \partial_z u + \lambda(x)u/j(\lambda) = f(x)/j(\lambda) \). At this point, an Euler factor can be introduced, decoupling this d-bar equation on two others

\[
\partial_z (ue^{\lambda}) = \frac{fe^{\lambda}}{j(\lambda)}, \quad \partial_z q = \frac{\lambda}{j(\lambda)}
\]

(12)
i. Given $\theta \in [0, 2\pi]$ and $m = \gamma_2 - \gamma_1$, define

$$L_0(x) = -\ln \left[ \frac{\omega_\mu(x, \theta)\omega_\mu(x, \theta + \pi)}{m^2} \right]$$

(13)

ii. Using $w(t) = \sqrt{1 - t^2}$, compute the sinogram

$$\mathcal{S}b(t, \theta) = \left\{ \begin{array}{ll}
\frac{\mathcal{S}L_0(t, \theta)}{\omega_\mu(t, \theta)}, & |t| < 1 \\
0, & |t| = 1
\end{array} \right.$$

(14)

iii. Set $p = \mathcal{S}\lambda + \mathcal{S}b$

iv. Define $\mathcal{H}$ the Hilbert transform and the operator $m: (r, d) \in V \times V \rightarrow m(r, d) \in V$ by:

\[ h_c(r) = \cos\left(\frac{1}{2}Hr\right), \quad h_s(r) = \sin\left(\frac{1}{2}Hr\right) \]

b) $m_c(r, d) = h_c(r)\mathcal{H}\left(h_s(r)e^{\frac{r}{d}}\right)$

c) $m_s(r, d) = h_a(r)\mathcal{H}\left(h_s(r)e^{\frac{r}{d}}\right)$

d) $m(r, d) = e^{-\frac{r}{d}}(m_c(r, d) + m_s(r, d))$

v. Compute the sinogram $n(t, \theta) = m(p, \mathcal{S}\sigma f)$ using step iv.

vi. Reconstruct $f$ using

a) $z(t, \theta) = \mathcal{S}\sigma f(t, \theta)n(t, \theta) \in V$

b) $f(x) = \frac{1}{4\pi} \int_0^{2\pi} \partial z(x \cdot \xi, \theta) d\theta$

It is worth noting that using $\mu = 0$, the above algorithm also includes Novikov’s [17] and Tretiak & Metz inversion [18]. Indeed, since $\omega_\mu = 1$, it follows by (13) that $L = 0$ and $\mathcal{S}b = 0$, giving Novikov’s formula.

4 Final Remarks

We have presented a brief overview of mathematical work for the inversion in fluorescence computed tomography. Further work should be done in order to solve some other important problems in XFCT like, for example: improving the photon flux by dealing with polychromatic data and faster data collection modes by using several detectors. In both cases, the mathematical model becomes more complicated introducing new challenges to obtain useful inversion results. Also, a large set experiments can be found in the references [11, 12, 14, 15, 16]. A public software package for the algorithms is described in [13].

Acknowledgments

We thank the Nuclear Instrumentation Laboratory (Federal University of Rio de Janeiro), Brazil, and the Brazilian Synchrotron Light Source (LNLS), that provided the real data. The first author was supported by CNPq grant No 301064/2009-1.

References

Minisymposium M6

Inverse Spectral Problems
Inverse scattering for energy-dependent Schrödinger equations

R O Hryniv\textsuperscript{1,2} and S S Manko\textsuperscript{1,3}

\textsuperscript{1}Pidstryhach Institute for Applied Problems of Mechanics and Mathematics, Ukraine
\textsuperscript{2}University of Rzeszów, Poland
\textsuperscript{3}Doppler Institute for Mathematical Physics and Applied Mathematics and Czech Technical University, Czech Republic

E-mail: rhryniv@iapmm.lviv.ua, stepan.manko@gmail.com

Abstract: The inverse scattering problem for energy-dependent Schrödinger equations on the half-line is studied. Under suitable conditions, the corresponding scattering function is shown to determine uniquely the potentials of the problem, and the reconstruction algorithm is justified. The approach is based on relations between this scattering problem and the one for non-canonical Dirac and ZS-AKNS systems.

Keywords: Inverse scattering, energy-dependent potential, Schrödinger equation

1 Introduction

The purpose of this paper is to discuss the scattering problem for one-dimensional energy-dependent Schrödinger equations of the form

\[-y'' + qy + 2kpy = k^2y\]  \hspace{1cm} (1)

on the half-line and subject to e.g. the Dirichlet boundary condition \(y(0) = 0\); here \(p\) and \(q\) are real-valued potentials, with \(p\) integrable and \(q\) highly singular. When \(p \equiv 0\), this becomes the standard Schrödinger equation with well understood scattering theory both for regular Faddeev–Marchenko potentials \([1, 2, 3, 4, 5, 6]\) and for a larger class of distributional potentials. Equations (refHrynoveq:EDSE) with nonzero \(p\) arise naturally in various models of quantum and classical mechanics (e.g. in Klein–Gordon model of interaction between spinless particles \([8, 9, 10, 11]\)), and the corresponding scattering problems have been studied since 1970-ies by Jaulent and Jean \([12, 13, 14, 15, 16]\); see also \([17, 18, 19, 20, 21, 22, 23, 24, 25, 26]\) for later progress. In most of these papers the potentials were assumed regular and to decay fast enough at infinity, \(q\) was real-valued but \(p\) took purely imaginary values. An interesting approach for the case of real-valued \(p\) and \(q\) was recently suggested by Kamimura \([27, 28, 29]\).

As in the standard case with regular \(q\) and \(p \equiv 0\), for real \(k\) equation (1) has the so-called Jost solution \(f(.,k)\) uniquely determined by the asymptotics \(f(x,k) = e^{ikx}(1 + o(1))\) as \(x \to \infty\). Moreover, for nonzero real \(k\) the solutions \(f(.,k)\) and \(f(.,k)\) are linearly independent, and there is a number \(S(k) \in \mathbb{C}\) such that the solution \(f(.,k) - S(k)f(.,k)\) satisfies the Dirichlet boundary condition \(y(0) = 0\). The function \(S\) is called the scattering function of the problem (1) and is a direct analogue of the scattering function for the classical Schrödinger scattering problem.

We introduce the Jost function

\[s(k) := f(0,k),\]  \hspace{1cm} (2)
in whose terms the scattering function $S$ can be expressed as

$$S(k) = \frac{s(k)}{s(k)} = f(0, k), \quad k \in \mathbb{R}. \quad (3)$$

The Jost solutions exist also for $k$ in the open upper half-plane $\mathbb{C}^+$, and the Jost function is analytic in $\mathbb{C}^+$ and continuous in the closure $\overline{\mathbb{C}^+}$. In general, the Jost function $s$ may have zeros in $\mathbb{C}^+$, but they need not belong to the imaginary axis. If $z_0 \in \mathbb{C}^+$ is a zero of $s$, then $f(\cdot, z_0)$ is a solution of (1) belonging to the Hilbert space $L^2(\mathbb{R}^+)$ and satisfying the Dirichlet boundary condition, and in that case $z_0$ is called an eigenvalue, or bound state of the problem (1). The scattering function $S$ and the eigenvalues together with the corresponding norming constants form the scattering data of the problem (1) subject to the boundary condition $y(0) = 0$.

We note that generically the eigenvalues may be non-real and non-simple, which raises the question on the proper definition of the norming constants and makes the scattering problem very involved. To avoid such complications and emphasize the new approach to the scattering problem for equation (1), we impose here the conditions that guarantee absence of the bound states; then the scattering data consists just of the scattering function $S$.

The direct scattering problem consists in finding the scattering data for given potentials $p$ and $q$. The inverse scattering problem is to construct $p$ and $q$ given the scattering data. Our aim here is to suggest a very simple reconstructing algorithm and to prove uniqueness in the inverse scattering problem for energy-dependent Schrödinger equation (1) under the weakest possible regularity assumptions on the potentials $p$ and $q$. In a subsequent paper we shall give complete proofs of all the statements and also characterize the set of scattering functions for a more general class of boundary conditions, thus completing the solution of the scattering problem for (1).

## 2 Preliminaries

### Miura potentials and their Riccati representatives

We shall consider the case where the potential $q$ admits a Riccati representation given by the Miura map [30]. Recall that the Miura map is the nonlinear mapping $B : u \mapsto u' + u^2$ acting from $L^2_{\text{loc}}(\mathbb{R})$ to $H^{-1}_{\text{loc}}(\mathbb{R})$; it relates the solutions for the KdV and mKdV equations. It is known [31] that a real-valued $q \in H^{-1}_{\text{loc}}(\mathbb{R})$ is in the range of the Miura map, i.e., $q = B(u)$ for a function $u \in L^2_{\text{loc}}(\mathbb{R})$, if and only if the quadratic form associated with the Schrödinger operator

$$-\frac{d^2}{dx^2} + q \quad (4)$$

is non-negative. We note that $u$ satisfying $B(u) = q$ need not be unique; any such $u$ is called a Riccati representative for $q$, and every real-valued $q \in H^{-1}_{\text{loc}}(\mathbb{R})$ possessing a Riccati representation $q = B(u)$ is called a Miura potential.

In this paper, we discuss the scattering problem for equations (refHrynoveq;EDSE) in which $q$ is a Miura potential possessing a Riccati representative $u \in X^+_2 := L^1(\mathbb{R}^+) \cap L^2(\mathbb{R}^+)$. Note that such $q$ may contain Coulomb-type singularities, Dirac delta-functions, or even be unbounded and rapidly oscillating and contain all Faddeev–Marchenko potentials without bound states [32]. Despite this fact, the classical scattering theory has recently been extended to the above class of singular Miura potentials, cf. [32], so that the conditions imposed on $u$ seem natural. Moreover, a Riccati representative $u$ from $X^+_2$ is necessarily unique, so we may parameterize the potentials $q$ by their Riccati representatives. The reason for considering $q$ for which (refHrynoveq;Schroed) is non-negative is that otherwise the spectral problem (1) may possess non-real and non-simple eigenvalues, and the scattering problem then becomes much more complicated.

### Transformation to a ZS-AKNS system

Next we establish relation between solutions to the energy-dependent Schrödinger equation (1) with Miura potentials $q = B(u)$ and those for some Dirac-type equation. Namely, take any solution $y(\cdot, k)$ of (1) for a
nonzero $k$ and set $y_2(\cdot, k) := y(\cdot, k)$ and
\begin{equation}
y_1(\cdot, k) = \frac{y_2'(\cdot, k) - uy_2(\cdot, k)}{k}, \tag{5}
\end{equation}
here $u$ is the unique Riccati representative in $X_2^+$ of the potential $q$. Then the column vector $y(\cdot, k) := (y_1(\cdot, k), y_2(\cdot, k))^t$ satisfies the non-canonical Dirac system
\begin{equation}
\sigma_2 y' + P(x)y = ky, \tag{6}
\end{equation}
where
\begin{equation*}
\sigma_2 := \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \quad \text{and} \quad P := \begin{pmatrix} 0 & -u \\ -u & 2p \end{pmatrix}.
\end{equation*}
The boundary condition for $y$ reads $y_2(0, k) = 0$.

Conversely, if $y'(\cdot, k) = (y_1(\cdot, k), y_2(\cdot, k))^t$ solves the above Dirac system, then the first component $y_1$ verifies (5), while the second component $y := y_2$ solves the equation
\begin{equation*}
-\left(\frac{d}{dx} + u\right)\left(\frac{d}{dx} - u\right)y + 2kpy = k^2y,
\end{equation*}
which coincides with the energy-dependent Schrödinger equation (1) with $q = B(u) = u' + u^2$.

Hence in order to construct the Jost solutions for the energy-dependent Schrödinger equation with a Miura potential $q = B(u)$, it suffices to study the Jost solutions for the corresponding Dirac system (6). The simplest way to do this is to transform this system to a canonical ZS-AKNS form as follows. We set
\begin{equation*}
U := \begin{pmatrix} i & -1 \\ 1 & 1 \end{pmatrix}, \quad \sigma_3 := \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}
\end{equation*}
and make a substitution $y = Ue^{i\gamma(x)}\sigma_3 z$ with a vector $z := (z_1, z_2)^t$ and $\gamma(x) := \int_0^x p(t) \, dt$. Then the straightforward calculations show that $z$ satisfies the equation
\begin{equation}
z' + V(x)z = ik\sigma_3 z \tag{7}
\end{equation}
with
\begin{equation}
V(x) = \begin{pmatrix} 0 & v(x) \\ \overline{v(x)} & 0 \end{pmatrix} \tag{8}
\end{equation}
and $v(x) := (-u(x) + ip(x))e^{-2i\gamma(x)}$. We observe that under the assumption made on $p$ and $u$ the function $v$ belongs to the space $X_2^+$. Set $\beta := \gamma(0) = \int_0^\infty p(t) \, dt$; then $z$ should satisfy the boundary condition
\begin{equation}
e^{i\beta}z_1(0, k) + e^{-i\beta}z_2(0, k) = 0. \tag{9}
\end{equation}

**Jost solutions**

Properties of the Jost solutions for the canonical ZS-AKNS system (7) with $v \in X_2^+$ were thoroughly studied in the papers [32, 33]. Namely, it was proved therein that for every real $k \in \mathbb{R}$ the equation
\begin{equation*}
\Psi' + V\Psi = ik\sigma_3 \Psi
\end{equation*}
with $V$ of (8) and a complex-valued $v \in X_2^+$ has a unique matrix $2 \times 2$ solution $\Psi$ obeying the asymptotics $\Psi(x, k) = e^{ikx\sigma_3}(1 + o(1))$. Tracing back the substitutions made in the previous subsection, we find a corresponding matrix-valued solution $Y(x, k)$ of the Dirac equation (6); its first column $y$ obeys the asymptotics
\begin{equation*}
y(x, k) = \begin{pmatrix} 1 \\ i \end{pmatrix} e^{ikx}(1 + o(1))
\end{equation*}
as $x \to \infty$. Recalling that the second component of $y = (y_1, y_2)^t$ solves the energy-dependent Schrödinger equation (1), we arrive at the following result:

**Theorem 2.1.** For every nonzero $k \in \mathbb{R}$, the Jost solution $y(\cdot, k)$ of the energy-dependent Schrödinger equation (1) with a Miura potential $q = B(u)$, $u \in X_2^+$, exists and is unique.
3  Direct scattering problem

By (3), the scattering function $S$ for the problem (1) and the boundary condition $y(0) = 0$ is equal to $s(k)/\overline{s(k)}$, where the Jost function $s$ of (2) in terms of the Jost solution $y$ of the Dirac system (6) reads $s(k) = y_2(0,k)$. The scattering function $S$ is therefore equal to $y_2(0,k)/\overline{y_2(0,k)}$; it is uniquely defined by the requirement that the function $y(\cdot,k) - S(k)y(\cdot,k)$ is a solution of the system (6) satisfying the boundary condition $y_2(0) = 0$.

In terms of the solution $z$ of (7) introduced above the Jost function $s$ and the scattering function $S$ read

$$s(k) = e^{i\beta}w_1(0,k) + e^{-i\beta}w_2(0,k),$$

$$S(k) = \frac{e^{i\beta}w_1(0,k) + e^{-i\beta}w_2(0,k)}{e^{-i\beta}w_1(0,k) + e^{i\beta}w_2(0,k)}.$$  

(10)

In view of [33], the above formula for the scattering function $S$ suggests that it is also a scattering function for some canonical ZS-AKNS system. It is proved in [33] that the scattering functions for canonical ZS-AKNS systems (7) with potentials (8) for $v \in X_2^+$ and subject to the boundary condition (9) belong to the following set $S_2$.

**Definition 3.1.** We say that a function $S: \mathbb{R} \to \mathbb{C}$ belongs to the class $S_2$ if and only if

1. there are $F \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ and $\delta \in (0, \pi)$ such that for all $k \in \mathbb{R}$ it holds

   $$S(k) = e^{2i\delta} + \int_{-\infty}^{\infty} F(\zeta)e^{2ik\zeta}d\zeta;$$  

   (11)

2. the function $S$ is unimodular, i.e. $(S(k))^{-1} = \overline{S(k)}$ for all real $k$;

3. the winding number $W(S)$ of the function $S$ is equal to zero.

One of the main results of [33] claims that the scattering function $S$ is the scattering function of the problem (7), (9) corresponding to some $(v, \beta) \in X_2^+ \times (0, \pi)$ if and only if $S$ belongs to $S_2$ and the number $\delta$ in its integral representation (11) is equal to $\beta$. Moreover, there is an algorithm based on the Marchenko equation that for a given function $S \in S_2$ constructs a ZS-AKNS system (i.e., the function $v$ in the potential $V$ and the number $\beta$ in the boundary condition) whose scattering function is this $S$. As a corollary, we get the following result:

**Theorem 3.1.** The scattering function $S$ for the problem (1) corresponding to $u, p \in X_2^+$ and subject to the Dirichlet boundary condition belongs to $S_2$.

4  Scattering for energy-dependent Schrödinger equations

The representation (10) of the scattering function for the energy-dependent Schrödinger equation (1) in terms of the companying Jost solution $z$ of the canonical ZS-AKNS system (7) reveals a close connection to the scattering functions of the latter.

We parametrise the set of problems (1) by the pairs $(u, p) \in X_2^+ \times X_2^+$, with $u$ being the Riccati representative of the potential $q$. Then the following uniqueness result holds:

**Theorem 4.1.** The scattering function $S$ for the problem (1) corresponding to $u, p \in X_2^+$ and subject to the Dirichlet boundary condition determines $u$ and $p$ uniquely.

The crucial observation is that for such an $S$ there is a unique canonical ZS-AKNS equation (7) with potential $V$ of (8), for which $S$ is its scattering function. This potential $V$ gives a complex-valued function $v \in X_2^+$; according to the results of subsection relHrynovssctransf, our task is to represent this function $v$ as $(-u + ip)e^{-2i\gamma}$ for some real-valued $X_2^+$-functions $u$ and $p$ and with $\gamma(x) := \int_x^\infty p(t)dt$. In other words,

$$-u(x) + ip(x) = v(x)e^{2i\gamma(x)},$$
and, equating the imaginary parts and recalling that $p = -\gamma'$, we get the differential equation

$$\gamma' = -(\Re v) \sin 2\gamma - (\Im v) \cos 2\gamma$$

for the function $\gamma$. By the definition of $\gamma$, it should vanish at $+\infty$. Using the Banach fixed point theorem, one can prove that the above equation has a unique solution vanishing at $+\infty$. This solution gives a unique $p := -\gamma'$ and a unique $u := i v(x)e^{2i\gamma(x)}$ and, therefore, provides a reconstruction algorithm.

We note that the problem of complete characterization of all scattering functions of (1) is much more difficult. Not every $S \in S_2$ is a scattering function for equation (1) subject to the Dirichlet boundary condition; indeed, to this end the solution $\gamma$ of (12) and the number $\delta$ in the integral representation (11) of $S$ must be related via $\delta = \gamma(0)$. This and some other related issues will be discussed in detail elsewhere.

References

[22] Nabiev A A 2006 Inverse Problems 22 2055–68
Minisymposium M7

Regularisation Methods – Theory
On stable parameter estimation in epidemiology using abstract discrepancy principle

A Smirnova¹, A Bakushinsky² and L deCamp³

¹Department of Mathematics and Statistics, Georgia State University, Atlanta, GA 30303, USA
²Institute of System Analysis, Russian Academy of Sciences, Moscow 117312, Russia
³Department of Mathematics and Statistics, Georgia State University, Atlanta, GA 30303, USA

E-mail: asmirnova@gsu.edu

Abstract: In this paper we aim to further theoretical and computational analysis of the discrepancy principle (DP) for linear irregular operator equations in Banach spaces. The driving force for this research was a set of preliminary numerical simulations for estimating the bird-to-human transmission rate in avian influenza. This inverse problem was cast as a Volterra integral equation of the first kind with noise both in the kernel and in the right-hand side. Given real data for human cases and poultry outbreaks, we applied a combination of general and "problem-oriented" stabilizing procedures in the attempt to obtain consistent reconstructions. Our experiment with selecting a regularization parameter by a discrepancy-type algorithm suggested that for a successful application of the DP the structure of a particular regularizing operator is not important. What is important is a reliable estimate of the overall noise level as well as a certain measure of how the accuracy and stability are balanced in a specific regularization algorithm. Based on this observation and the idea outlined by A. Bakushinsky [1, 2], we formulate and prove a convergence theorem for what we call the abstract discrepancy principle (ADP) that provides a unified approach to the implementation of the DP for arbitrary family of regularizing operators. In that sense, it generalizes prior results on the DP for linear ill-posed problems. For a noise-free operator and a special kind of the discrepancy function, the ADP was justified in [8].

Keywords: epidemiology, regularization, discrepancy principle.

1 Avian Influenza Model and Numerical Results

The study presented in this paper has been motivated by a parameter identification problem arising in epidemiology. The goal was to calculate a time-dependent bird-to-human transmission rate of the avian influenza virus. When it comes to wild birds, this disease is referred to as a low pathogenic avian influenza (LPAI), since it causes mild symptoms only. Once transmitted to domestic poultry, the virus emerges as highly pathogenic strain (HPAI). The mortality rate for domestic birds from HPAI of subtype H5N1 may reach 90-100% in 48 hours. Though currently in a zoonotic state, a dangerous mutability of HPAI coupled with nearly 60% in human death rate (WHO reports 384 deaths out of 648 cases as of December 2013 [11]) point to the need for a thorough investigation of HPAI-H5N1 models. A more accurate estimate of the transmission rate would enable government agencies to develop appropriate control strategies and safety measures, and to assess their effectiveness. For computing the unknown transmission rate, we process monthly data for human cases [11] and poultry outbreaks [3, 4] over the period from July 2008 until December 2013, and then find a regularized solution of the underlying inverse problem. As suggested in [10], we analyze the avian influenza dynamics by
using the SI (susceptible-infected) model with a variable transmission parameter. By combining the differential equation for the susceptible humans, \( S(t) \), with the expression for the cumulative number of HPAI-H5N1 human cases, \( D(t) \),

\[
\frac{dS}{dt} = \Lambda - \beta(t)I_b(t)S(t) - \mu S(t) \quad \text{and} \quad D(t) = \int_0^t \beta(\tau)S(\tau)I_b(\tau) \, d\tau, \quad t \in (0, T),
\]

we derive the Volterra integral equation of the first kind [7] for the reconstruction of bird-to-human transmission rate, \( \beta(t) \),

\[
\int_0^t K(D(\tau), \tau)\beta(\tau)d\tau = D(t), \quad K(D(\tau), \tau) := S(D(\tau), \tau)I_b(\tau), \quad t \in (0, T),
\]

where \( S(D(t), t) = \left( S_0 - \frac{\Lambda}{\mu} \right)e^{-\mu t} + \frac{\Lambda}{\mu} - D(t) + \mu \int_0^t D(\tau)e^{-\mu(t-\tau)}d\tau \). The goal of the inverse problem is to evaluate \( \beta(t) \), given the number of infected birds, \( I_b(t) \), the total number of confirmed H5N1 human cases, \( D(t) \), and the constant parameters \( \Lambda, \mu \) and \( S(0) \). Both, the right-hand side and the kernel are noise contaminated. The data for \( D(t) \) is collected from the WHO website [11], see Figure 1. For susceptible humans, \( \mu \) and \( \Lambda \) are the natural death and birth/recruitment rates respectively. Since data related to humans is given in units of \( 10^5 \) individuals, human population is set to 70,000 [10]. The average life expectancy is approximately 70 years and \( t \) is given in months, therefore \( \mu = 1/(70 \times 12) \) month\(^{-1} \) and \( \Lambda = 1000/12 \) births per month (in \( 10^5 \) individuals) [10]. We assume that initially all humans are susceptible so that \( S(0) = 70,000 \). For model (2), the number of infected poultry, \( I_b(t) \), is required, while FAO reports the total number of outbreaks [3, 4]. Thus, \( I_b(t) \) can be computed by estimating the average poultry farm size and multiplying this by the number of outbreaks. Worldwide farm size varies from hundreds to several thousands in the countries affected by the H5N1 virus. We take the average worldwide farm size to be 1000 [7], and data related to poultry is converted to units of \( 10^2 \) (see Figure 2).

In order to approximate the transmission rate stably, we utilize three regularization schemes: two general (Tikhonov’s [9] and Lavrentiev’s [6]) and one “problem-oriented” (the local regularization procedure [5]). In local regularization, the penalty term is extracted from the operator rather than added, and the regularization parameter stands for the length of the interval where \( \beta \) remains unchanged. Local regularization is particularly beneficial for Volterra-type equations. For each method employed, we target a relative discrepancy level of 0.005 in order to choose a regularization parameter \( \alpha \). To avoid solving an auxiliary nonlinear equation, we first pre-estimate \( \alpha \) on a coarse (with respect to \( \alpha \)) grid so that afterwards 10 – 15 “small” steps would bring the discrepancy to 0.005 mark. The last 5 steps of this process are shown in Table 1. It actually took 11 steps for Tikhonov’s algorithm, 13 for Lavrentiev’s and 12 for the local regularization. The results of these experiments were notable in that all methods improved the solely discretized solution and produced similar oscillations, which were expected from the ebb and flow of incidents (Figure 3). There is a significant discontinuity in
the real data, ascribed to Indonesia’s non-monthly reporting for 2009. As a result, all regularizing algorithms show a rapid increase in the bird-to-human transmission rate around December 2009 that is not entirely appropriate. Additionally, in 2013, the data for human infections is more sporadically reported, again resulting in rapid increases in the transmission rate that may not be entirely justified. The remaining oscillations can be attributed to seasonality due to temperature fluctuations, environmental changes, and other natural factors. The consistency of the results achieved when approaching this problem with multiple methods yet maintaining the same discrepancy level leads us to the formulation of the abstract discrepancy principle.

2 The Abstract Discrepancy Principle

In this section we study a general inverse problem in the form \( A\beta = f, \ A : \mathcal{X} \rightarrow \mathcal{Y}, \) on a pair of Banach spaces \( \mathcal{X} \) and \( \mathcal{Y}, \) given an approximate operator \( A_h, \) linear, bounded, with a dense range, and an element \( f_\delta, \) a noise-contaminated right-hand side, satisfying inequalities

\[
||A - A_h|| \leq h \quad \text{and} \quad ||f - f_\delta|| \leq \delta.
\]

Suppose that \( A \) is linear, bounded, and one-to-one, \( f \in R(A), \) \( ||f|| \neq 0, \) and \( \beta^* \) is the solution of interest. Let a family of linear bounded operators \( R_\alpha : \mathcal{L} \times \mathcal{X} \rightarrow \mathcal{Y}, \) \( \alpha > 0 \) generate a regularizing strategy such that

(a) \( R_\alpha(A_h) \) is strongly continuous with respect to \( \alpha; \)

(b) \( \lim_{\alpha \rightarrow 0^+} R_\alpha(A_h)A_h\beta = \beta \quad \text{for all} \quad \beta \in \mathcal{X}, \quad \text{and} \)

\[
\]
(c) \( \lim_{\alpha \to 0, h \to 0} R_\alpha(A_h)A_h\beta^* = R_\alpha(A)A\beta^* \) for any \( \alpha > 0 \). 

Assume that the regularization parameter \( \alpha = \alpha(\delta, f_\delta, h, A_h) \) is selected by the abstract discrepancy principle as a solution to the equation

\[
||A_h\beta_\alpha - f_\delta|| = \rho(\delta, h),
\]

where \( \beta_\alpha := R_\alpha(A_h)f_\delta \) and \( \rho = \rho(\delta, h) < ||f_\delta|| \) is an \textit{a priori} adjusted discrepancy level.

**Theorem 1.** Under the above assumptions, a parameter \( \alpha \) satisfying (6) exists and

\[
\lim_{\delta \to 0, h \to 0} ||R_\alpha(\delta, f_\delta, h, A_h)(A_h)f_\delta - \beta^*|| = 0,
\]

provided the following conditions hold

\[ \lim_{\alpha \to \infty} \sup_{\|A - A_h\| \leq h} ||A_h R_\alpha(A_h)|| = 0, \]

\[ \sup_{\alpha \geq 0, \|A - A_h\| \leq h} ||R_\alpha(A_h)|| < \infty, \]

\[ \rho(\delta, h) > c(\delta + h\|\beta^*\|), \]

\[ \lim_{\delta \to 0, h \to 0} \max\{\delta, h\} \rho(\delta, h) = 0. \]

**Proof of Theorem 1.** First, we verify solvability of equation (6). From (8) one concludes that \( \lim_{\alpha \to \infty} ||A_h\beta_\alpha - f_\delta|| = ||f_\delta|| \). On the other hand, due to the range of \( A_h \) being dense in \( Y \), for any \( \epsilon > 0 \) there is \( \tilde{\beta} = (\delta, \epsilon) \in \mathcal{Y} \) such that \( ||A_h\tilde{\beta} - f_\delta|| \leq \frac{\epsilon}{2c} \). By condition (4) for sufficiently small \( \alpha \), \( ||R_\alpha(A_h)A_h\tilde{\beta} - \tilde{\beta}|| \leq \frac{\epsilon}{2||A_h||} \). Hence

\[
||A_h\beta_\alpha - f_\delta|| = ||A_h R_\alpha(A_h)(f_\delta - A_h\tilde{\beta}) + A_h(R_\alpha(A_h)A_h\tilde{\beta} - \tilde{\beta} + f_\delta)|| \\
\leq ||A_h R_\alpha(A_h) - I|| ||f_\delta - A_h\tilde{\beta}|| + ||A_h|| ||R_\alpha(A_h)A_h\tilde{\beta} - \tilde{\beta}|| \\
\leq \epsilon,
\]

and \( \lim_{\alpha \to 0^+} ||A_h\beta_\alpha - f_\delta|| = 0 \). This, combined with the strong continuity of \( R_\alpha(A_h) \) with respect to \( \alpha \), implies that (6) is solvable. Now, we turn our attention to the main part of the theorem and prove relationship (7). One has

\[
||R_\alpha(\delta, f_\delta, h, A_h)(A_h)f_\delta - \beta^*|| \leq ||R_\alpha(\delta, f_\delta, h, A_h)(A_h)||\|\delta + h\|\|\beta^*\|| + ||R_\alpha(\delta, f_\delta, h, A_h)(A_h)A_h\beta^* - \beta^*||.
\]

Let us check that

\[
\lim_{\delta \to 0, h \to 0} ||R_\alpha(\delta, f_\delta, h, A_h)(A_h)||\|\delta + h\|\|\beta^*\|| = 0.
\]

Conditions (3), (6) and (8) yield

\[
||A_h\beta^* - A_h R_\alpha(A_h)A_h\beta^*|| \geq ||A_h R_\alpha(A_h)f_\delta - f_\delta|| - ||I - A_h R_\alpha(A_h)|| ||A - A_h|| ||\beta^*|| + ||f - f_\delta||
\]

\[ \geq \rho(\delta, h) - c(h\|\beta^*\| + \delta). \]

From (9) and (12), one arrives at the estimate

\[
||R_\alpha(\delta, f_\delta, h, A_h)(A_h)||\|\delta + h\|\|\beta^*\|| \leq d(\delta + h\|\beta^*\||)/\left[\rho(\delta, h) - c(h\|\beta^*\| + \delta)\right].
\]

By (10), this means that relationship (11) is satisfied. In order to show that

\[
\lim_{\delta \to 0, h \to 0} ||R_\alpha(\delta, f_\delta, h, A_h)(A_h)A_h\beta^* - \beta^*|| = 0,
\]

denote \( \bar{\alpha} := \lim_{\delta \to 0, h \to 0} \alpha(\delta, f_\delta, h, A_h) \). Suppose \( \bar{\alpha} = \infty \), and let \( \{\delta_m\} \), \( \lim_{m \to \infty} \delta_m = 0 \), and \( \{h_m\} \), \( \lim_{m \to \infty} h_m = 0 \), be such that \( \lim_{m \to \infty} \alpha_m = \infty \), where \( \alpha_m := \alpha(\delta_m, f_{\delta_m}, h_m, A_{h_m}) \). From (8), one derives

\[
\lim_{m \to \infty} ||A_{h_m}\beta_{\alpha_m} - f_{\delta_m}|| \geq \lim_{\delta \to 0} ||f_\delta|| \cdot \lim_{\alpha \to \infty} \left(1 - \sup_{||A - A_h|| \leq h} ||A_h R_\alpha(A_h)||\right) = ||f||.
\]
At the same time, according to (6) \( \lim_{n \to \infty} ||A_{h_n} \beta_{\alpha_n} - f_{\delta_n}|| = \lim_{h \to 0, \delta \to 0} \rho(\delta, h) = 0 \neq ||f|| \). This contradiction yields \( \alpha < \infty \). If it is actually zero, then (13) follows from (5). Otherwise, \( 0 < \alpha < \infty \). Consider arbitrary sequences \( \{\delta_n\} \), \( \lim_{n \to \infty} \delta_n = 0 \), and \( \{h_n\} \), \( \lim_{n \to \infty} h_n = 0 \). If the corresponding sequence \( \{\alpha_n\} \), \( \alpha_n := \alpha(\delta_n, h_n, A_{h_n}) \), converges then \( \lim_{n \to \infty} \alpha_n = \alpha \leq \alpha \). As it follows from (3), (5), and (8)

\[
\lim_{n \to \infty} ||A_{h_n} \beta_{\alpha_n} - f_{\delta_n}|| \geq \lim_{\alpha \to \alpha, h \to 0} ||A_h(R_{\alpha}(A_h - I))\beta^*|| - \lim_{\delta \to 0, h \to 0} c(\delta + h)||\beta^*||
\]

and \( \lim_{n \to \infty} ||A_{h_n} \beta_{\alpha_n} - f_{\delta_n}|| = ||A(R_{\alpha}(A - I))\beta^|| \), while equation (6) together with (10) imply

\[
\lim_{n \to \infty} ||A_{h_n} \beta_{\alpha_n} - f_{\delta_n}|| = \lim_{h \to 0, \delta \to 0} \rho(\delta, h) = 0. \]

Thus, \( ||A(R_{\alpha}(A - I))\beta^*|| = 0 \) and, since \( A \) is one-to-one, \( R_{\alpha} A \beta^* = \beta^* \). Applying (5) one more time, we get

\[
\lim_{n \to \infty} ||A_{h_n} (R_{\alpha_n}) (A_h \beta^* - \beta^*)|| = \lim_{\alpha \to \alpha, h \to 0} ||R_{\alpha}(A_h (A_h \beta^* - \alpha A(\beta^*)|| = 0,
\]

which proves (13). If \( \{\alpha_n\} \) is divergent, then (13) still holds. Indeed, assume the opposite:

\[
\limsup_{n \to \infty} ||R_{\alpha_n}(A_{h_n})A_h \beta^* - \beta^*|| = \eta, \quad 0 < \eta \leq \infty, \quad \text{and let} \quad \{\alpha_{n_k}\} \quad \text{be such that}
\]

\[
\lim_{k \to \infty} ||R_{\alpha_{n_k}}(A_{h_{n_k}})A_h \beta^* - \beta^*|| = \eta. \quad \text{If one takes} \quad \{\alpha_{n_{k_j}}\}, \quad \lim_{j \to \infty} \alpha_{n_{k_j}} = \tilde{\alpha} \leq \alpha, \quad \text{then by the above argument} \quad \lim_{j \to \infty} ||R_{\alpha_{n_{k_j}}}(A_{h_{n_{k_j}}})A_h \beta^* - \beta^*|| = 0 < \eta. \quad \text{This contradiction proves that (13) is fulfilled in that case as well. This completes the proof.}
\]

**Remark 2.** For equation (2), we have verified conditions of Theorem 1 when \( \mathcal{X} = \mathcal{Y} = L_2(0, T) \). The details of this analysis will be published in our future notes.

**Acknowledgments**

This work is supported by NSF under grant (DMS-1112897).

**References**


Minisymposium M8

Inverse Problems in Cell Biology
3D migration of cells solving an inverse problem

G. Vitale\textsuperscript{1,2}, L. Laforgue\textsuperscript{1,2}, R. Michel\textsuperscript{1,2}, J. Étienne\textsuperscript{1,2}, V. M. Laurent\textsuperscript{1,2}, A. Duperray\textsuperscript{3,4,5}, C. Verdier\textsuperscript{1,2}.

\textsuperscript{1}Univ. Grenoble Alpes, LIPHY, F-38000, Grenoble, France.
\textsuperscript{2}CNRS, LIPHY, F-38000, Grenoble, France.
\textsuperscript{3}INSERM, IAB, F-38000, Grenoble, France.
\textsuperscript{4}Univ. Grenoble Alpes, IAB, F-38000, Grenoble, France.
\textsuperscript{5}CHU de Grenoble, IAB, F-38000, Grenoble, France.

E-mail: guido.vitale@ujf-grenoble.fr, vitale.gui@gmail.com

Abstract: Traction Force Microscopy (TFM) is an inverse method that allows to obtain the stress field applied by a living cell on the environment on the basis of a pointwise knowledge of the displacement produced by the cell itself during its migration. This biophysical problem, usually addressed in terms of Green functions, can be alternatively tackled in a variational framework. In such a case, a suitable penalty functional has to be minimized. The resulting Euler-Lagrange equations include both the direct problem based on the linear elasticity operator as well as an equation built on its adjoint. Results from a two-dimensional model, i.e. where living cancer cells are migrating on a plane substrate, are briefly presented. While the mathematics is well established also in the three-dimensional case, i.e. where cells are completely embedded in the gel matrix, the experimental data needed are more difficult to obtain than the two-dimensional counterpart. First steps towards the complete three-dimensional traction reconstruction are reported.

Keywords: Traction Force Microscopy, Inverse Problem, Three-dimensional Cell Migration.

1 Introduction

Many living cells have the ability to migrate, both in physiological and pathological conditions; examples include wound healing, embryonic morphogenesis and the formation of new vessels in tumours. The motility of a cell is driven by the reorganization of its inner structure, the cytoskeleton, according to a complex machinery. The net effect of this process is that a cell is able to apply a stress on the environment, pulling the surrounding material to produce its own movement. Therefore, the determination of forces on the basis of measured displacement, has become a popular problem in the biophysical community.

The early idea to study the force applied by cells in their migration (so called Traction Force Microscopy, TFM) on flat substrate as an inverse problem dates back to the work of Harris and coworkers in the 1980s [6]. Afterwards, Dembo and Wang [4] came up with a new methodology: the living cell is again put on an elastic substrate and the displacement of fluorescent beads dispersed in such material is evaluated from different images. Finally, they solve the problem of elasticity in the substrate in terms of Green functions and then minimize the error between the measured and the calculated displacement under Tichonov regularization.

An alternative approach can be stated in a continuous variational framework [1]. Again, the starting point is a Tichonov functional defined as the displacement error norm plus a regularization. If a variation of such a functional is operated, the definition of an adjoint problem for the unknown force naturally arises. By doing this, two coupled elliptic partial differential equations are obtained and their solution can be approximated, for instance, using finite element.
Although that approach is less common than the seminal method based on Green functions, it has some attractive features that make it worth investigating further: it can be generalized easily to more complex geometry and material behaviour including non-linearities.

In this work we aim to recall the mathematical structure of the TFM problem. We then present some results of the well established two-dimensional theory for a living cancer cell migrating on a plane substrate. Finally, some early data corresponding to an instance of the transmigration of a living cell in a three-dimensional environment are shown.

2 Mathematical Setting

The Direct Problem of Linear Elasticity

The stationary force balance equations on a linear elastic continuum body \( \Omega \subset \mathbb{R}^3 \) writes:

\[
\begin{cases}
- \text{div} (A \nabla u) = 0, & \text{in } \Omega, \\
(A \nabla u) \mathbf{n} = T, & \text{on } \Gamma_N, \\
u = 0, & \text{on } \Gamma_D.
\end{cases}
\] (1)

where \( u \) is the displacement field. The domain \( \Omega \) represents the gel. The boundaries \( \Gamma_D, \Gamma_N \subset \partial \Omega \) are such that \( \Gamma_N \cap \Gamma_D = \emptyset \), \( \Gamma_N \cup \Gamma_D = \partial \Omega \) and \( \mathbf{n} \) is the unit outward normal of the boundary \( \partial \Omega \). On the Neumann boundary \( \Gamma_N \) a load is applied per unit surface (here referred as ‘traction’ and denoted with \( T \)) is applied. The region where the cell and the gel are in contact, say \( \Gamma_c \), is a subset of \( \Gamma_N \). The set \( \Gamma_c \) is indeed the support of the traction field \( T \). Last, \( A \) is the fourth order Hooke elasticity positive tensor.

For notational convenience, we define the solution operator \( S \), as the map that, for a given control \( T \), assigns the displacement field \( u \) that solves the elasticity problem:

\[ u = ST \quad \text{if and only if } \quad u \text{ solves } (1). \]

Two-dimensional approximations of the elasticity equation have already been succesfully employed in the field of TFM [1]. They rely on the existence of suitable averaging operators [8] whose action is denoted by an overbar (\( \bar{\cdot} \)). Let us consider \( \Omega \) being a cylinder with \( \Gamma_N \) the upper surface, i.e. the plane of cell migration. In [1] it is shown that the system (1) can be averaged in the direction orthogonal to \( \Gamma_N \) and that gives:

\[
\begin{cases}
- \text{div} (\tilde{A} \nabla \bar{u}) = T, & \text{in } \Gamma_N, \\
\bar{u} = 0, & \text{on } \partial \Gamma_N.
\end{cases}
\] (2)

where \( \tilde{A} \) denotes the depth-averaged elasticity tensor. Observe that on the \( \Gamma_N \) boundary, Dirichlet conditions apply.

Available Experimental Data

Standard imaging techniques are available in order to detect the cell, thus giving the region \( \Gamma_c \subset \Gamma_N \).

The experimental devices are able to give information on the actual gel displacement field. The position of small fluorescent beads embedded into the gel can be, indeed, tracked for the case of a plane substrate. When dealing with a three-dimensional environment, the position of the intersection of the gel fibers can be similarly obtained. This latter situation represents a novelty from the point of view of data acquisition in TFM and it is currently under development. In view of the smallness of both the beads and the fiber intersection area, we are basically dealing with pointwise measurements of the displacement field. That remark justifies the representation of the observation operator as a list of Dirac mass centered at bead positions. Then, the observation operator writes \( \mathcal{O} := (\delta_{x_1}, \ldots, \delta_{x_N}) \), where \( x_k \) denotes the position of the bead labeled with the number \( k \) and \( \delta_{x_k} \) is the Dirac mass located at \( x_k \).

Biomechanical Constraints

When dealing with applications, the need to include constraints into the model appears. It seems that this was not explicitly stated in the previous literature on TFM, except [10, 11, 8]. We note, indeed, that...
Inverse Problems – from Theory to Applications (IPTA2014)

- since some reasonably accurate images of the cell are available, the support of the traction field \( T \) can be determined, leading to the constraint \( \text{supp} \, T \subseteq \Gamma_c \);

- the mechanical balance of the cell requires \( \int_{\Gamma_c} T = 0 \) and \( \int_{\Gamma_c} x \times T = 0 \);

In the following, the space of tractions satisfying the above constraints will be called admissible and denoted with

\[
F_{\text{adm}} := \{ T : \Gamma_N \to \mathbb{R}^d \mid \int_{\Gamma_c} T = 0, \int_{\Gamma_c} x \times T = 0, T = 0 \text{ a.e. on } \Gamma_N \setminus \Gamma_c \}. \tag{3}
\]

Where \( d = 2 \) or \( 3 \) depending on the specific application. Note that \( F_{\text{adm}} \) is a linear subspace of the space of tractions [10]. The projection on the space \( F_{\text{adm}} \) is denoted by \( P \).

The Inverse Problem

The information provided experimentally to solve the inverse problem, i.e. the pointwise measurements of the displacement \( u \), are not sufficient to yield a unique traction field \( T \). The problem is therefore underdetermined and, as customary, we have to enforce a suitable minimization problem to circumvent this drawback.

Tichonov Functional

The Tichonov functional is defined as:

\[
\mathcal{J}(T) = \frac{1}{2} \| \mathcal{OS} T - \mathbf{u}_0 \|^2 + \frac{\varepsilon}{2} \| T \|^2. \tag{4}
\]

Here \( \mathbf{u}_0 = (u_{01}, \ldots, u_{0N}) \) are the known displacements at \( x_1, \ldots, x_N \) respectively. The Tichonov functional \( \mathcal{J} \) is the sum of two parts: the first term is the discrepancy between the measured displacement \( \mathbf{u}_0 \) and the calculated displacement for a given force \( T \) (i.e. \( \mathcal{OS} T \)), evaluated at the beads location (i.e. \( \mathcal{OS} T \)); the second one is the force magnitude. The two additive contributions are weighted by the positive constant \( \varepsilon \), the regularization parameter.

This traction field \( T \) is thus defined as the (unique) minimizer of the Tichonov functional \( \mathcal{J} \) in the set of the admissible tractions \( F_{\text{adm}} \). It therefore satisfies:

\[
P \mathcal{J}'(T) = \varepsilon T + P(\mathcal{OS})^T (\mathcal{OS} T - \mathbf{u}_0) = 0. \tag{5}
\]

Adjoint Equation

Although the stationarity condition (5) is in principle sufficient to define the optimal \( T \), it is convenient to reformulate the problem in terms of a differential equation. It turns useful to define the adjoint state \( p \) as

\[
\mathcal{S}^{-T} p = \mathcal{O}^T (\mathcal{O} u - \mathbf{u}_0), \tag{6}
\]

where \( \mathcal{S} \) is the solution operator and \( u \) is the displacement field, as defined in equation (1). Substituting (6) into (5), we find the relationship between the optimal force \( T \) and the adjoint state:

\[
\varepsilon T + P p = 0 \tag{7}
\]

The presence of the projection \( P \) requires, when solving the problem in practice, the introduction of Lagrange multipliers as done in [10, 8]. Since it can be shown that \( \mathcal{S} \) is self-adjoint [3], equation (6) can be rewritten in a more familiar way as follows:

\[
\begin{cases}
- \text{div}(\mathcal{A} \nabla p) = \mathcal{O}^T (\mathcal{O} u - \mathbf{u}_0), & \text{in } \Omega, \\
(\mathcal{A} \nabla p) n = 0, & \text{on } \Gamma_N, \\
u = 0, & \text{on } \Gamma_D.
\end{cases} \tag{8}
\]

The latter equation is specific for the three dimensional problem. It is easy to deduce the adjoint equation relative to the two-dimensional averaged case (2), which can be found in [1, 8].
Numerical Approximation

Summarizing, our differential problem in strong form is represented by the system of equations (1) and (8) supplemented by the definition of the adjoint state (7) and the constraints (3). Well-posedness of the aforementioned system and further details are available in [10]. The numerical implementation using Finite Elements is thus straightforward. The choice of the regularization parameter $\varepsilon$ is made by the use of the $L$-curve method [5]. A preliminary validation test in the three-dimensional case can be seen in [11].

Our approach differs from [7] where the three-dimensional TFM problem is tackled using numerically approximated Green functions. However, no details are provided on the mathematical well-posedness. Also, the latter method does not permit to incorporate the biomechanical constraints mentioned here. In addition, the Green function-based method is proved to be computationally more expensive as compared to ours [11].

3 Applications

The well established tool for evaluating traction forces on a plane substrate [1] has been successfully applied when studying the migration of a living cancer cell [2]. In Fig. 1a) a living cancer cell is shown while migrating on a plane substrate and is tagged with fluorescence. Fig. 1b) shows the corresponding traction map. Relevant biophysical facts that may be deduced from two-dimensional TFM are detailed in [9].

The more difficult case corresponding to migration in a three-dimensional environment is currently under study. An image of the cell embedded in a collagen matrix is shown in Fig. 2a). In this case, the displacement field is measured using an auto-correlation technique and can be reconstructed pointwise. An interpolation has been required in order to visualize the latter as a colormap on a section plane in Fig. 2b), whereas it is not necessary for the numerical calculations.

![Figure 1](image1.png)

Figure 1. a) GFP actin fluorescent living cancer cell during migration on a plane substrate; b) Traction map.

The force reconstruction is, therefore, possible since all the ingredients are available. Together with a discussion of the biological implications of such data, that will be the subject of forthcoming work by our lab.

References

Figure 2. a) A living cancer cell during its migration in a three dimensional collagen gel (confocal microscopy, z-slices); b) colormap of the magnitude of the displacement field (µm) induced by the cell in the z-plane (interpolated on the whole domain just for visualization).

An optimal control approach to cell tracking

K. Blazakis¹,*, C.C. Reyes-Aldasoro², V. Styles¹, C. Venkataraman¹ and A. Madzvamuse¹,*

¹University of Sussex, Department of Mathematics, Brighton, UK, BN1 9QH
²City University London, School of Mathematical Sciences, Computer Science and Engineering, Northampton Square, London, UK, EC1V OHB
*Corresponding authors

E-mail: K.Blazakis@sussex.ac.uk, reyes@city.ac.uk, V.Styles@sussex.ac.uk, C.Venkataraman@sussex.ac.uk and A.Madzvamuse@sussex.ac.uk

Abstract: Cell tracking is of vital importance in many biological studies, hence robust cell tracking algorithms are needed for inference of dynamic features from (static) in vivo and in vitro experimental imaging data of cells migrating. In recent years much attention has been focused on the modelling of cell motility from physical principles and the development of state-of-the art numerical methods for the simulation of the model equations. Despite this, the vast majority of cell tracking algorithms proposed to date focus solely on the imaging data itself and do not attempt to incorporate any physical knowledge on cell migration into the tracking procedure. In this study, we present a mathematical approach for cell tracking, in which we formulate the cell tracking problem as an inverse problem for fitting a mathematical model for cell motility to experimental imaging data. The novelty of this approach is that the physics underlying the model for cell migration is encoded in the tracking algorithm. To illustrate this we focus on an example of Zebrafish (Danio rerio’s larvae) Neutrophil migration and contrast an ad-hoc approach to cell tracking based on interpolation with the model fitting approach we propose in this study.

Keywords: cell tracking, optimal control of PDEs, chemotaxis, cell motility

1 Introduction

Cell migration is a fundamental process in cell biology and is tightly linked to many important physiological and pathological events such as the immune response, wound healing, tissue differentiation, metastasis, embryogenesis, inflammation and tumour invasion [2]. Advances in experimental techniques means that we now have access to both in vivo and in vitro imaging data of migrating cells. Cell tracking is concerned with the development of methods to track and analyse dynamic cell shape changes from static imaging data (see for example [10] for a review), with level set or electrostatic based methods among the most widely used. One feature of the aforementioned methods is that the trajectories they generate are not physical in nature rather they are designed with the goal of achieving nice geometric properties, e.g., equidistribution of vertices, smoothness of the trajectories, etc. On the other hand a major focus of current research is the derivation of mathematical models for cell migration based physical principles, e.g., [4]. Furthermore, such models appear to show good qualitative and quantitative agreement with experimental observations of migrating cells. Recently we investigated fitting parameters in such models to experimental imaging data of migrating cells where observations of both the position of the cells and the concentrations of cell-resident proteins related to motility were available [3].

In this study we focus on the setting, more prevalent in cell tracking problems, where only the position of the cell, specifically the cell membrane, at a series of discrete times is available and no further biological information
is given. To this end, we derive a mathematical model based on physical principles for cell migration and then formulate an inverse problem, which takes the form of a PDE constrained optimisation problem, for fitting the model to the experimental observations. To solve the optimisation problem we propose an algorithm based on previous studies on the optimal control of geometric evolution laws [5]. Finally, we apply the algorithm to some experimental data on the migration of neutrophils and contrast the results of our approach with a simple cell tracking algorithm based on interpolation.

Due to space constraints we eschew technical details in the subsequent discussion focussing on the general idea of our approach and contrasting it with a more standard cell tracking approach in a biological application of in vivo migration. For further details on the methodology see [1].

2 An optimal control approach to cell tracking

In this section we present a numerical method for cell tracking based on optimal control theory. Our method is inspired by the work of Hauer et al. [5], who use an optimal control approach to control the shape evolution of nanoscale islands with electric fields.

As previously stated our approach seeks to fit a mathematical model to experimental data. The model we consider describes the movement of the cell by proposing an evolution law for the motion of the cell membrane which we denote by \( \Gamma_t, t \in [0, T] \), where \( \Gamma_t \subset \mathbb{R}^n, n = 2, 3 \) is assumed to be a closed \( n-1 \) dimensional hypersurface. Given an initial surface \( \Gamma_0 \) our model for the evolution law takes the form of mean curvature with forcing

\[
\vec{v}(\vec{x}, t) = -\sigma H(\vec{x}, t) + \eta(\vec{x}, t) + \lambda(t), \quad \vec{x} \in \Gamma_t, t \in (0, T],
\]

where \( \vec{v} \) is the normal velocity, \( H \) is the sum of the principle curvatures, \( \lambda \) is a spatially uniform forcing term enforcing volume conservation and the forcing function \( \eta \) drives the movement of the cell. Such models may be derived by assuming a force balance on the cell membrane where the various force contributions accounted for are the resistance of the cell membrane to stretching, a hydrostatic pressure that conserves the enclosed volume, a viscous force opposing the motion and protrusive and retractive forces arising due to actin polymerisation or myosin mediated contraction (for further information on the derivation of the model and the inclusion of other forces see [4]).

We now assume we are given some experimental observations of the position of the cell membrane \( \{\hat{\Gamma}_i\}_{i=0,...,N} \) where \( N > 0 \). The idea of the cell tracking algorithm we propose in this study is to find a forcing function \( \eta \) such that the solution to the model equations is close (c.f., (3)) to the experimental data. The initial data for the model is given by the initial experimental observation i.e., \( \hat{\Gamma}_0 := \Gamma_0 \). We approximate the geometric evolution equation with a phase field model (conserved Allen-Cahn with forcing) of the form

\[
\varepsilon \partial_t \varphi(\vec{x}, t) = \varepsilon \Delta \varphi(\vec{x}, t) - \frac{1}{\varepsilon} G'(\varphi(\vec{x}, t)) + c_w(\eta(\vec{x}, t) + \lambda(t)),
\]

for the phase field \( \varphi \) of the curve with initial condition \( \varphi(\vec{x}, t) = \varphi_0(\vec{x}) \) and zero flux boundary conditions on a rectangular domain \( \Omega \) in \( \mathbb{R}^2 \). We take \( G(\varphi) = \frac{1}{4}(\varphi^2 - 1)^2 \) a double well potential, \( \varepsilon \) a small length parameter of the interfacial thickness and \( c_w \) a scaling constant dependent on the double well potential. We also approximate the sharp interface observations of the experimental data (\( \hat{\Gamma}_i \)'s) with diffuse interfaces defined on the same domain \( \Omega \) and using the same value of \( \varepsilon \) (see [3] for details).

As in [5] we formulate the problem of fitting to the experimental data as the following minimisation problem: given an initial phase field function \( \varphi(\vec{x}, 0) \) at time \( t = 0 \), and a desired final phase field \( \varphi_{des}(\vec{x}, T) \), find a suitable control function \( \eta(\vec{x}, t) : \Omega \times [0, T] \rightarrow \mathbb{R}^2 \) which minimises:

\[
J(\varphi(\vec{x}, t), \eta(\vec{x}, t)) = \frac{1}{2} \int_0^T \int_\Omega (\varphi(\vec{x}, t) - \varphi_{des}(\vec{x}, t))^2 \, d\vec{x} dt + \gamma \int_0^T \int_\Omega \eta^2(\vec{x}, t) \, d\vec{x} dt
\]

subject to the constraints given by (2), where \( \gamma \) is a positive regularisation parameter. For simplicity we have assumed \( N = 1 \), i.e., we are attempting to fit to a single observation using a previous observation as data. \( \hat{\phi}(\vec{x}, 0) \) and \( \hat{\phi}(\vec{x}, T) \) are the diffuse interface representations of the initial and final data respectively. We adopt an optimal control approach for the solution of the minimisation problem, formally deriving the first order necessary conditions and employing a simple gradient update scheme for the control. To solve the forward equation (2) and the adjoint equation (for efficient computation of the gradient) we employ a finite element method.
3 Application to experimental data

To contrast the cell tracking approach proposed above with a more standard algorithm, we apply the optimal control based algorithm to the problem of tracking in vivo migration of neutrophils in zebrafish larvae and compare it with an approach based on interpolation. Neutrophil migration as observed in zebrafish larvae is a popular model to study the cellular inflammatory response mainly due to the transparent nature of the zebrafish in the developmental stage and the capacity to reproduce in vivo conditions of inflammation similar to those in humans [9].

Experimental Setup

In the interests of space we only briefly state the details of the experimental setup. For the experimental methodology and that used to generate and segment the imaging data we refer to [6]. Zebrafish were maintained according to standard protocols, tail-fin transection was performed at 3 days post fertilisation, then mounted on melting point agarose, and images were captured using an UltraVIEWVoX spinning disk confocal microscope (PerkinElmer Life and Analytical Sciences) as previously described [9]. The biological data sets were acquired from transgenic Tg(mpx:eGFP)i114 zebrafish larvae in which neutrophils specifically express Green Fluorescent Protein (GFP). We selected eight observations of the migration of a single neutrophil as shown in Figure 1(a). The observations were obtained in MATLAB format from the open-source software PhagoSight [8].

The biological data sets were acquired from transgenic Tg(mpx:eGFP)i114 zebrafish larvae in which neutrophils specifically express Green Fluorescent Protein (GFP). We selected eight observations of the migration of a single neutrophil as shown in Figure 1(a). The orientation is such that tail-fin (which is the direction of expected migration) is to the right of the neutrophils.

A simple interpolation based approach to cell tracking

To contrast the optimal control approach with a cell tracking method that is purely geometric in nature (i.e., the trajectories are non-physical) we implemented a simple algorithm for cell tracking based on cubic spline interpolation. Here, following Madzvamuse et al. [7], given experimental observations of the cell membrane we use cubic splines to generate a series of intermediate cell-surface boundaries with the interpolation chosen such that the point trajectories are smooth ($C^2$). To do this given two successive observations of the cell membrane (that consist of a series of points on the cell membrane) we first manually select a single point on the first surface to be mapped to a specific point on the second surface which uniquely determines the entire mapping of every point as it is assumed the mapping preserves the connectivity (topology).

Results

In Figure 1 we report on the results of the two tracking algorithms applied to the experimental data. We see that the cubic spline interpolation algorithm (as expected) and the optimal control based algorithm both generate centroid trajectories that are significantly smoother than those obtained by linear interpolation of the cell centroids alone. The usual measures of chemotaxis persistence length, chemotactic indices, maximum velocities etc. have been computed and exhibit significant variation using the different approaches but are not reported on due to space constraints. For this example the area of the cell is not conserved in the experimental data therefore we take the linear interpolant of the area of the experimental data to be the target area and determine the $\lambda$ term in (2) such that it (weakly) penalises deviations from this target area (see [3] for details). In Figure 1(d) we clearly observe good agreement between the target area (green line) and the area of the cell using the optimal control approach while the area of the cell using the cubic spline interpolation approach is significantly smaller than the area of the cell away from the end points of the interval. In Figure 1(e) we plot one example of the optimal control based approach fitting between the fourth and fifth (from the left) snapshot of Figure 1(a). We observe good agreement between the final position with the computed optimal control and the experimental data we also above strong contractile forces in the rear of the cell away from the chemoattractant and strong protrusive forces in the front of the cell that points towards the chemoattractant (the chemoattractant concentration is expected to increase from left to right).
Figure 1. (a) Experimental data. (b)-(c) Experimental data and centroids computed with the different cell tracking algorithms. (d) area of the cell, the green line indicates the linear interpolant of the area of the observations, the blue line the area using cubic spline interpolation and the red line the area using the optimal control approach. (e) An example of intermediate morphologies obtained with the optimal control algorithm.
4 Conclusion

In recent years, the rapid development in microscopy and imaging techniques has generated a huge amount of data on migrating cells *in vivo* and *in vitro*. A key challenge is to study from discrete observed snapshots, quantities of interest, such as trajectories of material points on the membrane, velocities and geometric quantities [3]. In this study we have proposed an alternative to the purely geometric widely used cell tracking approaches for the estimation of such quantities in which the estimated trajectories correspond to those generated by a physical model for the evolution and thus may be more physically meaningful than those obtained previously.

While this proof of concept application to experimental data focused on 2-dimensional observations each of the algorithms is immediately applicable to 3-dimensional data sets. Moreover while we specifically focussed on the case where minimal information regarding the biology is available, in theory our approach is applicable to models where more biology is included. In particular models for the dynamics of actin and myosin as well as other motility related species within the cell and on the membrane may be included and the evolution law may be modified to take into account the dependence of the movement of the cell on these species, see for example [4]. We also believe the optimal control/inverse problem approach we present in this study could be a useful framework within which to investigate other biological questions beyond cell tracking such as the inference of chemotactic fields during *in vivo* chemotaxis [6]. Finally we remark that data on the morphology, curvature, normal velocity, etc., may be easily extracted using the optimal control based approach which may be useful in applications.

Acknowledgements

This work (AM, VS, CV) is supported by the Engineering and Physical Sciences Research Council, UK grant (EP/J016780/1) and the Leverhulme Trust Research Project Grant, UK (RPG-2014-149).

References

TGV-based flow estimation for 4D leukocyte transmigration

L Frerking\textsuperscript{1,4}, M Burger\textsuperscript{1,4}, D Vestweber\textsuperscript{2,4} and C Brune\textsuperscript{3,4}

\textsuperscript{1} Applied Mathematics Münster, University of Münster, 48149 Münster, Germany
\textsuperscript{2} Max Planck Institute for Molecular Biomedicine, 48149 Münster, Germany
\textsuperscript{3} Department of Applied Mathematics and MIRA – Institute for Biomedical Technology and Technical Medicine, University of Twente, 7500 AE Enschede, Netherlands
\textsuperscript{4} Cells in Motion – Cluster of Excellence, University of Münster, 48149 Münster, Germany

E-mail: lena.frerking@uni-muenster.de

Abstract: The aim of this paper is to track transmigrating leukocytes via TGV flow estimation. Recent results have shown the advantages of the nonlinear and higher order terms of TGV regularizers, especially in static models for denoising and medical reconstruction. We present TGV-based models for flow estimation with the goal to get an exact recovery of simple intracellular and extracellular flows, as well as its implication on realistic tracking situations for transmigration through barriers. To study and quantify different pathways of transmigrating leukocytes, we use large scale 4D fluorescence live microscopy data in vivo.

Keywords: TGV, Optical Flow, Tracking, Leukocytes, Transmigration.

1 Introduction

In the past few years, mathematical reconstruction and analysis of 4D data received great attention. In a medical context, its meaning for the recognition and control of a huge variety of diseases is outstanding. An example for the influence of such data analysis is the movement of leukocytes. These cells are part of the immune system and are largely involved in defeating infectious diseases. Leukocyte transmigration denotes the process where leukocytes leave the circulatory system through endothelial cells and move towards a damaged tissue. Nowadays, the process of capturing leukocytes to the endothelial layer is well understood. However, many aspects of the mechanisms of the transmigration are still unexplored like the fact that leukocytes use two different pathways for transmigration. These pathways are referred to as transcellular and paracellular. A leukocyte taking the transcellular pathway moves straight through the body of an endothelial cell. During this process, it forms a hole into the endothelial cell, which gets closed after the leukocyte has left. If a leukocyte takes the paracellular pathway, it moves through junctions of endothelial cells. In this case, the hole is not built into an endothelial cell but into the contact between the cells. Little is known about the mechanical constraints or the kinetic differences of leukocytes preferring one of the pathways. The usage of 4D data allows us to get detailed insight into these processes. We use 4D (3D + time) in vivo data of a fluorescence microscope that shows the leukocytes as well as the endothelial cell contacts. Figure 1 shows a time step of a data set where the leukocytes are colored in green and the endothelial cell contacts are colored in red.

We analyze the behavior of the leukocytes to detect differences between the two pathways concerning e.g. movement, velocity and shape. In this way, we try to get more information about the advantages and disadvantages of the single pathways and aim at discovering the reasons for a preference in specific situations. Since we are especially interested in intracellular movement, we use L1 optical flow models with a total generalized variation (TGV) regularizer. By using a TGV prior it is possible to detect smooth transitions within flow fields. These transitions approximate realistic intracellular flows better than the piecewise constant parts and discontinuities that are detected by TV models (staircasing effect). Moreover, the TGV regularizer is still able to keep sharp discontinuities at flow boundaries related to the contour of leukocytes.
2 Higher-order regularization for optical flow

The performance of flow estimation is based on the assumption of brightness constancy. If we regard the brightness \( f \) of two consecutive images at a distinct place \((x, y)\) and time \( t \), we assume
\[
\frac{\partial f}{\partial t} + \frac{\partial f}{\partial x} \frac{dx}{dt} + \frac{\partial f}{\partial y} \frac{dy}{dt} = 0.
\]
We are searching for a vector field \( u(x, y) = (u_1(x, y), u_2(x, y)) \) that satisfies (1), i.e. we want to solve the inverse problem
\[
Ku = g,
\]
with \( Ku = \nabla f \cdot u \) and \( g = -f_t \).

To approximate a solution, we use a variational model
\[
\min_u J(u) = \min_u \left\{ \lambda D(u, f) + R(u) \right\}
\]
with a positive regularization parameter \( \lambda \), a data fidelity term \( D \) and a regularization term \( R \). We use an L1 norm for the data term. Thus, from (2) we get
\[
D(u, f) = \int_{\Omega} |f_t + (\nabla f)^T \cdot u|.
\]

A variety of former studies (e.g. [1]) consider total variation (TV) as a regularizer for the optical flow model, i.e.
\[
R(u) = \sup_{p \in C^2_0(\Omega; \text{Sym}^2(\mathbb{R}^d)), \|p\|_{\infty} \leq 1} \int_{\Omega} u \text{div} p.
\]
Recently, other imaging topics showed that higher order regularization can have significant advantages for specific applications [2, 3]. For the case of denoising, Benning et al. [4] show that the TGV regularizer is able to reconstruct smooth transitions as well as sharp edges. This specialty is a consequence of the combination of first and second order derivatives in the TGV functional. A TGV regularizer, where \( \beta \) weights first versus second order components, has the following form:
\[
R(u) = \sup_{p \in C^2_0(\Omega; \text{Sym}^2(\mathbb{R}^d)), \|p\|_{\infty} \leq \beta, \|\nabla p\|_{\infty} \leq 1} \int_{\Omega} u \text{div}^2 p dx.
\]

3 Optimization and numerical realization

We adapt the idea of higher order regularization to perform flow estimation for transmigrating leukocytes. A primal realization of the TGV functional is given by
\[
R(u) = \min_{u,v} \int_{\Omega} \alpha_1 |\nabla u - v| + \alpha_0 |\nabla v|.
\]
Obviously, ignoring one term results in a primal realization of the TV functional. To take advantage of this characteristic, we split the functional into a TV equivalent part and a remaining part. Therefore, we introduce Lagrange multipliers $p$ and $q$ and obtain a saddle point formulation for the overall objective functional $J$:
\[
\min_{u \in \mathbb{R}^{2MN}, v \in \mathbb{R}^{2MN}} \max_{p \in P, q \in Q} \left\{ \alpha_1 \langle \nabla u - v, p \rangle + \alpha_0 \langle \nabla v, q \rangle + \lambda D(u, f) \right\},
\]
where the model parameters $\alpha_1$ and $\alpha_0$ are related to $\beta$, and the convex sets $P$ and $Q$ are given by $P = \{ p \in \mathbb{R}^{2MN}, \|p\|_\infty \leq 1 \}$ and $Q = \{ q \in \mathbb{R}^{2MN}, \|q\|_\infty \leq 1 \}$. To solve this equation, we use a primal-dual algorithm, which is mainly based on an algorithm of Ranftl et al. [5], who use TGV for a stereo matching problem:
\[
p_{n+1} = \mathcal{P}_P \left( p_n + \tau_p \alpha_1 (\nabla \hat{u}_n - \hat{v}_n) \right) \tag{3}
\]
\[
q_{n+1} = \mathcal{P}_Q \left( q_n + \tau_q \alpha_0 (\nabla \hat{v}_n) \right) \tag{4}
\]
\[
u_{n+1} = (I + \tau_u D)^{-1} \left( u_n + \tau_u \alpha_1 \nabla \cdot (p_{n+1}) \right) \tag{5}
\]
\[
v_{n+1} = v_n + \tau_v (\alpha_0 \nabla \cdot q_{n+1} + \alpha_1 p_{n+1}) \tag{6}
\]
\[
\bar{u}_{n+1} = 2u_{n+1} - u_n \tag{7}
\]
\[
\bar{v}_{n+1} = 2v_{n+1} - v_n \tag{8}
\]
with stability parameters $\tau_p, \tau_q, \tau_u$, and $\tau_v$. (3) minimizes the first order dual variable. (4) minimizes the second order dual variable. (5) minimizes the primal variables. (6) combines the first and the second order derivatives. (7) and (8) serve as relaxations. By ignoring (4),(5) and (8), and setting $\bar{v} = 0$ in (3), this algorithm reduces to the L1-TV algorithm of Chambolle and Pock [6]. The operators $\mathcal{P}_P$ and $\mathcal{P}_Q$ project the variables onto the sets $P$ and $Q$:
\[
\left( \mathcal{P}_P (\hat{p}) \right)_{i,j} = \frac{\hat{p}_{i,j}}{\max(1, \|\hat{p}_{i,j}\|)} \quad \left( \mathcal{P}_Q (\hat{q}) \right)_{i,j} = \frac{\hat{q}_{i,j}}{\max(1, \|\hat{q}_{i,j}\|)}
\]
The resolvent operator in (5) evaluated at $\hat{u} = u_n + \tau_u \nabla \cdot (p_{n+1})$ is explicitly given by
\[
(I + \tau_u D)^{-1}(\hat{u}) = \hat{u} + \begin{cases} 
\tau_u \lambda \nabla f & \text{if } f_t + (\nabla f)^T \cdot \hat{u} < -\tau_u \lambda \|\nabla f\|^2 \\
-\tau_u \lambda \nabla f & \text{if } f_t + (\nabla f)^T \cdot \hat{u} > \tau_u \lambda \|\nabla f\|^2 \\
-(f_t + (\nabla f)^T \cdot \hat{u}) \frac{\nabla f}{\|\nabla f\|^2} & \text{if } |f_t + (\nabla f)^T \cdot \hat{u}| \leq \tau_u \lambda \|\nabla f\|^2
\end{cases}
\]

4 Numerical results
To depict the differences between TV and TGV regularization for optical flow, we compare the results for both methods with the ground truth result of a 2D synthetic dataset from the IPOL database [7]. Figure 2 clearly shows the typical staircasing effect for TV regularization. In contrast, the regularization with TGV is able to approximate smooth transitions instead of piecewise constant parts. Additionally, the borders of the sphere are still visible in the TGV results. Table 1 verifies that the TGV regularizer reaches improved results compared to the TV regularizer for the chosen synthetic dataset. Both the average endpoint error and the average angular error are smaller for TGV.

Figure 3 shows the result of 3D flow estimation with TGV regularization for experimental leukocyte data. Since the cells are moving fast, we incorporated multiscale steps in the TGV algorithm. Otherwise, the large distances would prohibit a continuous flow inside a cell. To receive a better relation between regularization and resolution of the images, we adapted the regularization parameter in each multiscale step. The higher the resolution of the images the less we regularize. This avoids an enlargement of the influence of small fragments, especially in regions with only little flow. To utilize the whole 3D space of the data, we extended the dimension of the introduced algorithm. Similar to the synthetic data, the TGV approximation also achieves smooth transitions for the intracellular flows, while keeping the sharp edges at the borders of the moving areas. Moreover, the intracellular effect of transmigration becomes obvious through opposed directions of the vector field at the region of the barrier.
Figure 2. Vector field and color coding results for synthetic flows from [7]. The grey horizontal lines in the color coding results specify the position that was chosen for the 1D plots.

Table 1. Average endpoint error (EPE) and average angular error (AAE) of flow estimations for $\lambda = 15$.

<table>
<thead>
<tr>
<th></th>
<th>EPE</th>
<th>AAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>TV</td>
<td>0.1491 pixels</td>
<td>1.3647$^\circ$</td>
</tr>
<tr>
<td>TGV</td>
<td>0.0916 pixels</td>
<td>1.3512$^\circ$</td>
</tr>
</tbody>
</table>

5 Conclusion

Applying TGV to optical flow leads to results, which bear similar characteristics as TGV denoising results. It is possible to approximate smooth transitions in situations where flow estimation with TV offers piecewise constant parts. Furthermore, the TGV regularizer keeps discontinuities at flow boundaries similar to TV. Both characteristics are important to gain information about the intracellular dynamics of migrating leukocytes. Cells are not expected to have strong edges in intracellular flow. However, cell contours are often related to the boundary of the real cell flow. Thus, TGV seems to be an adequate regularizer for analyzing leukocyte behavior.
(a) Greyscale version of leukocyte channel from Figure 1. Right: overview; left: crop of region of interest of time frame 1 and frame 2.

(b) 3D TGV flow estimation.

(c) 2D slice of 3D TGV flow estimation.

Figure 3. 3D flow estimation with experimental data of transmigrating leukocytes.

The algorithm used in this paper can easily be improved by applying Bregman iterations to overcome a loss of contrast in the flow components, see [8] for further details. Another possible adaption is to improve the step sizes by line search. However, in this paper we focused on pointing out the results of applying TGV to a molecular cell biology problem.

Acknowledgments

This work was funded by the German Science Foundation DFG through the Cluster of Excellence 1003 Cells in Motion – CiM: Imaging to Understand Cellular Behaviour in Organisms.

References

Minisymposium M9
Physical Imaging
A CS-based approach for physical imaging

A Massa\textsuperscript{1}, G Oliveri\textsuperscript{1}, P Ding\textsuperscript{2}, and N Anselmi\textsuperscript{1}

\textsuperscript{1}ELEDIA Research Center@DISI, University of Trento, Italy
\textsuperscript{2}Laboratoire des signaux et systèmes, UMR8506 (CNRS-SUPELEC-Univ. Paris Sud), France

E-mail: andrea.massa@unitn.it

Abstract: The exploitation of physical information as a prior within the class of Compressive Sensing (CS) methods is discussed in this paper. More in detail, the formulation of CS inversion algorithms able to retrieve objects of arbitrary size and geometry is addressed by considering sparseness domains which are selected starting from the a-priori physical information available on the scattering scenario. A set of representative numerical examples is reported to assess the features, advantages and drawbacks of the obtained methodologies.

Keywords: Compressive Sensing, Sparse Reconstruction, Physical Imaging, Wavelets, Total Variation.

1 Motivation

The development of inversion techniques able to properly address or at least mitigate the theoretical issues associated to imaging problems \cite{1} is an active research topic in several applications including biomedical diagnostics, non-destructive testing, and subsurface prospecting \cite{2}-\cite{6}. Regularized formulations have been widely developed in the literature to achieve this goal \cite{5}-\cite{6}. In this line of reasoning, the Compressive Sensing (CS) \cite{7}-\cite{9} has been proposed as an effective approach to regularize inverse problems by means of sparseness constraints, and thus to deduce robust and efficient imaging strategies in several different contexts \cite{10}-\cite{16}. The success of these approaches is related both to the efficiency of CS solvers, and to their flexibility \cite{10}-\cite{16}. Indeed, the application of CS paradigm to inverse problem needs only that (a) the relation between the collected data vector \(y\) (e.g., the scattered field) and the vector of unknowns \(x\) (e.g., the target contrast) is linear (i.e., the observation equation is written as \(y = Ax\)), and that (b) \(x\) is sparse in a suitable basis. Nevertheless, most microwave imaging CS techniques have been formulated to effectively handle only “pixel-shaped” targets \cite{10}-\cite{16}. Such a limitation is not actually intrinsic to CS, but it is rather caused by the representation basis chosen for the unknowns, as well as to the domain in which the sparseness is enforced.

The use of physical information has been recently introduced as a powerful approach to generalize CS imaging methods and to enable their application in a significant wider set of applicative scenarios. Indeed, the information on the features of the class of scatterers at hand as well as on the physics of the associated inverse problem can enable the choice of the most proper basis to expand the unknown properties of \(x\), as well as the selection of the correct (direct or transformed) space in which to enforce the sparseness prior. This in turn can guarantee that both (a) and (b) are satisfied in a much wider set of scenarios with respect to pixel-based scatterers \cite{10}-\cite{16}.

This work is then aimed at presenting two innovative imaging methodologies developed within the physical CS imaging framework. More in detail, the a-priori knowledge on the nature and physics of the electromagnetic problem (i.e., field behaviour) or of the scatterer at hand (i.e., class of objects) will be exploited to perform the most proper choice of the expansion basis. Numerical results will be provided to remark the potentialities and limitations of the arising Physical CS Imaging approaches.
2 Problem Formulation and CS Physical Imaging Solution

Consider a 2D inverse scattering problem in which \( V \) transverse-magnetic time-harmonic waves \( E_v^v(r), v = 1, \ldots, V \) impinge on an unknown target placed within an investigation domain \( \Omega \), whose dielectric properties are defined by the contrast \( \tau(r) = \varepsilon_r(r) - 1 - \alpha \sigma(r)/\omega \varepsilon_0 \) where \( \varepsilon_r(r) \) and \( \sigma(r) \) are the relative dielectric permittivity and the electric conductivity, respectively. In this assumption, the First-Order Born Approximated data equation is

\[
E_v^v(r) = -k^2 \int_{\Omega} \tau(r') E_v^v g(r, r') dr', \quad r \in \Sigma^v
\]

where \( k = \omega (\varepsilon_0 \mu_0)^{-1/2} \) is the wavenumber, \( E_v^v(r) \) is the scattered field, and \( g(r, r') \) is the Green’s function. The CS inversion algorithm is aimed at retrieving by discretizing (1) through the following expansion

\[
\tau(r) = \sum_{n=1}^{N} \tau_n \varphi_n(r)
\]

where \( N \) is the number of basis functions \( \varphi_n(r) \), and \( \tau_n \) are the associated unknown coefficients. Accordingly, (1) can be directly written in the CS form \( y = Ax \) by using a point matching approach, where \( x = \{ \tau_n, n = 1, \ldots, N \} \), \( y = \{ E_v^v(\varphi_m), m = 1, \ldots, M \} \) and \( A \) is the observation matrix.

The introduction of physical information within the CS inversion process can be carried out so that either (i) suitable expansion bases \( \varphi_n(r) \), \( n = 1, \ldots, N \), for the unknowns \( x \) are considered, or (ii) an ad-hoc transformed domain is adopted to enforce sparseness (i.e., instead of enforcing it on \( r \)).

As concerns (i), let us firstly recall that the CS formulation of the time-harmonic two-dimensional transverse-magnetic imaging problem can be found in the Multi-Task Bayesian framework (MTBCS) \cite{sholkovitz2012multitask} as \( x^{BCS} = \arg\max_x (P(y|x)) \) such that \( x \) is sparse, where \( P(y|x) \) is the posterior probability of the data, and \( x \) is enforced to be sparse in a Bayesian sense by a proper definition of its prior \( P(x) \). Accordingly, the above formulation can be adopted whatever basis is considered in (2). Standard CS methods \cite{elad2010sparse}–\cite{ma2015optimal} adopt a pixel discretization, thus enabling the use of the arising method only for pixel-sparse methodologies. However, let us consider the case in which profiles comprising both “large” and “localized” targets are of interest [e.g., Fig. 1(a)]. In this scenario, the use of a different expansion functions, for instance belonging to the Haar family \cite{shalett2012adaptive}, should enable the arising Wavelet-MTBCS (W-MTBCS) approach to effectively handle the arising imaging problem.

On the other hand, with reference to (ii), the sparseness prior does not necessarily need to be enforced in the \( x \) domain \cite{zhou2013regularized}. For instance, the class of targets with piecewise-constant contrast profiles are known to have a sparse representation in the gradient domain \cite{zhou2013regularized}. In such a case, the associated imaging problem can be mathematically solved according to the Total-Variation Compressive Sensing (TVCS) paradigm as follows \cite{zhou2013regularized}

\[
x^{BCS} = \arg\min_x \{ ||\nabla x||_1 \} \text{s.t.} \ y = Ax
\]

where \( ||\nabla x||_1 \) is the \( L1 \)-norm of the gradient of the unknowns \cite{zhou2013regularized}, and (3) can be numerically computed by means of an Augmented Lagrangian technique \cite{boyd2011dual}.

3 Numerical Validation

The first numerical example is concerned with a non-homogeneous profile with \( \tau_{\text{max}} = 0.04 \) [Fig. 1(a)], which is imaged by the pixel-based MTBCS [Fig.1(b)], the TVCS [Fig.1(c)], and the W-MTBCS [Fig.1(d)] assuming a signal-to-noise ratio (on the scattered field) equal to SNR=5 [dB].

The retrieved profiles (Fig. 1) remark that the W-MTBCS method is able to exploit the physical information on the sparseness of the target in the Haar domain, as demonstrated by its imaging accuracy [see Fig. 1(d) vs. Fig. 1(a)]. On the contrary, the MTBCS reconstruction resembles a cluster of sparse pixels [Fig. 1(b)], while with the TVCS reconstruction turns out heavily blurred [Fig.1(c)]. These results, which are quantitatively confirmed by the associated integral errors (Tab. 1), are actually expected by the respective physical information which is encoded in the three methods. Indeed, the MTBCS (wrongly) expects the target to be faithfully expressed as a collection of isolated pixels \cite{elad2010sparse}, while the TVCS approach incorrectly enforces the sparseness of the profile gradient (i.e., requests the retrieved contrast to be piecewise constant).
Inverse Problems – from Theory to Applications (IPTA2014)

Figure 1. [Inhomogeneous Object, $\tau_{\text{max}} = 0.04$, SNR=5 dB] Actual (a) and retrieved profiles using MTBCS (b), TVCS (c), and W-MTBCS (d) approaches.

Table 1. [Inhomogeneous Object, $\tau_{\text{max}} = 0.04$, SNR=5 dB] Error figures.

<table>
<thead>
<tr>
<th>Method</th>
<th>Total Error</th>
<th>Internal Error</th>
<th>External Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>MTBCS</td>
<td>$5.1 \times 10^{-3}$</td>
<td>$2.3 \times 10^{-4}$</td>
<td>0.0</td>
</tr>
<tr>
<td>TVCS</td>
<td>$3.2 \times 10^{-3}$</td>
<td>$5.7 \times 10^{-3}$</td>
<td>$2.5 \times 10^{-3}$</td>
</tr>
<tr>
<td>W-MTBCS</td>
<td>$1.4 \times 10^{-3}$</td>
<td>$1.3 \times 10^{-3}$</td>
<td>$9.6 \times 10^{-4}$</td>
</tr>
</tbody>
</table>

However, it is worth remarking that no ideal choice exists concerning the selection of the basis to be used within the CS formulations, and that different approaches can be more or less suitable depending on the applicative domain. To further clarify this point, the retrieval of a T-shaped target characterized by $\tau_{\text{max}} = 0.05$ [Fig. 2(a)] has been considered next.

In this scenario, the piecewise-constant nature of the contrast profile enables the TVCS approach to yield the most accurate reconstruction [Fig. 2(c)]. On the contrary, both the pixel-based MTBCS [Fig. 2(b)] and the wavelet-based W-MTBCS [Fig. 2(d)] fail in terms of retrieval accuracy, because of the a-priori physical information encoded in their solution processes, which turn out incorrect in this case [i.e., the T-shaped scatterer...
Figure 2. \([T\text{-shaped Object, } \tau_{\text{max}} = 0.05 , \text{SNR}=5 \text{ dB}]\) Actual (a) and retrieved profiles using MTBCS (b), TVCS (c), and W-MTBCS (d) approaches.

Table 2. \([T\text{-shaped Object, } \tau_{\text{max}} =0.05 , \text{SNR}=5 \text{ dB}]\) Error figures.

<table>
<thead>
<tr>
<th>Method</th>
<th>Total Error</th>
<th>Internal Error</th>
<th>External Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>MTBCS</td>
<td>$3.0 \times 10^{-3}$</td>
<td>$5.6 \times 10^{-2}$</td>
<td>0.0</td>
</tr>
<tr>
<td>TVCS</td>
<td>$2.6 \times 10^{-3}$</td>
<td>$2.1 \times 10^{-2}$</td>
<td>$1.8 \times 10^{-3}$</td>
</tr>
<tr>
<td>W-MTBCS</td>
<td>$3.9 \times 10^{-3}$</td>
<td>$3.5 \times 10^{-2}$</td>
<td>$2.6 \times 10^{-3}$</td>
</tr>
</tbody>
</table>

is not sparse in the pixel nor in the wavelet domains – Fig. 2(a)]. These considerations are quantitatively supported by the values of the corresponding integral errors for the three methodologies, which are reported in Tab. 2, for completeness.

The numerical results reported in Fig 1 and Fig 2 therefore remark the potential impact and effectiveness of the use of physical \(a\text{-priori}\) information on the reliability and robustness of CS imaging algorithms, as well as the flexibility of this paradigm in the solution of complex inverse problems whenever prior information on the target nature is available.
References

[9] Li C 2009 An Efficient Algorithm for Total Variation Regularization with Applications to the Single Pixel Camera and Compressive Sensing Master Thesis, Rice University
Inverse Problems – from Theory to Applications (IPTA2014)

Differential equation based inversion method for solving inverse scattering problems

Yu Zhong\(^1\), and Xudong Chen\(^2\)

\(^1\)Department of Electronics and Photonics, Institute of High Performance Computing, A*STAR, 138632 Singapore.
\(^2\)Department of Electrical and Computer Engineering, National University of Singapore, 117583 Singapore.

E-mail: zhongyu@ihpc.a-star.edu.sg, and elechenx@nus.edu.sg

Abstract: A new differential equation based inversion method is proposed in this paper, which does not need to solve any forward problem. The finite difference scheme is used to discretized the Helmholtz equation, and the twofold subspace-based regularization scheme \([2]\) is applied in the inversion method. Representative numerical tests are carried out to verify the efficacy of the proposed method.

Keywords: Partial differential equation, inverse scattering, finite difference method

1 Introduction

When dealing with inverse scattering problems, there are usually two types of methods in terms of the equation used to describe the wave, i.e., the integral equation based method and the differential equation (DE) based method. The former one has been intensively studied and the forward solver is no longer a necessity in the inversion, such as in the well-known contrast source inversion (CSI) method \([1]\). Without the forward solver, the efficiency of the inversion is increased. However, when using the DE, forward solver is usually needed to calculate the mismatch between the estimated scattered field and the measured ones \([5]\), or to calculate the numerical Green’s function so as to use the integral equation based inversion method \([3]\) \([4]\).

This paper is to present a new inversion method based on DE to solve two-dimensional (2-D) inverse medium scattering problems, in which there is no forward problem solver or numerical Green’s function. The inversion method can be considered as the counterpart of the CSI method in the DE frame.

In the proposed inversion method, we use finite difference method to discretize the DE, and we adopt the alternative optimization scheme used in the CSI method \([1]\) to avoid solving the forward problem. We modify the alternative optimization scheme such that it can be feasibly applied into the DE frame. We also apply the FFT type twofold subspace-based regularization technique \([2]\) into the inversion so as to stabilize the solver.

With these new aspects, we can efficiently solve the inverse medium scattering problem using the DE modeling, and numerical tests verify the interest.

2 The wave equation

The 2-D problem is with some unknown inhomogeneous medium scatterers embedded in a known background medium with permittivity \(\epsilon_0\) and permeability \(\mu_0\). We denote a domain \(D\) that includes all the unknown scatterers, the permittivity of which can be described as \(\epsilon(\mathbf{r})\).
Supposing that there are \( N_{\text{inc}} \) incidences with angular frequency \( \omega \) illuminating the domain \( D \), the Helmholtz wave equation about the electric fields of the \( l \)th incidence reads as
\[
[\nabla^2 + k^2(r)] E_l(r) = 0, \quad (1)
\]
provided that the active sources are outside the domain \( D \), where \( k(r) = \omega \sqrt{\varepsilon(r)} \mu_0 \) is the wave number. This is an equation about the total electric fields, which can be decomposed into two parts, the incident fields \( E^i_l \) and the scattered fields \( E^s_l \), the former of which are directly produced by the active sources outside the domain whereas the latter are generated by the secondary sources induced in the medium scatterers in the domain.

Therefore, one can write another type of Helmholtz equation for the \( l \)th incidence as
\[
(\nabla^2 + k_0^2) E^s_l(r) = -k_0^2 \chi(r) (E^s_l(r) + E^i_l(r)), \quad (2)
\]
where \( k_0 = \omega \sqrt{\varepsilon_0 \mu_0} \) is the wave number for the background medium, and the contrast function is \( \chi(r) = [\varepsilon(r) - \varepsilon_0] / \varepsilon_0 \). From Eq. (1) to Eq. (2), we use the fact \( (\nabla^2 + k_0^2) E^i_l(r) = 0 \) for \( r \in D \).

The DE is coupled with a boundary condition for the calculation domain as
\[
E^s_l(r) = \int_{\partial D} \frac{\partial G_0(r, r')}{\partial n'} E^s_l(r') - G_0(r, r') \frac{\partial E^s_l(r')}{\partial n'} \, dl', \quad (3)
\]
where \( n \) denotes the outward normal direction of the boundary \( \partial D \) of domain \( D \). Equation (3) can also be used to calculate the scattered fields on the measurement domain in the far-field zone.

### 3 The inversion method

**Discretization of the differential equations**

In the inversion method, we use finite difference method to discretize the DE. We first discretize the rectangular calculation domain \( D \) into \( M \times N \) small cells, in which the contrast are considered to be constant. For scattered electric fields, we assume they are linearly changing from cell to cell, and the scattered fields at the centers of these small cells are denoted as \( E^s_{l,m,n} \), with \( 1 \leq m \leq M \) and \( 1 \leq n \leq N \). We also assume that in those cells on the boundary, the contrast there are zeros, meaning that there is no scatterer there. Due to the finite difference algorithm, we need to extend the computational domain to \( -x, +x, -y, \) and \( +y \) direction with one additional cells, the electric fields in the center of these cells are denoted as \( E^s_{l,m,n} \) with \( m = 0 \) or \( M + 1, n = 1, \ldots, N, \) or \( m = 1, \ldots, M, n = 0 \) or \( N + 1 \). With these preparation, one can apply the difference operator on to the scattered fields as
\[
P_{l,m,n} = (\nabla^2 E^s_l)_{l,m,n} = \frac{E^s_{l,m+1,n} + E^s_{l,m-1,n} - 2E^s_{l,m,n}}{\Delta x^2} + \frac{E^s_{l,m,n+1} + E^s_{l,m,n-1} - 2E^s_{l,m,n}}{\Delta y^2}, \quad (4)
\]
for \( m \in [1, M] \) and \( n \in [1, N] \), where \( \Delta x \) and \( \Delta y \) are the dimension of the small cells along \( x \) and \( y \) directions.

With such a discretization, the integral path in Eq. (3) now can be chosen as the outer boundaries (referring to the domain \( D \)) of cells with indexes \( m = 1, m = M, n = 1, n = N \). For scattered fields in the integrand of the boundary condition Eq. (3), we use the fields on both sides of the boundary to approximate them, such as \( (E^s_{l,0,n} + E^s_{l,1,n})/2 \) with \( n \in [1, N] \) on one of the boundary. For the normal derivative of the field, we apply the finite difference rule, such as \( (E^s_{l,0,n} - E^s_{l,1,n}) / \Delta x \) with \( n \in [1, N] \) on one of the boundary. If we place the observation point of the Eq. (3) on the centers of the extended cells, meaning that the left hand side of the equation are fields on the centers of those extended cells, then we obtain the relation between the scattered fields on the inner side of the boundary and those on the outer side, which can be written as
\[
E^{\text{sc}}_l = \tilde{n} \cdot \tilde{E}^{\text{br}}_l + \tilde{E}^{\text{br}}_l = \begin{bmatrix} \tilde{I}_{2M/\Delta y} & 0 \\ 0 & \tilde{I}_{2M/\Delta x} \end{bmatrix} (E^{\text{sc}}_l - \tilde{E}^{\text{br}}_l), \quad (5)
\]
where \( \tilde{E}^{\text{sc}}_l \) is a vector storing the scattered fields on the extended cells, and \( \tilde{E}^{\text{br}}_l \) storing the scattered fields on the cells on the inner side of the boundary in the domain. The arrangement of these fields is \( \tilde{E}^{\text{br}}_l = [E^s_{l,1,1}, \ldots, E^s_{l,1,M}, E^s_{l,1,1}, \ldots, E^s_{l,1,N}, E^s_{l,2,1}, \ldots, E^s_{l,2,M}, E^s_{l,2,1}, \ldots, E^s_{l,2,N}]^T \), and the elements in \( \tilde{E}^{\text{sc}}_l \).
are arranged accordingly. \( \tilde{\eta} \) is the matrix operator with elements of the normal derivative of the Green’s function in Eq. (3), and \( \tilde{A} \) is with elements of the Green’s function itself. With the help of Eq. (5), the scattered fields on those extended cells can be represented by the scattered fields on the cells on the inner side of the boundary, meaning that the unknown field variables now only consist of those in the domain \( D \).

To efficiently record the scattered fields on the boundaries, we use another variable \( \tilde{E}^{sb}_{l} \) to store them so as to remove the 4 repeating elements in \( \tilde{E}^{sb}_{l} \), which can be obtained by \( \tilde{E}^{sb}_{l} = \tilde{T} \cdot \tilde{E}^{b}_{l} \). Besides, from Eq. (5) and (3), one can write the scattered fields on the measurement domain as

\[
\tilde{E}^{sm}_{l} = \tilde{G}_{S} \cdot \tilde{E}^{b}_{l}, \quad \text{with} \quad \tilde{G}_{S} = \left[ \frac{1}{2} \tilde{\eta}_{S} - \tilde{B}_{S} \right] \cdot \tilde{Q} + \left[ \frac{1}{2} \tilde{\eta}_{S} + \tilde{B}_{S} \right] \cdot \tilde{T}, \tag{6}
\]

where \( \tilde{\eta}_{S} \) is the same operator as \( \tilde{\eta} \) but with the observation points in the measurement domain instead of cells on the boundary (so does for \( \tilde{A}_{S} \)). \( \tilde{H} \tilde{S} = \tilde{A}_{S} \cdot \text{diag} \left[ \tilde{I}_{2M}/\Delta y, \tilde{I}_{2M}/\Delta x \right] \), and \( \tilde{Q} = \left( \tilde{I}_{2M+2N} - 0.5 \tilde{\eta} + \tilde{B} \right)^{-1} \cdot (0.5 \tilde{\eta} + \tilde{B}) \) with \( \tilde{B} = \tilde{A} \cdot \text{diag} \left[ \tilde{I}_{2M}/\Delta y, \tilde{I}_{2M}/\Delta x \right] \) and \( \tilde{I}_{2M+2N} \) the \( 2M+2N \) dimensional identity matrix.

Since the measurement of the scattered fields is in the far-field zone, the scattering operator \( \tilde{G}_{S} \) is compact. As what we have been doing in the integral equation based method in [2], we can determine part of the scattered fields on the boundary by using the singular value decomposition (SVD) of the scattering operator \( \tilde{G}_{S} \), and the remaining part of the scattered fields on the boundary are to be optimized within a field subspace. To do so, we have \( \tilde{G}_{S} = \sum_{j=1}^{\sigma_{L}} \tilde{u}_{j} \sigma_{j}^{1/2} \tilde{v}_{j}^{*} \) from the SVD (the superscript * denotes the complex conjugate transpose operation), and the deterministic part and the ambiguous part of the scattered field on the boundary can be expressed as

\[
\tilde{E}^{sbd}_{l} = \sum_{\sigma_{j} \geq \sigma_{L}} \tilde{E}^{b}_{l} \frac{\tilde{G}^*_j \cdot \tilde{E}^{sm}_{l}}{\sigma_{j}}, \quad \text{and} \quad \tilde{E}^{sba}_{l} = \left( \tilde{I}_{2(M+N)-4} - \sum_{\sigma_{j} \geq \sigma_{L}} \tilde{v}_{j} \tilde{v}_{j}^{*} \right) \cdot \tilde{a}_{l}, \tag{7}
\]

where \( \sigma_{L} \) is a chosen singular value that is large enough such that the subspaces with singular values larger than it are barely compromised by the noise in \( \tilde{E}^{sm}_{l} \) [2]. \( \tilde{I}_{2(M+N)-4} \) is a \((2M+N)-4\) dimensional identity matrix, and \( \tilde{a}_{l} \) is a \((2M+N)-4\) dimensional unknown vector, which is to be optimized through the inversion.

The DE-based inversion method

Without solving any forward problem, the new inversion method adopts the alternative optimization scheme as used in [1] and [2], where the scattered fields and the contrast function are considered as unknowns and are alternatively optimized.

First we denote that the scattered fields as a two-dimensional \( M \times N \) tensor for the \( l \)-th incidence as \( \tilde{E}^{s}_{l} \) with its elements \( \tilde{E}^{s}_{l,m,n} \), \( 1 \leq m \leq M \) and \( 1 \leq n \leq N \). So do \( \tilde{E}^{s}_{l} \) and \( \tilde{P}_{l} \). For each incidence, part of the scattered fields on the boundary cells are known as analyzed in the previous subsection, and the remaining cells are totally unknown, which we denote as \( \tilde{E}^{s}_{l} \). Then the scattered field tensor is a function of \( \tilde{E}^{s}_{l} \) and \( \tilde{a}_{l} \), i.e., \( \tilde{E}^{s}_{l} \left( \tilde{E}^{s}_{l}, \tilde{a}_{l} \right) \).

Given such a discretization and notations, we can construct the objective function for the inversion with the Helmholtz equation Eq. (2) as

\[
f \left( \tilde{E}^{s}_{1}, \ldots, \tilde{E}^{s}_{N_{inc}}, \tilde{a}_{1}, \ldots, \tilde{a}_{N_{inc}}, \tilde{\chi} \right) = \sum_{i=1}^{N_{inc}} \Delta_{\text{Helm}} \left[ \tilde{E}^{s}_{l} \left( \tilde{E}^{s}_{l}, \tilde{a}_{l} \right), \tilde{\chi} \right] / \left\| k_{0} \tilde{E}^{s}_{l} \right\|^{2} + \Delta_{\text{res}} \left( \tilde{a}_{l} \right) / \left\| \tilde{E}^{sm}_{l} \right\|^{2}, \tag{8}
\]

where \( \Delta_{\text{res}} \left( \tilde{a}_{l} \right) = \left\| \tilde{E}^{sm}_{l} - \tilde{G}_{S} \cdot \left[ \tilde{E}^{sba}_{l} \left( \tilde{a}_{l} \right) + \tilde{E}^{sbd}_{l} \right] \right\|^{2} \)

and

\[
\Delta_{\text{Helm}} \left[ \tilde{E}^{s}_{l} \left( \tilde{E}^{s}_{l}, \tilde{a}_{l} \right), \tilde{\chi} \right] = \left\| \tilde{P}_{l} \left( \tilde{E}^{s}_{l}, \tilde{a}_{l} \right) + k_{0}^{2} \tilde{E}^{s}_{l} \left( \tilde{E}^{s}_{l}, \tilde{a}_{l} \right) \right\| \left\| \tilde{E}^{s}_{l} \left( \tilde{E}^{s}_{l}, \tilde{a}_{l} \right) \right\|^{2}, \tag{9}
\]

where \( \otimes \) denotes the element-wise multiplication.
As in the integral equation based method \cite{2}, the scattered fields and the contrast function are alternatively optimized in every iteration of the inversion. That is to say, when optimizing the scattered fields ($E_i$ and $\bar{\alpha}_i$), the contrast function ($\bar{\chi}$) is considered to be known, and likewise for the contrast function. The update of the scattered fields uses the Polak-Ribiére conjugate gradient method \cite{1}, and the detail is omitted here for the sake of conciseness. Since we are using the DE to describe the field behaviors, the scope of such description is a local one, i.e., there is no long distance interaction as those by the Green’s function in the integral equation method. Therefore, when updating the scattered fields, we will increase the number of iterations of the CG algorithm in every inversion iteration, such that the fields can be updated in a large domain instead of the local updating. The number of the CG iteration should be at the level of the largest number of small cells in one direction.

Besides, we apply the FFT type twofold subspace-based regularization scheme \cite{2} into the method, and the key to do so is to filter out the high frequency components of the induced secondary sources, which is the right hand side of the Eq. \eqref{eq:2}. This is accomplished when calculating the residue of the DE before every CG iteration to update the scattered fields.

Similarly as given in the integral equation based inversion method \cite{2}, the contrast function $\bar{\chi}$ is updated by a least squares solution in every iteration of the optimization, which is

$$\chi_{m,n} = \frac{\sum_l \left[ -k_0^2 \left( E_{l,m,n} + E_{s,l,m,n} \right) \right]^* \left[ P_{l,m,n} + k_0^2 E_{s,l,m,n} \right]}{\sum_l \left| -k_0^2 \left( E_{l,m,n} + E_{s,l,m,n} \right) \right|^2}. \quad (10)$$

This means that in each iteration of the inversion, there are a number of iterations of CG method for the update of the scattered fields, and after the update of the scattered fields, there is a least squares type update for the contrast function.

4 Numerical tests

In this section, we will repeat the test given in \cite{2} with the well-known Austria profile. All the physical setup can be found in \cite{2}. All synthetic data are added with additive white Gaussian noise at level of 10\%. In the test, the domain is discretized into $60 \times 60$ meshes, and $L = 15$. The termination conditions are set to be $\delta^{2D} = 10^{-4}$ for the first round optimization with smaller $M_F$, and $\delta^{2D} = 10^{-5}$ for the second round optimization with

---

Figure 1. Reconstructed results for Austria type scatterers with $\epsilon_r = 2$. The first and the second column are the real parts and imaginary part, respectively, of the reconstructed results.

Figure 2. Reconstructed results for Austria type scatterers with $\epsilon_r = 2.2$. The first and the second column are the real parts and imaginary part, respectively, of the reconstructed results.
larger $M_F$ (the definition of this parameter can be found in [2]). The number of CG iterations is chosen 50 in every iteration of the inversion to update the fields.

We first use the standard Austria profile to verify the proposed inversion method. The inversion results show in Fig. 1, from which we see that the inversion successfully reconstruct the scatterer profile after 100 iterations with $M_F$ chosen as 5 and 10. Then we test the proposed inversion method with an Austria profile with relative permittivity $\epsilon_r = 2.2$. The nonlinearity of the problem for such a scatterer is larger compared to the one in the first case, and we are not able to reconstruct the correct profile by using the inversion method (integral equation based one) proposed in [2]. Using the proposed DE based method, the results after 184 iterations with $M_F = 5$ and 10 are shown in Fig. 2, where we find quite satisfactory reconstruction.

5 Conclusion

In this paper, we propose a new inversion method based on the differential equation. Using the alternative optimization scheme as in the integral equation based inversion method, there is no need to solve the forward problem in the inversion. The proposed inversion method is fully within the differential equation method frame, and therefore there is no numerical Green’s function involved. The FFT type twofold subspace-based regularization technique is applied into the inversion to stabilized the solver. Numerical tests verify the interest of the proposed method. Such a method provides another option to efficiently handle inverse medium scattering problems, especially when the background medium in the domain of interest is not a homogeneous one.

Acknowledgments

This work was supported by the Singapore Temasek Defence Systems Institute under grant TDSI/11-009/1A.

References

Minisymposium M10
Inverse Scattering
Imaging of anomalous components in unknown background

L Audibert\textsuperscript{1,2} A Girard\textsuperscript{1} and H Houssem\textsuperscript{2}

\textsuperscript{1}EDF R\&D, Departement STEP, 6 quai Watier, 78401, Chatou, France
\textsuperscript{2}INRIA, CMAP Ecole Polytechnique, Route de Saclay, 91128, Palaiseau, France

E-mail: lorenzo.audibert@edf.fr

Abstract: We introduce a qualitative method capable of imaging defects in an unknown complex environment using differential measurements. The main difficulty is that the background medium is unknown and too complex to obtain a reliable estimate of the associated Green function. To overcome this difficulty our approach exploits two measurements of the farfield operators, one without defects and one with defects. The analysis of our method relies on the recently introduced Generalized Linear Sampling Methods (GLSM) and the link to the solutions of the interior transmission problems. We give numerical examples related to non destructive testing in concrete-like materials, illustrating the performance of our method.

Keywords: Inverse scattering, Sampling methods, differential measurement

1 Introduction

We are interested in the imaging of defects inside unknown heterogeneous medium from multistatic measurements of waves at a fixed frequency. The main difficulty of our setting is that the background medium is unknown and complex, which means that one cannot obtain a good estimate for the background Green function. This discards the possibility of using classical (qualitative) imaging methods [2]. We assume that two sets of measurements, one for the defect free and one for the defect containing medium, are at our disposal and we design a numerical inversion algorithm that exploits those supplementary data to visualize the defect location. The analysis of our method relies on the GLSM and the link it has with the solution to the interior transmission problems (see [4] and [5]). Ones ends up with an indicator function that combines the indicator functions from the GLSM and a filtered difference term computed without additional significant numerical cost. We shall briefly introduce the scattering problem by a heterogeneous medium and the needed notations in Section 2. In Section 3 we review the main theoretical results of [5]. We give in Section 4 some numerical experiments on concrete like materials, illustrating the performance of our method.

2 Scattering by an inhomogeneous medium

We restrict ourselves to the case of scalar time harmonic waves and we focus on full aperture farfield measurements associated to incident plane waves. For a wave number $k > 0$, the total field solves the Helmholtz equation:

$$\Delta u + k^2 n u = 0 \text{ in } \mathbb{R}^d$$

for $d = 2$ or $3$ and $n$ the refractive index, where $\Im(n) \geq 0$. We denote by $\bar{D}$ the support of $n - 1$ and assume that $D$ is a bounded domain with Lipschitz boundary and connected complement. We are interested by the
where the operator $T$ all

We denote by $n$ and there exist a constant $\gamma$.

It is well known that the far-field operator admits two factorisations $F = GH = H^*TH$. The compact operator $H : L^2(S^{d-1}) \to L^2(D)$ is defined by:

$$Hg := \int_{S^{d-1}} e^{ikx \cdot \theta} g(\theta) d\sigma(\theta), \quad g \in L^2(S^{d-1}), \quad x \in D,$$

and is dense in $\{ v \in L^2(D) : \Delta v + k^2 v = 0 \text{ on } D \}$. Its adjoint $H^* : L^2(D) \to L^2(S^{d-1})$ is defined by:

$$H^* \varphi(x) := \int_D e^{-iky \cdot \hat{x}} \varphi(y) dy, \quad \varphi \in L^2(D), \quad \hat{x} \in S^{d-1}.$$

We define the compact operator $G : \mathcal{R}(H) \subset L^2(D) \to L^2(S^{d-1})$ defined by: $Gv := w^\infty$ where $w^\infty$ is the farfield of $w \in H^1_{\text{loc}}(\mathbb{R}^d)$ associated to the incident wave $v$:

$$\begin{cases} 
\Delta w + nk^2 w = k^2(1-n)v \text{ in } \mathbb{R}^d, \\
\lim_{r \to \infty} \int_{|x| = r} \left| \frac{\partial w}{\partial r} - ikw \right|^2 ds = 0.
\end{cases}$$

Finally we define $T : L^2(D) \to L^2(D)$ by:

$$Tv := -k^2(1-n)(v + w).$$

In the following we will use the operator $F_\# = |\Re(F)| + |\Im(F)|$, which can be factorised as $F_\# = H^*T_\#H$, where the operator $T_\#$ is a real selfadjoint operator [1] that satisfies (under hypothesis 2.1):

$$\|T_\# h, h\| = \left\|(T_\#)^{1/2} h\right\|^2 \geq \mu \|h\|^2 \quad \forall h \in \mathcal{R}(H),$$

where $\mu > 0$ is a constant independent of $h$.

**Hypothesis 2.1.** The index of refraction $n$ and the domain $D$ satisfy $n \in L^\infty(\mathbb{R}^d)$, supp$(n - 1) = \overline{D}$, $\Im(n) \geq 0$ and there exist a constant $n_\ast > 0$ such that $\Re(n(x) - 1) \geq n_\ast$ for a.e. $x \in D$.

## 3 Theoretical Results

The GLSM relies on the solvability of the so-called interior transmission problem defined by $(u, v) \in L^2(D) \times L^2(D)$ such that $u - v \in H^2(D)$ and

$$\text{ITP}(D, f, g, n, n') = \begin{cases} 
\Delta u + k^2 nu = 0 & \text{in } D, \\
\Delta v + k^2 n' v = 0 & \text{in } D, \\
(u - v) = f & \text{on } \partial D, \\
\frac{\partial}{\partial n}(u - v) = g & \text{on } \partial D,
\end{cases}$$

We denote by $\sigma(D, n, n')$ the set of wave number $k \in \mathbb{R}$ for which the ITP$(D, f, g, n, n')$ is not well posed for all $f \in H^2(\partial D)$ and $g \in H^{-2}(\partial D)$. 

---

[Creative Commons Attribution 3.0 licence](http://creativecommons.org/licenses/by/3.0/). Any further distribution of this work must maintain attribution to the author(s) and the title of the work, book title and publisher. Published under licence by IOP Publishing Ltd.
Hypothesis 3.1. We assumed that $k^2 \in \mathbb{R}_+$ and $n$ are such that for all $f \in H^\frac{1}{2}(\partial D)$ and $g \in H^{-\frac{1}{2}}(\partial D)$ the ITP($D,f,g,\nu$) has a unique solution $(u,v) \in L^2(D) \times L^2(D)$ such that $u - v \in H^2(D)$.

This hypothesis is verified [6] for all $k^2 \in \mathbb{R}_+$ except a countable set without finite point of accumulation if $n$ verifies $1/(n-1) \in L^\infty(D)$ and $\Re(n-1)$ is either positive or negative definite in the neighbourhood of $\partial D$.

We introduce the farfield pattern of the Green function:

$$\phi_z(\hat{x}) := e^{-ik\hat{x} \cdot z}$$

and the key ingredient of the GLSM:

Theorem 3.1. Assume that $k \notin \sigma(D,n,1)$. Then $G$ is compact, injective with dense range and $\phi_z \in \mathcal{R}(G)$ if and only if $z \in D$. Moreover, if $z \in D$ then $G(v) = \phi_z$ if and only if there exists $u \in L^2(D)$ such that $(u,v)$ is a solution of ITP($D,\Phi_z,\frac{\partial \Phi_z}{\partial n},n,1$).

We outline the main results of the GLSM in the case of noisy data (see [5]). The noisy operators, $F^\delta$ and $F^\delta_f$ are such that $\|F^\delta - F\| \leq c\delta$ and $\|F^\delta_f - F^\delta\| \leq \delta$ where $c$ is a real constant. Let $g_{z}^{\alpha,\delta} \in L^2(\mathbb{S}^{d-1})$ be the minimizer of

$$J^\delta(\phi_z;g) := \alpha(\|F^\delta g^{-\eta}\|g\|^2) + \|F^\delta g - \phi_z\|^2,$$

(7)

for $\alpha > 0$, $\delta > 0$, $\eta \in [0,1]$ and $\phi_z \in L^2(\mathbb{S}^{d-1})$. The functional

$$A^{\alpha,\delta}(g) := |(F^\delta g, g)| + \alpha^{-\eta}\|g\|^2$$

(8)

gives a characterisation of $D$ through the following result.

Theorem 3.2. Under hypothesis 3.1 and 2.1 we have:

- $z \in D$ implies $\lim_{\alpha \to 0} \limsup_{\delta \to 0} \sup A^{\alpha,\delta}(g_{z}^{\alpha,\delta}) < \infty$,
- $z \notin D$ implies $\liminf_{\alpha \to 0} \liminf_{\delta \to 0} A^{\alpha,\delta}(g_{z}^{\alpha,\delta}) = \infty$.

When we have two measurements campaigns, the same results applies to $\tilde{D}_0 = \text{supp}(n_0 - 1) \subset \tilde{D}$ where $A^{\alpha,\delta}_0$ is defined as above using $F^\delta_0$ (the farfield associated with $n_0$ and $D_0$) and

$$g_{0,z}^{\alpha,\delta} = \operatorname{arg} \min_{g \in L^2(\mathbb{S}^{d-1})} \alpha(\|F^\delta_0 g^{-\eta}\|g\|^2) + \|F^\delta_0 g - \phi_z\|^2.$$  

However we are interested in $\text{supp}(n - n_0)$. The filtered difference term defined by:

$$D^\delta(g,g_0) := |(F^\delta_0 g^{-\eta}(g - g_0), g - g_0)| + \delta \|g - g_0\|^2,$$

will be used to image the simply connected part of $D_0$ that have been modified between the two measurements. We denote this domain by $\tilde{D}_0$. Let $\Omega$ be the part of $\text{supp}(n - n_0)$ such that $\Omega \cap \tilde{D}_0 = \emptyset$. We introduce the following indicator function:

$$T^{\alpha,\delta}(g_{z}^{\alpha,\delta}, g_{0,z}^{\alpha,\delta}) = \frac{1}{\sqrt{A^{\alpha,\delta}(g_{z}^{\alpha,\delta}) \left(1 + A^{\alpha,\delta}(g_{z}^{\alpha,\delta}) D^{\delta}(g_{z}^{\alpha,\delta}, g_{0,z}^{\alpha,\delta})\right)^{-1}}},$$

(9)

which will image a domain larger than the defect as follows:

Theorem 3.3. If we assume that $k \notin \sigma(D,n,1) \cup \sigma(D_0,n_0,1) \cup \sigma(D,n,n_0)$ and, $n$ and $n_0$ verify hypothesis 2.1, then $z \in \tilde{D}_0 \cup \Omega$ if and only if $\lim_{\alpha \to 0} \liminf_{\delta \to 0} T^{\alpha,\delta}(g_{z}^{\alpha,\delta}, g_{0,z}^{\alpha,\delta}) > 0$. 

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, book title and publisher. Published under licence by IOP Publishing Ltd.
4 Numerical Simulations

In order to fix the ideas, we shall limit ourselves to the two dimensional case and will introduce the algorithms for the discrete setting. We identify $S^1$ with the interval $[0, 2\pi]$. In order to collect the data of the inverse problem we solve numerically (3) for $N$ incident fields $u^j(\frac{2\pi j}{N}, \cdot)$, $j \in \{0, \ldots, N-1\}$ using a finite element method implemented with Freefem++ [8]. The discrete version of $F$ is then the matrix $F := (u^j(\frac{2\pi j}{N}, \cdot))_{0 \leq j, k \leq N-1}$. We add some noise to the data to build a noisy far field matrix $F^\delta$ where $(F^\delta)_{j,k} = (F)_{j,k}(1 + \sigma N_{ij})$ for $\sigma > 0$ and $N_{ij}$ an uniform complex random variable in $[-1,1]^2$. We similarly construct $F^\delta_0$, the vector defined by $\phi_z(j) = \exp(-ik(z_1 \cos(\frac{2\pi j}{N}) + z_2 \sin(\frac{2\pi j}{N})))$ for $0 \leq j \leq N - 1$. The analysis of previous sections suggests to consider

$$g_z^{GLSM} := \arg\min_{g \in L^2(S^1)} \left( \alpha \left\| (F^\delta_z)^\frac{1}{2} g \right\|^2_{L^2(S^1)} + \alpha^{1-\eta} \left\| g \right\|^2_{L^2(S^1)} + \left\| F^\delta g - \phi_z \right\|^2_{L^2(S^1)} \right).$$

The minimizer is explicitly given by

$$g_z^{GLSM} = (\alpha F^\delta_z + \alpha^{1-\eta} \delta Id + F^\delta F^\delta)^{-1} F^\delta \phi_z.$$

We similarly construct $g_{0,z}^{GLSM}$ using $F^\delta_0$. In our numerical simulations we choose $\eta = 0$ (which corresponds to the one used in [4]) and set $\alpha$ with the same heuristic as in [4]. We then look at $z \mapsto \mathcal{I}_T^0(g_z^{\delta}, g_{0,z}^{\delta})$ as an indicator function.

All our experiments are conducted for the background medium $n_0$ shown in Figure 1. This background medium is a simplified numerical description of a concrete like material. The wave frequency is $150kH\tilde{z}$, the celerity of the medium is $4300m.s^{-1}$ (which means a wavelength of $2.87cm$) and the celerity inside the inclusion is $5700m.s^{-1}$.

![Figure 1. The background medium, $n = 0.57$ inside the yellow inclusions](image)

Our theoretical analysis is only valid for inhomogeneous perturbations $n$ and $n_0$. One example of this setting is shown in Figure 2 where we modified the celerity (3 times higher) in two of the inclusions between the two measurements.

The main concern with concrete non destructive testing is the case of cracks. Our analysis do not include this case and its extension to it is the subject of an ongoing work. However the results shown in Figure 3 for a crack being either inside or outside the inhomogeneity $n_0$ gives promising results. To obtain this results we do not test with $\phi_z$ but with its normal derivatives as explained in [7].

References

Figure 2. From left to right and up to down: The index $n$, $I_T^{\alpha,\delta}(g_z^{\alpha,\delta}, g_{0,z}^{\alpha,\delta})$, $A_0^{\alpha,\delta}(g_{0,z}^{\alpha,\delta})^{-1}$ and $A^{\alpha,\delta}(g_z^{\alpha,\delta})^{-1}$

Figure 3. Cracks: On the left the medium with the cracks and on the right the corresponding $I_T^{\alpha,\delta}(g_z^{\alpha,\delta}, g_{0,z}^{\alpha,\delta})$


Minisymposium M11

Inverse Problems for Wave Phenomena
Optimization approach to combined inverse tsunami problem

S Kabanikhin\textsuperscript{1,2} and O Krivorotko\textsuperscript{2}

\textsuperscript{1}Institute of Computational Mathematics and Mathematical Geophysics SB RAS, Russia
\textsuperscript{2}Novosibirsk State University, Russia

E-mail: krivorotko.olya@mail.ru

Abstract: We investigate two different inverse problems of determining the tsunami source using two different additional data, namely underwater measurements and satellite wave-form images, and combination of these two inverse problems. We investigate gradient-type methods for inverse problem solutions and show that combination of two types of data allows one to increase stability and convergence of numerical inverse problem solutions. Results of numerical experiments of the tsunami source reconstruction are presented and discussed.

Keywords: Shallow water equations, Dirichlet problem, Satellite data, DART data, Conjugate gradient method, Combined data.

1 Introduction

Most suitable physical models related to simulation of tsunamis are based on shallow water equations (1), [1, 2]. There exist many numerical approaches for solving shallow water equations such as finite-difference, finite volume, finite element, etc [3]. An overview of methodologies and techniques related to estimation of tsunami source characteristics are given in [4, 5, 6, 7]. The most of them consists in determining the tsunami source using additional measurements of a passing wave (this problem is often called inverse tsunami problem) such as DART (Deep-ocean Assessment and Reporting of Tsunamis) buoys positioned on the ocean floor, tide gauges measurements, satellite wave-form images, etc. Our goal is to reconstruct the tsunami source using a combination of two types of data: DART buoys and satellite wave-form image. We show that using a combination of two types of data allows one to increase the stability and efficiency of tsunami source reconstruction.

The paper is organized as follows. In Section 2 we describe the statement of three inverse problems. In Section 3 we consider a variational formulation of the inverse tsunami problem for two types of measured data: DART data (inverse problem 1) and satellite image data measured on the part of the water surface (inverse problem 2), and then we consider the combined inverse problem (inverse problem 3). We compare two inverse problems and their combination and show the benefits of usage of combined data. Results of numerical experiments of the tsunami source reconstruction are presented in Section 4.

2 Statement of the problems

The ocean domain being considered is bounded from above by the free water surface $\eta(x, y, t)$, and from below, by the bottom relief $H(x, y) > 0$. We assume that the computational time $T$ is not large enough for the wave to reach the edges of the domain, and therefore we can set homogeneous boundary conditions at the boundary of
the domain \( \Omega := (0, L_x) \times (0, L_y) \) (figure 1). We formulate the initial boundary-value problem in the Cartesian coordinate system

\[
\begin{align*}
\eta_{tt} &= \text{div}(g H(x,y) \text{grad} \eta), & t \in (0, T); \\
\eta|_{t=0} &= q(x,y), & (x,y) \in \Omega; \\
\eta|_{\partial \Omega_T} &= 0,
\end{align*}
\]

(1)

for the linear equations of shallow water theory in terms of the free surface without external forces, e.g. the Coriolis force and bottom friction [2]. Here \( H \in H^1(\Omega) \) is a known function describing the bottom relief (bathymetry), \( q \in H^2(\Omega) \) is a tsunami source which is supposed to have a compact support belonging to \( \Omega \), \( g = 9.8 \text{ [m/s}^2] \). Further, we will use notation \( c(x,y) = \sqrt{g H(x,y)} \) that describes the tsunami propagation velocity according to the long-wave theory.

![Figure 1. Domain of calculation of direct and inverse problems.](image1)

**Figure 1.** Domain of calculation of direct and inverse problems.

**The direct tsunami problem** (1) consists in determining of a function \( \eta \in C(\Omega_T; H^2(\Omega)) \) in the domain \( \Omega \) by known functions \( H(x,y) \) and \( q(x,y) \).

Let us consider three inverse problems for linear shallow water equations:

**Inverse problem 1 (IP 1):** find \( q(x,y) \) from (1) using function \( H(x,y) \) and data \( f_m^n(x,y,t) \) from underwater systems (DART buoys, tide gauges measurements) at \( \varepsilon \)-neighborhoods of points \( (x_m,y_m) \in \Omega \)

\[
\begin{align*}
\eta(x,y,t) &= f_m^n(x,y,t), & x \in (x_m-\varepsilon, x_m+\varepsilon), & y \in (y_m-\varepsilon, y_m+\varepsilon), & \varepsilon > 0, \\
& t \in (T_m^{(1)}, T_m^{(2)}), & m = 1, 2, \ldots, M.
\end{align*}
\]

(2)

**Inverse problem 2 (IP 2):** find \( q(x,y) \) from (1) using function \( H(x,y) \) and satellite altimeters data \( F_2(x,y) \)

\[
\eta(x,y,T) = F_2(x,y), \quad (x,y) \in \omega \subset \Omega, \quad T > 0.
\]

(3)

Here \( \omega := (l_1^{(1)}, l_2^{(1)}) \times (l_1^{(2)}, l_2^{(2)}) \) is a subset of \( \Omega \).

**Inverse problem 3 (combined IP 3):** find \( q(x,y) \) from (1)-(3) using function \( H(x,y) \), measured data \( f_m^n(x,y,t), m = 1, \ldots, M, \) and \( F_2(x,y) \).

Let us present inverse problems 1, 2 and 3 in the operator form: \( A_1 q = F_1, \) \( i = 1, 2, 3 \). Here \( A_1 : H^2(\Omega) \rightarrow C(\Omega_T; E^M), \) \( A_2 : H^2(\Omega) \rightarrow L_2(\Omega) \) and \( A_3 := (A_1, A_2)^T, \) \( F_1 := (f_1^1, f_2^1, \ldots, f_M^1), \) \( E^{M} \) is the vector of discrete output data depends on \( (x,y,t), F_3 = (F_1, F_2)^T, \) \( E^{M} \) is Euclidean space. The inverse problem 3 is ill-posed because \( A_3 \) is a compact operator [8]. The compactness of operators \( A_1 \) and \( A_2 \) is established in papers [7] and [9, 10, 11], respectively. We will find the solution \( q(x,y) \) of inverse problems in the class of functions \( q(x,y) = \sum_{k=1}^K q_k(x) \sin(2\pi ky/L_y) \) which means that we regularize our inverse problems using cut Fourier series [8].

![Figure 2. The 1D bottom relief H(x).](image2)

**Figure 2.** The 1D bottom relief \( H(x) \).
3 Variational formulation of inverse problems

Inverse problems $A_i q = F_i$ can be reduced to the minimization problems $\min_{q \in H^2(\Omega)} J_i(q)$, $i = 1, 2$ [8]. Here $J_i(q) = \|A_i q - F_i\|^2$ are cost functions, $i = 1, 2$.

In this section we find gradients of cost functions $J_i(q)$, $i = 1, 2$, and introduce a cost function $J_3^{(\beta)}(q)$ for the combined IP 3.

Inverse problem 1

The conditions of well-posedness of IP 1 in one-dimensional case are given in [12]. The algorithm of constructing function $q(x, y)$ in two-dimensional case based on truncated singular value decomposition is proposed in [4, 5, 13].

The cost function $J_1(q)$ for IP 1 has the form:

$$J_1(q) = \|A_1 q - F_1\|^2_{L^2(0,T)} := \sum_{m=1}^M \int_{T_m^{(2)}} \int_{T_m^{(1)}} \int \left[ \eta(x, y, t; q) - f_m(x, y, t) \right]^2 dy \, dx \, dt.$$

**Lemma 1** [7]. The gradient of the cost function $J_1(q)$ has the form $J_1'q = \psi_1(x, y, 0)$. Here $\psi_1 \in C(\Omega_T; H^2(\Omega))$ is the weak solution of the following problem:

\[\begin{align*}
\psi_{1t} - \text{div}(c^2(x, y) \text{grad} \psi_1) &= R_1(x, y, t), & (x, y) \in \Omega, t \in (0, T), \\
\psi_1(x, y, T) &= 0, & (x, y) \in \Omega, \\
\psi_1|_{\partial\Omega_T} &= 0, & t \in (0, T),
\end{align*}\]

where

$$R_1(x, y, t) = -2 \sum_{m=1}^M \{ [(\eta(x, y, t) - f_m(x, y, t)) \theta(x - x_m + \varepsilon) \theta(y - y_m + \varepsilon - x) \cdot \theta(y - y_m + \varepsilon - y) \theta(t - T_m^{(2)})] - \theta(T_m^{(2)} - t) \}.$$

Inverse problem 2

The cost function $J_2(q)$ for IP 2 has the form:

$$J_2(q) = \|A_2 q - F_2\|^2_{L^2(0,T)} := \int_{T_2^{(1)}} \int_{T_2^{(2)}} (\eta(x, y, T) - F_2(x, y))^2 dy \, dx.$$

**Lemma 2** [6, 14]. The gradient of the cost function $J_2(q)$ has the form $J_2'q = \psi_2(x, y, 0)$. Here $\psi_2 \in H^2(\Omega)$ is the weak solution of the following problem:

\[\begin{align*}
\psi_{2t} = \text{div}(c^2(x, y) \text{grad} \psi_2), & (x, y) \in \Omega, t \in (0, T); \\
\psi_2(x, y, T) &= 0, & (x, y) \in \Omega; \\
\psi_2|_{\partial\Omega_T} &= 0, & t \in (0, T)
\end{align*}\]

where

$$R_2(x, y) = 2 (\eta(x, y, T) - F_2(x, y)) \theta(x - t_2^{(1)}) \theta(t_2^{(2)} - x) \cdot \theta(y - y_3^{(1)}) \theta(t_3^{(2)} - y).$$

Inverse problem 3

We introduce the cost function $J_3^{(\beta)}(q)$ for IP 3 in the form: $J_3^{(\beta)}(q) = \beta J_1(q) + (1 - \beta)J_2(q)$, $\beta \in [0, 1]$. The gradient of a cost function $J_3^{(\beta)}(q)$ has the form: $J_3^{(\beta)'}q = \beta J_1'q + (1 - \beta)J_2'q$. 
4 Results of numerical calculations

We apply the conjugate gradient method [15, 8] for solving IP 1, IP 2 and IP 3 numerically.

We choose the following parameters for numerical experiments: $L_x = 50$ km, $L_y = 100$ km, $T = 60$ min, $\varepsilon = 125$ m, $N_x = 500$, $N_y = 500$, $N_t = 600$. The bottom is assumed to be one-dimensional (see figure 2) with the highest $H_{\text{max}} = 6$ km and lowest $H_{\text{min}} = 5$ m average depth of the ocean. We choose an exact solution $q_e(x, y)$ of inverse problems with a wave height $A = 8$ m (see figure 3). We use the explicit finite-difference conservative scheme of the second order approximation [7] with Courant condition $h_t = 0.8 \cdot h_x h_y (h_x^2 + h_y^2)^{-1/2} / \|c\|_C$. We set data $f_m$, $m = 1, 6$, and $F_2$ with "white" noise 1-7%, i.e. $f_m(x, y, t) = f_m(x, y, t) + \gamma \text{Random}(f_m) ||f_m||$, $\gamma \in (0.01, 0.07)$. Noise data for IP 1 is generated from the discrete numerical solution of the direct problem in six points $(x_m, y_m)$ equally-spaced on the interval $(40, 15); (47, 89))$. We choose an initial approximation $q_0 = H_{\text{max}}$ which corresponds to an unperturbed sea surface.

We use the stopping condition $J_i(q_n) < \varepsilon_s$, $i = 1, 2$, where $\varepsilon_s > 0$ is choosing according to [7]. Let us denote $q_n(i)$, $i = 1, 2$, is $n$-th approximation of the solution of IP 1 and 2. The reconstructed solution $q_n^{(1)}$ of IP 1 from the random noisy output data $\gamma = 3\%$ is demonstrated on figure 4.

For solving numerically IP 2 we put $\omega = (0, 25) \times (0, 50)$ km. The reconstructed solution $q_n^{(2)}$ of IP 2 from the random noisy output data $\gamma = 3\%$ is demonstrated on figure 5.

The reconstructed solution $q_n^{(3)}(x, y)$ of IP 3 from the random noisy output data with $\gamma = 3\%$, $\beta = 0.3$, is demonstrated on figure 6. Note, that the location of initial source as well as its amplitude is reconstructed better than in case of IP 1 and IP 2. The parameter $\beta$ in combined function $J_3(q)$ depends on sensitivity of the functional $J_1(q)$ and $J_2(q)$ (figure 7).

We compare relative accuracy error curves $E_i(n; q_n^{(i)} ; \gamma) = \|q_e - q_n^{(i)}\| / \|q_e\|$, $i = 1, 2$, for IP 1, 2 and $E_3(n; q_n^{(3)} ; \gamma)$ for IP 3. Figure 7 shows that using of combined underwater systems and satellite data allows...
one to increase the stability and efficiency of tsunami source reconstruction.

Note, that after reconstruction $q(x, y)$ we can calculate the amplitude of the tsunami wave using Airy-Green formula in case of 1D bottom profile (figure 2) [16]. In case of 2D bottom profile and linear source $q(x, y) = g(y)\delta(x)$ we can solve 2D direct problem for the amplitude $S(z, y)$

$$\begin{cases} 
    S_z + 0.5a_1S_y + 0.5a_2S = 0, & z > 0, y \in (-\infty, +\infty); \\
    S(0, y) = g(y), & y \in (-\infty, +\infty)
\end{cases}$$

which coincides with Airy-Green formula in 1D case [16]. Here new variable $z = \tau(x, y)$ denotes the solution of eikonal equation $\tau_x^2 + \tau_y^2 = (gH(x, y))^{-1}$, $a_1$ and $a_2$ depend on $H(x, y)$ [16].

Acknowledgments

This work is supported by the Ministry of Education and Science of the Russian Federation, by the Russian Foundation for Basic Research (projects No. 12-01-00773 and No. 14-01-31182).

References


---

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, book title and publisher. Published under licence by IOP Publishing Ltd.


Inverse force problems for the wave equation

S O Hussein and D Lesnic

University of Leeds, UK

E-mail: ml10soh@leeds.ac.uk, D.Lesnic@leeds.ac.uk

Abstract: We consider inverse force problems for the wave equation which consists of determining the unknown space or time-dependent force function from additional data. In the spacewise dependent case the additional data is Cauchy boundary data, whilst in the time-dependent case the additional data is represented by a time-dependent measurement of an integral space average of the displacement solution. The problems are linear, but ill-posed. The solution may exist and is unique, but it does not depend continuously on the input measurement data which is subject to noise. Numerically, the finite difference method combined with the Tikhonov regularization are employed in order to obtain a stable solution.

Keywords: Inverse force problem; Regularization; L-curve; Finite difference method; Wave equation.

1 Introduction

The wave equation governs many physical problems such as the vibrations of a spring or membrane, acoustic scattering, etc. It is the objective of this study to investigate inverse force problems for the hyperbolic wave equation. The forcing function is assumed to depend only upon the space or time variable in order to ensure uniqueness of the solution, [1, 2, 4, 6, 7]. These authors have given conditions to be satisfied by the data in order to ensure uniqueness and continuous dependence upon the data. However, no numerical results were presented and it is the main purpose of our study to develop an efficient numerical solution for this inverse linear, but ill-posed problem.

The mathematical formulation is given in Section 2. Numerical results are illustrated and discussed in Sections 3 and conclusions are provided in Section 4.

2 Mathematical Formulation

The governing equation for a vibrating bounded structure $\Omega \subset \mathbb{R}^n$, $n = 1, 2, 3$, acted upon by a force $F(\vec{x}, t)$ is given by the wave equation

$$u_{tt}(\vec{x}, t) = \nabla^2 u(\vec{x}, t) + F(\vec{x}, t), \quad (\vec{x}, t) \in \Omega \times (0, T),$$

where $T > 0$ is a given time and $u(\vec{x}, t)$ represents the displacement.

Equation (1) has to be solved subject to the initial conditions

$$u(\vec{x}, 0) = u_0(\vec{x}), \quad u_t(\vec{x}, 0) = v_0(\vec{x}), \quad \vec{x} \in \Omega,$$

where $u_0$ and $v_0$ represent the initial displacement and velocity, respectively. Let us consider, for the sake of simplicity, Dirichlet boundary conditions being prescribed, namely,

$$u(\vec{x}, t) = P(\vec{x}, t), \quad (\vec{x}, t) \in \partial \Omega \times (0, T).$$

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, book title and publisher. Published under licence by IOP Publishing Ltd.
If the force $F(x,t)$ is given, then equations above form a direct well-posed problem. However, if the force function $F(x,t)$ cannot be directly observed it hence becomes unknown and then clearly, the above set of equations is not sufficient to determine the pair solution $(u(x,t),F(x,t))$. Then, we consider the additional overdetermination measurement, given by the

$$\frac{\partial u}{\partial \nu}(x,t) = g(x,t), \quad (x,t) \in \Omega \times (0,T),$$

(4)

where $\nu$ is the outward unit normal to $\partial \Omega$, $(\Gamma \subset \partial \Omega)$, or by the integral

$$\int_{\Omega} u(x,t) \omega(x) d\mathcal{L}_x = \psi(t), \quad t \in [0,T],$$

(5)

where $\omega$ is a given weight function. In order to ensure a unique solution we further assume that

$$F(x,t) = f(x)h(x,t),$$

(6)

or

$$F(x,t) = g(t)h(x,t),$$

(7)

where $h(x,t)$ is a known function and $f(x)$ or $g(t)$ represents the unknown space- or time-dependent forcing function to be determined.

For the space-dependent force identification we have the following uniqueness theorems, see [2, Theorem 9], [4, Theorem 3.8] and [7].

**Theorem 1.** Assume that $\Omega \subset \mathbb{R}^n$ is a bounded star-shaped domain with sufficiently smooth boundary such that $T > \text{diam}(\Omega)$. Let $h \in H^2(0,T;L^\infty(\Omega))$ be such that $h(.,0) \in L^\infty(\Omega)$, $h(.,0) \in L^\infty(\Omega)$ and

$$H := \frac{\|h\|_{L^2(0,T;L^\infty(\Omega))}}{\inf_{x \in \Omega} |h(x,0)|} \quad \text{insufficiently small.}$$

(8)

If $\Gamma = \partial \Omega$, then the inverse problem (1)-(4) and (6) has at most one solution $(u(x,t),f(x))$ in the class of functions

$$u \in L^2(0,T;H^1(\Omega)), \quad u_t \in L^2(0,T;L^2(\Omega)), \quad u_{tt} \in L^2(0,T;(H^1(\Omega)'), \quad f \in L^2(\Omega),$$

(9)

where $(H^1(\Omega)')$ denotes the dual of $H^1(\Omega)$.

**Theorem 2.** Assume that $\Omega \subset \mathbb{R}^n$ is a bounded domain with piecewise smooth boundary. Let $h \in C^3(\overline{\Omega} \times [0,T])$ be such that

$$h(x,0) \neq 0 \quad \text{for} \quad x \in \overline{\Omega}.$$  

(10)

If $\Gamma = \partial \Omega$, then the inverse problem (1)-(4) and (6) has at most one solution $(u(x,t),f(x)) \in C^3(\overline{\Omega} \times [0,T]) \times C(\overline{\Omega})$.

**Theorem 3.** Assume that $\Omega \subset \mathbb{R}^n$ is a bounded star-shaped domain with smooth boundary such that $T > \text{diam}(\Omega)$. Let $h \in C^1[0,T]$ be independent of $x$ in (6) such that equation (1) becomes

$$u_{tt}(x,t) = \nabla^2 u(x,t) + f(x)h(t), \quad (x,t) \in \Omega \times (0,T),$$

(11)

and assume further that $h(0) \neq 0$. Then the inverse problem (2)-(4) and (11) has at most one solution in the class of functions

$$u \in C^1([0,T];H^1(\Omega)) \cap C^2([0,T];L^2(\Omega)), \quad f \in L^2(\Omega).$$

(12)
For the time-dependent force identification we assume, without loss of generality, that the Dirichlet boundary data in (3) is homogeneous, i.e. \( P = 0 \), and that the input data satisfy the regularity conditions
\[
u_0 \in W^1_2(\Omega), \quad v_0 \in L^2(\Omega), \quad \psi \in C^2[0, T], \quad h \in C([0, T]; L^2(\Omega)), \quad \omega \in W^1_2(\Omega),
\]
the compatibility conditions
\[
\int_{\Omega} u_0(\mathbf{x}) \omega(\mathbf{x}) d\mathbf{x} = \psi(0), \quad \int_{\Omega} v_0(\mathbf{x}) \omega(\mathbf{x}) d\mathbf{x} = \psi'(0),
\]
and the identifiably condition
\[
\int_{\Omega} h(\mathbf{x}, t) \omega(\mathbf{x}) d\mathbf{x} \neq 0, \quad t \in [0, T].
\]
Then from the corollary 9.2.1 of [6], we obtain the following unique solvability theorem of the inverse problem (1)-(3), (5) and (7).

**Theorem 4.** Let \( P = 0 \) in (3). If the input data satisfy conditions (13)-(15), then there exists a unique solution \((u(x, t), g(t))\) of the inverse problem (1)-(3), (5) and (7) in the class of functions
\[
u \in C([0, T]; W^1_2(\Omega)) \cap C^1([0, T]; L^2(\Omega)), \quad g \in C[0, T].
\]

### 3 Numerical Results and Discussion

For simplicity, we illustrate numerical results only for the inverse space-dependent force problem (2)-(4) and (11). Take the one-dimensional case, \( \Omega = (0, L) \) with \( L = 1, T = 1, h(x, t) = 1 + t \), and consider first the direct problem (2), (3) and (11) with the input data
\[
u(x, 0) = u_0(x) = 0, \quad u_t(x, 0) = v_0(x) = 0, \quad x \in [0, 1],
\]
\[
u(0, t) = p_0(t) = 0, \quad u(1, t) = p_L(t) = 0, \quad t \in (0, 1],
\]
\[
f(x) = \begin{cases} x & \text{if } 0 \leq x \leq \frac{1}{2}, \\ 1 - x & \text{if } \frac{1}{2} < x \leq 1. \end{cases}
\]

The numerical solutions for the flux tension \( q_0(t) = -\partial u/\partial x(0, t) \) at \( x = 0 \) obtained using the finite-difference method (FDM) for the direct problem with various \( N = M \in \{10, 20, 40, 80\} \), where \( M \) and \( N \) represent the number of space and time intervals in which the intervals \([0, L = 1]\) and \([0, T = 1]\) are uniformly discretised, are presented in Table 1. From this table it can be seen that the numerical solution is convergent, as the FDM mesh size decreases to zero.

<table>
<thead>
<tr>
<th>( l )</th>
<th>( N = M = 10 )</th>
<th>( N = M = 20 )</th>
<th>( N = M = 40 )</th>
<th>( N = M = 80 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>-0.00500</td>
<td>-0.02100</td>
<td>-0.031900</td>
<td>-0.35900</td>
</tr>
<tr>
<td>0.2</td>
<td>-0.02125</td>
<td>-0.03095</td>
<td>-0.34862</td>
<td>-0.37875</td>
</tr>
<tr>
<td>0.8</td>
<td>-0.03125</td>
<td>-0.30712</td>
<td>-0.34603</td>
<td>-0.37593</td>
</tr>
<tr>
<td>0.9</td>
<td>-0.10516</td>
<td>-0.02131</td>
<td>-0.30653</td>
<td>-0.34538</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Consider now the inverse problem given by equation (11) with \( h(x, t) = 1 + t \), equations (17), (18) and (4) with \( q_0(t) = -\partial u/\partial x(0, t) \) numerically simulated and given in Table 1 for \( N = M = 80 \). We perturb further this flux by adding to it some \( p \in \{1, 3, 5\}\)% random noise with mean zero and standard deviation equal to...
The numerical FDM solution for \( f(x) \) obtained with no regularization has been found highly oscillatory and unstable. In order to deal with this instability we employ the Tikhonov regularization of various orders such as zero, first and second, which yields the solution, [5],

\[
\hat{f}_\lambda = (A^T A + \lambda D_k^T D_k)^{-1} A^T b,
\]

where \( A \) represents the matrix resulting from the FDM discretisation of the inverse problem, \( b \) represents the right-hand side vector which is contaminated with noise, \( D_k \) is the regularization derivative operator of order \( k \in \{0, 1, 2\} \) and \( \lambda \geq 0 \) is the regularization parameter. The regularization derivative operator \( D_k \) imposes continuity, i.e. class \( C^0 \) for \( k = 0 \), first-order smoothness, i.e. class \( C^1 \) for \( k = 1 \), or second-order smoothness, i.e. class \( C^2 \) for \( k = 2 \). Thus \( D_0 = I \),

\[
D_1 = \begin{pmatrix} 1 & -1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & -1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 1 & -1 \end{pmatrix}, \quad D_2 = \begin{pmatrix} 1 & -2 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 1 & -2 & 1 \end{pmatrix}.
\]

The values of the regularization parameter \( \lambda \) are obtained by trial and error or by the L-curve criterion, [3].

Figure 1 shows the regularized numerical solutions (20) for various orders of regularization \( k \in \{0, 1, 2\} \) in comparison with the exact solution (19). From this figure it can be seen that the numerical results are stable and they become more accurate as the amount of noise \( p \) decreases. Also, the first- and second-order regularization methods produce smoother and more accurate numerical solutions than the zeroth-order regularization.

![Figure 1](image-url)

**Figure 1.** The exact solution (19) for the force \( f(x) \) in comparison with the numerical regularized solution (20) of various orders, for \( N = M = 80 \) and \( p \in \{1, 3, 5\} \% \) noise.
4 Conclusions

In this paper, the determination of space or time-dependent forces from boundary Cauchy data in the wave equation has been investigated. The solution of this linear inverse problem is unique, but is still ill-posed since small errors in the input data cause large errors in the output force. The problem is discretised numerically using the FDM, and in order to stabilise the solution, the Tikhonov regularization method has been employed. Numerical examples indicate that the method can accurately recover in a stable manner the unknown force. Future work will consist in investigating the nonlinear inverse problem in which the unknown force $f(u)$ depends on the displacement $u$.

Acknowledgments

S.O. Hussein would like to thank the Human Capacity Development Programme (HCDP) in Kurdistan for their financial support in this research.

References

Gelfand–Levitan, Marchenko and Krein equations. Theory, numerics and applications

S Kabanikhin¹,³ and M Shishlenin²,³

¹ Institute of Computational Mathematics and Mathematical Geophysics SB RAS, Russia
² Sobolev Institute of Mathematics SB RAS, Russia
³ Novosibirsk State University, Russia

E-mail: kabanikhin@sscc.ru, mshishlenin@ngs.ru

Abstract: We consider the method of regularization of two dimensional (2D) inverse coefficient problems based on the projection method and the approach of I.M. Gelfand, B.M. Levitan, M.G. Krein and V.A. Marchenko. We propose a method of reconstruction of the potential, density and velocity in 2D inverse coefficient problems. The 2D analogies of the I.M. Gelfand, B.M. Levitan and M.G. Krein method are established. Our approach can be easily applied to corresponding multidimensional inverse problems. The results of numerical calculations are presented.

Keywords: Gelfand-Levitan equation, Krein equation, inverse coefficient problems

1 Introduction

We consider the method of regularization of 2D inverse coefficient problems based on the projection method and the approach of I.M. Gelfand, B.M. Levitan, M.G. Krein and V.A. Marchenko.


One of the advantages of our approach (for 1D inverse coefficient problems see also [6, 7, 8]) is that it allows one to avoid multiple solution of 2D direct problem (see also the boundary control method proposed by M.I. Belishev [9, 10] and the globally convergent method proposed by M.V. Klibanov [11, 12]). In [13] we proved that boundary control method and the method by M.G. Krein are equivalent in 1D case.

2 2D analogy of Gelfand-Levitan equation

Let us consider the sequence of direct problems \((k = 0, ±1, ±2, \ldots)\)

\[
\begin{align*}
    u^{(k)}_{tt} &= u^{(k)}_{xx} + u^{(k)}_{yy} - q(x, y)u^{(k)}, & x \in \mathbb{R}, & y \in \mathbb{R}, & t > 0; \\
    u^{(k)}|_{t=0} &= 0, & u^{(k)}_{x}|_{t=0} &= \delta(x)e^{iky}, \\
    u^{(k)}|_{y=\pi} &= u^{(k)}|_{y=-\pi}.
\end{align*}
\]
Inverse problem 1: find function \( q(x, y) \) using additional information
\[ u^{(k)}|_{x=0} = f^{(k)}(y, t), \quad u^{(k)}|_{y=0} = 0, \quad k = 0, \pm 1, \pm 2, \ldots \]
The uniqueness of the inverse problem 1 can be proved using the technique in [23, 24]
Let us consider the sequence of the auxiliary problems \((m = 0, \pm 1, \pm 2, \ldots)\) [15, 16]
\[ w^{(m)}_t = w^{(m)}_{xx} + w^{(m)}_{yy} - q(x, y)w^{(m)}, \quad x > 0, \quad y \in \mathbb{R}, \quad t \in \mathbb{R}; \]
\[ w^{(m)}|_{x=0} = e^{imy} \delta(t), \quad w^{(m)}|_{y=0} = 0. \] (2)
It was proved in [15, 16] that the solution to the problem (1), (2) has the form
\[ \tilde{w}^{(m)}(x, y, t) = \frac{1}{4} e^{imy} \theta(x - |t|) \left[ x m^2 + \int_0^{x+t} q(\xi, y) d\xi + \int_0^{x-t} q(\xi, y) d\xi \right] + \frac{1}{2} \int_0^t \int_{x-t+\tau}^{x+t} [-\tilde{w}^{(m)}_{yy} + q(x, y)\tilde{w}^{(m)}](\xi, y, \tau) d\xi d\tau. \]
Therefore
\[ \tilde{w}^{(m)}(x, y, x - 0) = \frac{1}{4} e^{imy} \left[ x m^2 + \int_0^x q(\xi, y) d\xi \right]. \] (3)
The inverse problem 1 can be reduced to the system of integral equations \((k = 0, \pm 1, \pm 2, \ldots)\)
\[ \tilde{w}^{(k)}(x, y, t) + \int_{-x}^{x} \sum_m f^{(k)}_{m} (t-s) \tilde{w}^{(m)}(x, y, s) ds = -\frac{1}{2} \left[ f^{(k)}(y, t-x) + f^{(k)}(y, t+x) \right]. \] (4)
Here \(|t| < x, \ x \in \mathbb{R}\). The system (4) is 2D analogy of the Gelfand-Levitan equation.
Note that according to (3) \( q(x, y) \) can be calculated as follows
\[ q(x, y) = 4 \frac{d}{dx} \tilde{w}^{(0)}(x, y, x - 0). \]

3 2D analogy of M.G. Krein equation
Let us consider the sequence of direct problems \((k = 0, \pm 1, \pm 2, \ldots)\):
\[ u^{(k)}_{tt} = u^{(k)}_{xx} + u^{(k)}_{yy} - \nabla \ln \rho(x, y) \nabla u^{(k)}, \quad x > 0, \ y \in \mathbb{R}, \ t > 0; \]
\[ u^{(k)}|_{t<0} = 0, \quad u^{(k)}|_{y=0} = e^{iky} \delta(t); \quad u^{(k)}|_{y=0} = u^{(k)}|_{y=-\pi}. \]

Inverse problem 2: find function \( \rho(x, y) \) using additional information
\[ u^{(k)}(+0, y, t) = f^{(k)}(y, t), \quad k = 0, \pm 1, \pm 2, \ldots \]
The inverse problem 2 can be reduced to the 2D analogy of M.G. Krein equation [15, 16]
\[ 2\Phi^k(x, t) + \sum_m \int_{-x}^{x} f^{(k)}(t-s) \Phi^{(m)}(x, s) ds = -\int_{-\pi}^{\pi} \frac{e^{iky}}{\rho(0, y)} dy, \quad |t| < x, \quad k = 0, \pm 1, \pm 2, \ldots \] (5)
The inverse problem solution \( \rho(x, y) \) can be calculated by the formula
\[ \rho(x, y) = \frac{\pi^2}{\rho(0, y)} \left[ \sum_{m=-\infty}^{\infty} \Phi^{(m)}(x, x - 0) e^{-imy} \right]^2. \] (6)
For finding inverse problem solution \( \rho(x, y) \) in point \( x_0 > 0 \) we have to solve the system (5) with \( x = x_0 \) and calculate \( \rho(x_0, y) \) by formula (6). For numerical calculations (see figures 1–4) we use \( N \)-approximation [14, 20] of M.G. Krein equation [16] e.g. we cut the system (5) putting \( \Phi^k(x, t) \equiv 0 \) for all \( N < |k| \) [17].
Discrete analogies of the Gelfand–Levitan equation were considered in [18, 19, 22, 21].
4 Reconstruction of the velocity \(c(x, y)\)

**Inverse problem 3:** find the velocity \(c(x, y)\) from the sequence of relations \((k = 0, \pm 1, \pm 2, \ldots)\):

\[
c^{-2}(x, y)u^{(k)}_{tt} = u^{(k)}_{xx} + u^{(k)}_{yy}, \quad x \in \mathbb{R}, \quad y \in \mathbb{R}, \quad t > 0;
\]

\[
u^{(k)}|_{t=0} = 0, \quad u^{(k)}_t|_{t=0} = e^{iky} \delta(x).
\]

\[
u^{(k)}(0, y, t) = f^{(k)}(y, t), \quad u^{(k)}_x(+0, y, t) = 0.
\]

Let \(\tau(x, y)\) be a solution of Cauchy problem for the eikonal equation

\[
\tau_x^2 + \tau_y^2 = c^{-2}(x, y), \quad x > 0, \quad y \in \mathbb{R};
\]

\[
\tau|_{x=0} = 0, \quad \tau_x|_{x=0} = c^{-1}(0, y), \quad y \in \mathbb{R}.
\]

Let us introduce new variables \(z = \tau(x, y), y = y\) and new functions

\[
v^{(k)}(z, y, t) = u^{(k)}(x, y, t), \quad b(z, y) = c(x, y).
\]

Since the velocity is supposed to be strictly positive this change of variables is not degenerate at least in some interval \(x \in (0, h)\).

Let us consider the sequence of the auxiliary problems \((m = 0, \pm 1, \pm 2, \ldots) [15, 16]:(10)\):

\[
w^{(m)}_{tt} = w^{(m)}_{xx} + b^2 w^{(m)}_{yy} + qw^{(m)}_{yy} + pw^{(m)}_y, \quad z > 0, \quad y \in \mathbb{R}, \quad t \in \mathbb{R};
\]

\[
w^{(m)}(0, y, t) = e^{imy} \delta(t), \quad w^{(m)}(0, 0, t) = 0.
\]

Here

\[
q(z, y) = 2b^2 \tau_y, \quad p(z, y) = b^2(z, y)(\tau_{xx} + \tau_{zz}).
\]

We suppose that \(c(0, y) = b(0, y)\) is known and for simplicity \(b(0, y) \equiv 1\) for \(y \in \mathbb{R}\).

In the neighborhood of the plane \(t = z\) the solution of the direct problem \((10), (11)\) has the form \([15, 16]:\)

\[
w^{(m)}(z, y, t) = S^{(m)}(t, y)\delta(z - t) + Q^{(m)}(t, y)\theta(z - t) + \tilde{w}^{(m)}(z, y, t).
\]

Here \(\tilde{w}^{(m)}\) is continuous function and functions \(S^{(m)}\) and \(Q^{(m)}\) solve the following problems:

\[
2S^{(m)}_t + qS^{(m)}_y + pS^{(m)} = 0, \quad t > 0, \quad y \in \mathbb{R};
\]

\[
S^{(m)}|_{t=0} = \frac{1}{2} e^{imy}.
\]

\[
2Q^{(m)}_{tt} = S^{(m)}_{tt} - \left[qQ^{(m)}_y + b^2 S^{(m)}_{yy} + pQ^{(m)}\right], \quad t > 0, \quad y \in \mathbb{R};
\]

\[
Q^{(m)}|_{t=0} = 0.
\]

The 2D analogy of M.G. Krein equation follows from \((13) (m = 0, \pm 1, \pm 2, \ldots):\)

\[
\sum_m S^{(m)}(z, y)f^{(k)}_m(t - z) + \tilde{w}^{(k)}(z, y, t) + \sum_m \int_{-z}^{z} f^{(k)}_m(t - s)\tilde{w}^{(m)}(z, y, s)ds = 0, \quad |t| < z.
\]

So for solving the inverse problem 3 we can solve the system \((14) - (18)\), using the projection method and then find \(c(x, y)\) from the following iterative algorithm.

First, we introduce \(N\)-approximation of the system \((14) - (18)\), e.g. let \(\tilde{w}^{(m)}, S^{(m)}\) and \(Q^{(m)}\) be equal to 0 for all \(|m| > N\). Let us suppose that \(c_n(x, y)\) is known. Then we calculate \(\tau_n(x, y)\) from \((7), (8) [25], b_n(z, y)\).
from (9) and \( q_n(z, y) \) and \( p_n(z, y) \) from (12). Function \( S^{(m)}_n(t, y) \) is calculated from (14), (15). Then solving the 2D analogy of M.G. Krein equation (18) we find \( \tilde{w}^{(m)}_n(z, y, t) \) for \(|m| \leq N\). It follows from (13) that \( Q^{(m)}_n(t, y) = \tilde{w}^{(m)}_n(t + 0, y, t) \). Then from equations (14) and (16) we find function \( b_{n+1}(z, y) \) and after that new value \( c_{n+1}(x, y) \) is calculated.

In numerical experiments (see figures 1–4) 2D inverse problem 2 is approximated by the finite system of one dimensional inverse acoustic problems [16, 20, 17]. The inverse problem 2 is solved in the domain \( x \in (0, 1), y \in (-\pi, \pi) \) and \( t \in (0, 2) \). The number \( N \) is equal to 5 for figure 2 and the number \( N \) is equal to 10 for figures 3 and 4. The noisy data is taken as

\[
f^\varepsilon(t) = f(t) + \varepsilon \alpha(t)(f_{\max} - f_{\min}).
\]

Here \( \varepsilon \) is the level of noise, \( \alpha(t) \) is white noise for fixed \( t \), \( f_{\max} \) and \( f_{\min} \) are maximum and minimum values of exact data. The dimension of the space grid is equal to 100 \( \times \) 100.

**Figure 1.** The exact solution of the inverse problem 2.

**Figure 2.** The approximate solution of the inverse problem 2, \( N = 5, \varepsilon = 0 \).

**Figure 3.** The approximate solution of the inverse problem 2, \( N = 10, \varepsilon = 0 \).

**Figure 4.** The approximate solution of the inverse problem 2, \( N = 10, \varepsilon = 0.05 \).
Acknowledgments
The work was partially supported by the Ministry of Education and Science of the Russian Federation, interdisciplinary project SB RAS 14 “Inverse problems and applications: theory, algorithms and software” and RFBR grant 12-01-00773.

References
The methods of the field continuation from the part of the boundary

M Shishlenin$^{1,2}$

$^1$Sobolev Institute of Mathematics SB RAS, Russia
$^2$Novosibirsk State University, Russia

E-mail: mshishlenin@ngs.ru

Abstract: The field continuation problems from the part of the boundary (or the Cauchy Problem) are ill-posed problems. The similar problems can be found, for instance, in geophysics and tomography when the field continuation allows to detect the parameters of the medium outside the investigated domain. We reduce the ill-posed problem to the inverse problem and reformulate it in the operator equation $Aq = f$. For numerical solution of the continuation problem we apply singular value decomposition method and gradient methods. Theory and numerical methods are developed for the continuation problem. The formulae to calculate the singular values of the continuation problem operator was obtained in case of the simple geometry. The results of numerical calculations are presented.

1 Introduction. Problem formulation

We offer a unified approach to regularization of continuation problem for a hyperbolic, parabolic and elliptic equations of mathematical physics. For the first time a similar iteration approach was offered in the publication by V.A. Kozlov, V.G. Maz’ya and A.V. Fomin in 1991 [3].

The continuation problems are related to inverse problems of mathematical physics. Its theoretical framework has been set in publications of A.N. Tikhonov, M.M. Lavrentiev, V.K. Ivanov, as well as of their students and followers. In many inverse problems the sought heterogeneities are located at a certain depth beneath a layer of the medium with known parameters (in geophysics these are, as a rule, either homogeneous or layered media). In this case the problem of continuation of geophysical fields from the land surface in the direction of the heterogeneity position becomes one of the important tools is the hands of a practitioner.

We consider the scalar two-dimensional continuation problem

$$\varepsilon u_{tt} + \sigma u_t - \frac{1}{\mu}(u_{xx} + u_{yy}), \quad (x, y, t) \in \Omega,$$

$$u(0, y, t) = f(y, t), \quad y \in (0, L), \quad t \in (0, T).$$

$$u_x(0, y, t) = g(y, t), \quad y \in (0, L), \quad t \in (0, T).$$

Here $y$ is a horizontal variable, $x$ stands for the depth, $\Omega(h) = \{(x, y, t) : x \in (0, h), y \in (0, L), t \in (0, T)\}$; the positive functions $\varepsilon(x, y), \sigma(x, y)$ and $\mu(x, y)$ describe the permittivity, conductivity and permeability of the ground, respectively. The continuation problem (1)–(3) is to find function $u(x, y, t)$ inside the domain. This is a general statement of the continuation problem. For instance, if $\varepsilon = 0$ then we obtain Cauchy problem for parabolic equation, if $\sigma = 0$ we obtain Cauchy problem for acoustic equation, if $\varepsilon = 0$ and $\sigma = 0$ we obtain Cauchy problem for Laplace equation. The continuation problem (1)–(3) is ill-posed problem, its solution is unique, but it does not depend continuously on the Cauchy data [1, 15].
Physical meaning of problem (1)–(3) consists in the following. A source of electromagnetic waves (3) located on the surface \( x = 0 \) of the investigated domain \((0, h) \times (0, L)\), is switched on at the moment \( t = 0 \). The response of the medium (2) is measured on the surface \( x = 0 \) during \( t \in (0, T) \). We suppose the medium as being at rest for \( t < 0 \). Consequently, we can add the condition

\[
u(x, y, 0) = u_t(x, y, 0) = 0, \quad x \in (0, h), \quad y \in (0, L).
u\]

We assume also that the support of the source function lies inside \((0, L)\) and \( L \) is large enough to provide the homogeneous boundary conditions

\[
u(x, 0, t) = u(x, L, t) = 0, \quad x \in (0, h), \quad t \in (0, T).
u\]

Theoretical and numerical techniques for solving different statements of the continuation problems are presented, for instance, in [3, 2, 4, 8, 12, 16].

In this work we consider two methods for the field continuation for Helmholtz equation from the part of the boundary: the singular value decomposition [17] and optimization method [15].

2 Analysis of singular values

When studying acoustic or electrodynamic problems, in many cases one shifts to harmonic motions and the Helmholtz equation. In this section we are going to analyze the singular values of a continuation problem operator for a complex-valued formulation of the Helmholtz equation in a case of simple geometry.

A Cauchy problem for the Helmholtz equation is a well-known example of an ill-posed problem. Its solution isn’t stable relative to the small variations of the Cauchy data [1, 15].

In [6, 14] it has been demonstrated that the ill-posedness of the Cauchy problem for the Helmholtz equation depends on the wave number \( k \) and increases with its growth. The numerical calculations using different methods have been presented in [5, 6, 7, 11, 10, 13, 17].

Let us consider a continuation problem for the Helmholtz equation in a homogeneous medium for simple geometry [16]:

\[
\Delta u + k^2 u = 0, \quad x \in (0, h), \quad y \in (0, \pi),
\]

\[
u(x, y) = f(y), \quad y \in (0, \pi),
\]

\[
u(x, 0) = u(x, \pi) = 0, \quad x \in (0, h).
\]

Here \( k^2 = \varepsilon \omega^2 - i\sigma \omega \), \( \omega \) is a frequency, \( \varepsilon \) and \( \sigma \) are positive constants.

The continuation problem (4)–(7) is to find the function \( u(x, y) \) in the domain \( x \in (0, h), \ y \in (0, \pi) \) using the boundary conditions (5)–(7).

Now we formulate the continuation problem (4)–(7) as an inverse with respect to the direct problem (4), (6) and (7) with the boundary conditions

\[
u(h, y) = q(y), \quad y \in (0, \pi),
\]

The inverse problem is finding the function \( q(y) \) using the additional information (5). Let us formulate the inverse problem (4)–(8) as operator equation \( Aq = f \) and we will find the singular values of the operator \( A \) exactly.

To find the solution of the direct problem (4), (6)–(8) we assume that \( q(y) \) is expressed as a Fourier series.

In this case, the solution of the initial direct problem (4), (6), (7) and (8) is presented as a Fourier series:

\[
\sum_{m=1}^{\infty} \frac{\cosh(\lambda_m x)}{\cosh(\lambda_m h)} q^{(m)} \sin(my).
\]
Here $q^{(m)}$ is the Fourier coefficients of $q(y)$, $\lambda_m = \alpha_m + i\beta_m$ and

$$
\alpha_m = \sqrt{\frac{(m^2 - \varepsilon \omega^2)^2 + \sigma^2 \omega^2 + m^2 - \varepsilon \omega^2}{2}}, \quad \beta_m = \sqrt{\frac{(m^2 - \varepsilon \omega^2)^2 + \sigma^2 \omega^2 - m^2 + \varepsilon \omega^2}{2}}.
$$

Then the solution of the inverse problem (4)–(7) and (8) has the form

$$
q(y) = \sum_{m=1}^{\infty} f^{(m)} \cosh(\lambda_m h) \sin(my).
$$

Since the operator $A$ is diagonal, the singular values of electromagnetic field continuation problem operator ($\varepsilon \neq 0$, $\sigma \neq 0$) are expressed as:

$$
\sigma_m(A) = \frac{\sqrt{2}}{\cosh(2\alpha_m h) + \cos(2\beta_m h)}.
$$

Now, we consider a several particular cases of singular values of operator $A$. For the acoustic equation ($\varepsilon \neq 0$, $\sigma = 0$) the singular values formula is expressed as [14]:

$$
\sigma_m(A) = \begin{cases} 
\frac{1}{\cos(\sqrt{m^2 - \varepsilon \omega^2})}, & m^2 \leq \varepsilon \omega^2, \\
\frac{1}{\cosh(\sqrt{m^2 - \varepsilon \omega^2})}, & \varepsilon \omega^2 < m^2.
\end{cases}
$$

The singular values depend on the wave number $k_n^2 = \varepsilon \omega^2 - m^2$ and the ration of $m$, $\varepsilon$ and $\omega$ [14]. In the case of $m^2 \leq \varepsilon \omega^2$ the singular values of the operator $A$ are limited to 1 from below. At the same time in the case of $m^2 > \varepsilon \omega^2$ the singular values decay to zero exponentially.

Now, let's consider the parabolic equation ($\varepsilon = 0$ and $\sigma \neq 0$). In this case

$$
\sigma_m(A) = \frac{\sqrt{2}}{\cosh(2\alpha_m h) + \cos(2\beta_m h)},
$$

$$
\alpha_m = \sqrt{\frac{m^4 + \sigma^2 \omega^2 + m^2}{2}}, \quad \beta_m = \sqrt{\frac{m^4 + \sigma^2 \omega^2 - m^2}{2}}.
$$

In case of the Laplace equation ($\varepsilon = 0$, $\sigma = 0$) the singular values decay exponentially:

$$
\sigma_m(A) = \frac{1}{\cosh(m h)}.
$$

3 Optimization method

Let us consider the following continuation problem for an elliptic equation:

$$
u_{xx} + L(y)u = 0, \quad (x, y) \in \Omega, \quad (9)$$

$$u(0, y) = f(y), \quad y \in \mathcal{D}, \quad (10)$$

$$u_x(0, y) = 0, \quad y \in \mathcal{D}, \quad (11)$$

$$u(x, y) = 0, \quad x \in (0, h), \quad y \in \partial \mathcal{D} \quad (12)$$

with the matching conditions

$$f(y) = 0, \quad y \in \partial \mathcal{D}. \quad (13)$$
Here $\Omega = (0, h) \times D$, $D \in \mathbb{R}^n$ is the bounded domain with a Lipschitz boundary $\partial D$,
\[
L(y)u = \sum_{i,j=1}^{n} \frac{\partial}{\partial y_i} \left( a_{ij}(y) \frac{\partial u}{\partial y_j} \right) - c(y)u,
\]
\[
M_4 \sum_{j=1}^{n} \nu_j^2 \leq \sum_{i,j=1}^{n} a_{ij}(y) \nu_i \nu_j, \quad \forall \nu_i \in \mathbb{R}, \quad a_{ij} = a_{ji}, \quad i, j = 1, \ldots, n,
\]
\[
0 \leq c(y) \leq M_5, \quad a_{ij} \in C^1(\mathcal{D}), \quad c \in C(\mathcal{D}).
\]

Let us consider the ill-posed continuation problem (9)–(13) as the inverse problem of the following direct problem (9), (11)–(13) and boundary condition
\[
u_{n+1} = \nu_n - \alpha_n J'(\nu_n).
\]

Here $\alpha_n$ is descent parameter. Let us consider the conjugate problem:
\[
\psi_{xx} + L(y)\psi = 0, \quad (x, y) \in \Omega, \quad \psi_x(0, y) = 2\left[u(0, y) - f(y)\right] - \psi(h, y) = 0, \quad y \in \mathcal{D},
\]
\[
\psi(x, y) = 0, \quad x \in (0, h), \quad y \in \mathcal{D}.
\]

The gradient of the functional $J'q$ can be calculated by the formula:
\[
(J'q)(y) = \psi_x(h, y).
\]

It follows from [9, 15, 17] that the solution is unique and the estimate of the conditional stability holds true of the continuation problem (9)–(13). Moreover, the number of iteration $n$ is the regularization parameter.

## 4 Numerical experiment

Let us investigate numerically (see fig. 1 and 2) degree of ill-posedness of the continuation problem (4)–(7). Let $h_x = L_x/N_x$, $h_y = \pi/N_y$ and discretize the equations and boundary conditions
\[
\frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h_x^2} + \frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{h_y^2} + k^2u_{i,j} = 0, \quad i = 1, N_x - 1, \quad j = 1, N_y - 1,
\]
\[
u_{i,j} - u_{0,j} = 0, \quad u_{i,j} = f_j, \quad j = 0, N_y,
\]
\[
u_{i,0} = 0, \quad i = 0, N_x.
\]

Let us denote $a = h_x^2$, $b = h_x^2$ and $c = -2h_x^2 - 2h_y^2 + k^2h_x^2h_y^2$. Therefore
\[
a = b + cu_{i,j} + cu_{i,j+1} + cu_{i+1,j} + bu_{i,j+1} + bu_{i+1,j} = 0, \quad i = 1, N_x - 1, \quad j = 1, N_y - 1,
\]

Therefore we obtain the system of algebraic equations
\[
Aq = f.
\]

Here $A$ is a matrix of $(N_x + 1) \times (N_y + 1)^2$, $q$ is unknown vector
\[
q = (u_{0,0}, u_{0,1}, u_{0,2} \ldots u_{0,N}, u_{1,0}, u_{1,1}, u_{1,2} \ldots u_{1,N}, \ldots u_{N,0}, u_{N,1}, u_{N,2}, \ldots u_{N,N}),
\]
f is a vector of right hand side (boundary conditions and inverse problem data).
Figure 1. \( \log \) of singular values \( N_x = 50, \ N_y = 50 \).

Figure 2. \( \log \) of singular values \( N_x = 100, \ N_y = 100 \).

Acknowledgments

The work was partially supported by the Ministry of Education and Science of the Russian Federation and RFBR grant 14-01-00208.

References

Minisymposium M13

Regularisation Methods – Algorithms
ARTs penalized by non-smooth convex functions

Qinian Jin

Mathematical Sciences Institute, Australian National University, Canberra, ACT 0200, Australia
E-mail: Qinian.Jin@anu.edu.au

Abstract: Algebraic reconstruction techniques are a family of powerful iterative algorithms used in computed tomography. These algorithms, however, show smoothing effect which results in undesired artifacts and thus blurs the reconstruction results. To weaken the smoothing effect, we propose new iterative algorithms by incorporating non-smooth convex functions into the classical algorithms as penalty terms. We provide the convergence results and present numerical simulations on computed tomography to illustrate the performance.

Keywords: Algebraic reconstruction techniques, strongly convex functions, sparsity, total variation, computed tomography

1 Introduction

The problems of image reconstruction from projections arise in many fields of applications. After suitable discretization, such problems can be modelled by a system of linear equations

$$Ax = b,$$

where $A$ is an $m \times n$ system matrix, $b \in \mathbb{R}^m$ is the observed data, and $x \in \mathbb{R}^n$ is the image to be reconstructed. In practical applications, $A$ is in general an unstructured ill-conditioned matrix of huge size. Because $b$ always contains noise, the reconstruction of images from projections is a challenging subject.

For image reconstruction from projections, analytic methods form an important class of inversion methods among which the filtered back-projection algorithm [10] is the most prominent one. These methods are based on explicit inversion formulas and are therefore fast and inexpensive. However, analytic methods are suffered from the metal artifacts, scanner geometries, non-uniform attenuation and other physical factors, and are difficult to incorporate a prior information.

In contrast, algebraic reconstruction techniques (ART) do not involve explicit inversion, but use an iterative procedure to generate a convergent sequence. They only rely on matrix-vector multiplications and therefore are well suited for large-scale problems. Although they are slower than analytic methods, ART algorithms are generally simple and can be easily modified to different data geometries and to restricted angular coverage. Moreover, they are relatively easy to incorporate prior knowledge into the reconstruction process, and they can be used to decrease radiation exposure while maintaining image quality.

The first ART algorithm is the Kaczmarz’s algorithm introduced in [9] for solving system of consistent linear equations which was rediscovered in [6] for CT image reconstruction. This algorithm is sequential in nature because the data from projections are applied sequentially during an iteration cycle. In order to obtain a reconstructed image with high quality, it was advocated in [5] that, in each cycle of iterations, data from all the projections should be applied simultaneously. Such methods are called the simultaneous iterative reconstruction techniques (SIRT) which, mathematically, can be formulated as [11]

$$x_{k+1} = x_k + \mu_k V^{-1}A^T W(b - Ax_k), \quad k = 0, 1, \cdots$$
with relaxation parameters $\mu_k > 0$ and positive definite matrices $V$ and $W$. Under suitable choices of $V$ and $W$, one can reproduce many well-known algorithms including the Cimmino’s algorithm [4], the simultaneous algebraic iterative technique (SART) [1], the component averaging (CAV) algorithm [3], and the diagonally relaxed orthogonal projection (DROP) method [2]. When $V$ and $W$ are diagonal matrices these algorithms can be parallel implemented.

These classical algorithms have important application in computed tomography. However, they have the disadvantage of smoothing effect which produces undesired artifacts and thus blurs the reconstruction results. In order to weaken the smoothing effect, we propose new iterative algorithms by incorporating non-smooth convex functions into the classical algorithms as penalty terms. Our methods have the splitting character that each iteration consists of two separated steps: the first step involves only the classical ART procedures while the second step involves only the calculation of the proximal mapping of the convex penalty function. By using tools from convex analysis, we obtain the convergence of our methods. Numerical simulations on computed tomography are presented to illustrate the performance.

2 The methods

We first take two positive definite matrices $W$ and $V$ of size $m \times m$ and $n \times n$ respectively. Using the standard inner product $\langle \cdot, \cdot \rangle$ with the induced norm $\| \cdot \|_2$ on Euclidean spaces, we may introduce on $\mathbb{R}^n$ and $\mathbb{R}^m$ the inner products

\[ \langle x_1, x_2 \rangle_V := \langle x_1, Vx_2 \rangle, \quad \forall x_1, x_2 \in \mathbb{R}^n, \]
\[ \langle b_1, b_2 \rangle_W := \langle b_1, Wb_2 \rangle, \quad \forall b_1, b_2 \in \mathbb{R}^m. \]

The induced norms will be denoted by $\| \cdot \|_V$ and $\| \cdot \|_W$. If $A$ is considered as an operator from $(\mathbb{R}^n, \| \cdot \|_V)$ to $(\mathbb{R}^m, \| \cdot \|_W)$, its adjoint is given by $A^# = V^{-1}A^T W$.

To derive our method for solving (1), according to the available a priori information on the image to be reconstructed, we take a proper lower semi-continuous function $J : \mathbb{R}^n \to (-\infty, \infty]$ satisfying $J(x) \geq J(0) = 0$ that is strong convex in the sense that there is a constant $c_0 > 0$ such that

\[ J(\lambda x + (1 - \lambda)x_0) + c_0 \lambda (1 - \lambda)\|x - x_0\|_V^2 \leq \lambda J(x) + (1 - \lambda)J(x_0) \]  

(2)

for all $0 \leq \lambda \leq 1$ and $x, x_0 \in \mathbb{R}^n$, and determine the unique solution of (1) satisfying

\[ \min_{x \in \mathbb{R}^n} J(x) \quad \text{subject to} \quad Ax = b. \]  

(3)

The corresponding Lagrangian is $L(x, p) := J(x) + \langle p, b - Ax \rangle_W$, where $p \in \mathbb{R}^m$ denotes the Lagrange multiplier. Then the solution of (3) can be found by determining a saddle point of $L$ if exists. Let $(x_c, p_c)$ be a current guess of a saddle point of $L$, we may use the Uzawa procedure to get an update $(x_{+}, p_{+})$ by

\[ p_+ = \arg \max_{p \in \mathbb{R}^m} \left\{ L(x_c, p) - \frac{1}{2\mu} \|p - p_c\|_W^2 \right\} \quad \text{and} \quad x_+ = \arg \min_{x \in \mathbb{R}^n} L(x, p_+) \]

with a suitable relaxation parameter $\mu > 0$. Let $\xi_c = A^# p_c$ and $\xi_+ = A^# p_+$. By straightforward manipulation it follows

\[ \xi_+ = \xi_c + \mu A^# (b - Ax_c) \quad \text{and} \quad x_+ = \arg \min_{x \in \mathbb{R}^n} \left\{ J(x) - \langle \xi_+, x \rangle_V \right\}. \]

Because $A^# = V^{-1}A^T W$, this motivates us to determine the solution of (3) using the iterative scheme

\[ \xi_{k+1} = \xi_k + \mu_k V^{-1} A^T W (b - Ax_k), \]
\[ x_{k+1} = \arg \min_{x \in \mathbb{R}^n} \left\{ J(x) - \langle \xi_{k+1}, x \rangle_V \right\} \]  

(4)

with $x_0 = \xi_0 = 0$ and suitably chosen relaxation parameters $\mu_k > 0$.

The implementation of the method (4) requires to solving a minization problem related to $J$ in each iteration step. Although exact solvers are available for some $J$, the minimization problem in general can only be solved inexactly. This motivates us to propose the following algorithm with inexact solvers for the minimization problems.
Algorithm 2.1. Let $\xi_0 = x_0 = 0$, and let $\{\varepsilon_k\}_{k=1}^{\infty}$ be a non-increasing sequence of non-negative numbers satisfying $\sum_k \varepsilon_k < \infty$. For $k \geq 0$ we define
\[
\xi_{k+1} = \xi_k + \mu_k V^{-1} A^T W(b - Ax_k)
\]
with suitable relaxation parameters $\mu_k > 0$ and take $x_{k+1} \in \mathbb{R}^n$ to be any element such that
\[
J(x_{k+1}) - \langle \xi_{k+1}, x_{k+1} \rangle_v \leq \min_{x \in \mathbb{R}^n} \{J(x) - \langle \xi_{k+1}, x \rangle_v \} + \varepsilon_{k+1}.
\]

The following result shows that Algorithm 2.1 always produces a convergent sequence $\{x_k\}$ and gives the characterization of the limit even if (1) is inconsistent.

Theorem 2.1. Let $J : \mathbb{R}^n \to (-\infty, \infty]$ be a proper, lower semi-continuous function that is strongly convex in the sense of (2). Let $\rho = \|W^{1/2}AV^{-1/2}\|_2$ and $\sigma \leq \mu_k \leq (4c_0 - \sigma)/\rho$, where $\sigma > 0$ is an arbitrarily small but fixed number. Then for the sequence $\{x_k\}$ defined by Algorithm 2.1 there holds $\lim_{k \to \infty} \|x_k - x^1\| = 0$, where $x^1 \in \mathbb{R}^n$ is the unique element such that
\[
J(x^1) = \min \left\{ J(\hat{x}) : \hat{x} \in \arg \min_{x \in \mathbb{R}^n} \|Ax - b\|_W \right\}.
\]

We next develop the block-iterative version of Algorithm 2.1. The basic idea is to rewrite (1) into an equivalent linear system
\[
A_t x = b_t, \quad t = 1, \ldots, q
\]
consisting of $q$ equations, where $A_t : \mathbb{R}^n \to \mathbb{R}^{m_t}$, $t = 1, \ldots, q$ with $m_t < m$. By applying Algorithm 2.1 to each equation in (6) cyclically with a positive definite matrix $V$ of size $n \times n$ and $q$ positive definite matrices $W_t$ of size $m_t \times m_t$, it leads to the following algorithm.

Algorithm 2.2. Let $\xi_0 = x_0 = 0$, and let $\{\varepsilon_k\}_{k=1}^{\infty}$ be a non-increasing sequence of non-negative numbers satisfying $\sum_k \varepsilon_k < \infty$. For $k \geq 0$ we set $\xi_{k, 1} = \xi_k$ and $x_{k, 1} = x_k$ and for $t = 1, \ldots, q$ we define
\[
\xi_{k, t+1} = \xi_{k, t} + \mu_k V^{-1} A_t^T W_t(b_t - A_t x_{k, t})
\]
and take $x_{k, t+1} \in \mathbb{R}^n$ to be any element such that
\[
J(x_{k, t+1}) - \langle \xi_{k, t+1}, x_{k, t+1} \rangle_v \leq \min_{x \in \mathbb{R}^n} \{J(x) - \langle \xi_{k, t+1}, x \rangle_v \} + \varepsilon_{g_{k+1}}.
\]

We then set $\xi_{k+1} = \xi_{k, q+1}$ and $x_{k+1} = x_{k, q+1}$.

The following theorem shows that if the linear system (6) is consistent, i.e. it has a solution, then the sequence $\{x_k\}$ produced by Algorithm 2.2 is convergent and the limit is the solution of (6) with smallest Bregman distance to $x_0$ in the direction $\xi_0$.

Theorem 2.2. Let $J : \mathbb{R}^n \to (-\infty, \infty]$ be a proper, lower semi-continuous function that is strongly convex in the sense of (2). Assume that (6) is consistent and that $\|W_t^{1/2} A_t V^{-1/2}\|_2 \leq \rho$ for all $t = 1, \ldots, q$. If $\sigma \leq \mu_k \leq (4c_0 - \sigma)/\rho$, where $\sigma > 0$ is an arbitrarily small but fixed number, then for the sequence $\{x_k\}$ defined by Algorithm 2.2 there holds $\lim_{k \to \infty} \|x_k - x^1\| = 0$, where $x^1$ is such that
\[
J(x^1) = \min \left\{ J(x) : A_t x = b_t, t = 1, \ldots, q \right\}.
\]

The proofs of the above two theorems will be given in [8]. The main technical difficulties on convergence analysis come from non-smoothness of $J$ and the inexact resolution of the minimization problems. These difficulties can be conquered by using tools from convex analysis, in particular the Legendre-Fenchel transform and the $\varepsilon$-subdifferential calculus [12].
Examples

We give some examples covered by Algorithm 2.1; more examples can be found in [8]. We use $a^T_i$ and $a_{ij}$ to denote the $i$-th row and the $(i,j)$-entry of $A$ respectively. For a column vector $u$, $u[i]$ denotes its $i$-th component. We use $I_n$ to denote the $n \times n$ identity matrix.

Example 2.1. In Algorithm 2.1 we take $V = I_n$ and $W = \text{diag}(\frac{1}{w_1}, \cdots, \frac{1}{w_m})$ with $w_i = \sum_{j=1}^n s_j|a_{ij}|^2$, where $s_j$ denotes the number of nonzero elements in column $j$ of $A$, then the equation (5) becomes

$$
\xi_{k+1}[j] = \xi_k[j] + \mu_k \sum_{i=1}^m a_{ij}(b[i] - a^T_i x_k), \quad j = 1, \ldots, n.
$$

The corresponding method is convergent for $0 < \mu_k < 4c_0$ because $\|W^{1/2}A\|_2 \leq 1$. When $J(x) = \frac{1}{2}\|x\|_2^2$ and $\varepsilon_k = 0$ for all $k$, the method becomes the CAV algorithm from [3]. When taking $J(x) = \frac{1}{2}\|x\|_2^2 + \|x\|_1 + \lambda_B(x)$ and $\varepsilon_k = 0$, where $B := \prod_{j=1}^n [c_j, d_j]$ with $-\infty \leq a_j \leq b_j \leq \infty$ and $\lambda_B$ is the indicator function of $B$, we obtain the sparse version of the CAV algorithm with box constraints. We can also take $J$ to be the (high order) total variation (TV) like functions to improve image reconstruction.

Example 2.2. In Algorithm 2.1 we take $W = \text{diag} \left( \frac{1}{w_1}, \cdots, \frac{1}{w_m} \right)$ and $V = \text{diag} (v_1, \cdots, v_n)$, where

$$
w_i = \sum_{1 \leq j \leq n; a_{ij} \neq 0} |a_{ij}|^{2-\alpha}, \quad v_j = \sum_{1 \leq i \leq n; a_{ij} \neq 0} |a_{ij}|^{\alpha}
$$

with $0 \leq \alpha \leq 2$, then the equation (5) becomes

$$
\xi_{k+1}[j] = \xi_k[j] + \frac{\mu_k}{\sum_{i; a_{ij} \neq 0} |a_{ij}|^{\alpha}} \sum_{i=1}^m a_{ij}(b[i] - a^T_i x_k), \quad j = 1, \ldots, n.
$$

The convergence of the corresponding method is guaranteed for $0 < \mu_k < 4c_0$ as one can show that $\|W^{1/2}AV^{-1/2}\|_2 \leq 1$, see [11]. When $J(x) = \frac{1}{2}\|x\|_2^2$ and $\varepsilon_k = 0$ for all $k$, the method reduces to SART in [1] if $\alpha = 1$ and to DROP in [2] if $\alpha = 0$. Other choices of $J$ can significantly improve the quality of the reconstructed image.

3 Numerical simulations

We consider a test problem that models a standard 2D parallel-beam tomography. We use the function paralleltomo in the MATLAB package AIR TOOLS [7] to generate a sparse matrix $A$, an exact solution $x^\dagger$ which represents the Shepp-Logan phantom of size $200 \times 200$ and an exact data $b = Ax^\dagger$. We then add Gaussian white noise to $b$ to generate a noisy data $b^\delta$ with $\delta_{rel} := \|b^\delta - b\|_2/\|b\|_2 = 0.01$ and use it to reconstruct $x^\dagger$. During computation, we terminate the iterations by the discrepancy principle $\|Ax_k - b^\delta\|_2 \leq \tau \delta$ with $\delta = \delta_{rel}\|b\|_2$ and $\tau = 1.0001$.

We first consider a full angle problem using 90 projection angles evenly distributed between 0 and 179 degrees, with 256 lines per projection. The matrix $A$ has dimension size $m = 23040$ and $n = 40000$. In Figure 1(a) we plot the exact Shepp-Logan phantom. In (b) we plot the reconstruction result by CAV with nonnegativity constraint. In (c) we report the result using CAV penalized by the TV-like function

$$
J(x) = \frac{1}{2}\|x\|_2^2 + ||x||_{TV} + \lambda_{(0,\infty)}(x),
$$

where $\|x\|_{TV}$ denotes the total variation of $x$ considered as a two-dimensional image, see [13]. We then use the primal-dual hybrid gradient (PDHG) method ([13]) to solve the minimization problem associated with $J$; the PDHG procedure is terminated when the relative duality gap is $\leq k^{-1.2}$ at the $k$-th iteration of our method, we used $\mu_k = 2$; the result improves the one in (b) significantly.

We next consider a limited angle problem using 45 projection angles evenly distributed between 0 and 89 degrees, with 200 lines per projection. The matrix $A$ now has dimension size $m = 9000$ and $n = 40000$. In Figure 2 we report the computational results obtained by SART and its variant. Figure 2(a) is the reconstruction
result obtained by the classical SART algorithm with $\mu_k = 2$ and (b) presents the result by SART penalized by the TV-like penalty function

$$J(x) = \frac{1}{2\beta} \|x\|^2_V + \|x\|_{TV} + \epsilon_{[0,\infty)^n}(x)$$

with $\beta = 40$ and $\mu_k = 2/\beta$, where $V$ is the diagonal matrix defined in Example 2.2; the minimization problem associated with $J$ is solved by the PDHG algorithm which is terminated when the relative duality gap is $\leq 4k^{-1.2}$ at the $k$-th iteration. It is clear that the result in (b) is much better than the one in (a).

**Figure 1.** Reconstruction results by the CAV algorithm and its variant with TV-like penalty

**Figure 2.** Reconstruction results from limited angle projections using SART and its variant.

**References**

8. Jin Q 2014 Convex penalized algebraic reconstruction techniques *Preprint*
Minisymposium M15
Compressive Sensing
Non-linear compressed sensing and its application to beam hardening correction in x-ray tomography

Thomas Blumensath

ISVR, University of Southampton, SO17 1BJ, UK
E-mail: thomas.blumensath@soton.ac.uk

Abstract: Traditionally, compressed sensing assumes a linear, ill-posed or non-invertible forward model, which is inverted with the help of non-convex constraints. Recently these ideas have been extended to non-linear forward models. It could be shown that, under certain conditions, strong performance guarantees available for traditional compressed sensing also hold in the non-linear case.

In this paper I will present some initial results that explore these ideas empirically on an x-ray tomographic reconstruction problem. Using a non-linear x-ray transmission model which accounts for energy dependent absorption, it is demonstrated that non-linear compressed sensing methods can lead to superior tomographic reconstruction. However, it is also observed empirically that compressed sensing reconstruction not always finds the global optimum of the non-linear compressed sensing cost function, indicating that in some cases, the theoretical requirements are not always met.

Our results are based on an iterative reconstruction method that extends recently introduced conjugate gradient hard thresholding algorithms to the non-linear setting. Using an efficient line search procedure, the efficient computational method only requires the specification of the global, unconstrained cost function, its gradient and a non-linear compressed sensing projection operator, which, in the simplest case, boils down to the standard sparse projection operator used often in traditional compressed sensing. For x-ray tomographic reconstruction, we show that it is of advantage to adapt this operator to enforce more appropriate constraints.

1 Introduction

Compressed sensing [1, 2, 3] assumes that an unknown signal \( \mathbf{x} \in \mathbb{R}^N \) is measured using a linear system modelled with a matrix \( \Phi \). Observations \( \mathbf{y} \in \mathbb{R}^M \) are a noisy version of the measurements

\[
\mathbf{y} = \Phi \mathbf{x} + \mathbf{e},
\]

where \( \mathbf{e} \) accounts for the measurement uncertainty. Whilst we assume that we know \( \Phi \), the difficulty lies in the fact that we want to estimate \( \mathbf{x} \) from \( M \ll N \) measurements. This is only possible if we place additional constraints on the signal \( \mathbf{x} \). Compressed sensing developed around the idea of sparse models. In a sparse model, \( \mathbf{x} \) is assumed to have few non-zero (or few significant) coefficients. Alternatively, \( \mathbf{x} \) might be sparse in some transform domain, such as the wavelet domain or Fourier domain, in which case we can absorb the inverse of the sparsifying transform into the measurement system matrix \( \Phi \). Let us thus assume that \( \mathbf{x} \) is (approximately) sparse, that is, \( \mathbf{x} \) has no more than \( K \) non-zero entries where \( K \ll N \).

Let us also assume that we measure the error \( \mathbf{y} - \Phi \mathbf{x} \) using the squared Euclidean norm \( \| \mathbf{y} - \Phi \mathbf{x} \|_2^2 \). The compressed sensing problem can then be stated as the minimisation of \( \| \mathbf{y} - \Phi \mathbf{x} \|_2^2 \) subject to the constraint that \( \mathbf{x} \) has no more than \( K \) non-zero entries. In general, this is known to be an NP hard computational problem. It is thus customary to address this optimisation either, using relaxation of the non-linear sparsity...
constraint [1] or through the use of greedy algorithms (e.g. [4]). Surprisingly, for some measurement system \(\Phi\), these methods provide near optimal solutions, even in a regime where we have far fewer measurements than observations.

2 Non-linear compressed sensing

Non-linear observations

The application of compressed sensing recovery to non-linear observation models has previously been studied in [5], where some theoretical results were studied and recovery guarantees derived. We here build on these ideas and apply CS techniques to a more concrete application.

Let us write the non-linear observation model as

\[
y = \Phi(x) + e,
\]

where the notation \(\Phi(\cdot)\) reminds us that \(\Phi\) can be a non-linear map. In [5] it was shown that for non-linear maps \(\Phi(\cdot)\) that are not ‘too non-linear’, sparse recovery is still possible whenever a non-linear version of the method in [4] is used. However, the theoretical requirements are very strict and do not always hold in applications of interest. For example, for our particular application, it is difficult to verify if these conditions hold. We thus here concentrate on empirical evidence that supports the use of compressed sensing techniques for signal recovery from non-linear observations.

Conjugate gradient based iterative hard thresholding for non-linear observation models

We here use an advanced version of the recovery algorithm proposed in [5] for non-linear compressed sensing problems. Going back to ideas in [6], Blanchard et al. [7] recently studied conjugate gradient based variations of the IHT algorithm [4] and we here use a version of this for our non-linear observation setting.

The algorithm is developed so that it relies on two operations, for a given estimate \(\hat{x}\) it requires the calculation of a cost function \(f(\hat{x})\) and the evaluation of its gradient \(g(\hat{x})\). We here use

\[
f(\hat{x}) = \|y - \Phi(\hat{x})\|^2, \tag{3}
\]

\[
g(\hat{x}) = \nabla\|y - \Phi(\hat{x})\|^2. \tag{4}
\]

The algorithm is summarised as follows:

- Input: \(y\), \(\Phi(\cdot)\), \(x^0\), \(P\), \(f\), and \(g\)
- \(p^0 = 0\)
- : Iterate \((n = 1, n + +)\) until some stopping criterion is met:
  - \(T = \text{support}(x^{[n-1]})\)
  - if \(n = 1\) or \(T\) differs from support(\(x^{[n-2]}\))
    - \(\beta = 0\)
    - else
      - \(\beta = \frac{\|g(x^{[n-1]}(T))\|^2}{\|g(x^{[n-2]}(T))\|^2}\)
  - \(d = -0.5g(x^{[n]}) + \beta * p^{[n-1]}\)
  - \(a = \text{linesearch}(f, g, d, T)\)
  - \(x^{[n]} = P(x^{[n-1]} + ad)\)
  - \(p^{[n]} = d\).

where linesearch() is a function that performs a line search in direction \(d\) and where \(P(\cdot)\) is a projection operator, such as a hard thresholding operation that sets all but the largest \(K\) (in magnitude) elements of \(x\) to zero. Note that the notation \(g(x^{[n-1]}(T))\) is used here to denote the sub-vector of \(g(x^{[n-1]})\) with elements defined by the current support \(T\).
This algorithm differs from that in [7] in the line search used. For the linear model, the optimal step size $a$ can be found explicitly, whilst in our case, linesearch() uses a quadratic approximation to the cost function along direction $d$ restricted to the support $T$. The step size $a$ used is that which achieves the minimum of this quadratic function.

### 3 The non-linear x-ray model

We assume that x-rays are emitted from a point source, travel along a straight line, are attenuated by interactions with matter on their flight and have their intensity measured by a one or two dimensional sensor array. Keeping the x-ray source and detector fixed, but rotating the object under investigation, several of these x-ray ‘projections’ are acquired. The goal is the reconstruction of the spatial distribution of x-ray attenuation (see for example [8]).

For a line $r$ in space between the x-ray source and the location where we measure x-ray intensity, we can model x-ray attenuation using the model

$$I(r) = \int I_0(E) \exp \left( - \int_r \mu(z, E) dz \right) dE,$$

where the inner integral is along line $r$ and $\mu(z, E)$ is the x-ray attenuation at location $z$ and energy $E$. $I_0(E)$ is the intensity of the x-ray with energy $E$ emitted by the source along the path $r$.

If the x-ray source emitted photons at a single energy, then the model becomes linear after a simple transformation. This linear model is often a good first approximation and most x-ray tomography reconstruction algorithms are based on it. However, as most commercially used x-ray sources produce a spread of photon energies, the linear model is only approximate and its use can introduce artefacts which in some settings can lead to poor image quality. A standard approach to address these issues is to model the non-linear effect by quantisation of the x-ray photon energies which allows us to replace the integration with a summation.

$$I(r) = \sum_i I_0(E_i) \exp \left( - \int_r \mu(z, E_i) dz \right).$$

One of the drawbacks of this approximation is that now, source intensity within an energy band $I_0(E_i)$ is much harder to estimate. The other issue is that the model is now non-linear and that attenuation $\mu(z, E_i)$ needs to be estimated for each energy band.

### Spatial discretisation

For numerical purposes, a further discretisation is required, namely the discretisation of the inner integral along the different x-ray paths. This is done typically using a spatial basis function model to approximate $\mu(z, E_i)$. We here use the simplest approach in which we approximate $\mu(z, E_i)$ using pixels (in 2D) or voxels (in 3D) that have constant attenuation. To be consistent with the notation in our discussion of compressed sensing, we use the vector $x(E_i)$ to describe the pixel/voxel attenuation (at energy level $i$) in the image or volume (where we stack all pixels/voxels into a vector using some pre-specified order). The integral over each line $r$ can then be approximated as an inner product $a^T x(E_i)$, where the vector $a$ accounts for the length of the ray through each of the pixels/voxels. Finally, assuming that we stack all the intensities observed at the detector and at all rotation angles into a vector $y$, we can write the model as

$$y = \log \sum_i I_0(E_i) \exp (-Ax(E_i)).$$

In the single energy model, we write the observations in a slightly different form, using

$$y = -\ln \frac{I(r)}{I_0} = Ax.$$
Constraints

In the non-linear model, the vectors $x(E_i)$ can be stacked into a data matrix $X$. Different constraints on $X$ are then feasible and can be enforced (approximately) using different "projections". We explore several constraints below.

1. Positivity: Elements in $X$ are constrained to be positive. $P(x)$ sets all negative values in $X$ to zero.

2. Wavelet sparsity: Columns in $X$ represent images, which are transformed using the 2 dimensional wavelet transform, whose coefficients are stacked into a matrix $W$. Columns in $W$ are the wavelet coefficients for each energy level. $P(X)$ sets entire rows in $W$ to zero depending on the mean squared sum of the row's coefficients. After an inverse wavelet transform of each column in $X$, $X$ is furthermore restricted to be non-negative.

3. Low-rank: Matrix $X$ is modelled as a low rank matrix, where each row is modelled as the sum of few material absorption "fingerprints". $P(X)$ is a singular value decomposition followed by a thresholding of the singular values.

4 Empirical study

To evaluate the use of the non-linear compressed sensing approach, we generated simulated x-ray projections using the non-linear model above, with the difference that the data was generated on a different spatial quantisation grid from that used in the reconstruction (we rotated the original image and x-ray source and detector location by $\pi/4$ relative to the spatial grid). We used the Shepp-Logan phantom in 2 dimensions, quantised to a 128 by 128 pixel grid, with each grey level defining a different mixture of three materials. We quantised the energy levels into ten bands and assumed knowledge of $I_0(E_i)$, which we generated randomly as a decreasing vector with the steps between energy bands drawn from a uniform distribution. For each of the three materials, material absorption within each of the energy bands was generated in the same way. We did not add observation noise.

A fan beam geometry with linear line array was simulated with 32 rotations. We compared different reconstruction approaches based on the linear equation (8) and the non-linear equation (7). For the linear model we used the Algebraic Reconstruction Technique (ART), our conjugate gradient solver with sparsity in the Haar wavelet domain (WAVELET) and a TV regularised reconstruction [9]. (ART) was initialised with an empty image, (WAVELET) was initialised with the ART reconstruction, and all other algorithms were initialised with the (WAVELET) reconstruction. All reconstructions were constrained to be positive.

All non-linear reconstructions used our conjugate gradient based algorithm, but differed in the constraints used (i.e. we used different projection operators $P$). The first two methods did not enforce sparsity but used positivity (positive) and an additional low-rank constraint on the matrix $X$ (low-rank). We also used a combination of wavelet sparsity and positivity (wavelet) and a combination of low-rank, wavelet sparsity and positivity (wavelet + lr).

The results are analysed in terms of their SNR. For the reconstruction based on the linear models, we used a single estimate of $x$ that was compared to the mean (over energy levels) of the true absorption coefficients, whilst for the non-linear reconstruction, we got an estimate for each energy level, so compared these directly to the attenuation coefficients for that energy level.

The results for 10 independent runs of the experiment are shown in the box plots in Figure 1. Two main observations are in order. Firstly, the linear reconstruction is generally worse than the reconstruction based on the non-linear model with similar constraints, e.g. the two wavelet sparsity constraint reconstructions differ by about 2.5dB, with the non-linear model based reconstruction outperforming the linear model. Secondly, wavelet sparsity clearly improves performance, but so does the low-rank constraint used in the non-linear model, which basically enforces each of the regions to be a mixture of three different absorption profiles.

References

Figure 1. Comparision of the different methods in terms of reconstruction SNR for linear (names CAPITALISED, first three methods) and non-linear (names in lower-case, four methods on the right) models. Linear reconstruction with ART, and conjugate gradient reconstruction imposing wavelet sparsity (WAVELET) and total variation regularisation (TV). Non-linear reconstruction, also with a conjugate gradient solver, imposing positivity (positive), wavelet sparsity (wavelet), and/or low-rank (low-rank or +lr) constraints. Non-linear reconstruction clearly outperforms comparable linear methods.

[9] Li C, Yin W, Jiang H and Zhang Y 2013 Computational Optimization and Applications 56(3) 507–530

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, book title and publisher. Published under licence by IOP Publishing Ltd.
In this paper, we show how sparsity can be used to solve inverse problems. We first recall that many inverse problems involve the reconstruction of continuous-time or continuous-space signals from discrete measurements and show how to relate the discrete measurements to some properties of the original signal (e.g., its Fourier transform at specific frequencies). Given this partial knowledge of the original signal, we then solve the inverse problem using sparsity. We focus on two specific problems which have important practical implications: localisation of diffusion sources from sensor measurements and reconstruction of planar domains from samples. First, we show how to reconstruct specific planar domains which are driven by sparsity models, then we localise diffusion sources using a variation of the ‘reciprocity gap’ method and use it also to estimate the activation time of the source.

**Keywords:** Prony’s method, sparsity, analytic functions

## 1 Introduction

The notion of sparsity, namely the idea that signals can be modelled using a small number of free parameters has proved useful in many signal processing applications and recently sparsity has been successfully used in sampling. In these new sampling methods, the prior that the signal is sparse in a basis or in a parametric space is used to perfectly reconstruct classes of non-bandlimited signals from a set of suitable measurements. Depending on the set-up and reconstruction method involved, the above sparse sampling problem goes under different names like compressed sensing, compressive sampling [1, 2] or sampling signals with finite rate of innovation (FRI) [3, 4].

Sampling can be seen as a particular type of inverse problem where one tries to reconstruct a certain phenomenon or function from a set of discrete measurements. There are two types of inverse problem of this nature that we consider in this paper.

We first put ourselves in the typical sampling setup depicted in Fig. 1 where the original continuous-time signal \( x(t) \) is filtered before being (uniformly) sampled with sampling period \( T \). If we call \( y(t) = h(t) * x(t) \) the filtered version of \( x(t) \), the samples \( y_n \) are given by \( y_n = \left<x(t), \varphi(t/T-n)\right> \) where the sampling kernel \( \varphi(t) \) is the scaled and time-reversed version of \( h(t) \). In this paper we discuss extensions of this framework to the two-dimensional (2-D) case, we thus assume that the incoming signal is a 2-D function \( f(x, y) \) and try to reconstruct it from the discrete measurements \( y_{m,n} = \left<f(x, y), \varphi(x/T-m, y/T-n)\right> \).

The second inverse problem we consider is depicted in Fig. 2. Here a sensor network is monitoring a diffusion field inside a region \( \Omega \) and the task is to reconstruct the entire field from the spatio-temporal measurements given by the sensors under the assumption that the field is driven by \( M \) localised diffusion sources.

In both cases we solve the inverse problem by retrieving some continuous full-field information about the original signal/phenomenon and then reconstruct them using proper sparsity priors. Our methods are heavily influenced by the theory of sampling FRI signals introduced in [3, 4] and extended more recently.
Inverse Problems – from Theory to Applications (IPTA2014)

Figure 1. Sampling set-up. Here, \( x(t) \) is the continuous-time signal, \( h(t) \) the impulse response of the acquisition device and \( T \) the sampling period. The measured samples are \( y_n = \langle x(t), \varphi(t/T - n) \rangle \).

Figure 2. Estimation of diffusion fields driven by localised sources using a sensor network.

in [5, 6, 7, 8, 9, 10, 11]. FRI sampling theory has also had impact in other applications such as image super-resolution [12], for depth sensing [13], for calcium transient detection [14] and in compression [15, 16, 17].

In what follows, we first discuss the problem of reconstructing 2-D domains from samples then in Section 3 we provide an overview on how to reconstruct diffusion field from sensor measurements.

2 Reconstructing classes of 2-D domains from discrete measurements

For the sake of clarity, we begin by considering the 1-D case and the sampling set-up of Fig. 1. We want to show how we can retrieve some information about the Fourier transform of \( x(t) \) from the samples \( y_n \). The acquisition device or sampling kernel plays a central role in this context and a family of kernels that we will be considering is the family of exponential reproducing functions. A function \( \varphi(t) \) is an exponential reproducing function of order \( P \), if together with its shifted versions, it is able to reproduce exponentials

\[
\sum_{n \in \mathbb{Z}} c_{m,n} \varphi(t - n) = e^{j\omega_n t},
\]

for proper coefficients \( c_{m,n} \), with \( m = 0, 1, \ldots, P \) and \( \omega_m \in \mathbb{R} \). For the sake of argument, we are restricting our discussion to exponentials with purely imaginary exponents, however, the analysis can be extended to exponentials with arbitrary complex exponents. It is possible to show that a function satisfies (1) if and only if it meets the generalised Strang-Fix conditions [18]:

\[
\hat{\varphi}(j\omega_m) \neq 0 \quad \text{and} \quad \hat{\varphi}(j\omega_m + j2\pi l) = 0 \quad l \in \mathbb{Z} \setminus \{0\}
\]

where \( \hat{\varphi}(\cdot) \) is the Fourier transform of \( \varphi(t) \).

Exponential reproducing kernels are important because they allow us to map the samples \( y_n \) with the Fourier transform of \( x(t) \) at \( j\omega_m \) \( m = 0, 1, \ldots, P \) and this independently of the property of the incoming signal.
Figure 3. Sampling and reconstruction of classes of 2-D domains. Here, in (a) the original domain is acquired using exponential reproducing kernels. This leads to the samples of part (b). From these samples the Fourier transform of the original signal at specific frequencies is obtained and then the domain is perfectly reconstructed as shown in part (c). We refer to [19] for more details.

For the sake of clarity, assume that the signal $x(t)$ has compact support such that it is characterised by only $N$ non-zero samples. Moreover, assume that $T = 1$. We thus have that the $N$ samples are of the form $y_n = \langle x(t), \varphi(t-n) \rangle$, $n = 0, 1, \ldots, N - 1$.

We now linearly combine the samples $y_n$ using the coefficients $c_{n,m}$ of Eq. (1) to obtain:

$$s_m = \sum_{n=0}^{N-1} c_{m,n} y_n$$

(a) $\langle x(t), \sum_{n=0}^{N-1} c_{m,n} \varphi(t-n) \rangle$

(b) $\int_{-\infty}^{\infty} x(t) e^{j\omega_m t} dt$, $m = 0, 1, \ldots, P$,

where (a) follows from the linearity of the inner product and (b) is due to Eq. (1) and to the fact that $x(t)$ has compact support.

We note that $\int_{-\infty}^{\infty} x(t) e^{j\omega_m t} dt = \hat{x}(j\omega_m)$ is precisely the Fourier transform of $x(t)$ evaluated at $j\omega_m$, $m = 0, 1, \ldots, P$.

The above derivation, therefore, shows that it is possible to obtain a partial knowledge of the continuous-time Fourier transform of the original signal from proper discrete samples. A similar derivation can be applied to the 2-D scenario showing that a partial knowledge of the Fourier transform of $f(x, y)$ can be obtained from the samples $y_{m,n}$.

We are now faced with the more traditional problem of estimating the entire signal from this partial knowledge. This can be achieved by assuming that the original signal is sparse in a proper domain. In [19], we introduced a class of 2-D domains whose contour can be described using a small number of parameters. These domains are therefore sparse and an example of how perfect reconstruction can be achieved from the samples is shown in Fig. 3.

3 Inversion of Diffusion Fields

The propagation of a diffusion field follows the diffusion equation:

$$\frac{\partial}{\partial t} u(x, t) = \mu \nabla^2 u(x, t) + f(x, t),$$

(2)

where $u(x, t)$ is the field and $\mu$ is the diffusivity of the medium through which the field propagates. We assume that the field is monitored by a sensor network over a 2-D region $\Omega$ as shown in Fig. 2. Moreover we assume the field is generated by $M$ sources localised in space and time. Therefore the source can be written as follows:

$$f(x, t) = \sum_{m=1}^{M} c_m \delta(x - \xi_m, t - \tau_m).$$

(3)
Given the above assumption, the inversion problem reduces to the problem of retrieving the location and activation time of the sources from the sensor measurements. This is because once \( f(x,t) \) has been estimated, the field \( u(x,t) \) can be obtained by convolving \( f \) with the heat kernel. The problem is therefore sparse, because the whole field is driven by a small number of free parameters.

As in the previous application, we want to estimate some full-field measurements of the diffusion field from the spatio-temporal sensor readings. In the previous case, we used the exponential reproduction formula to have an exact mapping between the samples and the Fourier transform of the original signal at specific frequencies. In this new case, this is not possible and we can only obtain approximate full-field measurements. The aim is to estimate the following generalized measurements:

\[
Q(k) = \langle \Psi_k(x)W(t) , f \rangle = \int_{\Omega} \int_{t} \Psi_k(x)W(t)f(x,t)dVdt,
\]

where \( \Psi_k(x) = e^{-k(x_1+jx_2)} \) and \( W(t) \) is a properly chosen window. By replacing (3) into (4), we obtain:

\[
Q(k) = \sum_{m=1}^{M} c_m W(\tau_m)e^{-k(\xi_{1,m}+j\xi_{2,m})}.
\]

This is a sum of exponentials and the source locations can then be estimated from this sum using Prony’s method - a method commonly used in array signal processing [20]. The activation times are estimated in a similar way [21]. It is possible to show [21] that these generalised samples can be obtained from the boundary and interior sensor measurements when \( \Psi_k \) is analytic, which is the case here. We refer to [21] for further details.

Acknowledgments

The research leading to these results has received funding from the European Research Council under the European Unions Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement Nr. 277800 (RecoSamp).

The research in this paper was done in collaboration with T. Blu and H. Pan from Chinese University of Hong Kong and J. Murray-Bruce from Imperial College London.

References


Minisymposium M16
Seismic Imaging
A Lie group associated to seismic velocity estimation

Fons ten Kroode

Shell Global Solutions International, Kessler park 1, 2288 GS Rijswijk, the Netherlands
E-mail: a.tenkroode@shell.com

Abstract: Asymptotic solution operators for the extended linearized acoustic wave equation and their inverses play a pivotal role in seismic imaging. These operators are parameterized by a velocity function and it is natural to study compositions of solution and inverse solution operators associated to different velocity functions. These compositions generate a Lie group of invertible Fourier Integral Operators acting on distributions on space-shift-extended image space. I will show that its infinitesimal generators are order one pseudo-differential operators, anti-symmetric up to highest order, whose symbols are related to travel time perturbations calculated routinely in reconnection travel time tomography. Using these generators, I will argue that the classical Differential Semblance Optimization functional used for seismic velocity estimation is asymptotically positive semi-definite in a neighborhood of the true velocity.

Keywords: Seismic inverse problem, velocity estimation, migration, demigration.

1 Introduction

Seismic inversion aims at reconstructing material parameters describing the earth’s interior from seismic data, which consist of measurements of seismic waves at the surface of the earth. The simplest model for seismic wave propagation is the constant density acoustic wave equation

\[ \frac{1}{c^2(\vec{x})} \frac{\partial^2 p(\vec{x}, t)}{\partial t^2} - \Delta p(\vec{x}, t) = w(\vec{x}, t). \]  

(1)

Here \( w(\vec{x}, t) \) is a source function. Choosing \( w(\vec{x}, t) = \delta(\vec{x} - \vec{x}_s)\delta(t) \) and imposing causality, \( p = 0 \) for \( t < 0 \), leads to the causal fundamental solution \( G(\vec{x}, \vec{x}_s, t) \), which describes the response of the earth due to a point source at \( \vec{x} = \vec{x}_s \) at \( t = 0 \). In this simple setting the task at hand is to reconstruct the velocity function \( c(\vec{x}) \) from measurements of the pressure field \( p(\vec{x}_r, \vec{x}_s, t) \) depending on sources \( \vec{x}_s \) and receivers \( \vec{x}_r \) placed at the surface of the earth, i.e. for \( z_s = z_r = 0 \), at time \( t > 0 \). For \( w(\vec{x}, t) = \delta(\vec{x} - \vec{x}_s)w(t) \) the pressure field can be written as \( p = w * G \).

A fruitful approach is to assume a scale decomposition for the velocity, i.e.

\[ c(\vec{x}) = c_0(\vec{x}) + \delta c(\vec{x}), \]

(2)

where the so-called background velocity \( c_0(\vec{x}) \) is a smooth function which explains the kinematics of wave propagation and where \( \delta c(\vec{x}) \) represents the singularities of the velocity function, which are responsible for scattering of the wave field. Writing \( G_0 \) for the fundamental solution of the wave equation with velocity \( c_0 \) and defining \( \delta G := G - G_0 \), leads to the linearized wave equation

\[ \frac{1}{c^2(\vec{x})} \frac{\partial^2 \delta G}{\partial t^2} - \Delta \delta G = \frac{2\delta c(\vec{x})}{c_0^2(\vec{x})} \frac{\partial^2 G_0}{\partial t^2}. \]

(3)
In this set-up seismic imaging boils down to two steps, namely a non-linear one of estimating \( c_0 \) from the reflection data and a linear one of retrieving the singular perturbation \( \delta c \) by inverting the solution operator of equation (3) \( B[c_0] : V(\vec{x}) = 2\delta c(\vec{x})/c_0^2(\vec{x}) \rightarrow \delta G(\vec{x}_r, \vec{x}_s, t) \). This operator is called the Born operator. Its integral kernel is given by

\[
B[c_0] = \frac{1}{2\pi} \int d\omega \ e^{-i\omega t} (-i\omega)^2 \ G_0(\vec{x}_r, \vec{x}, \omega)G_0(\vec{x}, \vec{x}_s, \omega),
\]

where I have abused notation by using the same symbol for Green’s function and its temporal Fourier transform. The Born operator has two key properties, which hold under rather general conditions, namely (1) \( B[c_0] \) is a Fourier Integral Operator, see [8], and (2) its normal operator \( B^*[c_0]B[c_0] \) is an elliptic pseudo-differential operator, see [1, 6]. For Born data \( d = B[c_0]\delta c \) the solution of the linear problem is therefore simply given by

\[
\delta c = (B^*[c_0]B[c_0])^{-1}B^*[c_0]d.
\]

The much harder, non-linear problem of estimating \( c_0 \) is referred to as the velocity estimation problem. This problem has been studied for decades and all successful solution strategies have exploited the redundancy of the data, which, depending on acquisition, is parameterized by up to five variables \( x_r, y_r, x_s, y_s \) and \( t \), whereas the velocity function only depends on 3 variables \( x, y \) and \( z \). In these strategies one partitions the data into 3-dimensional subsets labeled by one or two redundant variables and inverts the linear problem for each such subset. This will in general lead to solutions which depend on the redundant variable(s), unless the background velocity model is correct. Differential Semblance Optimization as originally formulated by Symes (see [4] and references therein) poses the velocity estimation problem as least squares minimization of the derivative of the solution of the linear inverse problem with respect to the redundant variable(s).

A more recent formulation of this idea, aiming at making it robust for complex media in which wave fields may develop caustics, is based on extending \( B[c_0] \) to an order 0 Fourier Integral Operator \( \tilde{B}[c_0] \) [3] acting on so-called space-shift extended images \( i(\vec{x}, \vec{h}) \) with integral kernel

\[
\tilde{B}[c_0] = \frac{1}{2\pi} \int d\omega \ e^{-i\omega t} (-i\omega)^2 \ G_0(\vec{x}_r, \vec{x} + \vec{h}, \omega)G_0(\vec{x} - \vec{h}, \vec{x}_s, \omega),
\]

Here \( \vec{h} = (h_1, h_2, 0)^t \) is called a horizontal subsurface offset vector. It can be shown that \( \tilde{B}[c_0] : i(\vec{x}, \vec{h}) \rightarrow d(\vec{x}_r, \vec{x}_s, t) \) is invertible. This means that for \( d = B[c_0]V \) we have \( \tilde{B}^{-1}[c_0]d = V(\vec{x})\delta(\vec{h}) \), i.e. the inverse of the extended Born operator acting on data focuses at \( \vec{h} = 0 \) and this wouldn’t have been the case if we had applied \( B^{-1}[c] \) on the data for \( c \neq c_0 \). This simple observation leads to a strategy for finding the background velocity \( c \) from the data: it should be such that \( \tilde{B}[c]^{-1}d \) focuses at \( \vec{h} = 0 \). This can be cast into an optimization problem by introducing an order 0 pseudo-differential operator \( A \) which annihilates \( \delta(\vec{h}) \) and minimizing the functional

\[
J[c] := \frac{1}{2} \ |A\tilde{B}^{-1}[c]d|^2.
\]

In this paper, I will formulate a number of results relevant for this problem. First of all, I will provide an explicit formula for the inverse of the extended Born operator \( \tilde{B}^{-1}[c] \). My second result pertains to the perturbation \( \delta \tilde{B}[c] \) under smooth perturbations of the background velocity. It states that \( \tilde{B}^{-1}[c]\delta \tilde{B}[c] \) is an order 1 pseudo-differential operator and provides an explicit expression for its principal symbol. This will enable me to describe the group generated by all operators of the form \( \tilde{B}^{-1}[c_1]B[c_0] \) in a neighborhood of the identity and to demonstrate that the functional (6) is positive semi-definite in a neighborhood of the true velocity. Finally, I will introduce a differential semblance form of (6) by changing coordinates from \( h_1, h_2 \) to subsurface reflection angles \( \theta, \phi \).

2 Explicit inverse of the extended Born operator

The inverse of the extended Born operator can easily be constructed from the inverse of the extended Kirchhoff operator \( \tilde{K} \), which I have introduced in [7]. Its integral kernel is given by

\[
\tilde{K} := \frac{1}{2\pi} \int d\omega \ e^{-i\omega t} G(\vec{x}_r, \vec{x} + \vec{h}, \omega)G(\vec{x} - \vec{h}, \vec{x}_s, \omega) \frac{\partial}{\partial z}.
\]
From equations (7) and (5) one sees that $\tilde{K}$ and $\tilde{B}$ are related through $\tilde{B} = \partial_t^2 \tilde{K} \partial_t^{-1}$. $\tilde{K}$ and $\tilde{B}$ are Fourier Integral operators of order -1 and 0 respectively. In [7] I have shown that the inverse of $\tilde{K}$ is to highest order given by $\tilde{K}^{-1} = \psi^{-1} I$, where $I : d(\vec{x}_r, \vec{x}_s, t) \mapsto r(\vec{x}, \vec{h})$ is given by

$$I := \frac{32}{pc^2(x)} \int d\omega (-i\omega) e^{i\omega t} G^s_{\alpha}(\vec{x} + \vec{h}, \vec{x}_r, \omega) G^*_{\alpha}(\vec{x}_s, \vec{x} - \vec{h}, \omega)$$

and $\psi$ is an order zero pseudo-differential operator with principal symbol

$$\text{sym}(\psi)(\vec{x}, \vec{h}, \vec{p}, \vec{q}) := \frac{s^2}{s_+ s_-} \left( s_+ \epsilon_{\alpha,3} + s_- \epsilon_{\beta,3} \right)$$

Here $s := 1/c$, $s_+ := s(\vec{x} + \vec{h})$, $s_- := s(\vec{x} - \vec{h})$ and $\epsilon_{\alpha,3}, \epsilon_{\beta,3}$ are the z-components of unit tangents $\vec{e}_\alpha$ and $\vec{e}_\beta$ to rays from $\vec{x} + \vec{h}$ and $\vec{x} - \vec{h}$ to receivers $\vec{x}_r$ and sources $\vec{x}_s$ at the surface $z = 0$ respectively. These rays are well defined through an invertible map $M : (\vec{x}_r, \vec{x}_s, t, k_r, k_s, \omega) \mapsto T^* T(\vec{x} + \vec{h}, \vec{p}, \vec{q}) \in T^* X$ and $\vec{e}_\alpha$ and $\vec{e}_\beta$ can be expressed as functions of $\vec{x}, \vec{h}, \vec{p}$ and $\vec{q}$ by

$$\vec{e}_{\alpha, h} := \frac{\vec{p}_h + \vec{q}}{2 \omega s_+}, \quad \vec{e}_{\beta, h} := \frac{\vec{p}_h - \vec{q}}{2 \omega s_-}$$

$$\epsilon_{\alpha, 3} := \sqrt{1 - (\vec{e}_{\alpha, h})^2}, \quad \epsilon_{\beta, 3} := \sqrt{1 - (\vec{e}_{\beta, h})^2},$$

where

$$\omega(\vec{x}, \vec{h}, \vec{p}, \vec{q}) := \frac{\text{sgn} p_3}{\sqrt{a}} \sqrt{b - \sqrt{b^2 - ac}},$$

$$a(\vec{x}, \vec{h}, \vec{p}, \vec{q}) := (s_+^2 - s_-^2)^2,$$

$$b(\vec{x}, \vec{h}, \vec{p}, \vec{q}) := (s_+^2 + s_-^2) p_3^2 + (s_+^2 - s_-^2) (\vec{p}_h \cdot \vec{q}),$$

$$c(\vec{x}, \vec{h}, \vec{p}, \vec{q}) := p_3^4 + p_3^2 (p_h^2 + q^2) + (\vec{p}_h \cdot \vec{q})^2.$$ (11)

\textbf{Result 2.1.} The inverse of the extended Born operator $\tilde{B}$ is given by

$$\tilde{B}^{-1} = SB^*(-\partial_z, \partial_x \partial_x^{-1}),$$

where $S$ is a symmetric, positive definite pseudo-differential operator of order 2 with symbol

$$\text{sym}(S) = \frac{32}{p_3} \omega \sqrt{b^2 - ac}.$$ (13)

\section{Infinitesimal perturbations and a Lie group}

From here on I will use slowness $s(\vec{x}) := 1/c(\vec{x})$ rather than velocity to parameterize operators. I will consider the operator $\delta B[s]^{-1} \delta B[s]$, where $\delta B[s]$ is the first order perturbation of $B[s]$ under a smooth perturbation $\delta s(\vec{x})$ of the background slowness model $s(\vec{x})$.

\textbf{Result 3.1.} The operator $P[s, \delta s] := \delta B[s]^{-1} \delta B[s]$ is a pseudo-differential operator of order 1 with symbol

$$\text{sym}(P)(\vec{x}, \vec{h}, \vec{p}, \vec{q}) = i \omega(\vec{x}, \vec{h}, \vec{p}, \vec{q}) \delta T(\vec{x}, \vec{h}, \vec{p}, \vec{q}),$$

where $\omega(\vec{x}, \vec{h}, \vec{p}, \vec{q})$ is defined in (11) and $\delta T(\vec{x}, \vec{h}, \vec{p}, \vec{q})$ is the travel time perturbation resulting from the smooth slowness perturbation $\delta s(\vec{x})$ along rays from $\vec{x} + \vec{h}$ to a receiver $\vec{x}_r$ and from $\vec{x} - \vec{h}$ to a source $\vec{x}_s$ determined by their unit tangents $\vec{e}_\alpha(\vec{x}, \vec{h}, \vec{p}, \vec{q})$ at $\vec{x} + \vec{h}$ and $\vec{e}_\beta(\vec{x}, \vec{h}, \vec{p}, \vec{q})$ at $\vec{x} - \vec{h}$ defined in equations (10). Moreover, $P$ is to highest order an anti-symmetric operator.

\textbf{Remark 3.1.} Notice that the operator $P[s, \delta s]$ is non-linear in $s$ and linear in $\delta s$.
Now consider the group of operators generated by products of the form \( \tilde{B}^{-1}[s_1] \tilde{B}[s_2] \), where \( \Delta s := s_2 - s_1 \) is a finite slowness perturbation. Using terminology from seismic imaging, I will refer to this group as the migration-demigration group. The curve
\[
g_{12}(t) := \tilde{B}[s_1 + t\Delta s]^{-1} \tilde{B}[s_1], \quad 0 \leq t \leq 1,
\]
connects the element \( \tilde{B}[s_2]^{-1} \tilde{B}[s_1] = g_{12}(1) \) of this group to the identity \( 1 = g_{12}(0) \). Using result 3.1 and remark 3.1, one easily derives
\[
\frac{dg_{12}(t)}{dt} = -\tilde{B}[s_1 + t\Delta s]^{-1} \left( \frac{d}{dt} \tilde{B}[s_1 + t\Delta s] \right) \tilde{B}[s_1 + t\Delta s]^{-1} \tilde{B}[s_1]
\]
\[
= -\tilde{B}[s_1 + t\Delta s]^{-1} \lim_{\epsilon \to 0} \epsilon^{-1} \left( \tilde{B}[s_1 + (t + \epsilon)\Delta s] - \tilde{B}[s_1 + t\Delta s] \right) g_{12}(t)
\]
\[
= -P(s_1 + t\Delta s, \Delta s) g_{12}(t).
\]

The second argument of \( P \) in this formula is a finite slowness perturbation. The symbol of \( P(s_1 + t\Delta s, \Delta s) \) is given by formula (14) with \( \delta T \) replaced by \( \Delta T \), the travel time perturbation due to the slowness perturbation \( \Delta s \). This travel time perturbation is calculated by integrating \( \Delta s \) over ray trajectories in the slowness model \( s_1 + \Delta s \).

Writing \( P_{12}(t) := P(s_1 + t\Delta s, \Delta s) \), the differential equation for \( g_{12}(t) \) takes the compact form
\[
\frac{dg_{12}(t)}{dt} = -P_{12}(t) g_{12}(t).
\]

This shows that the operators \( P_{12}(0) = P(s_1, \Delta s) \) are infinitesimal generators for the migration-demigration group. The following theorem is a classical result by Magnus [2].

**Theorem 3.1.** Let \( g_{12}(t) \) satisfy (17), then \( \forall t \in [0, T) \) such that \( \int_0^T dt' ||P_{12}(t')|| < \pi \) there exists an order 1 pseudo-differential operator \( \Omega_{12}(t) \), anti-symmetric up to highest order, such that
\[
g_{12}(t) = e^{\Omega_{12}(t)}.
\]

Explicitly, \( \Omega_{12} = \sum_{n \geq 1} \Omega_{12}^{(n)} \), where \( \Omega_{12}^{(n)} \) is defined recursively using the Bernoulli numbers \( B_j \):
\[
\Omega_{12}^{(1)}(t) = -\int_0^t d\tau P_{12}(\tau),
\]
\[
\Omega_{12}^{(n)}(t) = -\sum_{j=1}^{n-1} \frac{B_j}{j!} \sum_{\substack{k_1 + k_2 + \ldots + k_j = n-1 \cr k_1 \geq 1, k_2 \geq 2, \ldots, k_j \geq 1}} \int_0^t d\tau \left[ \Omega_{12}^{(k_1)}(\tau), \Omega_{12}^{(k_2)}(\tau), \ldots \right]
\]
\[
\ldots \left[ \Omega_{12}^{(k_j)}(\tau), P_{12}(\tau) \right], \quad (n \geq 2).
\]

**Remark 3.2.** Using result 2.1 it is clear that the modification \( \tilde{B}[s] := \tilde{B}[s] S[s]^{1/2} \) leads to a group of unitary operators generated by \( \tilde{B}^{-1}[s_1 + t\Delta s] \tilde{B}[s_1] \), which can be written as the exponential of an order 1 pseudo-differential operator \( \Omega_{12}(t) \), which is antisymmetric for all orders.

## 4 Asymptotic convexity of velocity estimation

I now return to the functional (6), assuming data of the form \( d(\vec{x}, \vec{x}; t) := (w \ast \tilde{B}[s_0] r_0)(\vec{x}, \vec{x}; t), r_0(\vec{x}, \vec{h}) := V(\vec{x}) \delta(\vec{h}) \). The connection with the migration-demigration group is immediate as
\[
\tilde{B}[s]^{-1} d = \tilde{B}[s]^{-1} \tilde{B}[s_0] \tilde{B}[s_0]^{-1} w \ast \tilde{B}[s_0] r_0 = \tilde{B}[s]^{-1} \tilde{B}[s_0] w r_0.
\]

It is easy to see that \( \hat{w} := \hat{B}[s_0]^{-1} w \ast \hat{B}[s_0] \) is given by
\[
(\hat{w} \hat{r})(\vec{q}, \vec{h}) = \int d^3 p d^2 q e^{-i\langle \vec{p}, \vec{q} \rangle} e^{-i\vec{q} \cdot \vec{h}} \hat{w}(\omega(\vec{x}, \vec{h}, \vec{p}, \vec{q})) \hat{r}(\vec{p}, \vec{q}),
\]
where \( \hat{w} \) is the Fourier transform of \( w \). Notice that for band limited signals \( \hat{w} \) is a pseudo-differential operator of order \( -N \) for all \( N > 0 \). Consequently, \( \hat{w} \hat{r} \) is a square integrable function.
Result 4.1. In a neighborhood of the true model \( s_0 \), the functional (6) can be approximated up to second order in \( \Delta s := s - s_0 \) by

\[
J[s] = \frac{1}{2} \left\| Q(s_0, \Delta s) \hat{w}r_0 \right\|^2 + \cdots ,
\]

(22)

where \( Q(s_0, \Delta s) := [\mathcal{A}, P(s_0, \Delta s)] \) is a symmetric, order 0 pseudo-differential operator and where the dots represent terms bounded by \( C \left| \mathcal{A} \hat{w}r_0 \right| \) for some \( C > 0 \). The latter become insignificant if we increase the bandwidth of our source function, because then \( \hat{w}r_0 \to r_0 = V(\tilde{x})\delta(\tilde{h}) \), so \( \mathcal{A} \hat{w}r_0 \to 0 \). Consequently, \( J[s] \) is asymptotically positive semi-definite.

5 A differential semblance form

Let \( R : i(\tilde{x}, \tilde{h}) \to r(x, \theta, \phi) \) be the Fourier Integral Operator with integral kernel

\[
R := \frac{1}{(2\pi)^3} \int d^3p \frac{p_3}{|\tilde{p}|} e^{-ip_3\tilde{k}(\tilde{p}, \theta, \phi)\cdot\tilde{h}},
\]

(23)

where

\[
\tilde{k}(\tilde{p}, \theta, \phi) := \frac{\tan \theta}{\sqrt{p_1^2 + p_2^2}} \left[ \cos \phi \left( \begin{array}{c} p_1 \\ p_2 \end{array} \right) + \frac{|\tilde{p}|}{p_3} \sin \phi \left( \begin{array}{c} -p_2 \\ p_1 \end{array} \right) \right].
\]

(24)

It is easy to check that \( R \) is invertible and by Egorov’s theorem \( RAR^{-1} \) is an order 0 pseudo-differential operator, acting on angle-azimuth dependent image gathers. I now choose a special vector-valued annihilator \( \mathcal{A} \), which in case of a smooth reflector passing through \( \tilde{x} \) corresponds to projecting \( \tilde{h} \) on its tangent plane at \( \tilde{x} \) and scaling it by the inverse of the cosine of the dip angle. This annihilator has a simple form in \( \theta, \phi \) coordinates.

Result 5.1. Let \( \mathcal{A} \) be defined by

\[
\mathcal{A} := \frac{1}{(2\pi)^5} \int d^3p d^2qe^{ip\cdot(x' - x)} e^{i\tilde{q}\cdot(\tilde{k}' - \tilde{k})} \hat{f}(\tilde{p}, \tilde{h}),
\]

\[
\hat{f}(\tilde{p}, \tilde{h}) := \frac{1}{\sqrt{p_1^2 + p_2^2}} \left( \begin{array}{cc} 1 & 0 \\ 0 & |\tilde{p}|/p_3 \end{array} \right) \left( \begin{array}{cc} p_1 \\ p_2 \\ -p_2 \\ p_1 \end{array} \right) \left( \begin{array}{cc} h_1 \\ h_2 \end{array} \right).
\]

Then

\[
RAR^{-1} = \left( \begin{array}{cc} \cos \phi & -\sin \phi \\ \sin \phi & \cos \phi \end{array} \right) \left( \begin{array}{cc} \cos^2 \theta & 0 \\ 0 & \cot \theta \end{array} \right) \left( \begin{array}{cc} \partial_\theta \partial_\phi^{-1} \\ \partial_\phi \partial_\phi^{-1} \end{array} \right).
\]

(26)

Acknowledgements

I would like to thank Bill Symes for inspiring discussions during which we discovered we had each formulated result 3.1 independently, see his paper in these proceedings [5].

References

Minisymposium M17

Inverse Problems in Astronomy
Alternating minimization for Poisson blind deconvolution in astronomy

M Prato\textsuperscript{1}, S Bonettini\textsuperscript{2}, A La Camera\textsuperscript{3}, and S Rebegoldi\textsuperscript{1}

\textsuperscript{1}Dipartimento di Scienze Fisiche, Informatiche e Matematiche, Università di Modena e Reggio Emilia, Via Campi 213/b, 41125 Modena, Italy
\textsuperscript{2}Dipartimento di Matematica e Informatica, Università di Ferrara, Via Saragat 1, 44122 Ferrara, Italy
\textsuperscript{3}Dipartimento di Informatica, Bioingegneria, Robotica e Ingegneria dei Sistemi, Università di Genova, Via Dodecaneso 35, 16145 Genova, Italy

E-mail: marco.prato@unimore.it

Abstract: Although the continuous progresses in the design of devices which reduce the distorting effects of an optical system, a correct model of the point spread function (PSF) is often unavailable and in general it has to be estimated manually from a measured image. As an alternative to this approach, one can address the so-called blind deconvolution problem, in which the reconstruction of both the target distribution and the model is performed simultaneously by considering the minimization of a fit-to-data function in which both the object and the PSF are unknown. Due to the strong ill-posedness of the resulting inverse problem, suitable a priori information are needed to recover a meaningful solution, which can be included in the minimization problem under the form of constraints on the unknowns. In this work we consider a recent optimization algorithm for the solution of the blind deconvolution problem from data affected by Poisson noise, and we propose a strategy to automatically select its parameters based on a measure of the optimality condition violation. Some numerical simulations on astronomical images show that the proposed approach allows to provide reconstructions very close to those obtained by manually optimizing the algorithm parameters.

Keywords: blind deconvolution, gradient projection methods, astronomical imaging

1 Blind deconvolution

Blind deconvolution occurs when the measured data $g$ is the convolution between an unknown target $x$ and an unknown point spread function (PSF) $h$ \cite{1}. If object and PSF are two-dimensional distributions (stacked in two $n$-vectors) and the registration step introduces Poisson noise on the measured image, blind deconvolution can be reformulated as the constrained optimization problem

$$\min_{x \in \Omega_x, h \in \Omega_h} KL(x, h),$$

where $\Omega_x$ and $\Omega_h$ are the feasible sets of $x$ and $h$, respectively, $KL(x, h)$ is the generalized Kullback–Leibler divergence \cite{2} defined as

$$KL(x, h) = \sum_{i=1}^{n} \left\{ g_i \log \left( \frac{g_i}{(x \otimes h + b)_i} \right) + (x \otimes h + b)_i - g_i \right\},$$

$b$ is a (known) background radiation and $\otimes$ denotes the convolution operator.

Problem (1) is particularly challenging, since the objective function is nonconvex and the large size of the images makes the use of second order methods impractical. In a recent paper \cite{3} we adapted an alternating minimization method proposed by Bonettini \cite{4} for a general smooth objective function under separable
constraints to the blind deconvolution problem and we applied the resulting scheme to the reconstruction of astronomical images. In particular, the sets of constraints driven by the application were

\[ \Omega_x = \{ x \in \mathbb{R}^n \mid x \geq 0 \}, \]
\[ \Omega_h = \{ h \in \mathbb{R}^n \mid 0 \leq h \leq s, \sum_{i=1}^{n} h_i = 1 \}, \]

being \( s \) an upper bound on the PSF deduced by the so-called Strehl ratio (SR) of a given adaptive optics (AO) system, i.e. the ratio of peak diffraction intensity of an aberrated versus perfect waveform which can be estimated from the telescope and image features (see e.g. [5]).

The alternating minimization strategy consists in starting from an initial guess \( (x^{(0)}, h^{(0)}) \) and defining \( (x^{(k)}, h^{(k)}) \) \( (k = 1, 2, \ldots) \) as approximate solutions of the two subproblems

\[
\min_{x \in \Omega_x} K L(x, h^{(k-1)}) \tag{3}
\]

\[
\min_{h \in \Omega_h} K L(x^{(k)}, h) \tag{4}
\]

obtained by performing \( N_{x}^{(k)} \leq N_{x} \) and \( N_{h}^{(k)} \leq N_{h} \) iterations of the scaled gradient projection (SGP) algorithm \( (N_{x}, N_{h} \in \mathbb{N} \text{ fixed}) \) [6, 7, 8, 9] - see Algorithm 1. Since SGP is exploited for the inexact solution of the subproblems, the algorithm has been called cyclic scaled gradient projection (CSGP) method. The proposed strategy resulted to be particularly attractive in the blind deconvolution of stellar fields since a) several numerical experiments showed that for that kind of images the choice \( N_{x}^{(k)} = 50 \) and \( N_{h}^{(k)} = 1 \) for all \( k \) leads to very accurate reconstructions of both the stars and the PSF, and b) the addition of a regularization term or the introduction of an early stopping rule is not required thanks to the sparseness of the KL minimizers [10, 11].

As concerns diffuse objects, both advantages fails, since the optimal inner iteration numbers are problem-dependent and an early stopping of the alternating procedure is mandatory to avoid an increase of the reconstruction errors. In the next sections we propose a possible rule to automatically select the inner \( (N_{x}^{(k)}, N_{h}^{(k)}) \) and outer \( (k) \) iteration numbers and we test the effectiveness of the resulting method in some of the numerical experiments performed in [3].

---

**Algorithm 1 Cyclic scaled gradient projection (CSGP) method**

Choose the starting point \( x^{(0)}, h^{(0)} \) and the inner iterations numbers \( N_{x}, N_{h} \geq 1 \).

**FOR** \( k = 1, 2, \ldots \) **DO THE FOLLOWING STEPS:**

**STEP 1.** Compute \( x^{(k)} \) with \( N_{x}^{(k)} \leq N_{x} \) SGP iterations applied to subproblem (3) starting from the point \( x^{(k-1)} \)

**STEP 2.** Compute \( h^{(k)} \) with \( N_{h}^{(k)} \leq N_{h} \) SGP iterations applied to subproblem (4) starting from the point \( h^{(k-1)} \)

**END**

---

2 Projected gradient norm decrease

The rule we adopted for the inner iterations is the one proposed in [4] in the non-negative matrix factorization (NMF) framework and is similar to that suggested in [12]. The result exploited to design the stopping rule is the fact that any limit point \( (x, h) \) of the sequence generated by CSGP is stationary, i.e. is a point in which the projected gradient

\[
\nabla^{P} KL(x, h) = (P_{\Omega_x}(x - \nabla_x KL(x, h)) - x, P_{\Omega_h}(h - \nabla_h KL(x, h)) - h)
\]
Table 1. Reconstruction errors with manual (M) and automatic (A) choice of the iteration numbers.

<table>
<thead>
<tr>
<th>Image</th>
<th>SR</th>
<th>ObjM</th>
<th>ObjA</th>
<th>PSF_M</th>
<th>PSF_A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crab</td>
<td>0.67</td>
<td>11%</td>
<td>13%</td>
<td>6.7%</td>
<td>11%</td>
</tr>
<tr>
<td>Galaxy</td>
<td>0.40</td>
<td>12%</td>
<td>14%</td>
<td>12%</td>
<td>18%</td>
</tr>
</tbody>
</table>

vanishes (here \(P_\Omega\) denotes the Euclidean projection operator onto \(\Omega\)) [4, 13]. It follows that a possible stopping rule for the inner iterations can be designed as a function of the relative projected gradient norm decrease, e.g. by choosing as an approximate solution of subproblem (3) the first inner iteration \(\ell\) for which \(x^{(k-1,\ell)}\) satisfies

\[
\|\nabla^{P}\text{KL}(x^{(k-1,\ell)}, h^{(k-1)})\| \leq \eta_x^{(k)},
\]

where the sequence \(\{\eta_x^{(k)}\}\) is initialized as \(\eta_x^{(0)} = 10^{-3}\|\nabla^{P}\text{KL}(x^{(0)}, h^{(0)})\|\) and defined by

\[
\eta_x^{(k)} = \begin{cases} 
0.1 \cdot \eta_x^{(k-1)} & \text{if } \eta_x^{(k-1)} \geq \|\nabla^{P}\text{KL}(x^{(k-1,1)}, h^{(k-1)})\|, \\
\eta_x^{(k-1)} & \text{otherwise}.
\end{cases}
\]

The stopping rule for the inner iterations \(\ell\) of subproblem (4) is analogous and is given by

\[
\|\nabla^{h}\text{KL}(x^{(k)}, h^{(k-1,\ell)})\| \leq \eta_h^{(k)},
\]

where again the sequence \(\{\eta_h^{(k)}\}\) is initialized as \(\eta_h^{(0)} = 10^{-3}\|\nabla^{h}\text{KL}(x^{(0)}, h^{(0)})\|\) and defined by

\[
\eta_h^{(k)} = \begin{cases} 
0.1 \cdot \eta_h^{(k-1)} & \text{if } \eta_h^{(k-1)} \geq \|\nabla^{h}\text{KL}(x^{(k)}, h^{(k-1,1)})\|, \\
\eta_h^{(k-1)} & \text{otherwise}.
\end{cases}
\]

In few words, the rationale behind the choice of the adaptive parameters \(\eta_x^{(k)}\) and \(\eta_h^{(k)}\) is to decrease the tolerance for the stopping criterion if satisfied at the first inner iteration, thus forcing SGP to perform at least two steps in each subproblem.

As concerns the number of outer iterations, we followed the suggestion in [4] and stopped the alternating algorithm when

\[
\|\nabla^{P}\text{KL}(x^{(k)}, h^{(k)})\| \leq 10^{-4}\|\nabla^{P}\text{KL}(x^{(0)}, h^{(0)})\|.
\]

3 Numerical experiments

We considered four datasets described in [3] and created by using the HST images of two astronomical objects (the crab nebula NGC 1952 and the spiral galaxy NGC 6946) and two AO-corrected PSFs with different SR (0.67 and 0.40) (see [3] for more details on the images generation). The original and corrupted images are shown in Figure 1. We initialized the alternating algorithm by choosing a constant image \(x^{(0)}\) with total flux equal to the background-subtracted measured image, and the autocorrelation of the ideal PSF of the telescope as \(h^{(0)}\). The relative Euclidean error \(\|h^* - h^{(0)}\|/\|h^*\|\) between the true PSF \(h^*\) and the initial one \(h^{(0)}\) is 32\% when SR=0.67 and 54\% when SR=0.40. The best reconstruction errors obtained by CSGP have been computed in [3] by manually tuning the inner and outer iterations and are reported in Table 1 (suffix ‘M’) together with those provided by the automatic rules described in the previous section (suffix ‘A’). Moreover, in Figure 2 we show the reconstructed images for all the four datasets.

From both the numerical values of the reconstruction errors and the pictures of the restored images we can conclude that the proposed rules for the early stopping of the inner subproblems and the outer iterations succeed in providing satisfactory performances, with resulting objects very close to the “optimal” ones. As concerns the PSFs, the reconstruction errors are slightly higher than those achieved with the optimal, manually tuned parameters choice, even if notable decreases with respect to the initial values are nevertheless obtained.
Conclusions

In this paper we extend a recent work on blind deconvolution from Poisson data by introducing an automatic tuning of the CSGP algorithm in the case of diffuse objects. The proposed rule is based on the decrease of the norm of projected gradient and has been already used in NMF problems. Some numerical tests showed that the automatic rule succeeds in providing reconstructions close to the optimal ones, even if with slightly worse PSFs. Future work will involve the introduction of suitable regularization terms and boundary effect corrections [14, 15], as well as the generalization of the approach to blind deconvolution problems with multiple images provided by a Fizeau interferometer [16].

Acknowledgments

This work has been partially supported by MIUR (Italian Ministry for University and Research), under the projects FIRB - Futuro in Ricerca 2012, contract RBFR12M3AC, and PRIN 2012, contract 2012MTE38N. The Italian GNCS - INdAM (Gruppo Nazionale per il Calcolo Scientifico - Istituto Nazionale di Alta Matematica) is also acknowledged.

References


Figure 1. Crab nebula (top row) and spiral galaxy (bottom row): true objects (left column) and blurred and noisy images (middle column: SR=0.68, right column: SR=0.40).
Figure 2. Crab nebula (top row) and spiral galaxy (bottom row): true objects (left column) and reconstructed images (middle column: SR=0.68, right column: SR=0.40).

[16] Bertero M, Boccacci P, La Camera A, Olivieri C and Carlbillet M 2011 Imaging with LINC-NIRVANA, the Fizeau interferometer of the Large Binocular Telescope: state of the art and open problems Inverse Probl. 27(11) 113011
Restoration of astrophysical images, statistical influence of the camera

Céline Theys and Henri Lantéri

Laboratoire Lagrange, CNRS/UMR 7293, Université de Nice Sophia, Observatoire de la Côte d’Azur

E-mail: celine.theys@unice.fr

Abstract: The restoration of astrophysical images is a classical application of inverse problems. The astronomical object is first blurred by the Point Spread Function of the instrument-atmosphere set. The resulting convolved image is corrupted by a Poissonian noise in reason to low light intensity, then a Gaussian white noise is added during the electronic read-out operation by the Charge Coupled Device (CCD) camera, leading to a "Poisson Gaussian" density. Two basic cases with positivity constraints have been first considered, either a pure Poisson noise leading to the so-called Richardson Lucy algorithm or a pure Gaussian additive noise corresponding mainly to infrared experiments with a high intensity level, leading to the so-called ISRA algorithm. The complete model corresponding to the Poisson Gaussian process has been also developed. A more recent technology proposes to acquire astrophysical data with Low Light Level CCD (L3CCD) cameras in order to avoid the read-out noise. The physical process leading to the data has been previously described by a "Poisson Gamma" density. We propose to discuss the statistical models including the read-out noise in the CCD and L3CCD cases. Results are given on synthetic astrophysical data chosen to discuss the interest of one technology over the other according to the considered experiments.

Keywords: L3CCD detectors, statistical models, deconvolution

1 Introduction

This paper deals with the restoration of astrophysical images. Generally and until recently, the astrophysical objects are acquired using charge coupled devices (CCD) cameras, leading to a read-out noise additive Gaussian on the data after photoconversion. For very low intensities data, the variance of the read-out noise is of the same order as most of the intensities of the image leading to unrecoverable distortion. Recently, the low light level charge coupled devices (LLLCCD or L3CCD) have been developed. The L3CCD's amplify the signal prior to the readout process resulting in a subelectron effective readout noise. This combined with high quantum efficiency make their use very interesting for interferometric detectors. Here we propose to study the effect that the use of an L3CCD would have on image restoration. A probability density model has been proposed in ([Basd04, Basd03]) which is the one commonly used until now. A slightly modified version has been proposed in ([They06]) and used for image restoration. However both of them neglect the residual read-out noise and we propose another model in order to include it, leading to a Poisson-Gamma-Gaussian (PGG) model. The objective of this paper is to study and compare qualitatively and quantitatively image restoration in the CCD acquisition versus the L3CCD. For the comparison, the complete model including the read-out noise will be used to construct the algorithm in both cases.

A brief description of the physical process is given in section 2. A PGG model for the output of the L3CCD is discussed in section 3. Section 4 resumes the iterative algorithm used and its application to the model of interest. Finally, some numerical results are shown in the section 5.
2 Optical astronomy imagery

The light emanating from the object of interest is propagated through a turbulent atmosphere and is focused onto the Charge Coupled Device (CCD or L3CCD) by an imperfect optical system that limits the resolution and introduces aberrations. The detector is read by an electronic process that adds a white Gaussian read-out noise:

$$x \xrightarrow{\text{convolution}} Hx \xrightarrow{\text{Poisson}} n \xrightarrow{\text{Gaussian}} z,$$

with

$$z = n + b, \quad n \sim \mathcal{P}(Hx), \quad b \sim \mathcal{N}(0, \sigma_{\text{CCD}}^2).$$

In the case of L3CCDs, the Poisson process $n$ is amplified, in a stochastic way, giving a process $y$ prior to read-out, giving a supplementary step in the previous scheme:

$$x \xrightarrow{\text{convolution}} Hx \xrightarrow{\text{Poisson}} n \xrightarrow{\text{Avalanche}} y \xrightarrow{\text{Gaussian}} z,$$

For more details of the data model of the CCD camera, see ([Nuñez93, Snyd85, Bert04, Lant05, They05, Benv08]).

3 Statistical description

The L3CCD camera has the same architecture than the CCD one but with an extended output register that creates avalanche multiplication, at each step an electron creates another electron with a probability $p$, $p$ is very small, between 0.01 and 0.02 but the number of elements of the register is very high leading to a mean gain between 400 and 1000. To obtain a model of the data in output the probability density of the avalanche process is needed. Basden and Haniff (2003) propose a Gamma distribution to describe it, in the following we discuss and justify this choice.

The avalanche process can be viewed as a kind of branching process, ([Harr72]). A branching process is a Markov process that models a population in which each particle in step $k$ produces some random number of particles in generation $k + 1$, according to a fixed probability distribution that does not vary from particle to particle. In our case, the number of particles at step 0, $X_0$, corresponds to the Poisson process $n$ and the number of particles at the output $X_K$ corresponds to the process $y$. At the intermediate step $k + 1$, the number of particles $X_{k+1}$ is equal to the number of particles at step $k$, $X_k$, plus as much Bernoulli variables as $X_k$, this can be mathematically described by:

$$X_{k+1} = X_k + B_1 + \ldots + B_{X_k}, \quad X_0 = n,$$

with $B_k$, Bernoulli r.v iid, $B_k \sim \mathcal{B}(p)$ where $p$ is the probability to create another particle. From this equation, the mean and the variance of $X_{k+1}$ can be computed. The mean is given by:

$$\mathbb{E}[X_{k+1}] = (p + 1)^{k+1} n = G(p + 1)n.$$

Then, we see clearly that the process $n$ is multiplied "in mean" by a gain $G = (p + 1)^k$. In the same way, we can compute the variance of $X_{k+1}$, given in the following equation:

$$\text{var}[X_{k+1}] = nq \left((p + 1)^{2k+1} - (p + 1)^k \right),$$

with $q = 1 - p$. For a large number of stages $k$ and $p$ very small ($q \approx 1$) the mean and the variance can be approximated by:

$$\mathbb{E}[X_{k+1}] \approx Gn \quad \text{var}[X_{k+1}] \approx nG^2.$$

Expressions of the mean and the variance allow to fit the output $y$ with a Gamma law of parameters $n$ and $G$:

$$y \sim \mathcal{G}(n, G) = \frac{y^{n-1}}{G^n \Gamma(n)} \exp(-y/G).$$

In order to verify numerically the validity of this proposition, we have plotted the Gamma probability and the result of the avalanche process for $p = 0.01$ and $K = 400$ for $n = 1, 2, 10$ on Fig. 1. Correspondance between
the avalanche process and the Gamma law is relatively good. In the experiments, data will be simulated with an avalanche process while the algorithm will be derived by using the Gamma law. Usually, the Gaussian noise is neglected in both the CCD and L3CCD cases. Here we propose to take it into account. In the following, subscript $i$ denotes the pixel $i$ of the image lexicographically ordered. The density probability of $z_i$ is:

$$p(z_i) = \sum_{y_i} p(z_i|y_i)p(y_i),$$  \hspace{1cm} (9)$$

with

$$p(y_i) = \sum_n p(y_i|n)p(n) = \sum_{n=1} p(y_i|n)p(n) + p(y_i|n = 0)p(n = 0),$$  \hspace{1cm} (10)$$

$p(y_i|n)$ is a Gamma distribution and $p(n)$ is a Poisson distribution, Eq. 3 then $p(y_i)$ is a "Poisson Gamma" distribution:

$$p(y_i) = \sum_{n=1} \frac{y_i^{n-1}\exp(-y_i/G)}{\Gamma(n)G^n} \frac{\exp(-(Hx)_i)(Hx)^n_i}{n!} + \exp(-Hx)\delta_{y_i,0}.$$  \hspace{1cm} (11)$$

The first term of Eq. 11 is the one proposed by Basden et al., ([Basd04]), the second term completes the distribution for $n = 0$. $p(z_i|y_i)$ is a Gaussian distribution, $p(z_i|y_i) = \mathcal{N}(y_i, \sigma^2)$. Finally $p(z_i)$ is a "Poisson Gamma Gaussian" distribution:

$$p(z_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-(Hx)_i) \left( \sum_{y_i} \exp\left(-\frac{(z_i - y_i)^2}{2\sigma^2}\right) \frac{y_i^{n-1}}{\Gamma(n)G^n} \sum_{n=1}^{\infty} \frac{(Hx)^n_i y_i^{n-1}}{n!}\right) + \exp\left(-\frac{z_i^2}{2\sigma^2}\right) \delta_{y_i,0}.$$  \hspace{1cm} (12)$$

4 Deconvolution algorithm

The problem is to restore the object $x$ from the data $z$ with the constraint $x \geq 0$ and the total intensity conservation, $H$ being generally obtained via separated calibration measurements. A classical solution is to derive an iterative algorithm founded on the Maximum Likelihood Estimation (MLE). From Eq. (12) and with assumption of independence between pixels, the negative log-likelihood for the image is:

$$J(x) = -\sum_i \log p(z_i; x).$$  \hspace{1cm} (13)$$

Then the MLE is obtained by minimizing $J(x)$ versus $x$ with constraints:

$$\min_x J(x) + \gamma J_2(x) \hspace{1cm} s.t \hspace{1cm} x_i \geq 0 \text{ and } \sum_i x_i = \sum_i z_i,$$  \hspace{1cm} (14)$$

where $J_2(x)$ is a possible regularization term. A solution is to construct an iterative algorithm from the gradient, the used method is the so-called SGM. The algorithm cannot be detailed here by lack of space but it has been developed and used in numerous papers, see for example ([Lant02, Lant01, Lant05, They05]).
After computations, the gradient of $J$, for the pixel $i$ can be written following eq. 15:

$$\nabla J(x)_i = \sum_j (h_{ji} - h_{ji}r_j), \quad r_j = \frac{p_j}{q_j}$$

(15)

with

$$p_j = \sum_{y_i} \left( \exp\left(-\frac{(z_j - y_i)^2}{2\sigma^2} - \frac{y_i}{G}\right) \sum_n \frac{n(Hx)^{n-1}y_i^{n-1}}{n!G^n} \right)$$

(16)

and

$$q_j = \sum_{y_i} \exp\left(-\frac{(z_j - y_i)^2}{2\sigma^2} - \frac{y_i}{G}\right) \sum_n \frac{(Hx)^n y_i^{n-1}}{n!G^n} + \exp\left(-\frac{(z_j)^2}{2\sigma^2}\right) \delta_{y_i,0}.$$  

(17)

Numerical computations show that limiting the sum over $y_i$ around to $z_i \pm 3\sigma$ is sufficient and in this case, an approached expression for $r_j$ can be found:

$$r_j \approx \frac{0F1[1, \frac{(Hx)_j z_j}{G}]}{0F1[2, \frac{(Hx)_j z_j}{G}]}$$

(18)

where $0F1()$ is the confluent hypergeometric function.

5 Numerical illustrations

and conclusion The proposed algorithm has been illustrated on a picture taken from the Hubble Space Telescope (HST) site, http://hubblesite.org/gallery/. It is a sun-like star nearing the end of its life, Fig. 2(a). The data have been blurred with a synthetic but realistic normalized space invariant PSF, Fig. 2(b). To stop the iterative procedure before noise amplification and/or to check the quality of the restoration process, we use the normalized squared Euclidean distance. Fig. 3 compares reconstructions by using a CCD detector, $\sigma^2 = 4$, versus a L3CCD detector, $\sigma^2 = 25$ for 1000 photons in the whole image. Fig. 3(a) is the CCD raw picture, Fig. 3(b) is the L3CCD raw picture. Figs. 3(c) and 3(d) are the best reconstructions respectively for the CCD and the L3CCD acquisitions. The mean and the standard deviation of the reconstruction error has been computed over 50 noise realizations as a function of the number of photons for CCD and L3CCD cases, fig. (4) and fig. (5) shows a result with a $L_1$ regularization term.

**Figure 2.** (a) Galaxy (b) Normalized PSF (c) Object (d) Result of the convolution  

**Figure 3.** (a) CCD detection, $\sigma^2 = 4$ (b) L3CCD detection, $\sigma^2 = 25$ (c) $k_{min} = 6$, $\epsilon_{min} = 0.70$ (d) $k_{min} = 1$, $\epsilon_{min} = 0.59$
6 Conclusion

An iterative reconstruction algorithm has been proposed to deconvolve Poisson-Gamma-Gaussian data, statistics derived from a L3CCD acquisition. First results on deconvolution of astrophysical images acquired with L3CCDs are given. They show mainly the effectiveness of such detectors for imaging in very low light level situations. These results compared with those obtained using data acquired with classical CCDs, emphasize the interest of such cameras in the case of very low intensity imagery.

References


Sparse point source removal for full sky CMB missions

F C Sureau, J-L Starck, J Bobin, P Paykari, A Rassat
Laboratoire AIM, UMR CEA-CNRS-Paris 7, Irfu, SAP/SEDI, Service d’Astrophysique, CEA Saclay, F-91191
GIF-SUR-YVETTE Cedex, France
E-mail: florent.sureau@cea.fr

Abstract: Planck and WMAP missions measure temperature fluctuations on the microwave and far infra-red range with high resolution, providing crucial information over the full sky on galactic, extra-galactic and cosmological signals. Source separation is required to disentangle these emissions and is one of the major scientific challenge in these missions. We focus in this work in estimating point-source contribution to the microwave data using morphological information. We propose a new method (SPSR, for Sparse Point Source Removal) to estimate their flux and subtract their contribution to the data. As in morphological component analysis (MCA), the sky on the sphere is modeled as a superposition of sparse signals in different bases or frames, and at different locations in the sky: background (diffuse) emissions, point sources, and galactic compact sources in the galactic plane. Additional constraints are also enforced such as positivity of the fluxes and background band-limited. Source separation is achieved by solving a sparse convex problem using a primal-dual algorithm. We compare this approach to a standard local low-order polynomial fitting on realistic simulations of the sky and show results on WMAP-9-year data.

Keywords: Cosmology : CMB, Data Analysis, Methods : Primal-dual

1 Introduction

Planck and WMAP missions measure temperature fluctuations on the microwave and far infra-red range with high resolution. Their primary goal is to measure the full-sky anisotropies of the Cosmic Microwave Background (CMB) so as to provide key information on the birth and evolution of our universe. They also provide crucial full-sky information on all emissions in wavelengths that have not been probed before, either originating from our galaxy (synchrotron, free-free, dust, galactic compact sources for instance) or emitted by other galaxies (extra-galactic sources), or of cosmological origin. Source separation is required to disentangle these signals and is one of the major scientific challenge in these missions.

In particular, bright point-source emissions can be detected in a significant fraction of the sky [1, 2], even in regions where the CMB is dominating other diffuse signals at high galactic latitudes. These emissions therefore need to be separated from diffuse emissions for accurate data analysis. Contrary to diffuse emissions, these sources display spectral variability and therefore classical source separation techniques, based on factoring spatial and multichannel information, cannot be used. A significant fraction of the radiosources also varies with time [2] and so their flux needs to be assessed independently for each dataset. The most discriminative information for these sources is based on morphology: point-sources are unresolved and the shape of there emission is given by the Point Spread Function (PSF) of the instrument at the wavelength considered. This information is already employed for source detection and flux estimation techniques [3, 4, 2].

In this work, we propose a new strategy to estimate the flux of the detected point-sources for point source removal, which is based on a morphological source separation method. In section 2, we describe this method along with the proposed primal-dual algorithm. Results on compact source removal in full-sky WMAP simulations and data are presented in section 3, where our approach is compared to the standard flux fitting or low-order polynomial background fitting as performed in WMAP and Planck [1, 2].
2 Sparse Point Source Removal

We consider each microwave channel data over the full sky \( y \) as composed of three components \( \{x_1, x_2, x_3\} \) all being square-integrable fields on the unit sphere \( S^2 \): the point-sources, the extended compact sources and the diffuse emissions, gathering CMB, synchrotron, free-free and dust emissions as well as low flux point-sources that cannot be detected. In the rest of the paper, we consider a finite dimensional setting: \( \{y, x_1, x_2, x_3\} \) are considered as column vectors in \( \mathbb{R}^{N_x} \). The forward model is therefore:

\[
y = x_1 + x_2 + x_3 + n,
\]

where \( n \in \mathbb{R}^{N_x} \) is an additive noise. The associated ill-posed inverse problem, composed of three unknowns for one equation, needs regularization to be solved. We first assume that point-sources have been detected and their position are available in catalogs (as in WMAP or Planck data). \( x_1 \) is modeled as \( x_1 = Bf \) where \( B \in \mathcal{M}_{N_x \times N_p}(\mathbb{R}) \) is the operator implementing the local projection of the beams at the position of the \( N_p \) sources, and \( f \in \mathbb{R}_+^{N_p} \) is the vector of fluxes enforced to be positive.

Morphological information is weaker for extended compact sources and we adopt a multi-scale model in the galactic region for them, as typically performed for detection. More precisely, we assume these emissions are sparse in an undecimated nearly-isotropic spherical wavelet dictionary\[5\]: \( x_2 = MWw \), where \( M \in \mathcal{M}_{N_x \times N_x}(\mathbb{R}) \) is a galactic mask, \( W \in \mathcal{M}_{N_x \times (N_x, N_N)}(\mathbb{R}) \) indicates the wavelet reconstruction operator and \( w \) is the vector of wavelet coefficients.

The background emission \( x_3 \) is usually assumed to be locally smooth, and modeled as a local low-order polynomial background (either baseline, or first order). In this work we consider a more flexible model to capture its fluctuation on \( S^2 \): we assume it is sparse in spherical harmonics (see \[6\]). Furthermore we assume \( x_3 \in D \), the set of band-limited function on the sphere with a maximal non-zero multipole \( \ell_m \). In the following, \( S \) denotes the spherical harmonic transform up to \( \ell_m \), and \( a = Sx_3 \) contains the \( N_\ell \) complex spherical harmonic coefficients of \( x_3 \).

The noise \( n \) is assumed to be a centered real-valued Gaussian random field (but non-necessarily stationary), with covariance matrix \( \Sigma \). From all these constraints, the resulting inverse problem is:

\[
\min_{f \in \mathbb{R}_+^{N_x}, w \in \mathbb{R}^{N_w}, a \in D} \gamma ||a||_1 + \beta ||w||_1 \; \text{s.t.} \; ||y - (Bf + MWw + Sa)||_{2, \Sigma} < \epsilon,
\]

where \( ||x||_{2, \Sigma} = x^T \Sigma^{-1} x \) denotes the square of the \( \ell_2 \) norm weighted by \( \Sigma^{-1} \). The reconstructed point source-free map \( \tilde{y} \) can be obtained by \( \tilde{y} = y - Bf \).

The problem described in Equation 2 is a convex problem, and is related to a constrained morphological component analysis (MCA) \[7\] or a basis pursuit denoising problem with a deconvolution step. We chose the constrained form of the inverse problem so as to have only a few hyperparameters to set: the noise level \( \epsilon \) which can be derived from noise statistics ; the trade-off between sparsity in wavelet and spherical harmonic domains (\( \gamma = \beta \) in this work as in classical MCA).

Convex optimization with a constrained formulation as described by Equation 2, can be performed using primal-dual approaches \[?\], e.g.]Chambolle2011,Becker11,Combettes12]. Our SParse Source Removal algorithm (SPSR) was derived from \[8\]: such scheme was chosen because it only requires one application of the costly spherical harmonic and wavelet transforms and one application of their adjoint per iteration and does not require sub-iterations. The algorithm reads as follows:
Morphological Component Analysis with a Primal-Dual approach (SPSR)

1- Choose \((a^0, f^0, w^0, t^0) \in \mathbb{C}^{N_1} \times \mathbb{R}^{N_p} \times \mathbb{R}^{N_e} \times \mathbb{R}^{N_r}\), with \(\tilde{a}^0 = a^0, \tilde{f}^0 = f^0, \tilde{w}^0 = w^0\). Set the parameters

\[
\gamma = \beta \text{ and } \tau, \sigma \text{ s.t. } \tau \sigma < \frac{1}{3} (||S||_2 = ||MW||_2 = ||B||_2 = 1).
\]

2- Iterate \((n \geq 0)\):

\[
\begin{cases} 
\mathbf{r}_d^n = \mathbf{t}^n + \sigma \Sigma^{-1/2}(B\tilde{f}^n + S\tilde{a}^n + MW\tilde{w}^n - y) \\
\mathbf{t}^{n+1} = \begin{cases} 
0 & \text{if } ||\mathbf{r}_d^n||_2 \leq \sigma \epsilon \\
(1 - \frac{\epsilon \sigma}{||\mathbf{r}_d^n||_2})\mathbf{r}_d & \text{otherwise}
\end{cases} \\
\mathbf{a}^{n+1} = ST_{\tau \beta}(P_D(a^n - \tau S^1\Sigma^{-1/2}t^{n+1})) \\
\mathbf{f}^{n+1} = P_{R_{N_p}}(\mathbf{f}^n - \tau B^1\Sigma^{-1/2}(t^{n+1})) \\
\mathbf{w}^{n+1} = ST_{\tau \beta}(\mathbf{w}^n - \mathbf{W}^1\mathbf{M}\Sigma^{-1/2}t^{n+1}) \\
\mathbf{f}^{n+1} = 2\mathbf{f}^{n+1} - \mathbf{f}^n, \tilde{a}^{n+1} = 2a^{n+1} - a^n, \tilde{w}^{n+1} = 2w^{n+1} - w^n
\end{cases}
\]

where \(P_{R_{N_p}}\) is the projection onto the positive orthant and \(P_D\) is the projection onto \(D\) the set of considered band-limited signals; \(ST_{\tau \beta}\) is the standard soft-thresholding operator applied component-wise:

\[
[ST_{\tau \beta}\mathbf{x}]_i = x_i \left[1 - \frac{\tau \beta}{|x_i|}\right]_+, \quad (4)
\]

where \(|x_i|\) is the complex modulus of \(x_i\) for complex vectors. From \([8]\), \((\mathbf{f}^n, \mathbf{a}^n, \mathbf{w}^n, \mathbf{t}^n)\) converges to a saddle point of the primal-dual problem with a restricted dual gap decreasing as \(O(1/n)\) (first order method). We initialise the algorithm with null images, set \(\tau = \sigma\) for the algorithm parameters and \(\epsilon\) correspondsto the 95-th percentile of a \(\chi^2(N_r)\) distribution according to the whitened noise statistics. We evaluate our proposed approach by comparing the flux estimates with SPSR to the low-order polynomial fitting approach \([9, 1, 10]\). Fluxes (enforced to be positive) were estimated in a 3.5\(\sigma\) region as previously recommended in \([2]\) using a Levenberg-Marquardt algorithm with local \(\chi^2\) minimization (using the C++ library ALGLIB (www.alglib.net)), with background modeled either as a baseline or a first order polynomial (respectively named FIT-C and FIT-L in the following).

3 Results on WMAP simulations and real data

The Planck Sky Model (PSM) software \([11]\) was employed to simulate WMAP-like data. For each of the five WMAP channels, we simulated diffuse components (Gaussian CMB, generated from a 6-parameter Λ-CDM model, synchrotron, free-free, thermal and spinning dust emissions), a compact component (mainly radiosources and infrared sources) and noise modeled as a non-stationary Gaussian random field, with variance in each pixel derived from the WMAP 9-year hit maps. The instruments were assumed to have a dirac bandpass and Gaussian beams with full-width at half maximum reflecting the WMAP instrument. An example of a point-source region in these full-sky simulations is shown in Figure 1. We also generated a catalog of point-sources by applying a matched filter in spherical harmonics to each channel, considering second order statistics of CMB and noise \([9, 12]\). A 5\(\sigma\) local threshold was applied in the filtered map for source detection, resulting in 689 (respectively 442, 415, 248, 172) sources for channel K (respectively Ka, Q, V and W). After merging, the final multichannel catalog contained 724 sources.

These simulations were first employed to set the value of \(\beta\), which has a strong influence on the convergence speed of SPSR as illustrated in Figure 2: large values lead to spend many iterations to reach the required residual level in the data fidelity constraint, and the sparsity penalty term is decreasing very slowly with low values. We chose for SPSR the intermediate value \(\beta = 0.01M\) that gave in this test the best convergence properties among the tested values. The convergence speed of the algorithm was also varying with the resolution in the channels; we chose 13350 iterations of SPSR for channel K and 9750 iterations for all other channels, as a compromise between processing time and level of convergence. We also ensured that the statistics measured were marginally changing with respect to the relative difference in between methods when reaching this number of iterations.
Figure 1. Patch of the simulated sky centered on two detected point sources at the WMAP wavelengths.

Figure 2. Normalized $\ell_2$ norm of the residuals (left) and normalized $\ell_1$ norm of $x_2$ and $x_3$ for channel Ka and various values for $\beta$. $M$ is the maximal modulus of the input image in spherical harmonics.

We then assessed the relative performance of SPSR in flux recovery. The statistics obtained are presented in Table 1 and illustrate that in terms of bias and MSE, for all flux bands, SPSR performs better than local polynomial fitting, in particular for channels K, Ka and Q where diffuse component fluctuations are more important than in the higher noise channels V and W. SPSR is also more robust than FIT-C or FIT-L to large flux errors obtained in particular for large flux sources in the first channels.

Table 1. Estimates of flux for various flux bands (in mJy). The best results are displayed in bold.

<table>
<thead>
<tr>
<th>Channel</th>
<th>Flux &lt; 1 Jy</th>
<th>Flux &gt; 5 Jy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>Bias RMSE</td>
<td>Bias RMSE</td>
</tr>
<tr>
<td>FIT-C</td>
<td>312.2 435.7</td>
<td>311.7 336.3</td>
</tr>
<tr>
<td>FIT-L</td>
<td>106.5 226.6</td>
<td>97.5 472.6</td>
</tr>
<tr>
<td>SPSR</td>
<td>70.3 837.6</td>
<td>257.7 364.3</td>
</tr>
<tr>
<td>FIT-C</td>
<td>115.5 381.9</td>
<td>50.4 392.7</td>
</tr>
<tr>
<td>FIT-L</td>
<td>315.5 371.7</td>
<td>296.2 407.3</td>
</tr>
<tr>
<td>SPSR</td>
<td>49.3 208.3</td>
<td>210.7 231.9</td>
</tr>
<tr>
<td>FIT-C</td>
<td>75.4 354.1</td>
<td>267.1 356.9</td>
</tr>
<tr>
<td>FIT-L</td>
<td>79.1 354.7</td>
<td>285.5 356.9</td>
</tr>
<tr>
<td>SPSR</td>
<td>30.4 277.2</td>
<td>177.1 285.5</td>
</tr>
<tr>
<td>FIT-C</td>
<td>27.3 286.4</td>
<td>284.1 300.0</td>
</tr>
<tr>
<td>FIT-L</td>
<td>32.8 318.8</td>
<td>288.7 314.7</td>
</tr>
<tr>
<td>SPSR</td>
<td>17.5 259.6</td>
<td>202.8 274.0</td>
</tr>
<tr>
<td>FIT-C</td>
<td>53.8 260.1</td>
<td>301.8 357.2</td>
</tr>
<tr>
<td>FIT-L</td>
<td>52.5 266.6</td>
<td>290.7 353.8</td>
</tr>
<tr>
<td>SPSR</td>
<td>30.2 277.7</td>
<td>213.2 293.0</td>
</tr>
</tbody>
</table>

The proposed approach was then applied to subtract point source emission from WMAP 9-year data with symmetrized beams obtained as the average axisymetric beam provided for the differential assemblies at the frequency considered. Figure 3 illustrates the most extreme difference in the approaches for channel Ka. The positive ring around the fitted regions as well as a negative peak inside these regions seem to indicate that FIT-C or FIT-L are here overfitting. The same phenomenon is not present for SPSR.

In conclusion, in a realistic PSM simulation SPSR outperformed FIT-C and FIT-L in terms of biases and RMSE, in particular in the first channels. SPSR seems more robust to large flux errors as illustrated in both simulation results and WMAP real data. We are currently working on improvements of SPSR, focused on the choice of the value of $\beta$ without simulations, and on the model for extended compact sources too simple for these components to be adequately subtracted from the data.
Figure 3. Patch with two fitted sources in WMAP 9-year data. From the left to the right: WMAP data, FIT-C, FIT-L and SPSR.

Acknowledgment

This work was funded by the European Research Council grant SparseAstro (ERC-228261), and the Swiss National Science Foundation (SNSF).

References

Minisymposium M18
Inverse Statistical Methods
An ensemble variational filter for sequential inverse problems

C L Farmer

Mathematical Institute, University of Oxford, UK
E-mail: farmer@maths.ox.ac.uk

Abstract: Given a model dynamical system, a model of any measuring instrument relating states to measurements, and a prior assessment of uncertainty, the probability density of subsequent system states, conditioned upon the history of the measurements, is of some practical interest. When measurements are made at discrete times, it is known that the evolving probability density is a solution of the discrete Bayesian filtering equations. This paper describes the difficulties in approximating the evolving probability density using a Gaussian mixture (i.e. a sum of Gaussian densities). In general this leads to a sequence of optimisation problems and high-dimensional integrals. Attention is given to the necessity of using a small number of densities in the mixture, the requirement to maintain sparsity of any matrices and the need to compute first and second derivatives of the misfit between predictions and measurements. Adjoint methods, Taylor expansions, Gaussian random fields and Newton’s method can be combined to, possibly, provide a solution.

Keywords: Filtering, inverse problem, ensemble, variational, Bayesian, data assimilation.

1 The sequential filtering problem

The forward model

Let \( t^n = n\tau \) be a sequence of times separated by a time step \( \tau \) and consider a time-dependent system for a \( K \)-dimensional state vector \( \varphi^n \) with components \( \varphi^n_k \). In general the state vector consists of both time-independent parameters and time-dependent variables. Assume that the dynamical system is defined by a sequence of algebraic problems,

\[
G_k^n(\varphi^n;\varphi^{n+1}) = 0
\]

where the functions \( G^n \) are known in closed form (as is almost always the case). The functions \( G_k^n(\varphi^n;\varphi^{n+1}) \) are usually functions of just a few components of the state vector, as the functions often arise from discretising partial differential equations. The structure of the equations is such that the time-independent components, the parameters, are necessarily constant through time.

Assume that the dynamical system implies the existence of a unique sequence of vector valued functions \( F^n \) such that \( \varphi^{n+1} = F^n(\varphi^n) \). Also assume that the functions \( F^n \) have a unique differentiable inverse \( F^{-n} \) such that \( \varphi^n = F^{-n}(\varphi^{n+1}) \).

Further suppose that measurements are made on the dynamical system at discrete times also separated by the constant interval, \( \tau \). Introduce a model of any measuring apparatus such that at each time-\( t^n \), the value of the \( m \)-th of \( M \) measurements is given by,

\[
s_m^n = h_m(\varphi^n) + \sigma_m \xi_m^n
\]

where the \( h_m \) are known functions and the measurement noise \( \xi_m^n \) is a random variable sampled independently from the standard zero-mean, unit-variance normal density and \( \sigma_m \) is a known positive constant which is the
variance of the noise in the $m$-th measurement at each time-$t^n$. For later convenience in writing equations, define the *precisions*, $p_m$ by $p_m = (\sigma_m)^{-2}$.

At each time-$t^n$ let $s^n$ be the set of all the measurements at time-$t^n$ and let $S^n$ be the entire set of all the measurements up to and including those made at time-$t^n$. For definiteness we do not make measurements at time-$0$.

Note that in the following superscripts such as $n$, refer to a time index, subscripts such as $i, j, k, l$, label components of state vectors, the subscript $r$ labels different state vectors as in an ensemble of state vectors, and the subscript $m$ labels different measurements.

### The Bayesian filtering equations

Suppose that at time-$t^n$ the conditional probability density function (pdf), $\pi(\varphi^n|S^n)$ is known. (For background on Bayesian statistics and the notion of conditional probability density see, for example [5]. References [1] and [2] discuss relevant background with regard to filtering.) It follows from the definition of conditional probability, the dynamical equations and the measurement model that

$$\pi(\varphi^{n+1}, \varphi^n, s^{n+1}|S^n) = \pi(s^{n+1}|\varphi^{n+1}) \pi(\varphi^{n+1}|\varphi^n) \pi(\varphi^n|S^n). \quad \text{(2)}$$

Thus,

$$\pi(\varphi^{n+1}, \varphi^n, s^{n+1}|S^n) = z \exp[-\sum_{m=1}^{M} \frac{p_m}{2} (h_m(\varphi^{n+1}) - s_m^{n+1})^2] \delta(\varphi^{n+1} - F^n(\varphi^n)) \pi(\varphi^n|S^n) \quad \text{(3)}$$

where $z$ is a generic normalisation constant and the statistical convention, of using the symbol $\pi$ for each pdf such that different pdfs are indicated by their different arguments, is followed. Equation (3) follows from equation (2) since (i) the measurements are normally distributed around $h_m(\varphi^{n+1})$ and (ii) the new state $\varphi^{n+1}$ is known exactly given the previous state $\varphi^n$. An equation for $\pi(\varphi^{n+1}, s^{n+1}|S^n)$ can now be derived by integrating over the state vector $\varphi^n$. Using the standard transformation of variables theorem in a multivariate integral and the properties of the $\delta$-function, it follows that

$$\pi(\varphi^{n+1}, s^{n+1}|S^n) = z \exp[-\sum_{m=1}^{M} \frac{p_k}{2} (h_m(\varphi^{n+1}) - s_m^{n+1})^2] |A^n(\varphi^{n+1})| \pi(F^{-n}(\varphi^{n+1})|S^n) \quad \text{(4)}$$

where $|A|$ denotes the determinant of a matrix $A$ and the Jacobian matrix $A^n$ is given by

$$A^n_{i,j}(\varphi^{n+1}) = \frac{\partial F^{-n}_{i}(\varphi^{n+1})}{\partial \varphi^{n+1}_j}. \quad \text{(5)}$$

Finally the function $\pi(\varphi^{n+1}|S^{n+1})$ is obtained by an application of Bayes’ theorem. In other words, substitute the numerical values of the measurements into the function $\pi(\varphi^{n+1}, s^{n+1}|S^n)$ which is then proportional to $\pi(\varphi^{n+1}|S^{n+1})$.

The problem of sequential Bayesian filtering is to build convergent approximations to the sequence of functions $\pi(\varphi^n|S^n)$. One might refer to this process as ‘solving’ the sequential filtering equations (4). This is the topic of the next section.

### 2 Numerical Scheme

#### Gaussian mixture approximation for the density functions

To find an approximate solution of the Bayesian filtering equations, the first step is to approximate the pdf by a weighted sum of $R$ multivariate Gaussian densities. Each density has an associated mean $\varphi^n_i$, which will be called a ‘centre’ and also a symmetric precision matrix $L^n_i$ which is the inverse of a conventional covariance matrix. Thus the aim of the method is to construct a sequence of Gaussian mixture approximations [1]

$$\pi(\varphi^n|S^n) = \sum_{r=1}^{R} a^n_r \exp \left[ -\frac{1}{2} \sum_{i,j=1}^{K} (\varphi^n_i - \varphi^n_{r,i}) L^n_{r,i,j} (\varphi^n_i - \varphi^n_{r,i}) \right] \quad \text{(6)}$$
where the \( a_r^n \) are positive weights that sum to unity and the \( z_r^n \) are normalisation factors so that each component of the Gaussian mixture has a unit integral with respect to the state vector \( \varphi^n \).

To initialise the Gaussian mixture at \( t = 0 \) requires a pdf that summarises the initial information. For definiteness let us consider the case where the initial pdf is a multivariate Gaussian density

\[
\pi(\varphi^0) = z \exp \left[ -\frac{1}{2} \sum_{i,j=1}^{K} (\varphi_i^0 - \overline{\varphi}_j^0) L_{i,j}^0 (\varphi_j^0 - \overline{\varphi}_j^0) \right]
\]

for an initial mean state \( \overline{\varphi}^0 \) and precision matrix \( L^0 \). In applications where the state vector dimension is large it is necessary to specify a sparse precision matrix. One way to obtain a sparse matrix is to use a precision function. Then we expand the misfit functions in Taylor series up to second order about the minima. In general, however, explicit computation of the Hessian matrices is not feasible. To avoid computing Hessians directly, an indirect method involving further heuristic approximations is required, as is now described.

Maintaining a Gaussian mixture at all times

On substitution of equation (6) into the filtering equations (4) one finds in general that the left hand side is no longer a sum of Gaussian densities. The next step therefore, is to seek quadratic approximations to the ‘misfit functions’, \( J_r^n(\varphi^{n+1}) \) defined for \( n \geq 0 \) by

\[
J_r^n(\varphi^{n+1}) = -\ln |A^n(\varphi^{n+1})| + \sum_{m=1}^{M} \frac{p_m}{2} (h_m(\varphi^{n+1}) - s_m^n)^2 + \frac{1}{2} \sum_{i,j=1}^{K} (F_i^n(\varphi^{n+1}) - \varphi_{r,i}^n) L_{r,i,j}^n (F_j^n(\varphi^{n+1}) - \varphi_{r,j}^n).
\]

These quadratic approximations will be constructed by solving optimisation problems for the new centres which will be local minima of the misfit functions. Then we expand the misfit functions in Taylor series up to second order about the minima. In general, however, explicit computation of the Hessian matrices is not feasible. To avoid computing Hessians directly, an indirect method involving further heuristic approximations is required, as is now described.

Computing the action of the precision matrices

Equation (8) requires a definition of each precision matrix \( L_r^n \) or at least a prescription of its action on a given vector. At time \( n = 0 \) set \( L_r^0 = w(R)L^0 \) as stated earlier. For later times, \( n > 0 \), we suggest introducing an auxiliary misfit function,

\[
\mathcal{J}_r^{n-1}(\varphi^n) = -\ln |A^{n-1}(\varphi^n)| + \sum_{m=1}^{M} \frac{p_m}{2} (h_m(\varphi^n) - s_m^{n-1})^2 + \frac{1}{2} \sum_{k,l=1}^{K} (F_k^{-(n-1)}(\varphi^n) - \varphi_{r,k}^{n-1}) L_{r,k,l}^{n-1} (F_l^{-(n-1)}(\varphi^n) - \varphi_{r,l}^{n-1})
\]

where the matrix \( \mathcal{L}^{n-1} \) is defined by \( \mathcal{L}_{i,j}^{n-1} = w(R)L_{i,j}^0 + a_r^{n-1} \delta_{i,j} \) and where \( a_r^{n-1} \) is the reciprocal of the diagonal term of the ensemble covariance matrix. That is let \( a_r^{-1} = \sum_{r=1}^{R} a_r^{n-1} \varphi_r^{-1} \) and then \( \frac{1}{a_r^{n-1}} = \sum_{r=1}^{R} a_r^{n-1} (\varphi_{r,i}^{n-1} - \overline{\varphi}_i^{n-1})^2 \).

Then define the precision matrices in equation (8) using the expression

\[
L_{r,i,j}^n = \left. \frac{\partial^2 \mathcal{J}_r^{n-1}(\varphi^n)}{\partial \varphi_{r,i}^n \partial \varphi_{r,j}^n} \right|_{\varphi_r^n}.
\]
In some cases equation (10) can be evaluated in closed form but in general an adjoint method will be needed to compute the action of $L^n$ on any given vector. Note that the preceding definitions maintain sparse approximations when the initial precision matrix and the Jacobians of the dynamical system are sparse.

**Computing the new centres**

To find the new centres one needs to solve the optimisation problems

$$\phi^{n+1}_r = \arg \min_{\phi^{n+1}_r} J^n_r(\phi^{n+1}) \quad (11)$$

Suppose that the optimisation problems (11), which in many applications approximate variational problems, are solved using, for example, a modified Newton method or similar method that only requires the values of the gradient of the misfit function at particular values of the state vector. Thus the gradient $\frac{\partial J^n_r(\phi^{n+1})}{\partial \phi^{n+1}_i}$, evaluated at various values of the argument $\phi^{n+1}$, is needed. This is composed of three parts. First we need to compute the gradient of the determinant. By differentiation of the discrete equations (1) and taking determinants, we find that the logarithm of the determinant of the Jacobian is a difference of the logarithms of the determinants of two matrices that can be evaluated in closed form (by virtue of the sparsity of the equations). The two logarithms of the determinants will require an approximation, such as using the sum of the logarithms of the diagonal components. Secondly we need the gradient of the measurement misfit. In most cases this can be evaluated by differentiation of the known sparse functions $h_k$. The third term requires the product of the Jacobian matrix of the inverse forward function with vectors which can be calculated. This product can be computed by first solving the forward model equation (1) for $\phi^n$ given $\phi^{n+1}$ and then using differentiation with respect to $\phi^{n+1}$. Then using an adjoint matrix method one can compute the required products in a time and memory efficient manner. We note that in evaluating the third term the forward model is actually solved backwards in time, and the adjoint step (when needed) is the inverse of the usual adjoint method.

Later on, as an example, it is shown that in a large class of special cases, these calculations can be simplified so that adjoints are not required. Note that recent developments in software greatly assist in implementing methods that use adjoints [4]. However, further work is required on approximating the logarithms of the various determinants.

**Computing new weights and normalisation factors**

It follows from the quadratic approximation that the weight associated with each centre changes according to the equation

$$a^{n+1}_r = a^n_r \exp \left[ -J^n_r(\phi^{n+1}_r) \right] \frac{z^n_r}{z^{n+1}_r} \quad .$$

To avoid problems with rounding and other numerical error in evaluating exponentials with large negative arguments, particularly during the early stages of data assimilation, following [7] the modified update

$$a^{n+1}_r = \frac{\epsilon}{R} + (1 - \frac{\epsilon}{R}) \frac{\epsilon_0 + \tilde{a}^{n+1}_r}{\epsilon_0 + \sum_{r=1}^R \tilde{a}^{n+1}_r} \quad ,$$

is suggested where $\epsilon$ is a small positive constant and $\epsilon_0$ is a very small positive constant. As $R$ increases the effect of this modified update for the weights is reduced, but it prevents problems when the ensemble is small.

Regarding the normalisation coefficients, $z^n_r$, a convenient heuristic is to assume that their values are all the same, and do not change in time. Better approximations for the normalisation factors $z^n_r$ might be possible and this is a topic for future work.

### 3 An outline example

As an example, but one of considerable generality, consider a dynamical system defined by a system of ordinary differential equations that, when discretised using an implicit Euler method, has the discrete dynamics

$$\phi^{n+1}_i = \phi^n_i + \tau f^n_i(\phi^{n+1})$$


where the $f^n_i$ are components at time $t^{n+1}$ of the vector field for the dynamical system. With this form of discrete dynamics, the inverse map may be constructed directly by noticing that

$$F^{-n}_i(\varphi^{n+1}) = \varphi^{n+1}_i - t f^n_i(\varphi^{n+1}).$$

From this result one finds that for many dynamical systems the Jacobian of the inverse map, equation (5) can be found by exact differentiation and is generally a sparse matrix. This greatly simplifies the implementation of the method in this particular case. Indeed one can see, for example in the special case that the measurement model is $h_k(\varphi^{n+1}) = \varphi^{n+1}_k$, that the updated precision matrices have the explicit formulae

$$L^{n+1}_{r,i,j} = -\left[ \frac{\partial^2}{\partial \varphi^{n+1}_i \partial \varphi^{n+1}_j} \sum_{k=1}^{K} \ln(1 + \frac{\partial f^n_k}{\partial \varphi^{n+1}_k}) \right] + p_i \delta_{i,j} + \sum_{k,l=1}^{K} (\delta_{i,k} - \tau \frac{\partial f^n_k}{\partial \varphi^{n+1}_k}) L^n_{r,k,l} (\delta_{j,l} - \tau \frac{\partial f^n_l}{\partial \varphi^{n+1}_l}),$$

which follows from equation (9) and where (i) a diagonal approximation for the determinant has been used and (ii) a Gauss-Newton approximation has been made in neglecting second derivatives of $f^n$ and (iii) the precisions, $p_i$ are zero for state components that are not measured.

4 Concluding remarks

In lieu of a detailed description of numerical experiments let us conclude by remarking that (i) matrix inversions are not required by the ensemble variational filter, as only the actions of the precision matrices are needed and (ii) by maintaining sparsity, the method scales with increasing dimension of the dynamical system. Our numerical experiments indicate, even in systems with chaotic behaviour, that useful probabilistic estimates of parameter values can be obtained.

Clearly much further work is needed to evaluate the performance of this variant of ensemble method. However, the presence of the increasing function $w(R)$ in the initial Gaussian mixture, and the consistent method for computing the ensemble weights makes it (i) plausible that the ensemble variational filter converges in the limit of increasing ensemble size while (ii) the optimisation steps make the method practical even with a small ensemble.

Acknowledgments

This work was supported by the Oxford Martin School Programme on Modelling and Predicting Climate. The author thanks colleagues, in Oxford and elsewhere, for their criticism and encouragement.

References

Author index

Anselmi, N, 85
Audibert, L, 96
Bakushinsky, A, 63
Belina, L, 30
Binder, A, 40
Blazakis, K, 74
Blumensath, T, 131
Bobin, J, 158
Bonettini, S, 148
Brune, C, 79
Burger, M, 79
Carpio, A, 24
Chen, X, 90
de Camp, L, 63
De Pierro, A R, 51
Ding, P, 85
Dragotti, P L, 136
Duperray, A, 69
Étienne, J, 69
Farmer, C L, 164
Frerking, L, 79
Girard, A, 96
Hahn, B N, 46
Houssem, H, 96
Hryniv, R O, 57
Hussein, S O, 108
Jiao, Y, 7
Jin, B, 7
Jin, Q, 124
Kabanikhin, S, 102, 113
Klibanov, M V, 30
Krivorotko, O, 102
La Camera, A, 148
Laforgue, L, 69
Lantéri, H, 153
Laurent, V M, 69
Lesnic, D, 108
Louis, A K, 17
Lu, X, 7
Madzvamuse, A, 74
Malmberg, J B, 30
Manko, S S, 57
Massa, A, 85
Michel, R, 69
Miqueles, E X, 51
Oliveri, G, 85
Paykari, P, 158
Prato, M, 148
Rapun, M L, 24
Rassat, A, 158
Rebegoldi, S, 148
Reyes-Aldasoro, C C, 74
Sabatier, P C, 12
Shishlenin, M, 113, 118
Smirnova, A, 63
Starck, J-L, 158
Styles, V, 74
Sureau, F C, 158
Symes, W, 2
Takiguchi, T, 35
ten Kroode, F, 142
Thành, N T, 30
Theys, C, 153
Venkataraman, C, 74
Verdier, C, 69
Vestweber, D, 79
Vitale, G, 69
Yudytskiy, M, 40
Zhong, Y, 90

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, book title and publisher. Published under licence by IOP Publishing Ltd.