Springer Proceedings in Mathematics & Statistics

Larisa Beilina · Maïtine Bergounioux ·
Michel Cristofol · Anabela Da Silva ·
Amelie Litman *Editors*

Mathematical and Numerical Approaches for Multi-Wave Inverse Problems

CIRM, Marseille, France, April 1–5, 2019

Springer

Springer Proceedings in Mathematics & Statistics

Volume 328

Springer Proceedings in Mathematics & Statistics

This book series features volumes composed of selected contributions from workshops and conferences in all areas of current research in mathematics and statistics, including operation research and optimization. In addition to an overall evaluation of the interest, scientific quality, and timeliness of each proposal at the hands of the publisher, individual contributions are all refereed to the high quality standards of leading journals in the field. Thus, this series provides the research community with well-edited, authoritative reports on developments in the most exciting areas of mathematical and statistical research today.

More information about this series at http://www.springer.com/series/10533

Larisa Beilina · Maïtine Bergounioux ·
Michel Cristofol · Anabela Da Silva ·
Amelie Litman
Editors

Mathematical and Numerical Approaches for Multi-Wave Inverse Problems

CIRM, Marseille, France, April 1–5, 2019

Springer

Editors
Larisa Beilina
Department of Mathematical Sciences
Chalmers University of Technology and
University of Gothenburg
Gothenburg, Sweden

Maïtine Bergounioux
Département de Mathématiques - MAPMO
University of Orléans
Orleans, France

Michel Cristofol
Institut de Mathématiques de Marseille
Aix-Marseille University
Marseille, France

Anabela Da Silva
Institut Fresnel
Marseille, France

Amelie Litman
Institut Fresnel
Marseille, France

ISSN 2194-1009 ISSN 2194-1017 (electronic)
Springer Proceedings in Mathematics & Statistics
ISBN 978-3-030-48633-4 ISBN 978-3-030-48634-1 (eBook)
https://doi.org/10.1007/978-3-030-48634-1

Mathematics Subject Classification: 62P10, 65N21, 65N12, 65N15, 78M10, 65F22, 65R20, 65R32, 65J22, 65M32, 78A46

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

In this volume are collected some papers related to the conference *Mathematical and Numerical Approaches for Multi-Wave Inverse Problems* which took place from 1 to 5 April, 2019 at *Centre International de Rencontres Math'ematiques*, CIRM (https://www.cirm-math.com/), Marseille, France. One of the main objectives of this conference was exchange of ideas and tools between different scientific communities, specially to favor the discussions between researchers more involved in theoretical aspects of inverse problems with the ones more interested in numerical implementation of these problems.

Inverse problems have been a topic of interdisciplinary interest for many years for mathematicians, applied mathematicians, physical scientists or engineers. Considerable progress has been achieved in recent years in the development of innovative techniques, new theoretical tools, new approximations, as well as new optimization techniques to improve solution of multi-wave inverse problems.

In the inverse problems community, the current scientific interests are focused towards achieving better contrasted images, with higher spatial resolution and with quantitative contents, such as functional imaging in biomedical applications.

As a consequence, more and more multi-modal, multi-wave or hybrid systems are currently being proposed and/or being used routinely. Mathematically, solving these inverse problems is even more complicated because they require a coupling, that can be soft or hard, between a set of partial differential equations not necessarily of the same nature (elliptic, hyperbolic or parabolic). This leads to problems in terms of uniqueness, stability but also in terms of control which remain still little explored. From numerical point of view, these systems of PDEs may lead to large discretized domains, with large number of degrees of freedom, which must be adapted according to the wavelength of the considered waves.

This book gathered scientific works from a number of researchers strongly involved in multi-modal applications. The volume highlights new theoretical and numerical tools for the solution of real-life problems as well as proposes new systems for their solution on the basis of theoretical understanding.

For most of papers of this book, the reader will find the complete description of a new technique or numerical method for solution of some inverse problem which is supported by numerical simulations. The intended audience of the book is undergraduate and graduate university students, Ph.D. students specialized in applied mathematics, electrical engineering and physics, researchers and university teachers and R&D engineers with interest in applied mathematics.

Gothenburg, Sweden Larisa Beilina
Orleans, France Maïtine Bergounioux
Marseille, France Michel Cristofol
Marseille, France Anabela Da Silva
Marseille, France Amélie Litman

Contents

Thermoacoustic Applications 1
S. K. Patch

On the Transport Method for Hybrid Inverse Problems 15
Francis J. Chung, Jeremy G. Hoskins, and John C. Schotland

Stable Determination of an Inclusion in a Layered Medium with Special Anisotropy 21
Michele Di Cristo

Convergence of Stabilized *P1* Finite Element Scheme for Time Harmonic Maxwell's Equations 33
M. Asadzadeh and Larisa Beilina

Regularized Linear Inversion with Randomized Singular Value Decomposition 45
Kazufumi Ito and Bangti Jin

Parameter Selection in Dynamic Contrast-Enhanced Magnetic Resonance Tomography 73
Kati Niinimäki, M. Hanhela, and V. Kolehmainen

Convergence of Explicit P_1 Finite-Element Solutions to Maxwell's Equations 91
Larisa Beilina and V. Ruas

Reconstructing the Optical Parameters of a Layered Medium with Optical Coherence Elastography 105
Peter Elbau, Leonidas Mindrinos, and Leopold Veselka

The Finite Element Method and Balancing Principle for Magnetic Resonance Imaging. 127
Larisa Beilina, Geneviève Guillot, and Kati Niinimäki

Thermoacoustic Applications

S. K. Patch

Abstract This conference paper provides a cursory overview of thermoacoustic phenomena and attempts to highlight aspects that may be unfamiliar to the mathematical community or could benefit from more rigorous mathematical analysis. A new clinical application, thermoacoustic range verification during particle therapy, is presented. The goal is to ground expectations and generate further interest in thermoacoustics within the mathematical community.

Keywords Thermoacoustic · Spherical radon transform · Time reversal

MSC: 62P10

1 Introduction and Overview

By 1881 Alexander Graham Bell noted the photoacoustic effect, by which electromagnetic energy in the optical regime is converted to mechanical energy [1]. More generally, thermoacoustics refers to generation of acoustic signals by heating. In the early 1960's thermoacoustics was attributed to audible detection of microwave pulses, even by deaf study volunteers [2]. By the early 1980's, photoacoustic spectroscopy and pulse oximetry were used in commercial [3] and medical applications, respectively. Image reconstruction was first considered by physicists who developed series solutions [4]. Decades later, filtered backprojection type inversion formulae were first developed by mathematicians for specific measurement surfaces [5, 6]. Engineers quickly improved upon these results and developed analytic inversion formulae for arbitrary measurement surfaces [7]. These early results assumed ideal experimental conditions, and mathematicians have since contributed many results to expand upon them. For instance, too many results to cite have been published on image reconstruction in the presence of acoustic heterogeneity. Rather than recite results that are

S. K. Patch (✉)
Department of Physics, UW-Milwaukee and Acoustic Range Estimates, Milwaukee, WI 53211, USA
e-mail: patchs@uwm.edu

L. Beilina et al. (eds.), *Mathematical and Numerical Approaches for Multi-Wave Inverse Problems*, Springer Proceedings in Mathematics & Statistics 328,
https://doi.org/10.1007/978-3-030-48634-1_1

well known to the mathematical community, this manuscript attempts to highlight aspects that could benefit from more nuanced mathematical analysis. In particular, different signal generation techniques, i.e. object heating, are discussed in Sect. 2. A few experimental results that may impact approaches to image reconstruction are presented in Sect. 3.

It is hardly surprising that thermoacoustic signals are generated during rapid energy deposition, since the units of pressure and energy density are the same, $[p] = N/m^2 = J/m^3$. The conversion factor is the dimensionless Grueneisen parameter, $\Gamma = B\beta/C\rho$. Thermal expansion coefficient, β, and specific heat capacity, C, are thermal parameters, whereas bulk modulus, B, and density are mechanical parameters. Assuming instantaneous energy deposition pressure changes according to

$$\delta p(x) = \Gamma(x, T)S(x) \tag{1}$$

where T denotes temperature and $S(x)$ represents the applied energy density.

Regardless of the type of energy used to heat the object under test, recovering useful information can be challenging. A perfect reconstruction of thermoacoustic pressure increase, $\delta p(x)$, does not ensure perfect understanding of the applied energy density, the Grüneisen, or for that matter, any of the parameters upon which the Grüneisen depends. Parameters of interest vary depending upon application and fortunately, assumptions can be made to reduce the number of unknowns. For instance, in most soft tissues, $B \sim 2.2$ GPa and $\rho \sim 1000$ kg/m^3.

2 Different Signal Generation Techniques

Signal generation parameters of three well-known versions of electromagnetically induced thermoacoustics are compared and contrasted with ion-induced thermoacoustics in Table 1, where PAT, MITAT, and TCT refer to photoacoustic tomography utilizing near-infrared radiation (NIR), microwave-induced thermoacoustic tomography and very high frequency (VHF) thermoacoustic computerized tomography. Ionoacoustics indicates that heating is due to energetic ions (positively or negatively charged particles). For each of the tomographic techniques non-ionizing electromagnetic *photons* with insufficient energy to break a DNA strand are applied to generate thermoacoustic signal. Ionoacoustic signal is created by charged particles delivered with the goal of breaking DNA. Brief descriptions of the electromagnetic heating modalities are followed by a more detailed description of ion-induced thermoacoustics in Sect. 2.3.

The last row of Table 1 compares typical heating pulse durations. As a rule of thumb, driving amplifier tubes from 0 to kW power levels in less than a dozen cycles is challenging. We note that solid state amplifiers are fast, and prohibitively expensive for most research groups. Therefore, typical pulse durations for VHF and MITAT tend to be proportional to the period of the applied electromagnetic field. Additionally,

Table 1 Thermoacoustic signal generation

	TCT	MITAT	PAT	Ionoacoustic
Energy type	VHF	Microwave	NIR	Energetic ion
Frequency range	30–300 MHz	300 MHz–300 GHz	2-4 THz	NA
Goal to recover	Ionic content	Relaxation + absorption	Absorption	Dosimetry range verif.
Governing Eq.	Maxwell/wave	Maxwell/wave	Transport/diffusion	Bethe-Bloch
Energy transport	Wave propagation	Wave propagation	Photon scattering	Ion scattering
Polarization important?	YES	YES	NO	NA
Heating pulse duration	O(1 μs)	O(100 ns)	1–10 ns	O (1 ns–10 μs)

different types of particle accelerators have pulse durations that vary by orders of magnitude. This may be why experimentalists tend to model thermoacoustic wave propagation using homogeneous initial conditions and place the source term in the wave equation,

$$\Box p(x,t) = \frac{\partial}{\partial t}\left[\Gamma(x,T)S(x)I'(t)\right] \tag{2}$$

where $S(x)$ represents energy density, $I(t)$ is a dimensionless approximate delta function. Representative values of full width at half maximum for $I(t)$ are listed in Table 1. $S(x)I'(t)$ is the rate of energy deposition in W/m^3. Although Du Hamel showed the equivalence of treating the source term as an initial condition, homogeneous initial conditions simplifies analysis of heating pulse duration [8].

Signal generation by applying electromagnetic waves is discussed briefly in Sect. 2.1, whereas heating due to photon migration is briefly mentioned in Sect. 2.2. Finally, a new thermoacoustic application for particle therapy is described more fully in Sect. 2.3.

2.1 *Propagating Electric Fields Induce Thermoacoustic Pulses*

For electromagnetically induced heating, dielectric properties of the material govern electromagnetic energy penetration into the object and loss within the object. Following the notation used in reports on dielectric properties of tissue [9–11], we express permittivity as a complex number.

$$\varepsilon = \varepsilon' - i\varepsilon'' \tag{3}$$

where $\varepsilon^{'}$ is the relative permittivity of the material and $\varepsilon'' = \frac{\sigma}{\varepsilon_o \omega}$ governs energy loss. Here, σ is the *total* conductivity of the material which accounts for loss due to frequency-dependent relaxation effects and also frequency-independent ionic conductivity, σ_i. A few facts are listed below to provide intuition regarding ionic content and thermoacoustic imaging:

- Deionized water nearly zero ionic content and $\mathrm{\sigma_i} = 5.5\ \mu\mathrm{S/m}$ whereas in blood $\sigma_i \sim 1$ S/m.
- Permittivity in a vacuum is very small, $\varepsilon_o = 8.85e - 12$ F/m, so the impact of ionic loss becomes large at VHF frequencies. For instance, at the frequency of modern MRI systems (128 MHz), $\frac{\sigma_i}{\varepsilon_o \omega} \sim 140\sigma_i$ whereas at the frequency of most microwave ovens (2.45 GHz) $\frac{\sigma_i}{\varepsilon_o \omega} \sim 7\sigma_i$. Most cell phones operate at intermediate frequencies (0.9–2 GHz).
- Many organs are designed to secrete physiologic fluids with varying degrees of ionic content. Furthermore, health of the organ may be reflected by the ionic content of secreted fluid. For instance, ionic content of prostatic fluid is correlated to disease state [12]. Ion content in prostatic fluid from healthy and cancerous glands is qualitatively displayed in Fig. 1. For a more quantitative analysis, see Fig. 3 in [13].

Components of a plane wave propagating along the z-axis in free space decay according to

$$E(x, y, z, \omega) = E_o e^{-i\varepsilon z} = E_o e^{-i\varepsilon' z} e^{\frac{-\sigma}{\varepsilon_o \omega} z} \tag{4}$$

Here heating is due primarily to the E field, and the specific absorption rate (SAR) of nonionizing VHF and microwave energy is given by

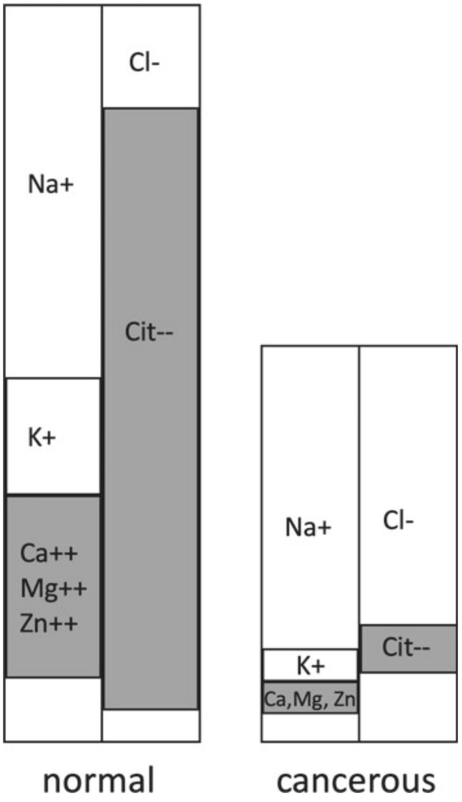
$$SAR = \sigma/\rho |E|^2 \tag{5}$$

When low frequency electric fields are used, relaxation effects can be dominated by ionic content and quantitative reconstructions can yield images with the potential to differentiate healthy from diseased organs.

2.1.1 Very High Frequencies (VHF)

Very high frequency (30–300 MHz) electromagnetic fields are used by FM radio stations and also MRI scanners. Therefore, SAR of VHF energy is strictly regulated to avoid patient overheating [14]. In the VHF regime, loss due to ionic content is at least as significant as relaxation effects. Therefore, the VHF-induced contrast mechanism may show changes in ionic [15, 16] and/or fat content.

Fig. 1 Qualitative Gamble plots depicting ion concentrations in prostatic fluids



Maxwell's equations are well approximated by wave equations in the VHF regime, with wavelengths ranging from 1 to 10 m in vacuum. Wavelengths in media are reduced by a factor of $1/\sqrt{\varepsilon_r}$, and ε_r is less than 10 in fatty tissue, 40–60 in muscle and fat, and close to 80 in water. At 100 MHz, $\lambda_{vac} = 3$ m and $\lambda_{water} = 33$ cm $< \lambda_{organ}$, so whole organs in adults can be heated without suffering hot or cold spots due to standing waves.

Polarization [17] and diffraction of the applied electromagnetic field affect energy deposition. Applying a variably polarized field will heat more uniformly. For instance, circularly polarized fields are applied by MRI scanners.

2.1.2 Microwave Induced Thermoacoustic Imaging

Microwaves have become ubiquitous in our kitchens and next to our ears. SAR of microwaves concentrated near a cell phone has come under close scrutiny and is tightly regulated. Microwave ovens heat even deionized water due entirely to relaxation effects, which are minimal in the VHF regime.

The microwave band ranges from 300 MHz to 300 GHz, with wavelengths from 1 m to 1 mm in vacuum, respectively. As with VHF, wavelengths are reduced in tissue. The impact of polarization and diffraction [18] is even more pronounced than in the VHF regime. Back of the envelope analysis for simple waveguides and chambers yield standing waves. Indeed, early microwave ovens lacked turntables and mode mixers and induced hot and cold spots in solid foods. Intuition for standing waves can be developed easily using a cheap microwave oven that lacks a mode mixer by disabling the rotation stage. Simply remove the glass plate and cover it with an overturned flat-bottomed container. Place a large block of butter on the container and observe where microwave heating melts the butter—and where it does not.

Although poor depth penetration and standing waves are disadvantageous compared to VHF, microwaves have some advantages. Sub-microsecond pulse durations are more easily achieved by microwave than VHF amplifiers, and energy loss and heating is greater in the imaging depth which is limited relative to VHF.

2.2 Photon Migration/Diffusion

Photoacoustic imaging is perhaps the best-known application of thermoacoustics, so just a few sentences on photoacoustics follow.

Depth penetration of near infrared (NIR) photons limits imaging depth even more severely than microwave and VHF. High photon fluence within a small volume near the optical source and very short (<10 ns) pulse durations generate photoacoustic pulses with high SNR and bandwidth. The result can be exquisitely high-resolution reconstructions, for instance of microvasculature. A photon transport model is required in the NIR regime, but diffusion models are appropriate for large numbers of photons and scattering events. Therefore, cold spots due to standing waves are not a concern.

A strong advantage of photoacoustics over VHF and MITAT is that multiple wavelengths can be used to reconstruct oxygen saturation, much like a pulse oximeter. Quantitative photoacoustics has been developed to handle non-uniform photon fluence and extract the optical attenuation coefficient from the reconstructed image. Photoacoustic imaging systems have been marketed by at least six vendors, and a journal dedicated to the field attest to its success. For a mathematical review of reconstruction techniques see [19], for a broader overview see the chapter devoted to photoacoustics in [20] and review paper [21].

2.3 Thermoacoustic Signals Generated by Charged Particles

Thermoacoustic emissions generated by protons stopping in large water baths were first detected in the 1970s [22, 23]. The results were recognized as potentially useful for particle therapy by the mid-1990s [24]. A thorough approach dosimetry utilized a cylindrical measurement surface in a simulation study [25]. Progress on range verification stalled, however, for several reasons. One of the limiting factors was likely inadequate acoustic hardware.

The Bethe-Bloch equation models energy loss by energetic charged particles,

$$\frac{dE}{dx} \propto -\frac{nz^2}{c^2\beta^2}\left[\ln\left(\frac{2m_e c^2\beta^2}{10eVz\left(1-\beta^2\right)}\right) - \beta^2\right] \tag{6}$$

where n is electron density of the target material, z is charge of the particle, m_e is electron mass, $\beta = v/c$, where v and c are speed of the particle in the target and light in a vacuum, respectively. High energy particles travel fast and have little time to interact with atoms near their trajectory. As they slow interaction times become longer and more energy is lost per unit pathlength. Within nanoseconds high energy particles decelerate from relativistic speeds and come to a stop.

2.3.1 Monte Carlo Simulation

Energy maps, $S(x)$, are computed using Monte Carlo simulations. Figure 2a shows proton trajectories through layered targets as computed using TRIM software [26]. Layers consisted of tissue mimicking gelatin, bone, water, and air. All protons had the same initial conditions, starting at the origin with 50 MeV energy directed along the vertical axis. Figure 2b. shows a dose map that more realistically mimics a high energy pencil beam as delivered clinically. The point at which the dose map achieves its maximum is referred to as the Bragg peak.

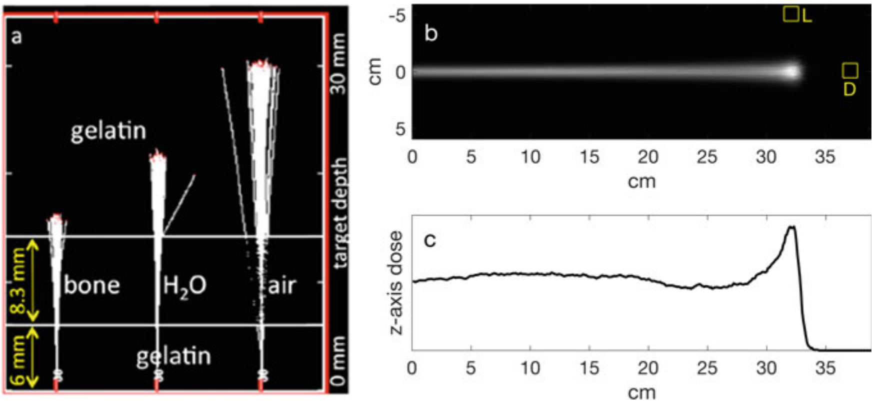


Fig. 2 Monte Carlo simulations of proton transport. **a** Tracks of 50 MeV protons through layered phantoms. **b**–**c** Results from 90 k proton simulation mimicking a clinical beam in water. **b** Dose map representing $S(x)$. **c** Dose profile along the z-axis, i.e. center of the beam

Rapid falloff of dose beyond the Bragg peak as shown in Fig. 2c allows for precise treatment of a tumor while sparing radiosensitive tissue distal to the Bragg peak. However, if the patient moves or is misaligned, then radiosensitive tissue could receive the maximal dose while the tumor is underdosed. Accurate and fast range verification is therefore needed to confirm that dose is being delivered to the intended target.

90 k protons were tracked in using TRIM software. Each proton had initial velocity parallel to the horizontal axis, but initial energies had a normal distribution with mean 230 MeV and standard deviation of 920 keV. Initial positions were also randomly distributed so that the radially symmetric beam had FWHM of 7.6 mm upon entry into the target. Viewed at low resolution the cross section of the dose map in Fig. 2b appears smooth, but a plot along the central beam axis in Fig. 2c is not smooth. The synchrocyclotron that this simulation was meant to model delivers O(1 pC) of charge or O(10^7) protons per pulse. Increasing the number of particles in the Monte Carlo simulation would smooth the dose map, but at prohibitive computational cost. Simply filtering in the spatial domain smooths a dose map far more efficiently than increasing the number of particles.

2.3.2 Synchrocyclotron Pulse Envelopes

We measured temporal profiles, $I(t)$, of proton pulses delivered by three synchrocyclotrons, two manufactured by IBA one by Mevion. In all cases, the pulse envelopes were approximately Gaussian, but had varying pulse durations, with FWHM ranging from 5 μs to nearly 10 μs.

2.3.3 Acoustic Simulation

Thermoacoustic pulses were simulated by taking the dose map as instantaneously deposited, or equivalently using it to compute initial pressure. k-Wave software [27] was run with receivers positioned 5 cm distal and lateral to the Bragg peak, indicated by yellow squares in Fig. 2b. To account for ion pulse durations, k-Wave time series were convolved by Gaussians with FWHM of 5 and 10 μs. Results are plotted in Fig. 3. Instantaneous deposition resulted in thermoacoustic emissions bandlimited below 150 kHz, simply because the dose maps are smooth. Convolving with $I(t)$ further reduced bandwidth below 100 kHz.

Applying the standard half-wavelength resolution limit of inverse scattering problems would yield unacceptable range errors. Fortunately, range verification requires recovering only one number, location along the beamline, rather than a three-dimensional array of voxel values. Therefore, we developed a method for leveraging a priori information from the patient's planning CT and treatment plan to improve accuracy of range estimates and beat the diffraction limit [28].

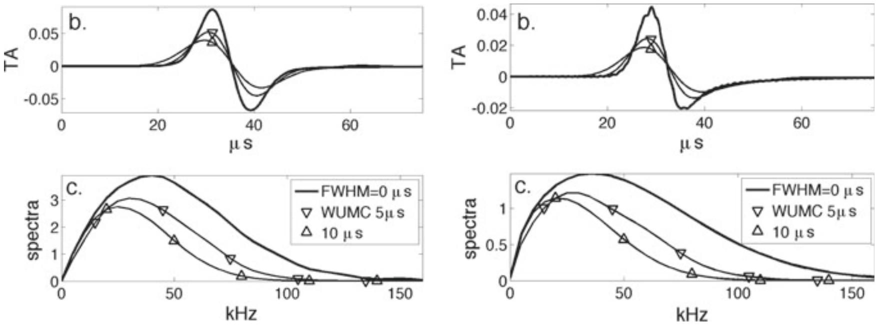


Fig. 3 Simulated thermoacoustic pulses (top) and spectra (bottom). Signals are measured by receivers positioned 5 cm distal (right) and lateral (left) to the Bragg peak

2.3.4 Measurements

Custom acoustic hardware is required to develop a thermoacoustic range verification system that provides clinical utility. Not only are pulses bandlimited well below the sensitivity bands of ultrasound imaging arrays, but pulses are weak. Thermoacoustic pressure induced by a dose map like that shown in Fig. 2b is on the order of 1 Pa/pC of charge—if delivered instantaneously. Therefore, a clinical system that delivers 1 pC over 5 μs increases pressure at the Bragg peak by less than one Pascal. Pulses attenuate as they travel to remotely positioned receivers, so measured pulses have millipascal amplitudes.

Many groups have measured thermoacoustic emissions with hydrophones and single element detectors, most recently using a therapeutic system [29] in research mode. More recently, we collected thermoacoustic emissions during delivery of a clinical treatment plan. For each treatment depth, or proton energy, required by the plan, dose was delivered by an IBA S2C2 synchrocyclotron as it would be during treatment. However, the system paused between initial energy levels as data was downloaded.

A custom system with 4 thermoacoustic channels at the corners of a 34 × 52 cm rectangle surrounding a wireless ultrasound array (Clarius L7) was positioned to generate an ultrasound image of the lesion *and* detect thermoacoustic emissions.

Channels 1–2 were located distal to the treatment volume and measured characteristic (broadband) “N” shapes; laterally offset channels 3–4 often measured (bandlimited) ringing (Fig. 4). Ringing is almost surely due to multiple reflections off of ribs, lung and rigid sidewalls of the 10 cm thick abdominal phantom. Ringing in channels 3–4 confounds straightforward range estimation and a priori information from the planning CT may be required to overlay the Bragg peak location onto the ultrasound image [28].

For a more comprehensive overview of thermoacoustic range verification, the reader is referred to the review article [30].

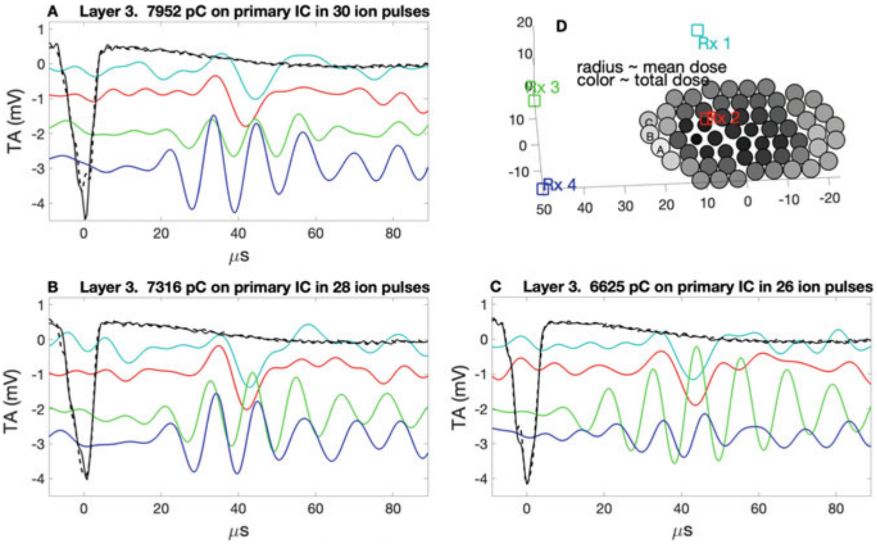


Fig. 4 Layer 3 emissions. **a**–**c** Thermoacoustic and gamma emissions generated at three neighboring spots. Gamma emissions represented $-I(t)$, time intensity of the proton beam, are plotted in black. Distal receivers 1–2 detected bandlimited "N" waves, plotted in cyan and red, respectively. Laterally positioned receivers 3–4 (green and blue, respectively) detected ringing, first in channel 4 (spot A), then moderate amplitude in both channels (spot B), and finally strong ringing in channel 3 (spot C). **d** Spot locations are denoted by circles with radius and fill color corresponding to average and total charges, respectively. Spots "A," "B" and "C" received 30, 28 and 26 proton pulses respectively

3 Experimental Observations that Impact Data Analysis/Reconstruction

3.1 Co-locating Ultrasound Pulse-Echo and Thermoacoustic Receive Elements

Ultrasound imaging is an inverse scattering problem, typically with a limited measurement surface and resolution is far worse than the diffraction limit. Additionally, ultrasound image formation relies upon an assumed soundspeed throughout the field of view. However, contrast in ultrasound images requires differences in acoustic impedance so it is unlikely that soundspeed is constant in an interesting ultrasound image. Even if soundspeed were constant, assuming the incorrect soundspeed in reconstruction would dilate the image. Variable soundspeeds deform images further. Nevertheless, ultrasound imaging is successful in most soft tissues. Whenever thermoacoustic receivers can be co-located with ultrasound arrays so that flight paths from the Bragg peak are essentially the same, then errors in range estimates due to acoustic heterogeneity will be almost identical to errors in the ultrasound image [31].

Therefore, range estimates can be inherently—and very accurately—co-registered to ultrasound images. Although range estimates may be inaccurate in absolute coordinates, they are very accurate relative to ultrasound images of the underlying anatomy. Commercial packages can quickly co-register ultrasound images to CT volumes which have submillimeter geometric accuracy, and range estimates could be easily carried along. The idea is depicted in Fig. 5, where a manual co-registration of CT and ultrasound is displayed using 3D Slicer.

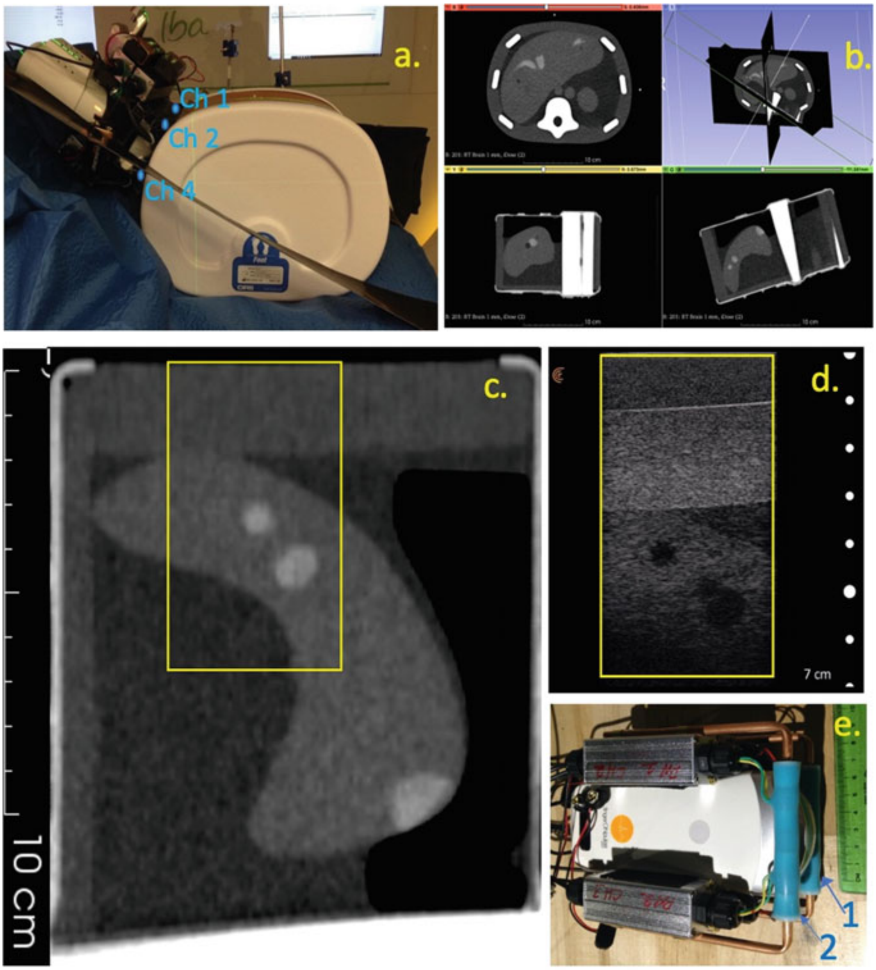


Fig. 5 **a** Photo of phantom and TA system with locations of channels 1, 2, 4 labeled. **b** Planning CT displayed in 3D Slicer. **c** Multi-planar reformat of CT corresponding to ultrasound image in **d**. **e** 8-channel prototype, with four-channel arrays on either side of the abdominal probe labeled "1" and "2"

3.2 Time Reversal

Reconstruction by time reversal has been well studied by mathematicians, primarily to handle acoustic heterogeneity and reflecting boundary conditions. Recent results by experimentalists in photoacoustics and ion-induced thermoacoustic range verification use multiply reflected signals, sometimes acquired over times exceeding the flight time across the field of view. These multiple reflections can effectively increase the measurement surface [32, 33] and improve resolution derived from limited angle measurements. For handheld probes this may prove important. Consider a wireless ultrasound array onto which eight thermoacoustic receivers are mounted, as depicted in Fig. 5e. Four thermoacoustic receivers on either side of the array prescribe a rectangle of 7 cm × 7 cm, and a solid angle of less than $2\pi(1 - \cos\theta_o) \sim \pi/10$ at an imaging depth of 10 cm.

Acknowledgements Many thanks to Rutherford Cancer Centres for access to their proton therapy equipment as well as Rutherford and IBA staff for support collecting data presented in Figs. 4, 5. In particular, thanks to Dr. Rudi Labarbe of IBA for facilitating those experiments.

References

1. Bell, A.: Production of sound by radiant energy. Manuf. Build. **13**, 156–158 (1881)
2. Frey, A.: Human auditory system response to modulated electromagnetic energy. J. Appl. Physiol. **17**, 689–692 (1962)
3. Heiser, D.: An Investigation of Photoacoustic Spectroscopy as a Technique for Measuring Diesel Particulate Emissions. US Environmental Protection Agency (1980)
4. Norton, S., Linzer, M.: Ultrasonic reflectivity imaging in three dimensions: reconstruction with spherical transducer arrays. Ultrason. Imaging **1**, 210–239 (1979)
5. Finch, D., Patch, S.K.: Rakesh: determining a function from its mean values over a family of spheres. SIAM J. Math. Anal. **35**, 1213–1240 (2003)
6. Patch, S.K.: Thermoacoustic tomography-consistency conditions and the partial scan problem. Phys. Med. Biol. **49**, 2305–2315 (2004)
7. Xu, M., Wang, L.V.: Universal back-projection algorithm for photoacoustic computed tomography. Phys. Rev. E **71**, 016706 (2005). https://doi.org/10.1103/PhysRevE.71.016706
8. John, F.: Duhamel's principle and the general Cauchy problem. In: Partial Differential Equations, pp. 135–138. Springer, New York, NY (1981)
9. Gabriel, C., Gabriel, S., Corthout, E.: The dielectric properties of biological tissues: I. literature survey. Phys. Med. Biol. **41**, 2231–2249 (1996)
10. Gabriel, S., Lau, R.W., Gabriel, C.: The dielectric properties of biological tissues: II. measurements in the frequency range 10 Hz to 20 GHz. Phys. Med. Biol. **41**, 2251–2269 (1996)
11. Gabriel, S., Lau, R.W., Gabriel, C.: The dielectric properties of biological tissues: III. parametric models for the dielectric spectrum of tissues. Phys. Med. Biol. **41**, 2271–2293 (1996). https://doi.org/10.1088/0031-9155/41/11/003
12. Kavanagh, J.P., Darby, C., Costello, C.B.: The response of seven prostatic fluid components to prostatic disease. Int. J. Androl. **5**, 487–496 (1982)
13. Costello, L.C., Franklin, R.B.: Prostatic fluid electrolyte composition for the screening of prostate cancer: a potential solution to a major problem. Prostate Cancer Prostatic Dis. **12**, 17–24 (2009). https://doi.org/10.1038/pcan.2008.19

14. NEMA: NEMA Standards Publication MS 10-2006. Determination of Local Specific Absorption Rate (SAR) in Diagnostic Magnetic Resonance Imaging (MRI). (2006)
15. Patch, S., Hull, D., Thomas, M., Griep, S., Jacobsohn, K., See, W.: Thermoacoustic contrast of prostate cancer due to heating by very high frequency irradiation. Phys. Med. Biol. **60**, 689–708 (2015). https://doi.org/10.1088/0031-9155/60/2/689
16. Patch, S.K., Hull, D., See, W.A., Hanson, G.W.: Toward quantitative whole organ thermoacoustics with a clinical array plus one very low-frequency channel applied to prostate cancer imaging. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **63**, 245–255 (2016). https://doi.org/10.1109/TUFFC.2015.2513018
17. Patch, S.K., Yan, L.: Object orientation in RF field determines thermoacoustic contrast. Presented at the Proc. SPIE February 24 (2009)
18. Li, C., Pramanik, M., Ku, G., Wang, L.V.: Image distortion in thermoacoustic tomography caused by microwave diffraction. Phys. Rev. E **77**, 031923 (2008). https://doi.org/10.1103/PhysRevE.77.031923
19. Kuchment, P., Kunyansky, L.: Mathematics of photoacoustic and thermoacoustic tomography. In: Scherzer, O. (ed.) Handbook of Mathematical Methods in Imaging. Springer, New York (2011)
20. Wang, L.V., Wu, H.-I.: Biomedical optics: principles and imaging. (2008)
21. Beard, P.: Biomedical photoacoustic imaging. Interface Focus **1**, 602–631 (2011). https://doi.org/10.1098/rsfs.2011.0028
22. Askariyan, G.A., Dolgoshein, B.A., Kalinovsky, A.N., Mokhov, N.V.: Acoustic detection of high energy particle showers in water. Nucl. Instrum. Methods **164**, 267–278 (1979). https://doi.org/10.1016/0029-554X(79)90244-1
23. Sulak, L., Armstrong, R., Baranger, H., Bregman, M., Levi, M., Mael, D., Strait, J., Bowen, T., Pifer, A., Polakos, P., Bradner, H., Jones, W., Learned, J.: Experimental studies of the acoustic signature of proton beams traversing fluid media. Nucl. Instrum. Methods **161**, 203–217 (1979)
24. Hayakawa, Y., Tada, J., Arai, N., Hosono, K., Sato, M., Wagai, T., Tsuji, H., Tsujii, H.: Acoustic pulse generated in a patient during treatment by pulsed proton radiation beam. Radiat. Oncol. Invest. **3**, 42–45 (1995)
25. Alsanea, F., Moskvin, V., Stantz, K.M.: Feasibility of RACT for 3D dose measurement and range verification in a water phantom. Med. Phys. **42**, 937–946 (2015). https://doi.org/10.1118/1.4906241
26. Ziegler, J., Ziegler, M., Biersack, J.: SRIM: the stopping and range of ions in matter (2010). Nucl. Instrum. Methods Phys. Res. Sect. B **268**, 1818–1823 (2010). https://doi.org/10.1016/j.nimb.2010.02.091
27. Treeby, B.E., Cox, B.T.: k-Wave: MATLAB toolbox for the simulation and reconstruction of photoacoustic wave-fields. J. Biomed. Opt. **15**, 021314 (2010)
28. Patch, S.K., Hoff, D.E.M., Webb, T.B., Sobotka, L.G., Zhao, T.: Two-stage ionoacoustic range verification leveraging Monte Carlo and acoustic simulations to stably account for tissue inhomogeneity and accelerator-specific time structure: a simulation study. Med. Phys. **45**, 783–793 (2018). https://doi.org/10.1002/mp.12681
29. Lehrack, S., Assmann, W., Bertrand, D., Henrotin, S., Herault, J., Heymans, V., Stappen, F.V., Thirolf, P.G., Vidal, M., Van de Walle, J., Parodi, K.: Submillimeter ionoacoustic range determination for protons in water at a clinical synchrocyclotron. Phys. Med. Biol. **62**, L20 (2017)
30. Hickling, S., Xiang, L., Jones, K.C., Parodi, K., Assmann, W., Avery, S., Hobson, M., El Naqa, I.: Ionizing radiation-induced acoustics for radiotherapy and diagnostic radiology applications. Med. Phys. **45**, e707–e721 (2018). https://doi.org/10.1002/mp.12929
31. Patch, S.K., Santiago-Gonzalez, D., Mustapha, B.: Thermoacoustic range verification in the presence of acoustic heterogeneity and soundspeed errors: robustness relative to ultrasound image of underlying anatomy. Med. Phys. (2018). https://doi.org/10.1002/mp.13256

32. Da Silva, Anabela, Handschin, Charles, Metwally, Khaled, Garci, Houssem, Riedinger, Christophe, Mensah, Serge, Akhouayri, Hassan: Taking advantage of acoustic inhomogeneities in photoacoustic measurements. J. Biomed. Opt. **22**, 1–10 (2017)
33. Yu, Y., Li, Z., Zhang, D., Xing, L., Peng, H.: Simulation studies of time reversal-based protoacoustic reconstruction for range and dose verification in proton therapy. Med. Phys. **46**, 3649–3662 (2019). https://doi.org/10.1002/mp.13661

On the Transport Method for Hybrid Inverse Problems

Francis J. Chung, Jeremy G. Hoskins, and John C. Schotland

Abstract There are several hybrid inverse problems for equations of the form

$$\nabla \cdot D(x)\nabla u - \sigma(x)u = 0$$

in which we want to obtain the coefficients D and σ on a domain Ω when the solutions u are known. One approach is to use two solutions u_1 and u_2 to obtain a transport equation for the coefficient D, and then solve this equation inward from the boundary along the integral curves of a vector field X defined by u_1 and u_2. Bal and Ren have shown that for any nontrivial choices of u_1 and u_2, this method suffices to recover the coefficients almost everywhere on a dense set in Ω Bal and Ren in (Inv Prob 075003 [3]). This article presents an alternate proof of the same result from a dynamical systems point of view.

Keywords Hybrid inverse problems · Transport equation · Reconstruction of coefficients

MSC: 65N21 · 65N12

F. J. Chung
Department of Mathematics, University of Kentucky, Lexington, KY, USA
e-mail: fj.chung@uky.edu

J. G. Hoskins
Department of Mathematics, Yale University, New Haven, CT, USA
e-mail: jeremy.hoskins@yale.edu

J. C. Schotland (✉)
Department of Mathematics, University of Michigan, Ann Arbor, MI, USA
e-mail: schotland@umich.edu

L. Beilina et al. (eds.), *Mathematical and Numerical Approaches for Multi-Wave Inverse Problems*, Springer Proceedings in Mathematics & Statistics 328,
https://doi.org/10.1007/978-3-030-48634-1_2

1 Introduction

Suppose Ω is a smooth bounded domain in $\mathbb{R}^n$ and $f \in C^\infty(\partial\Omega)$. Let D be a uniformly positive function on Ω, σ be a nonnegative function on Ω, and consider the time-independent diffusion equation

$$\begin{aligned} \nabla \cdot D(x)\nabla u - \sigma(x)u = 0 \text{ on } \Omega \\ u|_{\partial\Omega} = f. \end{aligned} \tag{1}$$

For this paper we will take $D \in C^1(\Omega)$ and $\sigma \in C(\Omega)$.

In several hybrid inverse problems, we can take advantage of physical phenomena to recover the solution u to (1) for a given boundary condition f, without a priori knowledge of D and σ [2–5, 8, 10]. To complete these problems, we need a method of recovering D and σ from the solutions u.

One approach [3, 6, 11] is to note that the equation in (1) can be written out as

$$D\Delta u + \nabla D \cdot \nabla u - \sigma u = 0. \tag{2}$$

If u is known, this can be viewed as a transport equation for D with coefficients determined by u. Indeed, if we have two known solutions u_1 and u_2 to (2), we can multiply the equation for u_1 by u_2 and vice versa, and subtract the two to obtain

$$D\left(u_2\Delta u_1 - u_1\Delta u_2\right) + \nabla D \cdot (u_2\nabla u_1 - u_1\nabla u_2) = 0. \tag{3}$$

This eliminates σ to provide a transport equation for D with known coefficients. Assuming we can measure $D|_{\partial\Omega}$, then it follows from the basic theory of transport equations ([9], Ch. 3) that we can solve (3) to obtain D on all of the integral curves of the vector field

$$X := u_2\nabla u_1 - u_1\nabla u_2 \tag{4}$$

that intersect the boundary of Ω. Once $D(x)$ is known, we can solve for $\sigma(x)$ using (2). Note that the maximum principle implies that if u is positive on the boundary, then u must be positive inside the domain, eliminating the possibility of difficulties if $u(x) = 0$.

The major potential problem with the transport method is the possibility that not every point in Ω can be reached from the boundary by following an integral curve of X. In [6], the authors use the existence of complex geometrical optics (CGO) solutions to (2) to show that there exist boundary conditions f_1 and f_2 for which the corresponding solutions u_1 and u_2 give rise to a vector field X whose boundary-intersecting integral curves cover Ω. However, the rapid exponential decay of CGOs can be difficult to work with in practice. Fortunately, it turns out that any non-trivial positive boundary conditions yield a pair of solutions u_1, u_2 whose corresponding vector field X lets us recover the coefficients on a dense set in Ω. This follows from the argument given in the proof of Theorem 2.2 in [3]; a version of this same argument is used to analyze the stability of the reconstruction in [7]. This article presents an alternate proof of this result by considering the flow on $\varOmega$ generated by X and applying a dynamical systems point of view. More precisely, we prove the following.

Theorem 1 *Suppose $f_1, f_2 \in C(\partial\varOmega)$ with f_2 positive and f_1/f_2 not constant. Let u_1 and u_2 be the solutions to* (2) *with $u_1 = f_1$ and $u_2 = f_2$ on $\partial\varOmega$, and let X be the vector field defined by* (4)*. Then the union of the integral curves of X that intersect the boundary of $\varOmega$ is dense in $\varOmega$.*

In other words, given continuous, positive, linearly independent boundary conditions, we can get arbitrarily close to any point in $\varOmega$ from the boundary by following an integral curve of X. It follows that the transport method allows us to recover D and σ on a dense set without special care in selecting the boundary conditions f_1 and f_2. Note that if D and σ are a priori continuous, then we can recover D and σ on all of $\varOmega$ by continuity.

2 Proof of Theorem 1

To begin, we will fix some notation. Let u_1, u_2, and X be as in the statement of Theorem 1, and make the following definitions.

Definition 1 Let $x, y \in \bar{\varOmega}$. We say that $x \sim y$ if there exists an integral curve $\gamma : [0, b] \to \bar{\varOmega}$ defined by $\dot{\gamma}(t) = X(\gamma(t))$ such that both x and y lie in the image of γ.

Definition 2 For a set $A \subset \bar{\varOmega}$, define

$$\varSigma_A = \{y \in \bar{\varOmega} | y \sim x \text{ for some } x \in A\}.$$

In other words, $\varSigma_A$ is the union of all integral curves of X that intersect A.

With this notation, the statement of Theorem 1 is that $\varSigma_{\partial\varOmega}$ has full measure in $\varOmega$:

$$m(\varOmega \setminus \varSigma_{\partial\varOmega}) = 0.$$

Before beginning the proof of Theorem 1, we make the following remark. Since D is uniformly positive, we can replace X by DX in Definition 1. That is, the following definition is equivalent to Definition 1.

Definition 3 Let $x, y \in \bar{\varOmega}$. We say that $x \sim y$ if there exists an integral curve $\gamma : [0, b] \to \bar{\varOmega}$ defined by $\dot{\gamma}(t) = DX(\gamma(t))$ such that both x and y lie in the image of γ.

Indeed, if we have an integral curve $\gamma : [0, b] \to \bar{\varOmega}$ defined by the equation $\dot{\gamma}(t) = X(\gamma(t))$, we can define a function g by the ODE

$$\dot{g}(t) = D(\gamma(g(t))) \quad \text{and} \quad g(0) = 0.$$

Since D is uniformly positive, g is increasing, so there exists b' such that $g(b') = b$. Then we can define a new curve $\tilde{\gamma} : [0, b'] \to \bar{\Omega}$ by reparameterizing γ with g:

$$\tilde{\gamma}(t) = \gamma(g(t)).$$

Now $\tilde{\gamma}([0, b']) = \gamma([0, b])$ and

$$\dot{\tilde{\gamma}}(t) = DX(\tilde{\gamma}(t)).$$

Therefore if $x \sim y$ according to Definition 1 then $x \sim y$ according to Definition 3, and the converse follows similarly. With this in mind, we turn to the proof of Theorem 1.

Proof (*Proof of Theorem* 1) Suppose that $\Omega \setminus \Sigma_{\partial\Omega}$ has positive measure. Then there exists an open set U in Ω which is disjoint from $\Sigma_{\partial\Omega}$. Since no integral curve of the vector field DX joins any point of U to $\partial\Omega$, it follows that Σ_U is disjoint from $\partial\Omega$, and therefore $\Sigma_U \subset \Omega$.

Now the vector field DX gives a flow on Σ_U, defined for all time, that maps Σ_U to itself. Moreover,

$$\nabla \cdot DX = \nabla \cdot D(u_2 \nabla u_1 - u_1 \nabla u_2) = 0,$$

so the vector field DX is divergence free. This means that the flow of DX preserves volume, so the Poincaré Recurrence Theorem applies to maps defined by this flow. This gives us the following result, (see e.g. [1], pp. 71–72).

Proposition 1 (Poincaré Recurrence Theorem) *Let $W \subset \Sigma_U$ be open. For $x \in W$ and $k \in \mathbb{N}$, define*

$$x_k = \gamma_x(k),$$

where γ_x is the integral curve defined by $\dot{\gamma}_x(t) = DX(\gamma_x(t))$, with the initial condition $\gamma_x(0) = x$. Then for almost every $x \in W$, $x_k \in W$ for infinitely many k.

The basic idea of the proof of Theorem 1 is as follows. A short calculation shows that

$$X = u_2^2 \nabla u, \tag{5}$$

where $u = u_1/u_2$. The maximum principle, together with the positivity of f_2, guarantees that u_2 is uniformly positive, so u is well defined. Moreover the integral curves of X and DX are the same as the integral curves of ∇u, by the same logic used in the discussion of Definition 3. If any integral curve of X were closed, we could integrate ∇u along that curve and obtain two different values of u, which would be a contradiction. The main idea of the proof is to apply the Poincaré Recurrence Theorem to a well chosen subset $W \subset \Sigma_U$, to provide us with a trajectory that approximates a closed curve well enough to force a contradiction.

To obtain this subset W, define $u = u_1/u_2$. Since u is not constant at the boundary, unique continuation guarantees that u is not constant on Σ_U. Therefore there exists some point y in Σ_U such that

$$|\nabla u(y)| > 0.$$

Then the regularity of u_1 and u_2 guarantees that there exists an open set $V \subset \Sigma_U$ containing y and a positive constant c such that $|\nabla u| > c$ on V.

Now consider an open set W which contains y and is compactly contained in V. Applying the Poincaré Recurrence Theorem to W, we see that there exists $x_0 \in W$ such that $x_k \in W$ for infinitely many k.

Let $\{x_{k_j}\}$ denote the subsequence of $\{x_k\}$ such that $x_{k_j} \in W$, and let $\gamma^j : [k_j, k_{j+1}] \to \Omega$ be the integral curve of DX joining x_{k_j} to $x_{k_{j+1}}$. We can obtain $u(x_{k_{j+1}})$ from $u(x_{k_j})$ by integrating ∇u over γ^j; in other words

$$u(x_{k_{j+1}}) - u(x_{k_j}) = \int_{\gamma_j} \nabla u \cdot dr. \tag{6}$$

For each j, one of the following two things must happen:

Case I: the image of γ^j is entirely contained in V.
Case II: the image of γ^j contains points outside V.

In Case I, we can parametrize (6) to get

$$\begin{aligned} u(x_{k_{j+1}}) - u(x_{k_j}) &= \int_{k_j}^{k_{j+1}} \nabla u(\gamma^j(t)) \cdot \dot{\gamma}^j(t)\, dt \\ &= \int_{k_j}^{k_{j+1}} \nabla u(\gamma^j(t)) \cdot DX(\gamma^j(t))\, dt. \end{aligned}$$

Then (5) implies that

$$u(x_{k_{j+1}}) - u(x_{k_j}) = \int_{k_j}^{k_{j+1}} Du_2^2(\gamma^j(t))|\nabla u(\gamma^j(t))|^2\, dt.$$

Since the image of γ^j is entirely contained in V, and $k_{j+1} - k_j \geq 1$, we have

$$u(x_{k_{j+1}}) - u(x_{k_j}) \geq \min_{\Omega} Du_2^2 \cdot c^2 > 0.$$

In Case II, the length of the portion of γ^j contained in V must be at least twice the distance from W to the exterior of V, so (6) tells us that

$$u(x_{k_{j+1}}) - u(x_{k_j}) \geq 2c\, \mathrm{dist}(W, \mathrm{ext}\, V) > 0.$$

In both cases, $u(x_{k_{j+1}}) - u(x_{k_j})$ is bounded below uniformly in j. By setting q to be the minimum of the bounds in both cases, we see that $u(x_{k_{j+1}}) - u(x_{k_j}) \geq q$ for each $j \in \mathbb{N}$, and therefore u is unbounded in W. But this contradicts the continuity of u, which is guaranteed by the continuity and positivity of u_1 and u_2, and so our initial supposition is false. Therefore $\Omega \setminus \Sigma_{\partial\Omega}$ has measure zero as claimed.

As a final remark, note that if $\sigma \equiv 0$, we can take u_2 to be the identity function. Then (5) implies that $X = \nabla u_1$, and Theorem 1 gives us a useful corollary:

Corollary 1 *Suppose $u \in C^2(\Omega) \cap C^1(\bar{\Omega})$, and*

$$\nabla \cdot D(x)\nabla u = 0$$

in Ω. Then the set of integral curves of ∇u that intersect the boundary of Ω is dense in Ω.

Acknowledgements We are grateful to Guillaume Bal for valuable discussions. This work was supported in part by the NSF grant DMS-1912821 and the AFOSR grant FA9550-19-1-0320.

References

1. Arnold, V.I.: Mathematical Methods of Classical Mechanics, 2nd edn. Springer, Translated by K. Vogtmann and A. Weinstein (1978)
2. Bal, G., Ren, K.: On multi-spectral quantitative photoacoustic tomography. Inv. Prob. **28**, 025010 (2012)
3. Bal, G., Ren, K.: Multi-source quantitative PAT in diffusive regime. Inv. Prob. 075003 (2011)
4. Bal, G., Schotland, J.: Inverse scattering and acousto-optic imaging. Phys. Rev. Lett. **104**, 042902 (2010)
5. Bal, G., Uhlmann, G.: Reconstruction of coefficients in scalar second-order elliptic equations from knowledge of their solutions. Comm. Pure. App. Math. **66–10**, 1629–1652 (2013)
6. Bal, G., Uhlmann, G.: Inverse diffusion theory for photoacoustics. Inv. Prob. **26–8**, 085010 (2010)
7. Bonnetier, E., Choulli, M., Triki, F.: Stability for quantitative photoacoustic tomography revisited (2019). Preprint, arXiv 1905:07914
8. Chung, F.J., Hoskins, J., Schotland, J.: Coherent acousto-optic tomography with diffuse light (2019). Preprint
9. Evans, L.C.: Partial Differential Equations, 2nd edn. AMS (2010)
10. McLaughlin, J.R., Zhang, N., Manduca, A.: Calculating tissue shear modulus and pressure by 2D log-elastographic methods. Inv. Prob. **26**, 085007 (2010)
11. Ren, K., Gao, H., Zhao, H.: A hybrid reconstruction method for quantitative PAT. SIAM J. Im. Sci. **6**(1), 32–55 (2013)

Stable Determination of an Inclusion in a Layered Medium with Special Anisotropy

Michele Di Cristo

Abstract In this note we review some recent results concerning the inverse inclusion problem. In particular we analyze the stability issue for defect contained in layered medium where the conductivity is different in each layer. We consider conductivities with special anisotropy. The modulus of continuity obtained is of logarithmic type, which as shown in Di Cristo and Rondi (Inverse Prob 19:685–701 [13]) turns out to be optimal.

Keywords Inverse inclusion problem · Stability · Layered medium

MSC: 65N21 · 65N12

1 Introduction

In this note we review some recent results related to the inverse problem of determining an inclusion in a conductor body. This is a special instance of the well–known Calderon's inverse conductivity problem [5] and it has been studied by Isakov [15], who shows that the defect can be uniquely recovered through a knowledge of all possible boundary measurements. In this paper the author shows that the defect can be uniquely recovered through a knowledge of all possible electrostatic boundary measurements, making use of the Runge Approximation Theorem and solutions of the governing equation with Green's function type singularities. In 2005 Alessandrini and Di Cristo [2] have studied the stability issue, that is the continuous dependance of the inclusion from the given data. The approach proposed by the authors is to convert Isakov's idea in a quantitative form. Under mild a priori assumptions on the regularity and the topology of the inclusion, they show that the modulus of continuity of the stability issue is of logarithmic type.

M. Di Cristo (✉)
Politecnico di Milano, Milan, Italy
e-mail: michele.dicristo@polimi.it

L. Beilina et al. (eds.), *Mathematical and Numerical Approaches for Multi-Wave Inverse Problems*, Springer Proceedings in Mathematics & Statistics 328,
https://doi.org/10.1007/978-3-030-48634-1_3

Their result is proved for piecewise constant conductivities and for variable coefficients conductivities [7]. The argument proposed turns out to be extremely flexible and it has been extended to other physical situations governed by different differential equations. Logarithmic stability estimates hold true for the inverse problem of locating a scattered object by a knowledge of the near field data [8], or an inclusion in an elastic body by assuming the displacement and the traction on the boundary [3]. These papers are based on an accurate use of the fundamental solution of the differential operator involved and a precise and quantitative evaluation of unique continuation.

These arguments work well in different frameworks with isotropic conductivities in homogeneous conductors but they become more delicate when the physical phenomena take place in a layered medium with anisotropic conductivities. The key items, that cause the main difficulties, are the presence of an unknown boundary (the layer), when we apply the unique continuation technique, and the matrixes, that model the anisotropic conductivities that create big difficulties in estimating the fundamental solution. In several recent results [9–12, 14] these problems have been considered and some preliminary results in this direction are now available. In this paper we go through these results and summarize the situation showing the state of the art.

The paper is organized as follow. In the next Sect. 2 we define our notation and state the main theorem. Its proof is presented in Sect. 3 using some auxiliary results that are proved in Sect. 4.

2 Notations and Main Result

To begin with, let us premise some notations and definitions, we will use throughout the paper. Let Ω be a bounded open set in $\mathbb{R}^n$ and Σ a layer contained in it. The layer Σ will be a closed hyper-surface that separates Ω in to the union of three parts

$$\Omega = \Omega_+ \cup \Sigma \cup \Omega_-,$$

where $\Omega_\pm$ are open subsets such that $\partial\Omega_- = \partial\Omega \cup \Sigma$ and $\partial\Omega_+ = \Sigma$. We also denote by D a subset of Ω such that $D \subset \Omega_+ \subset \Omega$. We consider $\gamma(x)$ the conductivity if Ω of the form

$$\gamma_D(x) = c_1 A(x) + (c_2 - c_1)A(x)\chi_{\Omega_+} + (k - c_2 A(x))\chi_D,$$

where $A(x)$ is a known Lipischitz matrix valued function satisfying $\|A\|_{C^{0,1}(\Omega)} \leq \overline{A}$, where with k we mean the identity matrix multiplied by k and ellipticity condition with constant $\sigma > 0$, that is

$$\sigma^{-1}|\xi|^2 \leq A(x)\xi \cdot \xi \leq \sigma|\xi|^2, \quad \forall x \in \Omega, \xi \in \mathbb{R}^n,$$

c_1 and c_2 are given constants and k is an unknown constant.

For points $x \in \mathbb{R}^n$, we will write $x = (x', x_n)$, where $x' \in \mathbb{R}^{n-1}$ and $x \in \mathbb{R}$. Moreover, denoted by $\mathrm{dist}(\cdot, \cdot)$ the standard Euclidean distance, we define

$$B_r(x) = \{y \in \mathbb{R}^n | \mathrm{dist}(x, y) < r\}, \qquad B'_r(x') = \{y' \in \mathbb{R}^{n-1} | \mathrm{dist}(x', y') < r\}$$

as the open balls with radius r centered at x and x' respectively. We write $Q_r(x) = B'_r(x') \times (x_n - r, x_n + r)$ for the cylinder in $\mathbb{R}^n$. For simplicity, we use B_r, B'_r, Q_r instead of $B_r(0)$, $B'_r(0')$ and $Q_r(0)$ respectively. We shall also denote half domain, as well as its associated ball and cylinder

$$\mathbb{R}^n_+ = \{(x', x_n) \in \mathbb{R}^n | x_n > 0\}; \quad B_r^+ = B_r \cap \mathbb{R}^n_+; \quad Q_r^+ = Q_r \cap \mathbb{R}^n_+.$$

Definition 1 Let Ω be the bounded domain in $\mathbb{R}^n$. Given $\alpha \in (0, 1]$, we say a portion S of $\partial\Omega$ is of $C^{1,\alpha}$ class with constants r, $L > 0$ if for any point $p \in S$, there exists a rigid transformation $\varphi : \mathbb{R}^{n-1} \mapsto \mathbb{R}$ of coordinates under which we have $p = 0$ and

$$\Omega \cap B_r = \{(x', x_n) \in B_r | x_n > \varphi(x')\},$$

where $\varphi(\cdot)$ is a $C^{1,\alpha}$ function on B'_r, which satisfies

$$\varphi(0) = |\nabla\varphi(0)| = 0$$

and

$$||\varphi||_{C^{1,\alpha}(B'_r)} \leq Lr,$$

where the norm is defined as

$$||\varphi||_{C^{1,\alpha}(B'_r)} := ||\varphi||_{L^\infty(B'_r)} + r||\nabla\varphi||_{L^\infty(B'_r)} + r^{1+\alpha}|\nabla\varphi|_{\alpha,B'_r}$$

$$|\nabla\varphi|_{\alpha,B'_r} := \sup_{\substack{x',y' \in B'_r,\\ x' \neq y'}} \frac{|\nabla\varphi(x') - \nabla\varphi(y')|}{|x' - y'|^\alpha}.$$

For $f \in H^{1/2}(\partial\Omega)$, let u be the solution of the problem

$$\begin{cases} \mathrm{div}(\gamma_D(x)\nabla u) = 0 & \text{in } \Omega, \\ u = f & \text{on } \partial\Omega. \end{cases} \tag{1}$$

The inverse problem we addressed to is determine the anomalous region D when the Dirichlet-to-Neumann map Λ_D

$$\begin{array}{rcl} \Lambda_D : H^{1/2}(\partial\Omega) & \longrightarrow & H^{-1/2}(\partial\Omega) \\ f & \longrightarrow & \gamma(x)\frac{\partial u}{\partial u}, \end{array}$$

is given for any $f \in H^{1/2}(\partial\Omega)$. Here, ν denotes the outer unit normal to $\partial\Omega$, and $\frac{\partial u}{\partial\nu}_{|\partial\Omega}$ corresponds to the current density measured on $\partial\Omega$. Thus, the Dirichlet–to–Neumann map represents the knowledge of infinitely many boundary measurements.

Given constants $r_1, M_1, M_2, \delta_1, \delta_2 > 0$ and $0 < \alpha < 1$, we assume the domain $\Omega \subset \mathbb{R}^n$ is bounded

$$|\Omega| \leq M_2 r_1^n,$$

where $|\cdot|$ denotes the Lebesgue measure.

The interface Σ is C^2 and assumed to stay away from the boundary of the domain, as $\mathrm{dist}(\Sigma, \partial\Omega) \geq \delta_2$, and the inclusion D is assumed to stay away from Σ, as $\mathrm{dist}(D, \Sigma) \geq \delta_1$, and also $\Omega \backslash D$ is connected. Both ∂D and $\partial\Omega$ are of $C^{1,\alpha}$ class with constants r_1, M_1.

We refer to $n, r_1, M_1, M_2, \alpha, \delta_1, \delta_2$ as the **a priori data**. To study the stability, we denote by D_1 and D_2 two possible inclusions in Ω, which satisfy the above properties. The associated Dirichlet-to-Neumann maps are Λ_{D_1} and Λ_{D_2}. We also denote by $d_{\mathscr{H}}$ the Housdorff distance between closed sets.

Theorem 1 *Let $\Omega \subset \mathbb{R}^n$, $n \geq 2$ and we have two known constants c_1, c_2 and one unknown constant k, which are given. Let D_1, D_2 be two inclusions in Ω as above. If for any $\varepsilon > 0$ we have*

$$\|\Lambda_{D_1} - \Lambda_{D_2}\|_{\mathscr{L}(H^{1/2}, H^{-1/2})} \leq \varepsilon,$$

then

$$d_{\mathscr{H}}(\partial D_1, \partial D_2) \leq \omega(\varepsilon),$$

where ω is an increasing function on $[0, +\infty)$, which satisfies

$$\omega(t) \leq C|\log t|^{-\eta}, \quad \forall t \in (0, 1)$$

and $C > 0$, $0 < \eta \leq 1$ are constants depending on the a priori data only.

3 Proof of the Main Result

The proof of Theorem 1 is based on some auxiliary propositions whose proofs are collected in the next Sect. 4. We denote by $\mathscr{G}$ the connected component of $\Omega \backslash (D_1 \cup D_2)$, whose boundary contains $\partial\Omega$. $\Omega_D = \Omega \backslash \overline{\mathscr{G}}$, $S_{2r} := \{x \in \mathbb{R}^n | r \leq dist(x, \Omega) \leq 2r\}$, $S_r := \{x \in \mathscr{C}\Omega | dist(x, \Omega) \leq r\}$ and $\mathscr{G}^h := \{x \in \mathscr{G} | dist(x, \Omega_D) \geq h\}$, where $\mathscr{C}\Omega$ stands for the complement set of Ω. We recall that the layer Σ separates the domain into two parts known as Ω_- and Ω_+. We also define $\mathscr{F}^\lambda := \{x \in \Omega_- | dist(x, \Sigma) \geq \lambda\}$, and $\Sigma_\lambda := \{x \in \Omega_- | dist(x, \Sigma) = \lambda\}$

We introduce a variation of the Hausdorff distance called the *modified distance*, which simplifies our proof.

Definition 2 The modified distance between D_1 and D_2 is defined as

$$d_m(D_1, D_2) := \max \left\{ \sup_{x \in \partial \Omega_D \cap \partial D_1} dist(x, \partial D_2), \sup_{x \in \partial \Omega_D \cap \partial D_2} dist(x, \partial D_1) \right\}.$$

With no loss of generality, we can assume that there exists a point $O \in \partial D_1 \cap \partial \Omega_D$ such that the maximum of $d_m = d_m(D_1, D_2) = \mathrm{dist}(O, D_2)$ is attainted. We remark here that d_m is not a metric, and in general, it does not dominate the Hausdorff distance. However, under our *a priori* assumptions on the inclusion, the following lemma holds.

Lemma 1 *Under the assumptions of Theorem 1, there exists a constant $c_0 \geq 1$ only depending on M_1 and α such that*

$$d_{\mathscr{H}}(\partial D_1, \partial D_2) \leq c_0 d_m(D_1, D_2). \tag{2}$$

Proof See [2, Proposition 3.3]

Another obstacle comes from the fact that the propagation of smallness arguments are based on an iterated application of the three spheres inequality for solutions of the equation over chains of balls contained in $\mathscr{G}$. Therefore, it is crucial to control from below the radii of these balls. In the following Lemma 2 we treat the case of points of $\partial \Omega_D$ that are not reachable by such chains of balls. This problem was originally considered by [4] in the context of cracks detection in electrical conductors.

Let us premise some notations. Given $O = (0, \ldots, 0)$ the origin, v a unit vector, $H > 0$ and $\vartheta \in \left(0, \frac{\pi}{2}\right)$, we denote

$$C(O, v, \vartheta, H) = \left\{ x \in \mathbb{R}^n : \ |x - (x \cdot v)v| \leq \sin \vartheta |x|, \ 0 \leq x \cdot v \leq H \right\}$$

the closed truncated cone with vertex at O, axis along the direction v, height H and aperture 2ϑ. Given R, d, $0 < R < d$ and $Q = -de_n$, where $e_n = (0, \ldots, 0, 1)$, let us consider the cone $C\left(O, -e_n, \arcsin \frac{R}{d}, \frac{d^2 - R^2}{d}\right)$.

From now on, without loss of generality, we assume that

$$d_m(D_1, D_2) = \max_{x \in \partial D_1 \cap \partial \Omega_D} \mathrm{dist}(x, \partial D_2)$$

and we write $d_m = d_m(D_1, D_2)$.

We shall make use of paths connecting points in order that appropriate tubular neighborhoods of such paths still remain within $\mathbb{R}^n \setminus \Omega_D$. Let us pick a point $P \in \partial D_1 \cap \partial \Omega_D$, let ν be the outer unit normal to ∂D_1 at P and let $d > 0$ be such that the segment $[(P + d\nu), P]$ is contained in $\mathbb{R}^n \setminus \Omega_D$. Given $P_0 \in \mathbb{R}^n \setminus \Omega_D$, let γ be a path in $\mathbb{R}^n \setminus \Omega_D$ joining P_0 to $P + d\nu$. We consider the following neighborhood of

$\gamma \cup [(P + d\nu), P] \setminus \{P\}$ formed by a tubular neighborhood of γ attached to a cone with vertex at P and axis along ν

$$V(\gamma) = \bigcup_{S\in\gamma} B_R(S) \cup C\left(P, \nu, \arcsin\frac{R}{d}, \frac{d^2 - R^2}{d}\right). \tag{3}$$

Note that two significant parameters are associated to such a set, the radius R of the tubular neighborhood of γ, $\cup_{S\in\gamma} B_R(S)$, and the half-aperture $\arcsin\frac{R}{d}$ of the cone $C\left(P, \nu, \arcsin\frac{R}{d}, \frac{d^2-R^2}{d}\right)$. In other terms, $V(\gamma)$ depends on γ and also on the parameters R and d. At each of the following steps, such two parameters shall be appropriately chosen and shall be accurately specified. For the sake of simplicity we convene to maintain the notation $V(\gamma)$ also when different values of R, d are introduced. Also we warn the reader that it will be convenient at various stages to use a reference frame such that $P = O = (0, \dots, 0)$ and $\nu = -e_n$.

Lemma 2 *Under the above notation, there exist positive constants $\overline{d}$, c_1, where $\frac{\overline{d}}{\rho_0}$ only depends on M_1 and α, and c_1 only depends on M_1, α, M_2, and there exists a point $P \in \partial D_1$ satisfying*

$$c_1 d_m \leq dist(P, D_2),$$

and such that, giving any point $P_0 \in S_{2\rho_0}$, there exists a path $\gamma \subset (\overline{\Omega^{\rho_0}} \cup S_{2\rho_0}) \setminus \overline{\Omega_D}$ joining P_0 to $P + \overline{d}\nu$, where ν is the unit outer normal to D_1 at P, such that, choosing a coordinate system with origin O at P and axis $e_n = -\nu$, the set $V(\gamma)$ introduced in (3) *satisfies*

$$V(\gamma) \subset \mathbb{R}^n \setminus \Omega_D,$$

provided $R = \frac{\overline{d}}{\sqrt{1+L_0^2}}$, where L_0, $0 < L_0 \leq M_1$, is a constant only depending on M_1 and α.

Proof See [3, Lamma 4.2].

A crucial tool to get the stability estimates is the so called Alessandrini identity [1] the permits to relate the information provided by the boundary measurements with the unknown inclusion. Let $u_i \in H^1(\partial\Omega)$, $i = 1, 2$, solutions to (1) with conductivities

$$\gamma_{D_i}(x) = c_1 A(x) + (c_2 - c_1)A(x)\chi_{\Omega_+} + (k - c_2)\chi_{D_i}, \qquad i = 1, 2,$$

we have

$$\int_\Omega \left(\gamma_{D_1}\nabla u_1 \cdot \nabla u_2\right) - \int_\Omega \left(\gamma_{D_2}\nabla u_1 \cdot \nabla u_2\right) = \int_{\partial\Omega} c_1 A(x) u_1 [\Lambda_{D_1} - \Lambda_{D_2}] u_2. \tag{4}$$

Therefore, applying (4) replacing $u_i = \Gamma_{D_i}$, $i = 1, 2$, where Γ_{D_i} is the fundamental solution of the operator $\mathrm{div}(\gamma_i \nabla \cdot)$, we get

$$\int_{D_1} (k - c_2)\nabla \Gamma_{D_1}(\cdot, y) \cdot \nabla \Gamma_{D_2}(\cdot, z) - \int_{D_2} (k - c_2)\nabla \Gamma_{D_1}(\cdot, y) \cdot \nabla \Gamma_{D_2}(\cdot, z)$$
$$= \int_{\partial \Omega} c_1 A(\cdot)\Gamma_{D_1}(\cdot, y)[\Lambda_{D_1} - \Lambda_{D_2}]\Gamma_{D_2}(\cdot, z). \tag{5}$$

For $y, z \in \mathscr{G} \cap \mathscr{C}\Omega$, where $\mathscr{C}\Omega$ is the complementary set of Ω, we define

$$S_{D_1}(y, z) = (k - c_2) \int_{D_1} \nabla \Gamma_{D_1}(\cdot, y) \cdot \nabla \Gamma_{D_2}(\cdot, z)$$
$$S_{D_2}(y, z) = (k - c_2) \int_{D_2} \nabla \Gamma_{D_1}(\cdot, y) \cdot \nabla \Gamma_{D_2}(\cdot, z)$$
$$f(y, z) = S_{D_1}(y, z) - S_{D_2}(y, z).$$

Therefore (5) can be written as

$$f(y, z) = \int_{\partial \Omega} c_1 A(\cdot)\Gamma_{D_1}(\cdot, y)[\Lambda_{D_1} - \Lambda_{D_2}]\Gamma_{D_2}(\cdot, z), \quad \forall y, z \in \mathscr{C}\overline{\Omega}. \tag{6}$$

In what follows, we analyze the behavior of f and S_{D_i} as the singularities y and z get close to the inclusion D.

Proposition 1 *Let Ω, D_1, D_2 be open sets satisfying the above properties and let $y = h\nu(O)$. If, given $\varepsilon > 0$, we have*

$$||\Lambda_{D_1} - \Lambda_{D_2}||_{L(H^{1/2}, H^{-1/2})} < \epsilon, \tag{7}$$

then for every h where $0 < h < cr$, $0 < c < 1$, and c depends on M_1, we have

$$|f(y, y)| \leq C_0 \frac{\epsilon^{Bh^F}}{h^T}. \tag{8}$$

Here $0 < T < 1$ and $C_0, B, F > 0$ are constants that depend only on the a priori data.

Proposition 2 *Let Ω, D_1, D_2 be open sets satisfying the above properties and let $y = h\nu(O)$. Then for every $0 < h < r_0/2$*

$$|S_{D_1}(y, y)| \geq C_1 h^{2-n} - C_2 d_m^{2-2n} + C_3, \tag{9}$$

where $r_0 := \frac{r}{2} \min\left[\frac{1}{2}(8M_1)^{-1/\alpha}, \frac{1}{2}\right]$, and C_1, C_2, C_3 are positive constants depending only on the a priori data.

We can conclude this section proving our main theorem.

Proof *(Proof of Theorem 1)* We start from the origin of the coordinate system, point $O \in \partial D_1 \cap \partial \Omega_D$, for which the maximum in Definition 2 is attainted

$$d_m := d_m(D_1, D_2) = \operatorname{dist}(O, D_2).$$

By a transformation of coordinates, we can write $y = h\nu(O)$ where $0 < h < h_1, h_1 := \min\{d_m, cr, r_0/2\}, 0 < c < 1$, where c depends on M_1. By applying [2] Proposition 3.4 (i); i.e., $|\nabla_x \varGamma_{D_i}(x, y)| \le d_1|x - y|^{1-n}$, where $d_1 > 0$ depending only on k, n, α, M_1; we have

$$\begin{aligned} |S_{D_2}(y, y)| &= \left|(k - c_2)\int_{D_2} \nabla\varGamma_{D_1}(x, y)\nabla\varGamma_{D_2}(x, y)\right| \\ &\le d_2(k - c_2)\int_{D_2}(d_1|x - y|^{1-n})^2 \le d_1(k - c_2)d_1^2\int_{D_2}(|d_m - h|^{1-n})^2 \\ &\le d_1(k - c_2)d_1^2|d_m - h|^{2-2n}|D_2| \le C_4|d_m - h|^{2-2n}, \end{aligned} \tag{10}$$

where d_2, C_4 are constants depending on the a priori data only. Here $|D_2|$ is the measure of the inclusion D_2 which is bounded by $|D_2| \le |\varOmega| \le M_2 r_1^n$. If we apply the triangular inequality, we obtain

$$|S_{D_1}(y, y)| - |S_{D_2}(y, y)| \le |S_{D_1}(y, y) - S_{D_2}(y, y)| = |f(y, y)| \le C_0\frac{\epsilon^{Bh^F}}{h^T}. \tag{11}$$

Meanwhile, (9) gives us the lower bound of $S_{D_1}(y, y)$. Therefore, together with (10) and (11), we obtain

$$C_1h^{2-n} - C_2d_m^{2-2n} + C_3 \le C_4|d_m - h|^{2-2n} + C_0\frac{\epsilon^{Bh^F}}{h^T}$$

Rearranging terms we get

$$C_1h^{2-n} \le C_4|d_m - h|^{2-2n} + C_0\frac{\epsilon^{Bh^F}}{h^T}.$$

By setting $C_5 = C_4/C_0$ and $C_6 = C_1/C_0$

$$C_5|d_m - h|^{2-2n} \ge C_6h^{2-n} - \frac{\epsilon^{Bh^F}}{h^T} = C_6h^{2-n}(1 - \epsilon^{Bh^F}h^K),$$

where $0 < K = n - 2 - T$. Now let $h = h(\epsilon) = \min\left\{|\ln\epsilon|^{-\frac{1}{2F}}, d_m\right\}$, for $0 < \epsilon \le \epsilon_1, \epsilon_1 \in (0, 1)$ such that $\exp(-B|\ln\epsilon_1|^{1/2}) = 1/2$. It is easy to see if $d_m \le |\ln\epsilon|^{-\frac{1}{2F}}$, Theorem 1 is proved using Lemma 1. Indeed we can set $\eta = \frac{1}{2F} > 0$, then

$$d_{\mathscr{H}}(\partial D_1, \partial D_2) \le c_0 d_m \le c_0|\ln\epsilon|^{-\eta} = \omega(\epsilon) \tag{12}$$

In the other case if $d_m \ge |\ln\epsilon|^{-\frac{1}{2F}}$, it is easy to check

$$(d_m - h)^{2-2n} \geq \frac{C_6}{2C_5} h^{2-n} \quad \Longrightarrow \quad d_m \leq C_7 |\ln \epsilon|^{-\frac{n-2}{4F(n-1)}}.$$

Here we can solve d_m because here $h = h(\epsilon) = |\ln \epsilon|^{-\frac{1}{2F}}$, and C_7 depends only on the *a priori* data. Therefore we conclude the proof by setting $\eta = \frac{n-2}{4F(n-1)}$

$$d_{\mathscr{H}}(\partial D_1, \partial D_2) \leq c_0 d_m \leq c_0 C_7 |\ln \epsilon|^{-\eta} = \omega(\epsilon) \tag{13}$$

and for $\epsilon_1 \leq \epsilon$, we can also include the proof because $d_m \leq |\Omega| \leq M_2 r_1^n$.

$$d_{\mathscr{H}}(\partial D_1, \partial D_2) \leq c_0 d_m \leq c_0 M_2 r_1^n = \omega(\epsilon). \tag{14}$$

We can conclude the proof Theorem 1 by (12), (13) and (14)

$$d_{\mathscr{H}}(\partial D_1, \partial D_2) \leq C d_m = \omega(\epsilon),$$

where C only depends on the a priori data.

4 Proof of Propositions 1 and 2

In this section we prove the auxiliary propositions needed to prove our main theorem. The proofs are based on some quantitative estimates of unique continaution, which for this special context has been developed in [9] (see also [14]).

Proof *(Proposition 1)* Let us consider $f(y, \cdot)$ with a fixed $y \in S_{2r}$ then

$$\mathrm{div}_w(\gamma_D(x) \nabla f(y, w)) = 0 \quad \text{in } \mathscr{C}\overline{\Omega}_D. \tag{15}$$

For $x \in S_{2r}$, by (6) and (7), we have the smallness quantity

$$|f(y, x)| \leq C(r, M_1, M_2) ||\Gamma_{D_1} - \Gamma_{D_2}|| = \epsilon. \tag{16}$$

Also by y [Al-DC] Proposition 3.4, the uniform bound of f is given as

$$|f(y, x)| \leq c h^{2-2n}, \qquad \text{in } \mathscr{G}^h \cup \mathscr{F}^\lambda. \tag{17}$$

At this point the proposition can be obtained using iteratively the three sphere inequality derived in [12] for elliptic equation wtih coefficients with jump discontinuity (see also [6] for similar results) along the line of the proof of [2, Proposition 3.5].

Proof *(Proposition 2)* We write the upper bound of S_{D_1} as

$$\begin{aligned} |S_{D_1}(y,y)| &= \Big|(k-c_2)\int_{D_1}\nabla\Gamma_{D_1}(x,y)\nabla\Gamma_{D_2}(x,y)\mathrm{d}x\Big| \\ &\geq C\Big|\Big(\int_{D_1\cap B_r(O)\cap D_2}+\int_{D_1\cap B_\rho(O)\cap\mathscr{C}D_2}\Big)\nabla\Gamma_{D_1}\nabla\Gamma_{D_2}\Big| \\ &-C\Big|\int_{D_1\cap B_r(O)\cap\mathscr{C}B_\rho(O)\cap\mathscr{C}D_2}\nabla\Gamma_{D_1}\nabla\Gamma_{D_2}\Big| \\ &-C\Big|\int_{D_1\setminus B_r(O)}\nabla\Gamma_{D_1}\nabla\Gamma_{D_2}\Big| \end{aligned} \tag{18}$$

where C depends on $k, \overline{A}$ only, $r=|x-y|$, $0<r<r_0$, $0<\rho<\min\{d_m, r\}$. To explain the formula, notice we separate the integrand $\int_{D_1\cap B_r(O)}\nabla\Gamma_{D_1}\nabla\Gamma_{D_2}$ into two parts, because we don't have any information on x. So, either it can be $x\in D_1\cap B_r(O)\cap D_2$ or $x\in D_1\cap B_r(O)\cap\mathscr{C}D_2$. Then we separate the integrand again with respect to an even smaller ball $B_\rho(O)$.

If $x\in D_1\cap B_r(O)\cap D_2$, By [1, Lemma 3.1] and [10, Theorem 4.1], we get

$$\nabla\Gamma_{D_1}(x,y)\cdot\nabla\Gamma_{D_2}(x,y)\geq C_A|x-y|^{2-2n}=C_Ar^{2-2n}>0 \tag{19}$$

where C_A depends on the a priori data. If $x\in D_1\cap B_r(O)\cap\mathscr{C}D_2$, we consider in a smaller ball $B_\rho(O)$. In this case, we actually have $x\in D_1\cap B_\rho(O)\cap\mathscr{C}D_2$. By definition of d_m, $B_\rho(O)\cap D_2=\emptyset$, for $x,y\in B_\rho(O)$, we have

$$\begin{cases} \Delta\Big(\Gamma_{D_2}(x,y)-\Gamma(x,y)\Big)=0 \quad \text{in } B_\rho(O) \\ \Big(\Gamma_{D_2}(x,y)-\Gamma(x,y)\Big)|_{\partial B_\rho(O)}\leq C_K\rho^{2-n}, \end{cases}$$

where Γ denotes the standard fundamental solution of the Laplace operator. By the maximum principle, the value on interior is smaller than boundary

$$\Big|\Gamma_{D_2}(x,y)-\Gamma(x,y)\Big|\leq C_K\rho^{2-n}\quad\forall x,y\in B_\rho(O)$$

And by interior gradient bound, we have

$$\Big|\nabla\Gamma_{D_2}(x,y)-\nabla\Gamma(x,y)\Big|\leq C_{K_0}\rho^{1-n}\quad\forall x\in B_{\rho/2}(O);\forall y\in B_\rho(O)$$

Applying [A] Lemma 3.1 in $B_{\rho/2}(O)$, we have (notice $|x-y|=r>\rho$)

$$\nabla\Gamma_{D_1}(x,y)\cdot\nabla\Gamma_{D_2}(x,y)\geq C_A|x-y|^{2-2n}-C_K\rho^{2-2n}=C_Ar^{2-2n}-C_K\rho^{2-2n}>0 \tag{20}$$

Now we can bound the first term of (18) thanks to (19) and (20)

$$
\begin{aligned}
&\Big|\Big(\int_{D_1\cap B_r(O)\cap D_2} + \int_{D_1\cap B_\rho(O)\cap \mathscr{C} D_2}\Big)\nabla \Gamma_{D_1}\nabla \Gamma_{D_2}\Big| \\
\geq &\Big|\Big(\int_{D_1\cap B_r(O)\cap D_2} + \int_{D_1\cap B_\rho(O)\cap \mathscr{C} D_2}\Big)(C_A r^{2-2n} - C_K \rho^{2-2n})\Big| \qquad (21)\\
\geq &\Big|\Big(\int_{[D_1\cap B_r(O)\cap D_2]\cup[D_1\cap B_\rho(O)\cap \mathscr{C} D_2]}\Big)c_1 r^{2-2n}\Big| \geq c_1 h^{2-n}
\end{aligned}
$$

For the upper bounds of the second and third term, we can apply the natural bound of $\nabla \Gamma_{D_i}, i = 1, 2$. When $x \in D_1 \cap B_r(O) \cap \mathscr{C} B_\rho(O) \cap \mathscr{C} D_2$, we have

$$
\begin{aligned}
&\Big|\int_{D_1\cap B_r(O)\cap \mathscr{C}\, B_\rho(O)\cap \mathscr{C}\, D_2} \nabla \Gamma_{D_1}\nabla \Gamma_{D_2}\Big| \leq \Big|\int_{D_1\cap B_r(O)\cap \mathscr{C}\, B_\rho(O)\cap \mathscr{C}\, D_2} c_1|x-y|^{1-n}\cdot c_1|x-y|^{1-n}\Big| \\
\leq &\Big|\int_{D_1\cap B_r(O)\cap \mathscr{C}\, B_\rho(O)\cap \mathscr{C}\, D_2} c_1 r^{1-n}\cdot c_1 r^{1-n}\Big| \leq c_2 d_m^{2-2n} \qquad (22)
\end{aligned}
$$

$$
\begin{aligned}
\Big|\int_{D_1\setminus B_r(O)} \nabla \Gamma_{D_1}\nabla \Gamma_{D_2}\Big| &\leq \Big|\int_{D_1\setminus B_r(O)} c_1|x-y|^{1-n}\cdot c_1|x-y|^{1-n}\mathrm{d}x\Big| \\
&= \Big|\int_{D_1\setminus B_r(O)} c_1^2 r^{2-2n}\mathrm{d}x\Big| \qquad (23)\\
&= c_3
\end{aligned}
$$

Now we can plug (21), (22) and (23) into (18), we obtain the lower bound for $S_{D_1}(y, y)$

$$|S_{D_1}| \geq c_1 h^{2-n} - c_2 d_m^{2-2n} - c_3$$

where $c_i, i = 1, 2, 3$ depends only on the a prior data.

References

1. Alessandrini, G.: Singular solutions of elliptic equations and the determination of conductivity by boundary measurements. J. Differ. Eqn. **84**, 252–272 (1990)
2. Alessandrini, G., Di Cristo, M.: Stable determanation of an inclusion by boundary measurements SIAM. J. Math. Anal. **37**, 200–217 (2005)
3. Alessandrini, G., Di Cristo, M., Morassi, A., Rosset, E.: Stable determination of an inclusion in an elastic body by boundary measurements SIAM. J. Math. Anal. **46**, 2692–2729 (2014)
4. Alessandrini, G., Sincich, E.: Cracks with impedance, stable determination from boundary data Indiana Univ. Math. J. **62**(3), 947–989 (2013)
5. Calderon, A.: On an inverse boundary value problem. Seminar on Numerical Analysis and its Applications to Continuum Physics (Rio de Janeiro, 1980), pp. 65–73, Soc. Brasil. Mat., Rio de Janeiro (1980)
6. Carstea, A., Wang, J.N.: Propagation of smallness for an elliptic PDE with piecewise Lipschitz ceofficients. J. Differ. Eqn **268**, 7609-7628 (2020)
7. Di Cristo, M.: Stable determination of an inhomogeneous inclusion by local boundary measurements. J. Comput. Appl. Math. **198**, 414–425 (2007)
8. Di Cristo, M.: Stability estimates in the inverse transmission scattering problem. Inverse Probl. Imaging **3**, 551–565 (2009)

9. Di Cristo, M., Francini, E, Lin, C.L., Vessella, S., Wang, J.N.: Carleman estimate for second order elliptic equations with Lipschitz leading coefficients and jumps at an interface. J. Math. Pures Appl. **108**, 163–206 (2017)
10. Di Cristo, M., Ren, Y.: Stable determination of an inclusion for a class of anisotropic conductivities. Inverse Probl. **33**, 15 (2017)
11. Di Cristo, M., Ren, Y.: Stable determination of an inclusion in a layered medium. Math. Methods Appl. Sci. **41**, 4602–4611 (2018)
12. Di Cristo, M., Ren, Y.: Three sphere inequality for second order elliptic equations with coefficients with jump discontinuity. J. Diff. Eqn. **266**, 936–941 (2019)
13. Di Cristo, M., Rondi, L.: Examples of exponential instability for inverse inclusion and scattering problems. Inverse Prob. **19**, 685–701 (2003)
14. Francini, E., Lin, C.-L., Vessella, S., Wang, J.-N.: Three-region inequalities for the second order elliptic equation with discontinuous coefficients and size estimate. J. Diff. Eqn. **261**, 5306–5323 (2016)
15. Isakov, V.: On uniqueness of recovery of a discontinuous conductivity coefficient. Comm. Pure Appl. Math. **41**, 865–877 (1988)

Convergence of Stabilized $P1$ Finite Element Scheme for Time Harmonic Maxwell's Equations

M. Asadzadeh and Larisa Beilina

Abstract The paper considers the convergence study of the stabilized P1 finite element method for the time harmonic Maxwell's equations. The model problem is for the particular case of the dielectric permittivity function which is assumed to be constant in a boundary neighborhood. For the stabilized model a coercivity relation is derived that guarantee's the existence of a unique solution for the discrete problem. The convergence is addressed both in a priori and a posteriori settings. Our numerical examples validate obtained convergence results.

Keywords Time harmonic Maxwell's equations · P_1 finite elements · A priori estimate · A posteriori estimate · Convergence

MSC: 65N12 · 65N15 · 78M10

1 Introduction

In implementing the finite element methods for the Maxwell system, the divergence-free edge elements are the most advantageous from a theoretical point of view [12, 13]. On the other hand for the time-dependent problems, where a linear system of equations need to be solved at each iteration step, the divergence-free approach requires an unrealistic fine degree of time resolution. To circumvent this difficulty, it has been suggested to use the continuous P1 finite elements which provides inexpensive and reliable algorithms for the numerical simulations, in particular compared to $H(curl)$ conforming finite elements. Based on this fact, in this paper we consider stabilized P1 finite element for the approximate solution of time harmonic Maxwell's

M. Asadzadeh · L. Beilina (✉)
Department of Mathematical Sciences, Chalmers University of Technology
and University of Gothenburg, 412 96 Gothenburg, Sweden
e-mail: larisa.beilina@chalmers.se

M. Asadzadeh
e-mail: mohammad@chalmers.se

L. Beilina et al. (eds.), *Mathematical and Numerical Approaches for Multi-Wave Inverse Problems*, Springer Proceedings in Mathematics & Statistics 328,
https://doi.org/10.1007/978-3-030-48634-1_4

equations when the dielectric permittivity function is constant in a boundary neighborhood. This converts the Maxwell's equations into a set of time-independent wave equations on the boundary neighborhood.

An outline of this paper is as follows. In Sect. 2 we introduce a model problem for the time harmonic Maxwell's equations obtained through Laplace transform of the time-dependent equations. In Sect. 3 we introduce our finite element scheme, prove its well-posedness. As well as a optimal a priori and a posteriori error bounds which are derived in a, gradient dependent, triple norm. In the a posteriori case the boundary residual is in the form of a normal derivative and therefore is balanced by a multiplicative power of the mesh parameter h. Section 4 is devoted to implementations and justify the robustness of the approximation procedure. Finally, in Sect. 5 we conclude the results of the paper.

Throughout the paper C will denote a generic constant, not necessarily the same at each occurrence and independent of the mesh parameter and solution, unless otherwise specifically specified.

2 The Mathematical Model

We study the time-harmonic Maxwell's equations for electric field $\hat{E}\,(x, s)$, under the assumption of the vanishing electric charges, given by

$$\begin{aligned} s^2\varepsilon(x)\hat{E}(x,s) + \nabla\times\nabla\times\hat{E}(x,s) &= s\varepsilon(x)f_0(x), \quad x\in\mathbb{R}^d, \quad d=2,3 \\ \nabla\cdot(\varepsilon(x)\hat{E}(x,s)) &= 0 \end{aligned} \tag{1}$$

where $\varepsilon(x) = \varepsilon_r(x)\varepsilon_0$ is the dielectric permittivity, $\varepsilon_r(x)$ is the dimensionless relative dielectric permittivity and ε_0 is the permittivity of the free space. Furthermore

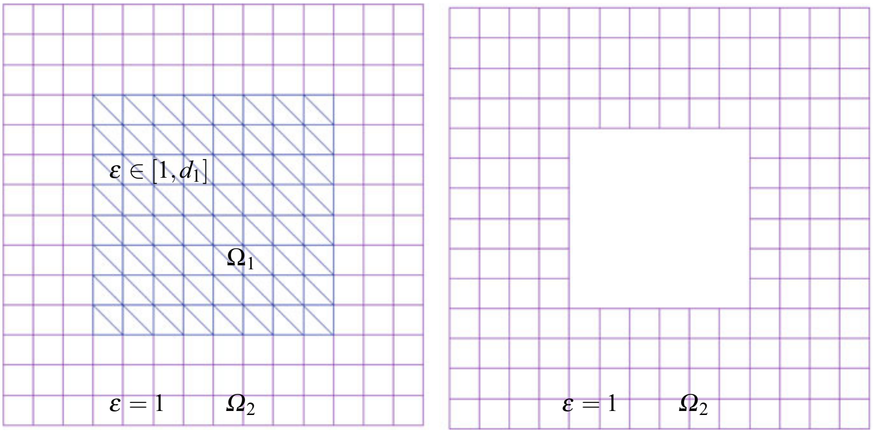
$$\nabla\times\nabla\times E = \nabla(\nabla\cdot E) - \nabla^2 E. \tag{2}$$

Equation (1) is obtained through the Laplace transformation in time

$$\widehat{E}(x,s) := \int\limits_0^{+\infty} E(x,t)e^{-st}dt, \qquad s = const. > 0 \tag{3}$$

where $E\,(x, t)$ is the the solution of time-dependent Maxwell's equations:

$$\begin{aligned} \varepsilon(x)\frac{\partial^2 E(x,t)}{\partial t^2} + \nabla\times\nabla\times E(x,t) = 0, \quad x\in\mathbb{R}^d, d=2,3, \quad t\in(0,T]. \\ \nabla\cdot(\varepsilon E)(x,t) = 0, \\ E(x,0) = f_0(x), \quad \frac{\partial E}{\partial t}(x,0) = 0, \quad x\in\mathbb{R}^d, \quad d=2,3. \end{aligned} \tag{4}$$



a) $\Omega = \Omega_1 \cup \Omega_2$

b) Ω_2

Fig. 1 Domain decomposition in Ω

Note that, since we have a non-zero initial condition: $E(x, 0) = f_0(x)$, the problem (4) is adequate as a coefficient inverse problem to determine the function $\varepsilon(x)$ in (4) through a finite number of observations E at the boundary [6].

To solve the problem (4) numerically, we consider it in a bounded convex and simply connected polygonal domain $\Omega \subset \mathbb{R}^d$, $d = 2, 3$ with boundary Γ: We define $\Omega_2 := \Omega \setminus \Omega_1$, where $\Omega_1 \subset \Omega$ has positive Lebesgue measure and $\partial\Omega \cap \partial\Omega_1 = \emptyset$. In this setting cutting out Ω_1 from Ω, the new subdomain Ω_2 shares the boundary with both Ω and Ω_1: $\partial\Omega_2 = \partial\Omega \cup \partial\Omega_1$, $\Omega = \Omega_1 \cup \Omega_2$, $\Omega_1 = \Omega \setminus \Omega_2$ and $\bar{\Omega}_1 \cap \bar{\Omega}_2 = \partial\Omega_1$, (see Fig. 1).

To proceed we assume that $\varepsilon(x) \in C^2(\mathbb{R}^d)$, $d = 2, 3$ satisfies

$$\begin{aligned} \varepsilon(x) &\in [1, d_1], && \text{for} \quad x \in \Omega_1, \\ \varepsilon(x) &= 1, && \text{for} \quad x \in \Omega \setminus \Omega_1, \\ \partial_\nu \varepsilon &= 0, && \text{for} \quad x \in \partial\Omega_2. \end{aligned} \tag{5}$$

Remark 1 Conditions (5) mean that, in the vicinity of the boundary of the computational domain Ω, the Eq. (4) transforms to a time-dependent wave equation.

At the boundary $\Gamma := \partial\Omega$ of Ω, we use the split $\Gamma = \Gamma_1 \cup \Gamma_2 \cup \Gamma_3$, so that Γ_1 and Γ_2 are the top and bottom sides, with respect to y- (in $2d$) or z-axis (in $3d$), of the domain Ω, respectively, while Γ_3 is the rest of the boundary. Further, $\partial_\nu(\cdot)$ denotes the normal derivative on Γ and ν is the outward unit normal to Γ.

Remark 2 In most estimates below, it suffices to restrict the Neumann boundary condition for the dielectric permittivity function to: $\partial_\nu \varepsilon(x) = 0$, on $\Gamma_1 \cup \Gamma_2$.

Now, using similar argument as in the studies in, e.g., [5] and by Remark 1, for the time-dependent wave equation, we impose first order absorbing boundary condition [11] at $\Gamma_1 \cup \Gamma_2$:

$$\partial_\nu E + \partial_t E = 0, \qquad (x,t) \in (\Gamma_1 \cup \Gamma_2) \times (0,T]. \tag{6}$$

To impose boundary conditions at Γ_3 we can assume that the surface Γ_3 is located far from the domain Ω_1. Hence, we can assume that $E \approx E^{inc}$ in a vicinity of Γ_3, where E^{inc} is the incident field. Thus, at Γ_3 we may impose Neumann boundary condition

$$\partial_\nu E = 0, \qquad (x,t) \in \Gamma_3 \times (0,T]. \tag{7}$$

Finally, using the well known vector-analysis relation (2) and applying the Laplace transform to the Eq. (4) and the boundary conditions (6)–(7) in the time domain, the problem (1) will be transformed to the following model problem

$$\begin{aligned} s^2\varepsilon(x)\hat{E}(x,s) + \nabla(\nabla \cdot \hat{E}(x,s)) - \triangle \hat{E}(x,s) &= s\varepsilon(x) f_0(x), \quad x \in \mathbb{R}^d, d = 2,3 \\ \nabla \cdot (\varepsilon(x)\hat{E}(x,s)) &= 0, \\ \partial_\nu \hat{E}(x,s) &= 0, \quad x \in \Gamma_3, \\ \partial_\nu \hat{E}(x,s) &= f_0(x) - s\hat{E}(x,s), \;\; x \in \Gamma_1 \cup \Gamma_2. \end{aligned} \tag{8}$$

3 Finite Element Method

We have the usual notation of the inner product in $[L_2(\Omega)]^d$: $(\cdot,\cdot)$, $d \in \{2,3\}$, and the corresponding norm $\|\cdot\|$, whereas $\langle \cdot,\cdot \rangle_\Gamma$ is the inner product of $[L_2(\Gamma)]^{d-1}$ and the associated $L_2(\Gamma)$-norm is denoted by $\|\cdot\|_\Gamma$. We define the L_2 scalar products

$$(u,v) := \int_\Omega u \cdot v \, d\mathbf{x}, \quad (u,v)_\omega := \int_\Omega u \cdot v \, \omega d\mathbf{x}, \quad \langle u,v \rangle_\Gamma := \int_\Gamma u \cdot v \, d\sigma,$$

and the ω-weighted $L^2(\Omega)$ norm

$$\|u\|_\omega := \sqrt{\int_\Omega |u|^2 \, \omega d\mathbf{x}}, \qquad \omega > 0, \quad \omega \in L^\infty(\Omega).$$

3.1 Stabilized Model

The stabilized formulation of the problem (8), with $d = 2, 3$, reads as follows:

$$\begin{aligned} s^2\varepsilon(x)\hat{E}(x,s) - \triangle\hat{E}(x,s) - \nabla(\nabla\cdot((\varepsilon-1)\hat{E}(x,s)) &= s\varepsilon(x)f_0(x) \quad x \in \mathbb{R}^d, \\ \partial_\nu\hat{E}(x,s) &= 0, \quad x \in \Gamma_3, \\ \partial_\nu\hat{E}(x,s) &= f_0(x) - s\hat{E}(x,s), \quad x \in \Gamma_1 \cup \Gamma_2, \end{aligned} \tag{9}$$

where the second equation of (8) is hidden in the first one.

3.2 Finite Element Discretization

We consider a partition of Ω into elements K denoted by $\mathscr{T}_h = \{K\}$, satisfying the minimal angle condition. Here, $h = h(x)$ is the mesh parameter defined as $h|_K = h_K$, representing the local diameter of the elements. We also denote by $\partial\mathscr{T}_h = \{\partial K\}$ a partition of the boundary Γ into boundaries ∂K of the elements K such that vertices of these elements lie on Γ.

To formulate the finite element method for (9) in Ω, we introduce the, piecewise linear, finite element space $W_h^E(\Omega)$ for every component of the electric field E:

$$W_h^E(\Omega) := \{w \in H^1(\Omega) : w|_K \in P_1(K), \ \ \forall K \in \mathscr{T}_h\},$$

where $P_1(K)$ denote the set of piecewise-linear functions on K. Setting $\mathbf{W}_h^E(\Omega) := [W_h^E(\Omega)]^3$ we define f_{0h} to be the $\mathbf{W}_h^E$-interpolant of f_0 in (9). Then the finite element method for the problem (9) is formulated as: *Find* $\hat{E}_h \in \mathbf{W}_h^E(\Omega)$ *such that* $\forall\mathbf{v} \in \mathbf{W}_h^E(\Omega)$

$$\begin{aligned} &(s^2\varepsilon\hat{E}_h, \mathbf{v}) + (\nabla\hat{E}_h, \nabla\mathbf{v}) + (\nabla\cdot(\varepsilon\hat{E}_h), \nabla\cdot\mathbf{v}) - (\nabla\cdot\hat{E}_h, \nabla\cdot\mathbf{v}) \\ &+ \langle s\hat{E}_h, \mathbf{v}\rangle_{\Gamma_1\cup\Gamma_2} = (s\varepsilon f_{0h}, \mathbf{v}) + \langle f_{0h}, \mathbf{v}\rangle_{\Gamma_1\cup\Gamma_2}. \end{aligned} \tag{10}$$

Theorem 1 (well-posedness) *Under the condition*

$$f_{0,h} \in L_{2,\varepsilon} \cap L_{2,1/s}(\Gamma_1 \cup \Gamma_2), \tag{11}$$

on the data, the problem (10) *has a unique solution* $\hat{E}_h \in \mathbf{W}_h^E(\Omega)$.

Proof See [1].

3.3 Error Analysis

In this subsection first we give a swift a priori error bound and then continue with a posteriori error estimates. For the sake of completeness, we set up an adaptive algorithm for the a posteriori setting. This, however, requires a thorough and lengthy implementations procedure which is beyond the scope of the present paper and may be considered in a future study.

3.3.1 A Priori Error Estimates

To derive a priori error estimates we consider the continuous variational formulation and define linear and bilinear forms in the finite element space $\mathbf{W}_h^E(\Omega)$:

$$\begin{aligned} a(\hat{E}, \mathbf{v}) =& (s^2 \varepsilon \hat{E}, \mathbf{v}) + (\nabla \hat{E}, \nabla \mathbf{v}) + (\nabla \cdot (\varepsilon \hat{E}), \nabla \cdot \mathbf{v}) \\ & - (\nabla \cdot \hat{E}, \nabla \cdot \mathbf{v}) + \langle s\hat{E}, \mathbf{v} \rangle_{\Gamma_1 \cup \Gamma_2}, \qquad \forall \mathbf{v} \in H^1(\Omega) \end{aligned} \tag{12}$$

and

$$\mathscr{L}^c(\mathbf{v}) := (s\varepsilon f_0, \mathbf{v}) + \langle f_0, \mathbf{v} \rangle_{\Gamma_1 \cup \Gamma_2}, \qquad \forall \mathbf{v} \in H^1(\Omega). \tag{13}$$

Hence we have the concise form of the variational formulation

$$a(\hat{E}, \mathbf{v}) = \mathscr{L}^c(\mathbf{v}), \qquad \forall \mathbf{v} \in H^1(\Omega). \tag{14}$$

This yields the *Galerkin orthogonality* [7] by letting, in (12) and (13), $\mathbf{v} \in \mathbf{W}_h^E(\Omega)$, as well as replacing f_0 by $f_{0,h}$ in (13). Subtracting from (14) its discrete version and letting $e(x, s) := \hat{E}(x, s) - \hat{E}_h(x, s)$ be the pointwise spatial error of the finite element approximation (10), we get

$$a(\hat{E} - \hat{E}_h, \mathbf{v}) = 0, \qquad \forall \mathbf{v} \in \mathbf{W}_h^E(\Omega), \qquad \text{(Galerkin orthogonality)}. \tag{15}$$

Now we are ready to derive the following theoretical error bound

Theorem 2 *Let $\hat{E}$ and $\hat{E}_h$ be the solutions for the continuous problem (9) and its finite element approximation, (10), respectively. Then, there is a constant C, independent of $\hat{E}$ and h, such that*

$$|||e||| \le C \parallel h\hat{E} \parallel_{H_w^2(\Omega)} .$$

where $w = w(\varepsilon(x), s)$ is the weight function which depends on the dielectric permittivity function $\varepsilon(x)$ and the pseudo-frequency variable s.

Proof See the proof of Theorem 1 in [1].

3.3.2 A Posteriori Error Estimates

For the approximate solution $\hat{E}_h = \hat{E}_h(x, s)$ of the problem (9), we define the residual errors

$$
\begin{aligned}
-\mathscr{R}(\hat{E}_h) :=& s^2\varepsilon(x)\hat{E}_h - \triangle\hat{E}_h - \nabla(\nabla\cdot((\varepsilon(x)-1)\hat{E}_h) - s\varepsilon(x)f_{0,h}(x), \quad \text{and}\\
-\mathscr{R}_\Gamma(\hat{E}_h) :=& h^{-\alpha}\Big(\partial_\nu\hat{E}_h + s\hat{E}_h - f_{0,h}(x)\Big), \quad \text{for} \quad x\in\Gamma_1\cup\Gamma_2, \quad 0<\alpha\le 1.
\end{aligned}
\tag{16}
$$

By the Galerkin orthogonality we have that $\mathscr{R}(\hat{E}_h) \perp \mathbf{W}_h^E(\Omega)$. Now the objective is to bound the triple norm of the error $e(x, s) := \hat{E}(x, s) - \hat{E}_h(x, s)$ by some adequate norms of $\mathscr{R}(\hat{E}_h)$ and $\mathscr{R}_\Gamma(\hat{E}_h)$ with a relevant, fast, decay. This may be done in a few, relatively similar, ways, e.g., one can use the variational formulation and interpolation in the error combined with Galerkin orthogonality. Or one may use a dual problem approach setting the source term (or initial data) on the right hand side as the error.

The proof of the main result relies on assuming a first order approximation for the initial value of the original field $f_0(x) := E(x, t)|_{t=0_-}$, for $\beta \approx 1$, viz,

$$
\| f_0 - f_{0,h} \|_\varepsilon \approx \| f_0 - f_{0,h} \|_{1/s,\Gamma} \approx \| f_0 - f_{0,h} \|_{(\varepsilon-1)^2/s,\Gamma} = \mathscr{O}(h^\beta). \tag{17}
$$

Theorem 3 *Let $\hat{E}$ and $\hat{E}_h$ be the solutions for the continuous problem* (9) *and its finite element approximation* (10)*, respectively. Further we assume that we have the error bound* (17) *for the initial field $f_0(x) := E(x, t)|_{t=0_-}$. Then, there exist interpolation constants C_1 and C_2, independent of h, and $\hat{E}$, but may depend on ε and s such that the following a posteriori error estimate holds true*

$$
|||e||| \le C_1\, h \,\| \mathscr{R} \| + C_2\, h^\alpha \,\| \mathscr{R}_\Gamma \|_{1/s,\, \Gamma_1\cup\Gamma_2} + \mathscr{O}(h^\beta), \tag{18}
$$

where $\alpha \approx \beta \approx 1$.

Proof See [1]

An adaptivity algorithm

Given an *admissible* small error tolerance $TOL > 0$, we outline formal adaptivity steps to reach

$$
|||e||| \le TOL. \tag{19}
$$

To this end we start with a course mesh with mesh size h and

Step 1. Compute the approximate solution $\hat{E}_h$ and its corresponding domain and boundary residuals $\mathscr{R}$ and $\mathscr{R}_\Gamma$, respectively.

Step 2. Check whether

$$
C_1\, h \,\| \mathscr{R} \| + C_2\, h^\alpha \,\| \mathscr{R}_\Gamma \|_{1/s,\, \Gamma_1\cup\Gamma_2} + \mathscr{O}(h^\beta) \le TOL? \tag{20}
$$

for $\alpha \approx \beta \approx 1$.

Step 3. If (20) is valid stop and accept the current h-function. Otherwise, refine in regions where the contribution to the right hand side in (18) is *large* (on each iteration step you need to choose a criterion for this *largeness*). Replace the h-function by the new refined one and go to Step 1.

4 Numerical Examples

We refer to [1] for complete description of numerical tests. Numerical tests are performed in the computational domain $\Omega = [0, 1] \times [0, 1]$. The source data $f(x)$, $x \in \mathbb{R}^2$ (the right hand side) in the model problem (8) for the electric field $\hat{E} = (\hat{E}_1, \hat{E}_2)$ is chosen such that the function

$$\begin{aligned} \hat{E}_1 &= \frac{2}{s^3\varepsilon}\pi \sin^2 \pi x \cos \pi y \sin \pi y, \\ \hat{E}_2 &= -\frac{2}{s^3\varepsilon}\pi \sin^2 \pi y \cos \pi x \sin \pi x. \end{aligned} \tag{21}$$

is the exact solution of the model problem (8).

We define the function ε as

$$\varepsilon(x, y) = \begin{cases} 1 + \sin^m \pi(2x - 0.5) \cdot \sin^m \pi(2y - 0.5) & \text{in } [0.25, 0.75] \times [0.25, 0.75], \\ 1 & \text{otherwise.} \end{cases} \tag{22}$$

for an integer $m > 1$.

The computational domain Ω is discretized into triangles K of sizes $h_l = 2^{-l}$, $l = 1, ..., 6$. Numerical tests are performed for different $m = 2, ..., 9$ in (22), $s = 20$ in (8), and the relative errors e_l^1, e_l^2 are measured in L_2-norm and the H^1-norms, respectively, which we compute as

$$e_l^1 = \frac{\|\hat{E} - \hat{E}_h\|_{L_2}}{\|\hat{E}\|_{L_2}}, \tag{23}$$

$$e_l^2 = \frac{\|\nabla(\hat{E} - \hat{E}_h)\|_{L_2}}{\|\nabla\hat{E}\|_{L_2}}. \tag{24}$$

Here,

$$\hat{E} := \sqrt{\hat{E}_1^2 + \hat{E}_2^2} \qquad \hat{E}_h := \sqrt{\hat{E}_{1h}^2 + \hat{E}_{2h}^2}. \tag{25}$$

Figure 2 presents convergence of P1 finite element scheme for $m = 2, 9$ in (22). Tables 1 and 2 present convergence rates q_1, q_2 for $m = 2, 9$ which we compute as

$$q_1 = \frac{\log\left(\frac{e^1_{l\,h}}{e^1_{l\,2h}}\right)}{\log(0.5)}, \quad q_2 = \frac{\log\left(\frac{e^2_{l\,h}}{e^2_{l\,2h}}\right)}{\log(0.5)}, \tag{26}$$

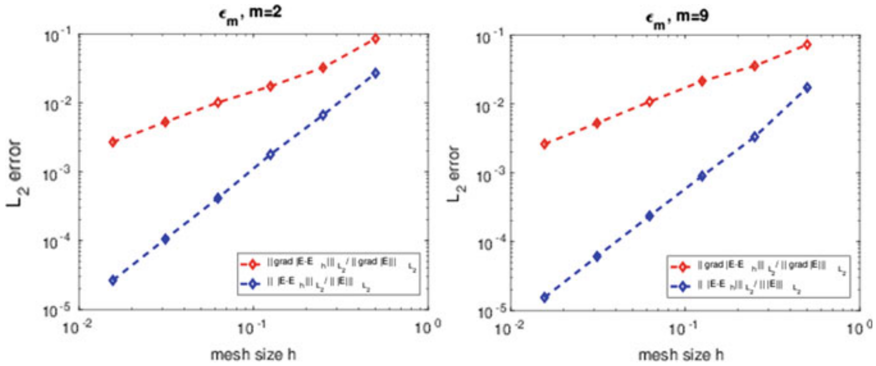


Fig. 2 Relative errors for $m = 2$ (left) and $m = 9$ (right)

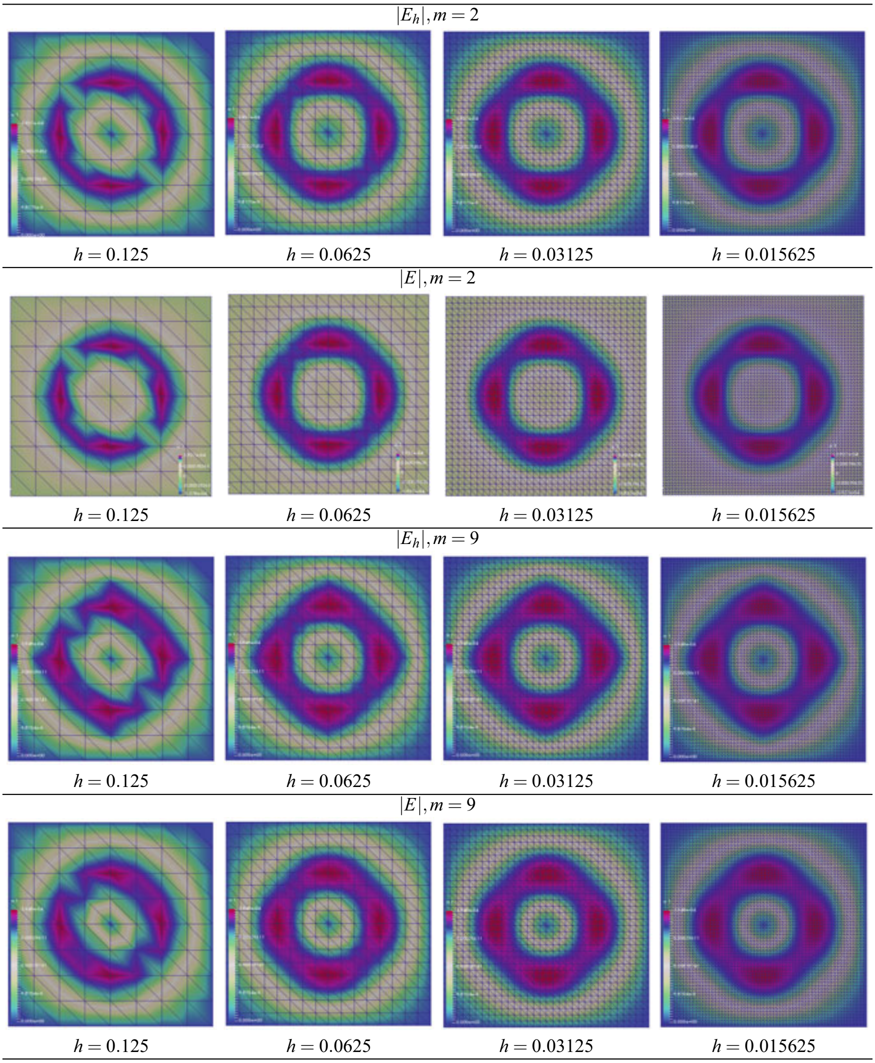
Table 1 Relative errors in the L_2-norm and in the H^1-norm for mesh sizes $h_l = 2^{-l}, l = 1, ..., 6$, for $m = 2$ in (22). Here, *nel* is number of elements and *nno* is number of nodes in the mesh

l	nel	nno	e_l^1	q_1	e_l^2	q_2
1	8	9	$2.71 \cdot 10^{-2}$		$8.60 \cdot 10^{-2}$	
2	32	25	$6.66 \cdot 10^{-3}$	2.02	$3.25 \cdot 10^{-2}$	1.40
3	128	81	$1.78 \cdot 10^{-3}$	1.90	$1.75 \cdot 10^{-2}$	$8.99 \cdot 10^{-1}$
4	512	289	$4.13 \cdot 10^{-4}$	2.11	$1.02 \cdot 10^{-2}$	$7.79 \cdot 10^{-1}$
5	2048	1089	$1.05 \cdot 10^{-4}$	1.97	$5.29 \cdot 10^{-3}$	$9.42 \cdot 10^{-1}$
6	8192	4225	$2.65 \cdot 10^{-5}$	1.99	$2.70 \cdot 10^{-3}$	$9.69 \cdot 10^{-1}$

Table 2 Relative errors in the L_2-norm and in the H^1-norm for mesh sizes $h_l = 2^{-l}, l = 1, ..., 6$, for $m = 9$ in (22). Here, *nel* is number of elements and *nno* is number of nodes in the mesh

l	nel	nno	e_l^1	q_1	e_l^2	q_2
1	8	9	$1.73 \cdot 10^{-2}$		$7.29 \cdot 10^{-2}$	
2	32	25	$3.33 \cdot 10^{-3}$	2.38	$3.57 \cdot 10^{-2}$	1.03
3	128	81	$8.98 \cdot 10^{-4}$	1.89	$2.15 \cdot 10^{-2}$	$7.33 \cdot 10^{-1}$
4	512	289	$2.36 \cdot 10^{-4}$	1.93	$1.08 \cdot 10^{-2}$	$9.94 \cdot 10^{-1}$
5	2048	1089	$6.09 \cdot 10^{-5}$	1.96	$5.26 \cdot 10^{-3}$	1.04
6	8192	4225	$1.55 \cdot 10^{-5}$	1.98	$2.62 \cdot 10^{-3}$	1.00

where $e^i_{l\,h}$, $e^i_{l\,2h}$, $i = 1, 2$, are computed relative norms e^i_l, $i = 1, 2$, on the finite element mesh with the mesh size h and $2h$, respectively. Similar convergence rates are obtained for $m = 3, 4, 5, 8$. Figure 3 shows computed and exact solutions on different finite element meshes for $m = 2$ and $m = 9$ in (22). We observe that our P1 finite element scheme behaves like a first order method for $H^1(\Omega)$-norm and second order method for $L^2(\Omega)$-norm.



$|E_h|, m = 2$

$h = 0.125$ $h = 0.0625$ $h = 0.03125$ $h = 0.015625$

$|E|, m = 2$

$h = 0.125$ $h = 0.0625$ $h = 0.03125$ $h = 0.015625$

$|E_h|, m = 9$

$h = 0.125$ $h = 0.0625$ $h = 0.03125$ $h = 0.015625$

$|E|, m = 9$

$h = 0.125$ $h = 0.0625$ $h = 0.03125$ $h = 0.015625$

Fig. 3 Computed versus exact solution for different meshes taking $m = 2$ *and* $m = 9$ *in* (22)

5 Conclusion

We presented convergence analysis for the stabilized P1 finite element scheme applied to the solution of time harmonic Maxwell's equations with constant dielectric permittivity function $\varepsilon(x)$ in a boundary neighborhood. For the convergence study of stabilized P1 finite element method for a time dependent problem for Maxwell's equations we refer to [2]. Optimal a priori and a posteriori error bounds are derived in weighted energy norms and numerical results validate obtained theoretical error bounds.

Proposed scheme can be applied for the solution of coefficient inverse problems with constant dielectric permittivity function in a boundary neighborhood, see [3–5, 8–10, 14, 15] for a such problems.

References

1. Asadzadeh, M., Beilina, L.: On stabilized P1 finite element approximation fortime harmonic Maxwell's equations. (2019). arXiv:1906.02089
2. Beilina, L., Ruas, V.: An explicit P1 finite element scheme for Maxwell's equations with constant permittivity in a boundary neighborhood. arXiv:1808.10720
3. Beilina, L., Klibanov, M.V.: Approximate Global Convergence and Adaptivity for Coefficient Inverse Problems. Springer, New York (2012)
4. Beilina, L., Th'anh, N.T., Klibanov, M., Malmberg, J.B.: Reconstruction of shapes and refractive indices from blind backscattering experimental data using the adaptivity. Inverse Probl. **30**, 105007 (2014)
5. Beilina, L., Th'anh, N.T., Klibanov, M.V., Malmberg, J.B.: Globally convergent and adaptive finite element methods in imaging of buried objects from experimental backscattering radar measurements. J. Comput. Appl. Math. (2015). Elsevier. https://doi.org/10.1016/j.cam.2014.11.055
6. Bellassoued, M., Cristofol, M., Soccorsi, E.: Inverse boundary value problem for the dynamical heterogeneous Maxwell's system. Inverse Probl. **28**, 095009 (188 pp.) (2012)
7. Brenner, S.C., Scott, L.R.: The Mathematical Theory of Finite Element Methods. Springer, Berlin (1994)
8. Bondestam Malmberg, J., Beilina, L.: An adaptive finite element method in quantitative reconstruction of small inclusions from limited observations. Appl. Math. Inf. Sci. **12**(1), 1–19 (2018)
9. Burov, V.A., Morozov, S.A., Rumyantseva, O.D.: Reconstruction of fine-scale structure of acoustical scatterers on large-scale contrast background. Acoust. Imaging **26**, 231–238 (2002)
10. Bulyshev, A.E., Souvorov, A.E., Semenov, S.Y., Posukh, V.G., Sizov, Y.E.: Three-dimensional vector microwave tomography: theory and computational experiments. Inverse Probl. **20**(4), 1239–1259 (2004)
11. Engquist, B., Majda, A.: Absorbing boundary conditions for the numerical simulation of waves. Math. Comp. **31**, 629–651 (1977)
12. Monk, P.B.: Finite Element Methods for Maxwell's Equations. Oxford University Press (2003)
13. Nédélec, J.-C.: Mixed finite elements in R3. Numer. Math. **35**, 315–341 (1980)
14. Thánh, N.T., Beilina, L., Klibanov, M.V., Fiddy, M.A.: Reconstruction of the refractive index from experimental backscattering data using a globally convergent inverse method. SIAMJ. Sci. Comput. **36**, B273–B293 (2014)
15. Thánh, N.T., Beilina, L., Klibanov, M.V., Fiddy, M.A.: Imaging of buried objects from experimental backscattering time-dependent measurements using a globally convergent inverse algorithm. SIAM J. Imaging Sci. **8**(1), 757–786 (2015)

Regularized Linear Inversion with Randomized Singular Value Decomposition

Kazufumi Ito and Bangti Jin

Abstract In this work, we develop efficient solvers for linear inverse problems based on randomized singular value decomposition (RSVD). This is achieved by combining RSVD with classical regularization methods, e.g., truncated singular value decomposition, Tikhonov regularization, and general Tikhonov regularization with a smoothness penalty. One distinct feature of the proposed approach is that it explicitly preserves the structure of the regularized solution in the sense that it always lies in the range of a certain adjoint operator. We provide error estimates between the approximation and the exact solution under canonical source condition, and interpret the approach in the lens of convex duality. Extensive numerical experiments are provided to illustrate the efficiency and accuracy of the approach.

Keywords Linear inverse problmes · Randomized singular value decomposition · Tikhonov regularization

MSC: 65F22

1 Introduction

This work is devoted to randomized singular value decomposition (RSVD) for the efficient numerical solution of the following linear inverse problem

$$Ax = b, \tag{1}$$

K. Ito
Department of Mathematics, North Carolina State University, Raleigh, NC 27607, USA
e-mail: kito@ncsu.edu

B. Jin (✉)
Department of Computer Science, University College London, Gower Street, London WC1E 6BT, UK
e-mail: b.jin@ucl.ac.uk; bangti.jin@gmail.com

L. Beilina et al. (eds.), *Mathematical and Numerical Approaches for Multi-Wave Inverse Problems*, Springer Proceedings in Mathematics & Statistics 328,
https://doi.org/10.1007/978-3-030-48634-1_5

where $A \in \mathbb{R}^{n\times m}$, $x \in \mathbb{R}^m$ and $b \in \mathbb{R}^n$ denote the data formation mechanism, unknown parameter and measured data, respectively. The data b is generated by $b = b^\dagger + e$, where $b^\dagger = Ax^\dagger$ is the exact data, and $x^\dagger$ and e are the exact solution and noise, respectively. We denote by $\delta = \|e\|$ the noise level.

Due to the ill-posed nature, regularization techniques are often applied to obtain a stable numerical approximation. A large number of regularization methods have been developed. The classical ones include Tikhonov regularization and its variant, truncated singular value decomposition, and iterative regularization techniques, and they are suitable for recovering smooth solutions. More recently, general variational type regularization methods have been proposed to preserve distinct features, e.g., discontinuity, edge and sparsity. This work focuses on recovering a smooth solution by Tikhonov regularization and truncated singular value decomposition, which are still routinely applied in practice. However, with the advent of the ever increasing data volume, their routine application remains challenging, especially in the context of massive data and multi-query, e.g., Bayesian inversion or tuning multiple hyperparameters. Hence, it is still of great interest to develop fast inversion algorithms.

In this work, we develop efficient linear inversion techniques based on RSVD. Over the last decade, a number of RSVD inversion algorithms have been developed and analyzed [10, 11, 20, 26, 31]. RSVD exploits the intrinsic low-rank structure of A for inverse problems to construct an accurate approximation efficiently. Our main contribution lies in providing a unified framework for developing fast regularized inversion techniques based on RSVD, for the following three popular regularization methods: truncated SVD, standard Tikhonov regularization, and Tikhonov regularization with a smooth penalty. The main novelty is that it explicitly preserves a certain range condition of the regularized solution, which is analogous to source condition in regularization theory [5, 13], and admits interpretation in the lens of convex duality. Further, we derive error bounds on the approximation with respect to the true solution $x^\dagger$ in Sect. 4, in the spirit of regularization theory for noisy operators. These results provide guidelines on the low-rank approximation, and differ from existing results [1, 14, 30, 32, 33], where the focus is on relative error estimates with respect to the regularized solution.

Now we situate the work in the literature on RSVD for inverse problems. RSVD has been applied to solving inverse problems efficiently [1, 30, 32, 33]. Xiang and Zou [32] developed RSVD for standard Tikhonov regularization and provided relative error estimates between the approximate and exact Tikhonov minimizer, by adapting the perturbation theory for least-squares problems. In the work [33], the authors proposed two approaches based respectively on transformation to standard form and randomized generalized SVD (RGSVD), and for the latter, RSVD is only performed on the matrix A. There was no error estimate in [33]. Wei et al [30] proposed different implementations, and derived some relative error estimates. Boutsidis and Magdon [1] analyzed the relative error for truncated RSVD, and discussed the sample complexity. Jia and Yang [14] presented a different way to perform truncated RSVD via LSQR for general smooth penalty, and provided relative error estimates. See also [16] for an evaluation within magnetic particle imaging. More generally, the idea of randomization has been fruitfully employed to reduce the computational cost associated with regularized inversion in statistics and machine learning, under the name of sketching in either primal or dual spaces [2, 22, 29, 34]. All these works also essentially exploit the low-rank structure, but in a different manner. Our analysis may also be extended to these approaches.

The rest of the paper is organized as follows. In Sect. 2, we recall preliminaries on RSVD, especially implementation and error bound. Then in Sect. 3, under one single guiding principle, we develop efficient inversion schemes based on RSVD for three classical regularization methods, and give the error analysis in Sect. 4. Finally we illustrate the approaches with some numerical results in Sect. 5. In the appendix, we describe an iterative refinement scheme for (general) Tikhonov regularization. Throughout, we denote by lower and capital letters for vectors and matrices, respectively, by I an identity matrix of an appropriate size, by $\|\cdot\|$ the Euclidean norm for vectors and spectral norm for matrices, and by $(\cdot,\cdot)$ for Euclidean inner product for vectors. The superscript * denotes the vector/matrix transpose. We use the notation $\mathscr{R}(A)$ and $\mathscr{N}(A)$ to denote the range and kernel of a matrix A, and A_k and $\tilde{A}_k$ denote the optimal and approximate rank-k approximations by SVD and RSVD, respectively. The notation c denotes a generic constant which may change at each occurrence, but is always independent of the condition number of A.

2 Preliminaries

Now we recall preliminaries on RSVD and technical lemmas.

2.1 SVD and Pseudoinverse

Singular value decomposition (SVD) is one of most powerful tools in numerical linear algebra. For any matrix $A \in \mathbb{R}^{n\times m}$, SVD of A is given by

$$A = U\Sigma V^*,$$

where $U = [u_1,\ u_2,\ \ldots,\ u_n] \in \mathbb{R}^{n\times n}$ and $V = [v_1, v_2, \ldots, v_m] \in \mathbb{R}^{m\times m}$ are column orthonormal matrices, with the vectors u_i and v_i being the left and right singular vectors, respectively, and V^* denotes the transpose of V. The diagonal matrix $\Sigma = \mathrm{diag}(\sigma_i) \in \mathbb{R}^{n\times m}$ has nonnegative diagonal entries σ_i, known as singular values (SVs), ordered nonincreasingly:

$$\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_r > \sigma_{r+1} = \cdots = \sigma_{\min(m,n)} = 0,$$

where $r = \mathrm{rank}(A)$ is the rank of A. Let $\sigma_i(A)$ be the ith SV of A. The complexity of the standard Golub-Reinsch algorithm for computing SVD is $4n^2m + 8m^2n + 9m^3$ (for $n \ge m$) [8, p. 254]. Thus, it is expensive for large-scale problems.

Now we can give the optimal low-rank approximation to A. By Eckhardt-Young theorem, the optimal rank-k approximation A_k of A (in spectral norm) is given by

$$\|A - U_k \Sigma_k V_k^*\| = \sigma_{k+1},$$

where $U_k \in \mathbb{R}^{n\times k}$ and $V_k \in \mathbb{R}^{m\times k}$ are the submatrix formed by taking the first k columns of the matrices U and V, and $\Sigma_k = \mathrm{diag}(\sigma_1, \ldots, \sigma_k) \in \mathbb{R}^{k\times k}$. The pseudoinverse $A^\dagger \in \mathbb{R}^{m\times n}$ of $A \in \mathbb{R}^{n\times m}$ is given by

$$A^\dagger = V_r \Sigma_r^{-1} U_r^*.$$

We have the following properties of the pseudoinverse of matrix product.

Lemma 1 *For any $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{n\times l}$, the identity $(AB)^\dagger = B^\dagger A^\dagger$ holds, if one of the following conditions is fulfilled:* (i) *A has orthonormal columns;* (ii) *B has orthonormal rows;* (iii) *A has full column rank and B has full row rank.*

The next result gives an estimate on matrix pseudoinverse.

Lemma 2 *For symmetric semipositive definite $A, B \in \mathbb{R}^{m\times m}$, there holds*

$$\|A^\dagger - B^\dagger\| \leq \|A^\dagger\|\|B^\dagger\|\|B - A\|.$$

Proof Since A is symmetric semipositive definite, we have $A^\dagger = \lim_{\mu\to 0^+}(A + \mu I)^{-1}$. By the identity $C^{-1} - D^{-1} = C^{-1}(D - C)D^{-1}$ for invertible $C, D \in \mathbb{R}^{m\times m}$,

$$\begin{aligned} A^\dagger - B^\dagger &= \lim_{\mu\to 0^+}[(A + \mu I)^{-1} - (B + \mu I)^{-1}] \\ &= \lim_{\mu\to 0^+}[(A + \mu I)^{-1}(B - A)(B + \mu I)^{-1}] = A^\dagger(B - A)B^\dagger. \end{aligned}$$

Now the estimate follows from the matrix spectral norm estimate. □

Remark 1 The estimate for general matrices is weaker than the one in Lemma 2: for general $A, B \in \mathbb{R}^{n\times m}$ with $\mathrm{rank}(A) = \mathrm{rank}(B) < \min(m, n)$, there holds [25]

$$\|A^\dagger - B^\dagger\| \leq \tfrac{1+\sqrt{5}}{2}\|A^\dagger\|\|B^\dagger\|\|B - A\|.$$

The rank condition is essential, and otherwise, the estimate may not hold.

Last, we recall the stability of SVs ([12, Corollary 7.3.8], [27, Sect. 1.3]).

Lemma 3 *For $A, B \in \mathbb{R}^{n\times m}$, there holds*

$$|\sigma_i(A + B) - \sigma_i(A)| \leq \|B\|, \quad i = 1, \ldots, \min(m, n).$$

2.2 Randomized SVD

Traditional numerical methods to compute a rank-k SVD, e.g., Lanczos bidiagonalization and Krylov subspace method, are especially powerful for large sparse or structured matrices. However, for many discrete inverse problems, there is no such structure. The prototypical model in inverse problems is a Fredholm integral equation of the first kind, which gives rise to unstructured dense matrices. Over the past decade, randomized algorithms for computing low-rank approximations have gained popularity. Frieze et al. [6] developed a Monte Carlo SVD to efficiently compute an approximate low-rank SVD based on non-uniform row and column sampling. Sarlos [23] proposed an approach based on random projection, using properties of random vectors to build a subspace capturing the matrix range. Below we describe briefly the basic idea of RSVD, and refer readers to [11] for an overview and to [10, 20, 26] for an incomplete list of recent works.

RSVD can be viewed as an iterative procedure based on SVDs of a sequence of low-rank matrices to deliver a nearly optimal low-rank SVD. Given a matrix $A \in \mathbb{R}^{n\times m}$ with $n \geq m$, we aim at obtaining a rank-k approximation, with $k \ll \min(m, n)$. Let $\Omega \in \mathbb{R}^{m\times(k+p)}$, with $k + p \leq m$, be a random matrix, with its entries following an i.i.d. Gaussian distribution $N(0, 1)$, and the integer $p \geq 0$ is an oversampling parameter (with a default value $p = 5$ [11]). Then we form a random matrix Y by

$$Y = (AA^*)^q A\Omega, \tag{2}$$

where the exponent $q \in \mathbb{N} \cup \{0\}$. By SVD of A, i.e., $A = U\Sigma V^*$, Y is given by

$$Y = U\Sigma^{2q+1}V^*\Omega.$$

Thus Ω is used for probing $\mathscr{R}(A)$, and $\mathscr{R}(Y)$ captures $\mathscr{R}(U_k)$ well. The accuracy is determined by the decay of σ_is, and the exponent q can greatly improve the performance when σ_is decay slowly. Let $Q \in \mathbb{R}^{n\times(k+p)}$ be an orthonormal basis for $\mathscr{R}(Y)$, which can be computed efficiently via QR factorization or skinny SVD. Next we form the (projected) matrix

$$B = Q^*A \in \mathbb{R}^{(k+p)\times m}.$$

Last, we compute SVD of B

$$B = WSV^*,$$

with $W \in \mathbb{R}^{(k+p)\times(k+p)}$, $S \in \mathbb{R}^{(k+p)\times(k+p)}$ and $V \in \mathbb{R}^{m\times(k+p)}$. This again can be carried out efficiently by standard SVD, since the size of B is much smaller. With $1 : k$ denoting the index set $\{1, \ldots, k\}$, let $\tilde{U}_k = QW(1 : n, 1 : k) \in \mathbb{R}^{n\times k}$, $\tilde{\Sigma}_k = S(1 : k, 1 : k) \in \mathbb{R}^{k\times k}$ and $\tilde{V}_k = V(1 : m, 1 : k) \in \mathbb{R}^{m\times k}$. The triple $(\tilde{U}_k, \tilde{\Sigma}_k, \tilde{V}_k)$ defines a rank-k approximation $\tilde{A}_k$:

$$\tilde{A}_k = \tilde{U}_k\tilde{\Sigma}_k\tilde{V}_k^*.$$

The triple $(\tilde{U}_k, \tilde{\Sigma}_k, \tilde{V}_k)$ is a nearly optimal rank-k approximation to A; see Theorem 1 below for a precise statement. The approximation is random due to range probing by Ω. By its very construction, we have

$$\tilde{A}_k = \tilde{P}_k A, \tag{3}$$

where $\tilde{P}_k = \tilde{U}_k \tilde{U}_k^* \in \mathbb{R}^{n\times n}$ is the orthogonal projection into $\mathscr{R}(\tilde{U}_k)$. The procedure for RSVD is given in Algorithm 1. The complexity of Algorithm 1 is about $4(q+1)nmk$, which can be much smaller than that of full SVD if $k \ll \min(m, n)$.

Algorithm 1 RSVD for $A \in \mathbb{R}^{n\times m}$, $n \geq m$.

```
Input matrix $A \in \mathbb{R}^{n\times m}$, $n \geq m$, and target rank $k$;
Set parameters $p$ (default $p = 5$), and $q$ (default $q = 0$);
Sample a random matrix $\Omega = (\omega_{ij}) \in \mathbb{R}^{m\times(k+p)}$, with $\omega_{ij} \sim N(0, 1)$;
Compute the randomized matrix $Y = (AA^*)^q A\Omega$;
Find an orthonormal basis $Q$ of range$(Y)$ by QR decomposition;
Form the matrix $B = Q^* A$;
Compute the SVD of $B = WSV^*$;
Return the rank $k$ approximation $(\tilde{U}_k, \tilde{\Sigma}_k, \tilde{V}_k)$, cf. (3).
```

Remark 2 The SV σ_i can be characterized by [8, Theorem 8.6.1, p. 441]:

$$\sigma_i = \max_{u\in\mathbb{R}^n, u\perp \mathrm{span}(\{u_j\}_{j=1}^{i-1})} \frac{\|A^* u\|}{\|u\|}.$$

Thus, one may estimate $\sigma_i(A)$ directly by $\tilde{\sigma}_i(A) = \|A^*\tilde{U}(:, i)\|$, and refine the SV estimate, similar to Rayleigh quotient acceleration for computing eigenvalues.

The following error estimates hold for RSVD $(\tilde{U}_k, \tilde{\Sigma}_k, \tilde{V}_k)$ given by Algorithm 1 with $q = 0$ [11, Corollary 10.9, p. 275], where the second estimate shows how the parameter p improves the accuracy. The exponent q is in the spirit of a power method, and can significantly improve the accuracy in the absence of spectral gap; see [11, Corollary 10.10, p. 277] for related discussions.

Theorem 1 *For $A \in \mathbb{R}^{n\times m}$, $n \geq m$, let $\Omega \in \mathbb{R}^{m\times(k+p)}$ be a standard Gaussian matrix, $k + p \leq m$ and $p \geq 4$, and Q an orthonormal basis for $\mathscr{R}(A\Omega)$. Then with probability at least $1 - 3p^{-p}$, there holds*

$$\|A - QQ^*A\| \leq (1 + 6((k+p)p\log p)^{\frac{1}{2}})\sigma_{k+1} + 3\sqrt{k+p}\left(\sum_{j>k}\sigma_j^2\right)^{\frac{1}{2}},$$

and further with probability at least $1 - 3e^{-p}$*, there holds*

$$\|A - QQ^*A\| \le \left(1 + 16\left(1 + \frac{k}{p+1}\right)^{\frac{1}{2}}\right)\sigma_{k+1} + \frac{8(k+p)^{\frac{1}{2}}}{p+1}\left(\sum_{j>k}\sigma_j^2\right)^{\frac{1}{2}}.$$

The next result is an immediate corollary of Theorem 1. Exponentially decaying SVs arise in, e.g., backward heat conduction and elliptic Cauchy problem.

Corollary 1 *Suppose that the SVs* σ_i *decay exponentially, i.e.,* $\sigma_j = c_0 c_1^j$*, for some* $c_0 > 0$ *and* $c_1 \in (0, 1)$*. Then with probability at least* $1 - 3p^{-p}$*, there holds*

$$\|A - QQ^*A\| \le \left[1 + 6((k+p)p\log p)^{\frac{1}{2}} + \frac{3(k+p)^{\frac{1}{2}}}{(1-c_1^2)^{\frac{1}{2}}}\right]\sigma_{k+1},$$

and further with probability at least $1 - 3e^{-p}$*, there holds*

$$\|A - QQ^*A\| \le \left[\left(1 + 16\left(1 + \frac{k}{p+1}\right)^{\frac{1}{2}}\right) + \frac{8(k+p)^{\frac{1}{2}}}{(p+1)(1-c_1^2)^{\frac{1}{2}}}\right]\sigma_{k+1}.$$

So far we have assumed that A is tall, i.e., $n \ge m$. For the case $n < m$, one may apply RSVD to A^*, which gives rise to Algorithm 2.

Algorithm 2 RSVD for $A \in \mathbb{R}^{n\times m}$, $n < m$.

```
Input matrix $A \in \mathbb{R}^{n\times m}$, $n < m$, and target rank $k$;
Set parameters $p$ (default $p = 5$), and $q$ (default $q = 0$);
Sample a random matrix $\Omega = (\omega_{ij}) \in \mathbb{R}^{(k+p)\times n}$, with $\omega_{ij} \sim N(0, 1)$;
Compute the randomized matrix $Y = \Omega A(A^*A)^q$;
Find an orthonormal basis $Q$ of range$(Y^*)$ by QR decomposition;
Find the matrix $B = AQ$;
Compute the SVD of $B = USV^*$;
Return the rank $k$ approximation $(\tilde{U}_k, \tilde{\Sigma}_k, \tilde{V}_k)$.
```

The efficiency of RSVD resides crucially on the truly low-rank nature of the problem. The precise spectral decay is generally unknown for many practical inverse problems, although there are known estimates for several model problems, e.g., X-ray transform [18] and magnetic particle imaging [17]. The decay rates generally worsen with the increase of the spatial dimension d, at least for integral operators [9], which can potentially hinder the application of RSVD type techniques to high-dimensional problems.

3 Efficient Regularized Linear Inversion with RSVD

Now we develop efficient inversion techniques based on RSVD for problem (1) via truncated SVD (TSVD), Tikhonov regularization and Tikhonov regularization with a smoothness penalty [5, 13]. For large-scale inverse problems, this can be expensive, since they either involve full SVD or large dense linear systems. We aim at reducing the cost by exploiting the inherent low-rank structure for inverse problems, and accurately constructing a low-rank approximation by RSVD. This idea has been pursued recently [1, 14, 30, 32, 33]. Our work is along the same line in of research but with a unified framework for deriving all three approaches and interpreting the approach in the lens of convex duality.

The key observation is the range type condition on the approximation $\tilde{x}$:

$$\tilde{x} \in \mathscr{R}(B), \tag{4}$$

with the matrix B is given by

$$B = \begin{cases} A^*, & \text{truncated SVD, Tikhonov,} \\ L^\dagger L^{*\dagger} A^*, & \text{general Tikhonov,} \end{cases}$$

where L is a regularizing matrix, typically chosen to the finite difference approximation of the first- or high-order derivatives [5]. Similar to (4), the approximation $\tilde{x}$ is assumed to live in $\mathrm{span}(\{v_i\}_{i=1}^k)$ in [34] for Tikhonov regularization, which is slightly more restrictive than (4). An analogous condition on the exact solution $x^\dagger$ reads

$$x^\dagger = Bw \tag{5}$$

for some $w \in \mathbb{R}^n$. In regularization theory [5, 13], (5) is known as source condition, and can be viewed as the Lagrange multiplier for the equality constraint $Ax^\dagger = b^\dagger$, whose existence is generally not ensured for infinite-dimensional problems. It is often employed to bound the error $\|\tilde{x} - x^\dagger\|$ of the approximation $\tilde{x}$ in terms of the noise level δ. The construction below explicitly maintains (4), thus preserving the structure of the regularized solution $\tilde{x}$. We will interpret the construction by convex analysis. Below we develop three efficient computational schemes based on RSVD.

3.1 Truncated RSVD

Classical truncated SVD (TSVD) stabilizes problem (1) by looking for the least-squares solution of

$$\min \|A_k x_k - b\|, \qquad \text{with } A_k = U_k \Sigma_k V_k^*.$$

Then the regularized solution x_k is given by

$$x_k = A_k^\dagger b = V_k \Sigma_k^{-1} U_k^* b = \sum_{i=1}^{k} \sigma_i^{-1} (u_i, b) v_i.$$

The truncated level $k \leq \mathrm{rank}(A)$ plays the role of a regularization parameter, and determines the strength of regularization. TSVD requires computing the (partial) SVD of A, which is expensive for large-scale problems. Thus, one can substitute a rank-k RSVD $(\tilde{U}_k, \tilde{\Sigma}_k, \tilde{V}_k)$, leading to truncated RSVD (TRSVD):

$$\hat{x}_k = \tilde{V}_k \tilde{\Sigma}_k^{-1} \tilde{U}_k^* b.$$

By Lemma 3, $\tilde{A}_k = \tilde{U}_k \tilde{\Sigma}_k \tilde{V}_k^*$ is indeed of rank k, if $\|A - \tilde{A}_k\| < \sigma_k$. This approach was adopted in [1]. Based on RSVD, we propose an approximation $\tilde{x}_k$ defined by

$$\tilde{x}_k = A^* (\tilde{A}_k \tilde{A}_k^*)^\dagger b = A^* \sum_{i=1}^{k} \frac{(\tilde{u}_i, b)}{\tilde{\sigma}_i^2} \tilde{u}_i. \tag{6}$$

By its construction, the range condition (4) holds for $\tilde{x}_k$. To compute $\tilde{x}_k$, one does not need the complete RSVD $(\tilde{U}_k, \tilde{\Sigma}_k, \tilde{V}_k)$ of rank k, but only $(\tilde{U}_k, \tilde{\Sigma}_k)$, which is advantageous for complexity reduction [8, p. 254]. Given the RSVD $(\tilde{U}_k, \tilde{\Sigma}_k)$, computing $\tilde{x}_k$ by (6) incurs only $O(nk + nm)$ operations.

3.2 Tikhonov Regularization

Tikhonov regularization stabilizes (1) by minimizing the following functional

$$J_\alpha(x) = \tfrac{1}{2}\|Ax - b\|^2 + \tfrac{\alpha}{2}\|x\|^2,$$

where $\alpha > 0$ is the regularization parameter. The regularized solution x_α is given by

$$x_\alpha = (A^*A + \alpha I)^{-1} A^* b = A^*(AA^* + \alpha I)^{-1} b. \tag{7}$$

The latter identity verifies (4). The cost of the step in (7) is about $nm^2 + \frac{m^3}{3}$ or $mn^2 + \frac{n^3}{3}$ [8, p. 238], and thus it is expensive for large scale problems. One approach to accelerate the computation is to apply the RSVD approximation $\tilde{A}_k = \tilde{U}_k \tilde{\Sigma}_k \tilde{V}_k^*$. Then one obtains a regularized approximation [32]

$$\hat{x}_\alpha = (\tilde{A}_k^* \tilde{A}_k + \alpha I)^{-1} \tilde{A}_k^* b. \tag{8}$$

To preserve the range property (4), we propose an alternative

$$\tilde{x}_\alpha = A^*(\tilde{A}_k\tilde{A}_k^* + \alpha I)^{-1}b = A^*\sum_{i=1}^{k}\frac{(\tilde{u}_i, b)}{\tilde{\sigma}_i^2 + \alpha}\tilde{u}_i. \tag{9}$$

For $\alpha \to 0^+$, $\tilde{x}_\alpha$ recovers the TRSVD $\tilde{x}_k$ in (6). Given RSVD $(\tilde{U}_k, \tilde{\Sigma}_k)$, the complexity of computing $\tilde{x}_\alpha$ is nearly identical with the TRSVD $\tilde{x}_k$.

3.3 General Tikhonov Regularization

Now we consider Tikhonov regularization with a general smoothness penalty:

$$J_\alpha(x) = \tfrac{1}{2}\|Ax - b\|^2 + \tfrac{\alpha}{2}\|Lx\|^2, \tag{10}$$

where $L \in \mathbb{R}^{\ell\times m}$ is a regularizing matrix enforcing smoothness. Typical choices of L include first-order and second-order derivatives. We assume $\mathcal{N}(A) \cap \mathcal{N}(L) = \{0\}$ so that J_α has a unique minimizer x_α. By the identity

$$(A^*A + \alpha I)^{-1}A^* = A^*(AA^* + \alpha I)^{-1}, \tag{11}$$

if $\mathcal{N}(L) = \{0\}$, the minimizer x_α to J_α is given by (with $\Gamma = L^\dagger L^{\dagger *}$)

$$\begin{aligned} x_\alpha &= (A^*A + \alpha L^*L)^{-1}(A^*y) \\ &= L^\dagger((AL^\dagger)^*AL^\dagger + \alpha I)^{-1}(AL^\dagger)^*b \\ &= \Gamma A^*(A\Gamma A^* + \alpha I)^{-1}b. \end{aligned} \tag{12}$$

The Γ factor reflects the smoothing property of $\|Lx\|^2$. Similar to (9), we approximate $B := AL^\dagger$ via RSVD: $\tilde{B}_k = U_k\Sigma_k V_k^*$, and obtain a regularized solution $\tilde{x}_\alpha$ by

$$\tilde{x}_\alpha = \Gamma A^*(\tilde{B}_k\tilde{B}_k^* + \alpha I)^{-1}b. \tag{13}$$

It differs from [33] in that [33] uses only the RSVD approximation of A, thus it does not maintain the range condition (19). The first step of Algorithm 1, i.e., $AL^{-1}\Omega$, is to probe $\mathcal{R}(A)$ with colored Gaussian noise with covariance Γ.

Numerically, it also involves applying Γ, which can be carried out efficiently if L is structured. If L is rectangular, we have the following decomposition [4, 32]. The A-weighted pseudoinverse $L^\#$ [4] can be computed efficiently, if $L^\dagger$ is easy to compute and the dimensionality of W is small.

Lemma 4 *Let W and Z be any matrices satisfying $\mathcal{R}(W) = \mathcal{N}(L)$, $\mathcal{R}(Z) = \mathcal{R}(L)$, $Z^*Z = I$, and $L^\# = (I - W(AW)^\dagger A)L^\dagger$. Then the solution x_α to (10) is given by*

$$x_\alpha = L^\# Z\xi_\alpha + W(AW)^\dagger b, \tag{14}$$

where the variable ξ_α *minimizes* $\frac{1}{2}\|AL^{\#}Z\xi - b\|^2 + \frac{\alpha}{2}\|\xi\|^2$.

Lemma 4 does not necessarily entail an efficient scheme, since it requires an orthonormal basis Z for $\mathscr{R}(L)$. Hence, we restrict our discussion to the case:

$$L \in \mathbb{R}^{\ell \times m} \quad \text{with } \operatorname{rank}(L) = \ell < m. \tag{15}$$

It arises most commonly in practice, e.g., first-order or second-order derivative, and there are efficient ways to perform standard-form reduction. Then we can let $Z = I_\ell$. By slightly abusing the notation $\Gamma = L^{\#}L^{\#*}$, by Lemma 4, we have

$$\begin{aligned} x_\alpha &= L^{\#}((AL^{\#})^*AL^{\#} + \alpha I)^{-1}(AL^{\#})^*b + W(AW)^{\dagger}b \\ &= \Gamma A^*(A\Gamma A^* + \alpha I)^{-1}b + W(AW)^{\dagger}b. \end{aligned}$$

The first term is nearly identical with (12), with $L^{\#}$ in place of $L^{\dagger}$, and the extra term $W(AW)^{\dagger}b$ belongs to $\mathscr{N}(L)$. Thus, we obtain an approximation $\tilde{x}_\alpha$ defined by

$$\tilde{x}_\alpha = \Gamma A^*(\tilde{B}_k\tilde{B}_k + \alpha I)^{-1}b + W(AW)^{\dagger}b, \tag{16}$$

where $\tilde{B}_k$ is a rank-k RSVD to $B \equiv AL^{\#}$. The matrix B can be implemented implicitly via matrix-vector product to maintain the efficiency.

3.4 Dual Interpretation

Now we give an interpretation of (13) in the lens of Fenchel duality theory in Banach spaces (see, e.g., [3, Chap. 2.4]). Recall that for a functional $F : X \to \overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$ defined on a Banach space X, let $F^* : X^* \to \overline{\mathbb{R}}$ denote the Fenchel conjugate of F given for $x^* \in X^*$ by

$$F^*(x^*) = \sup_{x \in X} \langle x^*, x\rangle_{X^*,X} - F(x).$$

Further, let $\partial F(x) := \{x^* \in X^* : \langle x^*, \tilde{x} - x\rangle_{X^*,X} \leq F(\tilde{x}) - F(x) \;\; \forall \tilde{x} \in X\}$ be the subdifferential of the convex functional F at x, which coincides with Gâteaux derivative $F'(x)$ if it exists. The Fenchel duality theorem states that if $F : X \to \overline{\mathbb{R}}$ and $G : Y \to \overline{\mathbb{R}}$ are proper, convex and lower semicontinuous functionals on the Banach spaces X and Y, $\Lambda : X \to Y$ is a continuous linear operator, and there exists an $x_0 \in W$ such that $F(x_0) < \infty$, $G(\Lambda x_0) < \infty$, and G is continuous at Λx_0, then

$$\inf_{x \in X} F(x) + G(\Lambda x) = \sup_{y^* \in Y^*} -F^*(\Lambda^* y^*) - G^*(-y^*),$$

Further, the equality is attained at $(\bar{x}, \bar{y}^*) \in X \times Y^*$ if and only if

$$\Lambda^*\bar{y}^* \in \partial F(\bar{x}) \quad \text{and} \quad -\bar{y}^* \in \partial G(\Lambda\bar{x}), \tag{17}$$

hold [3, Remark 3.4.2].

The next result indicates that the approach in Sects. 3.2–3.3 first applies RSVD to the dual problem to obtain an approximate dual $\tilde{p}_\alpha$, and then recovers the optimal primal $\tilde{x}_\alpha$ via duality relation (17). This connection is in the same spirit of dual random projection [29, 34], and it opens up the avenue to extend RSVD to functionals whose conjugate is simple, e.g., nonsmooth fidelity.

Proposition 1 *If $\mathscr{N}(L) = \{0\}$, then $\tilde{x}_\alpha$ in (13) is equivalent to RSVD for the dual problem.*

Proof For any symmetric positive semidefinite Q, the conjugate functional F^* of $F(x) = \frac{\alpha}{2}x^*Qx$ is given by $F^*(\xi) = -\frac{1}{2\alpha}\xi^*Q^\dagger\xi$, with its domain being $\mathscr{R}(Q)$. By SVD, we have $(L^*L)^\dagger = L^\dagger L^{\dagger *}$, and thus $F^*(\xi) = -\frac{1}{2\alpha}\|L^{\dagger *}\xi\|^2$. Hence, by Fenchel duality theorem, the conjugate $J_\alpha^*(\xi)$ of $J_\alpha(x)$ is given by

$$J_\alpha^*(\xi) := -\tfrac{1}{2\alpha}\|L^{\dagger *}A^*\xi\|^2 - \tfrac{1}{2}\|\xi - b\|^2.$$

Further, by (17), the optimal primal and dual pair (x_α, ξ_α) satisfies

$$\alpha L^*Lx_\alpha = A^*\xi_\alpha \quad \text{and} \quad \xi_\alpha = b - Ax_\alpha.$$

Since $\mathscr{N}(L) = \{0\}$, L^*L is invertible, and thus $x_\alpha = \alpha^{-1}(L^*L)^{-1}A^*\xi_\alpha = \alpha^{-1}\Gamma A^*\xi_\alpha$. The optimal dual ξ_α is given by $\xi_\alpha = \alpha(AL^\dagger L^{*\dagger}A^* + \alpha I)^{-1}b$. To approximate ξ_α by $\tilde{\xi}_\alpha$, we employ the RSVD approximation $\tilde{B}_k$ to $B = AL^\dagger$ and solve

$$\tilde{\xi}_\alpha = \arg\max_{\xi\in\mathbb{R}^n}\{-\tfrac{1}{2\alpha}\|\tilde{B}_k^*\xi\|^2 - \tfrac{1}{2}\|\xi - b\|^2\}.$$

We obtain an approximation via the relation $\tilde{x}_\alpha = \alpha^{-1}\Gamma A^*\tilde{\xi}_\alpha$, recovering (13). □

Remark 3 For a general regularizing matrix L, one can appeal to the decomposition in Lemma 4, by applying first the standard transformation and then approximating the regularized part via convex duality.

4 Error Analysis

Now we derive error estimates for the approximation $\tilde{x}$ with respect to the true solution $x^\dagger$, under sourcewise type conditions. In addition to bounding the error, the estimates provide useful guidelines on constructing the approximation $\tilde{A}_k$.

4.1 Truncated RSVD

We derive an error estimate under the source condition (5). We use the projection matrices $P_k = U_k U_k^*$ and $\tilde{P}_k = \tilde{U}_k \tilde{U}_k^*$ frequently below.

Lemma 5 *For any $k \le r$ and $\|A - \tilde{A}_k\| \le \sigma_k/2$, there holds $\|A^*(\tilde{A}_k^*)^\dagger\| \le 2$.*

Proof It follows from the decomposition $A = \tilde{P}_k A + (I - \tilde{P}_k)A = \tilde{A}_k + (I - \tilde{P}_k)A$ that

$$\begin{aligned}\|A^*(\tilde{A}_k^*)^\dagger\| = \|(\tilde{A}_k + (I - \tilde{P}_k)A)^*(\tilde{A}_k^*)^\dagger\| &\le \|\tilde{A}_k^*(\tilde{A}_k^*)^{-1}\| + \|A - \tilde{A}_k\|\|\tilde{A}_k^{-1}\| \\ &\le 1 + \tilde{\sigma}_k^{-1}\|A - \tilde{A}_k\|.\end{aligned}$$

Now the condition $\|A - \tilde{A}_k\| \le \sigma_k/2$ and Lemma 3 imply $\tilde{\sigma}_k \ge \sigma_k - \|A - \tilde{A}_k\| \ge \sigma_k/2$, from which the desired estimate follows. □

Now we can state an error estimate for the approximation $\tilde{x}_k$.

Theorem 2 *If Condition (5) holds and $\|A - \tilde{A}_k\| \le \sigma_k/2$, then for the estimate $\tilde{x}_k$ in (6), there holds*

$$\|x^\dagger - \tilde{x}_k\| \le 4\delta\sigma_k^{-1} + 8\sigma_1\sigma_k^{-1}\|A_k - \tilde{A}_k\|\|w\| + \sigma_{k+1}\|w\|.$$

Proof By the decomposition $b = b^\dagger + e$, we have (with $P_k^\perp = I - P_k$)

$$\begin{aligned}\tilde{x}_k - x^\dagger &= A^*(\tilde{A}_k\tilde{A}_k^*)^\dagger b - A^*(AA^*)^\dagger b^\dagger \\ &= A^*(\tilde{A}_k\tilde{A}_k^*)^\dagger e + A^*[(\tilde{A}_k\tilde{A}_k^*)^\dagger - (A_kA_k^*)^\dagger]b^\dagger - P_k^\perp A^*(AA^*)^\dagger b^\dagger.\end{aligned}$$

The source condition $x^\dagger = A^*w$ in (5) implies

$$\tilde{x}_k - x^\dagger = A^*(\tilde{A}_k\tilde{A}_k^*)^\dagger e + A^*[(\tilde{A}_k\tilde{A}_k^*)^\dagger - (A_kA_k^*)^\dagger]AA^*w - P_k^\perp A^*(AA^*)^\dagger AA^*w.$$

By the triangle inequality, we have

$$\begin{aligned}\|\tilde{x}_k - x^\dagger\| \le \|A^*(\tilde{A}_k\tilde{A}_k^*)^\dagger e\| &+ \|A^*[(\tilde{A}_k\tilde{A}_k^*)^\dagger - (A_kA_k^*)^\dagger]AA^*w\| \\ &+ \|P_k^\perp A^*(AA^*)^\dagger AA^*w\| := \mathrm{I}_1 + \mathrm{I}_2 + \mathrm{I}_3.\end{aligned}$$

It suffices to bound the three terms separately. First, for the term I_1, by the identity $(\tilde{A}_k\tilde{A}_k^*)^\dagger = (\tilde{A}_k^*)^\dagger\tilde{A}_k^\dagger$ and Lemma 5, we have

$$\mathrm{I}_1 \le \|A^*(\tilde{A}_k^*)^\dagger\|\|\tilde{A}_k^\dagger\|\|e\| \le 2\tilde{\sigma}_k^{-1}\delta.$$

Second, for I_2, by Lemmas 5 and 2 and the identity $(\tilde{A}_k\tilde{A}_k^*)^\dagger = (\tilde{A}_k^*)^\dagger\tilde{A}_k^\dagger$, we have

$$\begin{aligned}\mathrm{I}_2 &\le \|A^*[(\tilde{A}_k\tilde{A}_k^*)^\dagger - (A_kA_k^*)^\dagger]AA^*\|\|w\|\\ &\le \|A^*(\tilde{A}_k\tilde{A}_k^*)^\dagger(\tilde{A}_k\tilde{A}_k^* - A_kA_k^*)(A_kA_k^*)^\dagger AA^*\|\|w\|\\ &\le \|A^*(\tilde{A}_k^*)^\dagger\|\|\tilde{A}_k^\dagger\|\|\tilde{A}_k\tilde{A}_k^* - A_kA_k^*\|\|(A_kA_k^*)^\dagger AA^*\|\|w\|\\ &\le 4\tilde{\sigma}_k^{-1}\|A\|\|A_k - \tilde{A}_k\|\|w\|,\end{aligned}$$

since $\|\tilde{A}_k\tilde{A}_k^* - A_kA_k^*\| \le \|\tilde{A}_k - A_k\|(\|\tilde{A}_k\| + \|A_k^*\|) \le 2\|A\|\|\tilde{A}_k - A_k\|$ and $\|(A_k A_k^*)^\dagger AA^*\| \le 1$. By Lemma 3, we can bound the term $\|(\tilde{A}_k^*)^\dagger\|$ by

$$\|(\tilde{A}_k^*)^\dagger\| = \tilde{\sigma}_k^{-1} \le (\sigma_k - \|A - \tilde{A}_k\|)^{-1} \le 2\sigma_k^{-1}.$$

Last, we can bound the third term I_3 directly by $\mathrm{I}_3 \le \|P_k^\perp A^*\|\|w\| \le \sigma_{k+1}\|w\|$. Combining these estimates yields the desired assertion. □

Remark 4 The bound in Theorem 2 contains three terms: propagation error $\sigma_k^{-1}\delta$, approximation error $\sigma_{k+1}\|w\|$, and perturbation error $\sigma_k^{-1}\|A\|\|A - \tilde{A}_k\|\|w\|$. It is of the worst-case scenario type and can be pessimistic. In particular, the error $\|A^*(\tilde{A}_k\tilde{A}_k^*)^{-1}e\|$ can be bounded more precisely by

$$\|A^*(\tilde{A}_k\tilde{A}_k^*)^\dagger e\| \le \|A^*(\tilde{A}_k^*)^\dagger\|\|\tilde{A}_k^\dagger e\|,$$

and $\|\tilde{A}_k^\dagger e\|$ can be much smaller than $\tilde{\sigma}_k^{-1}\|e\|$, if e concentrates in the high-frequency modes. By balancing the terms, it suffices for $\tilde{A}_k$ to have an accuracy $O(\delta)$. This is consistent with the analysis for regularized solutions with perturbed operators.

Remark 5 The condition $\|A - \tilde{A}_k\| < \sigma_k/2$ in Theorem 2 requires a sufficiently accurate low-rank RSVD approximation $(\tilde{U}_k, \tilde{\Sigma}_k, \tilde{V}_k)$ to A, i.e., the rank k is sufficiently large. It enables one to define a TRSVD solution $\tilde{x}_k$ of truncation level k.

Next we give a relative error estimate for $\tilde{x}_k$ with respect to the TSVD approximation x_k. Such an estimate was the focus of a few works [1, 30, 32, 33]. First, we give a bound on $\|\tilde{A}_k\tilde{A}_k^*(A_k^*)^\dagger - A_k\|$.

Lemma 6 *The following error estimate holds*

$$\|\tilde{A}_k\tilde{A}_k^*(A_k^*)^\dagger - A_k\| \le \left(1 + \sigma_1\sigma_k^{-1}\right)\|A_k - \tilde{A}_k\|.$$

Proof This estimate follows by direct computation:

$$\begin{aligned}\|\tilde{A}_k\tilde{A}_k^*(A_k^*)^\dagger - A_k\| &= \|[\tilde{A}_k\tilde{A}_k^* - A_kA_k^*](A_k^*)^\dagger\|\\ &\le \|\tilde{A}_k(\tilde{A}_k^* - A_k^*)(A_k^*)^\dagger\| + \|(\tilde{A}_k - A_k)A_k^*(A_k^*)^\dagger\|\\ &\le \|\tilde{A}_k\|\|\tilde{A}_k^* - A_k^*\|\|(A_k^*)^\dagger\| + \|\tilde{A}_k - A_k\|\|A_k^*(A_k^*)^\dagger\|\\ &\le (\sigma_1\sigma_k^{-1} + 1)\|\tilde{A}_k - A_k\|,\end{aligned}$$

since $\|\tilde{A}_k\| = \|\tilde{P}_k A\| \leq \|A\| = \sigma_1$. Then the desired assertion follows directly. □

Next we derive a relative error estimate between the approximations x_k and $\tilde{x}_k$.

Theorem 3 *For any $k < r$, and $\|A - \tilde{A}_k\| < \sigma_k/2$, there holds*

$$\frac{\|x_k - \tilde{x}_k\|}{\|x_k\|} \leq 4\Big(1 + \frac{\sigma_1}{\sigma_k}\Big)\frac{\|A_k - \tilde{A}_k\|}{\sigma_k}.$$

Proof We rewrite the TSVD solution x_k as

$$x_k = A^*(A_k A_k^*)^\dagger b = A_k^*(A_k A_k^*)^\dagger b. \tag{18}$$

By Lemma 3 and the assumption $\|A - \tilde{A}_k\| < \sigma_k/2$, we have $\tilde{\sigma}_k > 0$. Then $x_k - \tilde{x}_k = A^*((A_k A_k^*)^\dagger - (\tilde{A}_k \tilde{A}_k^*)^\dagger)b$. By Lemma 2,

$$(A_k A_k^*)^\dagger - (\tilde{A}_k \tilde{A}_k^*)^\dagger = (\tilde{A}_k \tilde{A}_k^*)^\dagger(\tilde{A}_k \tilde{A}_k^* - A_k A_k^*)(A_k A_k^*)^\dagger.$$

It follows from the identity $(A_k A_k^*)^\dagger = (A_k^*)^\dagger A_k^\dagger$ and (18) that

$$\begin{aligned} x_k - \tilde{x}_k &= A^*(\tilde{A}_k \tilde{A}_k^*)^\dagger(\tilde{A}_k \tilde{A}_k^* - A_k A_k^*)(A_k A_k^*)^\dagger b \\ &= A^*(\tilde{A}_k \tilde{A}_k^*)^\dagger(\tilde{A}_k \tilde{A}_k^*(A_k^*)^\dagger - A_k)A_k^*(A_k A_k^*)^\dagger b \\ &= A^*(\tilde{A}_k^*)^\dagger \tilde{A}_k^\dagger(\tilde{A}_k \tilde{A}_k^*(A_k^*)^\dagger - A_k)x_k. \end{aligned}$$

Thus, we obtain

$$\frac{\|x_k - \tilde{x}_k\|}{\|x_k\|} \leq \|A^*(\tilde{A}_k^*)^\dagger\|\|\tilde{A}_k^\dagger\|\|\tilde{A}_k \tilde{A}_k^*(A_k^*)^\dagger - A_k\|.$$

By Lemma 3, we bound the term $\|\tilde{A}_k^\dagger\|$ by $\|\tilde{A}_k^\dagger\| \leq 2\sigma_k^{-1}$. Combining the preceding estimates with Lemmas 5 and 6 completes the proof. □

Remark 6 The relative error is determined by k (and in turn by δ etc). Due to the presence of the factor σ_k^{-2}, the estimate requires a highly accurate low-rank approximation, i.e., $\|A_k - \tilde{A}_k\| \ll \sigma_k(A)^{-2}$, and hence it is more pessimistic than Theorem 2. The estimate is comparable with the perturbation estimate for the TSVD

$$\frac{\|x_k - \bar{x}_k\|}{\|x_k\|} \leq \frac{\sigma_1\|A_k - \tilde{A}_k\|}{\sigma_k - \|A_k - \tilde{A}_k\|}\left(\frac{1}{\sigma_1} + \frac{\|Ax_k - b\|}{\sigma_k\|b\|}\right) + \frac{\|A_k - \tilde{A}_k\|}{\sigma_k}.$$

Modulo the α factor, the estimates in [30, 32] for Tikhonov regularization also depend on σ_k^{-2} (but can be much milder for a large α).

4.2 Tikhonov Regularization

The following bounds are useful for deriving error estimate on $\tilde{x}_\alpha$ in (9).

Lemma 7 *The following estimates hold*

$$\|(AA^* + \alpha I)(\tilde{A}_k\tilde{A}_k^* + \alpha I)^{-1} - I\| \leq 2\alpha^{-1}\|A\|\|A - \tilde{A}_k\|,$$
$$\|[(AA^* + \alpha I)(\tilde{A}_k\tilde{A}_k^* + \alpha I)^{-1} - I]AA^*\| \leq 2\|A\|(2\alpha^{-1}\|A\|\|A - \tilde{A}_k\| + 1)\|A - \tilde{A}_k\|.$$

Proof It follows from the identity

$$(AA^* + \alpha I)(\tilde{A}_k\tilde{A}_k^* + \alpha I)^{-1} - I = (AA^* - \tilde{A}_k\tilde{A}_k^*)(\tilde{A}_k\tilde{A}_k^* + \alpha I)^{-1}$$

and the inequality $\|\tilde{A}_k\| = \|\tilde{P}_k A\| \leq \|A\|$ that

$$\begin{aligned}\|(AA^* + \alpha I)(\tilde{A}_k\tilde{A}_k^* + \alpha I)^{-1} - I\| &\leq \alpha^{-1}(\|A\| + \|\tilde{A}_k\|)\|A - \tilde{A}_k\| \\ &\leq 2\alpha^{-1}\|A\|\|A - \tilde{A}_k\|.\end{aligned}$$

Next, by the triangle inequality,

$$\begin{aligned}&\|[(AA^* + \alpha I)(\tilde{A}_k\tilde{A}_k^* + \alpha I)^{-1} - I]AA^*\| \\ &\quad\leq \|AA^* - \tilde{A}_k\tilde{A}_k^*\|(\|(\tilde{A}_k\tilde{A}_k^* + \alpha I)^{-1}(AA^* + \alpha I)\| + \alpha\|(\tilde{A}_k\tilde{A}_k^* + \alpha I)^{-1}\|).\end{aligned}$$

This, together with the identity $AA^* - \tilde{A}_k\tilde{A}_k^* = A(A^* - \tilde{A}_k^*) + (A - \tilde{A}_k)\tilde{A}_k^*$ and the first estimate, yields the second estimate, completing the proof of the lemma. □

Now we can give an error estimate on $\tilde{x}_\alpha$ in (9) under condition (5).

Theorem 4 *If condition* (5) *holds, then the estimate $\tilde{x}_\alpha$ satisfies*

$$\|\tilde{x}_\alpha - x^\dagger\| \leq \alpha^{-\frac{3}{2}}\|A\|\|A - \tilde{A}_k\|\left(\delta + (2\alpha^{-1}\|A\|\|A - \tilde{A}_k\| + 1)\alpha\|w\|\right) + 2^{-1}\alpha^{\frac{1}{2}}\|w\|.$$

Proof First, with condition (5), $x^\dagger$ can be rewritten as

$$\begin{aligned}x^\dagger &= (A^*A + \alpha I)^{-1}(A^*A + \alpha I)x^\dagger = (A^*A + \alpha I)^{-1}(A^*b^\dagger + \alpha x^\dagger) \\ &= (A^*A + \alpha I)^{-1}A^*(b^\dagger + \alpha w).\end{aligned}$$

The identity (11) implies $x^\dagger = A^*(AA^* + \alpha I)^{-1}(b^\dagger + \alpha w)$. Consequently,

$$\begin{aligned}\tilde{x}_\alpha - x^\dagger &= A^*[(\tilde{A}_k\tilde{A}_k^* + \alpha I)^{-1}b - (AA^* + \alpha I)^{-1}(b^\dagger + \alpha w)] \\ &= A^*[(\tilde{A}_k\tilde{A}_k^* + \alpha I)^{-1}e + ((\tilde{A}_k\tilde{A}_k^* + \alpha I)^{-1} \\ &\quad - (AA^* + \alpha I)^{-1})b^\dagger - \alpha(AA^* + \alpha I)^{-1}w].\end{aligned}$$

Let $\tilde{I} = (AA^* + \alpha I)(\tilde{A}_k\tilde{A}_k^* + \alpha I)^{-1}$. Then by the identity (11), there holds

$$(A^*A + \alpha I)(\tilde{x}_\alpha - x^\dagger) = A^*[\tilde{I}e + (\tilde{I} - I)b^\dagger - \alpha w],$$

and taking inner product with $\tilde{x}_\alpha - x^\dagger$ yields

$$\|A(\tilde{x}_\alpha - x^\dagger)\|^2 + \alpha\|\tilde{x}_\alpha - x^\dagger\|^2 \leq \Big(\|\tilde{I}e\| + \|(\tilde{I} - I)b^\dagger\| + \alpha\|w\|\Big)\|A(\tilde{x}_\alpha - x^\dagger)\|.$$

By Young's inequality $ab \leq \frac{1}{4}a^2 + b^2$ for any $a, b \in \mathbb{R}$, we deduce

$$\alpha^{\frac{1}{2}}\|\tilde{x}_\alpha - x^\dagger\| \leq 2^{-1}(\|\tilde{I}e\| + \|(\tilde{I} - I)b^\dagger\| + \alpha\|w\|).$$

By Lemma 7 and the identity $b^\dagger = AA^*w$, we have

$$\begin{aligned}\|\tilde{I}e\| &\leq 2\alpha^{-1}\|A\|\|A - \tilde{A}_k\|\delta,\\ \|(\tilde{I} - I)b^\dagger\| &= \|(\tilde{I} - I)AA^*w\|\\ &\leq 2\|A\|(2\alpha^{-1}\|A\|\|A - \tilde{A}_k\| + 1)\|A - \tilde{A}_k\|\|w\|.\end{aligned}$$

Combining the preceding estimates yield the desired assertion. □

Remark 7 To maintain the error $\|\tilde{x}_\alpha - x^\dagger\|$, the accuracy of $\tilde{A}_k$ should be of $O(\delta)$, and α should be of $O(\delta)$, which gives an overall accuracy $O(\delta^{1/2})$. The tolerance on $\|A - \tilde{A}_k\|$ can be relaxed for high noise levels. It is consistent with existing theory for Tikhonov regularization with noisy operators [19, 21, 28].

Remark 8 The following relative error estimate was shown [32, Theorem 1]:

$$\frac{\|x_\alpha - \hat{x}_\alpha\|}{\|x_\alpha\|} \leq c(2\sec\theta\kappa + \tan\theta\kappa^2)\sigma_{k+1} + O(\sigma_{k+1}^2),$$

with $\theta = \sin^{-1}\frac{(\|b - Ax_\alpha\|^2 + \alpha\|x_\alpha\|^2)^{\frac{1}{2}}}{\|b\|}$ and $\kappa = (\sigma_1^2 + \alpha)(\frac{\alpha^{\frac{1}{2}}}{\sigma_n^2 + \alpha} + \max_{1\leq i\leq n}\frac{\sigma_i}{\sigma_i^2 + \alpha})$. κ is a variant of condition number. Thus, $\tilde{A}_k$ should approximate accurately A in order not to spoil the accuracy, and the estimate can be pessimistic for small α for which the estimate tends to blow up.

4.3 General Tikhonov Regularization

Last, we give an error estimate for $\tilde{x}_\alpha$ defined in (13) under the following condition

$$x^\dagger = \Gamma A^*w, \tag{19}$$

where $\mathscr{N}(L) = \{0\}$, and $\Gamma = L^\dagger L^{*\dagger}$. Also recall that $B = A\Gamma^\dagger$.

Theorem 5 *If Condition* (19) *holds, then the regularized solution $\tilde{x}_\alpha$ in* (13) *satisfies*

$$\|L(x^\dagger - \tilde{x}_\alpha)\| \leq \alpha^{-\frac{3}{2}}\|B\|\|B-\tilde{B}_k\|\left(\delta + (2\alpha^{-1}\|B\|\|B-\tilde{B}_k\|+1)\alpha\|w\|\right) + 2^{-1}\alpha^{\frac{1}{2}}\|w\|.$$

Proof First, by the source condition (19), we rewrite $x^\dagger$ as

$$\begin{aligned} x^\dagger &= (A^*A+\alpha L^*L)^{-1}(A^*A+\alpha L^*L)x^\dagger \\ &= (A^*A+\alpha L^*L)^{-1}(A^*b^\dagger + \alpha A^*w). \end{aligned}$$

Now with the identity $(A^*A+\alpha L^*L)^{-1}A^* = \Gamma A^*(A\Gamma A^* + \alpha I)^{-1}$, we have

$$x^\dagger = \Gamma A^*(A\Gamma A^* + \alpha I)^{-1}(b^\dagger + \alpha w).$$

Thus, upon recalling $B = AL^\dagger$, we have

$$\begin{aligned} \tilde{x}_\alpha - x^\dagger &= \Gamma A^*[(\tilde{B}_k\tilde{B}_k^* + \alpha I)^{-1}b - (BB^*+\alpha I)^{-1}(b^\dagger + \alpha w)] \\ &= \Gamma A^*[(\tilde{B}_k\tilde{B}_k^* + \alpha I)^{-1}e + ((\tilde{B}_k\tilde{B}_k^* + \alpha I)^{-1} \\ &\quad - (BB^*+\alpha I)^{-1})b^\dagger - \alpha(BB^*+\alpha I)^{-1}w]. \end{aligned}$$

It follows from the identity

$$(A^*A+\alpha L^*L)\Gamma A^* = (A^*A+\alpha L^*L)L^\dagger L^{\dagger *}A^* = A^*(BB^*+\alpha I),$$

that

$$(A^*A+\alpha L^*L)(\tilde{x}_\alpha - x^\dagger) = A^*[\tilde{I}e + (\tilde{I}-I)b^\dagger - \alpha w],$$

with $\tilde{I} = (BB^*+\alpha I)(\tilde{B}_k\tilde{B}_k^*+\alpha I)^{-1}$. Taking inner product with $x_\alpha - x^\dagger$ and applying Cauchy-Schwarz inequality yield

$$\|A(\tilde{x}_\alpha - x^\dagger)\|^2 + \alpha\|L(\tilde{x}_\alpha - x^\dagger)\|^2 \leq (\|\tilde{I}e\| + \|(\tilde{I}-I)b^\dagger\| + \|\alpha w\|)\|A(\tilde{x}_\alpha - x^\dagger)\|,$$

Young's inequality implies $\alpha^{\frac{1}{2}}\|L(\tilde{x}_\alpha - x^\dagger)\| \leq 2^{-1}(\|\tilde{I}e\| + \|(\tilde{I}-I)b^\dagger\| + \alpha\|w\|)$. The identity $b^\dagger = Ax^\dagger = AL^\dagger L^{\dagger *}A^*w = BB^*w$ from (19) and Lemma 7 complete the proof. □

5 Numerical Experiments and Discussions

Now we present numerical experiments to illustrate our approach. The noisy data b is generated from the exact data $b^\dagger$ as follows

$$b_i = b_i^\dagger + \delta \max_j(|b_j^\dagger|)\xi_i, \quad i = 1, \dots, n,$$

where δ is the relative noise level, and the random variables ξ_is follow the standard Gaussian distribution. All the computations were carried out on a personal laptop with 2.50 GHz CPU and 8.00G RAM by MATLAB 2015b. When implementing Algorithm 1, the default choices $p = 5$ and $q = 0$ are adopted. Since the TSVD and Tikhonov solutions are close for suitably chosen regularization parameters, we present only results for Tikhonov regularization (and the general case with L given by the first-order difference, which has a one-dimensional kernel $\mathcal{N}(L)$).

Throughout, the regularization parameter α is determined by uniformly sampling an interval on a logarithmic scale, and then taking the value attaining the smallest reconstruction error, where approximate Tikhonov minimizers are found by either (9) or (16) with a large k ($k = 100$ in all the experiments).

5.1 One-Dimensional Benchmark Inverse Problems

First, we illustrate the efficiency and accuracy of proposed approach, and compare it with existing approaches [32, 33]. We consider seven examples (i.e., `deriv2`, `heat`, `phillips`, `baart`, `foxgood`, `gravity` and `shaw`), taken from the popular public-domain MATLAB package **regutools** (available from http://www.imm.dtu.dk/~pcha/Regutools/, last accessed on January 8, 2019), which have been used in existing studies (see, e.g., [30, 32, 33]). They are Fredholm integral equations of the first kind, with the first three examples being mildly ill-posed (i.e., σ_is decay algebraically) and the rest severely ill-posed (i.e., σ_is decay exponentially). Unless otherwise stated, the examples are discretized with a dimension $n = m = 5000$. The resulting matrices are dense and unstructured. The rank k of $\tilde{A}_k$ is fixed at $k = 20$, which is sufficient to for all examples.

The numerical results by standard Tikhonov regularization and two randomized variants, i.e., (8) and (9), for the examples are presented in Table 1. The accuracy of the approximations, i.e., the Tikhonov solution x_α, and two randomized approximations $\hat{x}_\alpha$ (cf. (8), proposed in [32]) and $\tilde{x}_\alpha$ (cf. (9), the proposed in this work), is measured in two different ways:

$$\begin{aligned} \tilde{e}_{xz} &= \|\hat{x}_\alpha - x_\alpha\|, \quad \tilde{e}_{ij} = \|\tilde{x}_\alpha - x_\alpha\|, \\ e &= \|x_\alpha - x^\dagger\|, \quad e_{xz} = \|\hat{x}_\alpha - x^\dagger\|, \quad e_{ij} = \|\tilde{x}_\alpha - x^\dagger\|, \end{aligned}$$

where the methods are indicated by the subscripts. That is, $\tilde{e}_{xz}$ and $\tilde{e}_{ij}$ measure the accuracy with respect to the Tikhonov solution x_α, and e, e_{xz} and e_{ij} measure the accuracy with respect to the exact one $x^\dagger$.

The following observations can be drawn from Table 1. For all examples, the three approximations x_α, $\tilde{x}_\alpha$ and $\hat{x}_\alpha$ have comparable accuracy relative to the exact solution $x^\dagger$, and the errors e_{ij} and e_{xz} are fairly close to the error e of the Tikhonov solution

Table 1 Numerical results by standard Tikhonov regularization at two noise levels

example	δ	$\tilde{e}_{xz}$	$\tilde{e}_{ij}$	e	e_{xz}	e_{ij}
`baart`	1%	1.14e−9	1.14e−9	1.68e−1	1.68e−1	1.68e−1
	5%	5.51e−11	6.32e−11	2.11e−1	2.11e−1	2.11e−1
`deriv2`	1%	2.19e−2	2.41e−2	1.18e−1	1.20e−1	1.13e−1
	5%	1.88e−2	2.38e−2	1.59e−1	1.60e−1	1.62e−1
`foxgood`	1%	2.78e−7	2.79e−7	4.93e−1	4.93e−1	4.93e−1
	5%	1.91e−7	1.96e−7	1.18e0	1.18e0	1.18e0
`gravity`	1%	1.38e−4	1.41e−4	7.86e−1	7.86e−1	7.86e−1
	5%	1.83e−4	1.84e−4	2.63e0	2.63e0	2.63e0
`heat`	1%	1.33e0	1.13e0	9.56e−1	1.67e0	1.50e0
	5%	9.41e−1	9.45e−1	2.02e0	1.70e0	1.99e0
`phillips`	1%	5.53e−3	4.09e−3	6.28e−2	6.19e−2	6.24e−2
	5%	6.89e−3	7.53e−3	9.57e−2	9.53e−2	9.79e−2
`shaw`	1%	3.51e−9	3.49e−9	4.36e0	4.36e0	4.36e0
	5%	1.34e−9	1.37e−9	8.23e0	8.23e0	8.23e0

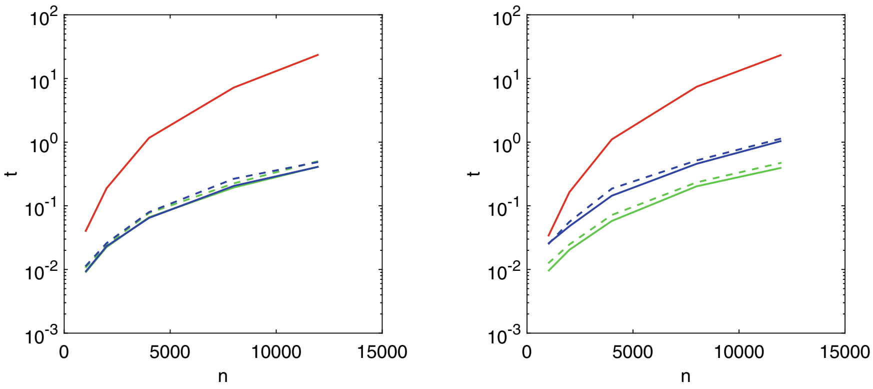
x_α. Thus, RSVD can maintain the reconstruction accuracy. For `heat`, despite the apparent large magnitude of the errors $\tilde{e}_{xz}$ and $\tilde{e}_{ij}$, the errors e_{xz} and e_{ij} are not much worse than e. A close inspection shows that the difference of the reconstructions are mostly in the tail part, which requires more modes for a full resolution. The computing time (in seconds) for obtaining x_α and $\tilde{x}_\alpha$ and $\hat{x}_\alpha$ is about 6.60, 0.220 and 0.220, where for the latter two, it includes also the time for computing RSVD. Thus, for all the examples, with a rank $k = 20$, RSVD can accelerate standard Tikhonov regularization by a factor of 30, while maintaining the accuracy, and the proposed approach is competitive with the one in [32]. Note that the choice $k = 20$ can be greatly reduced for severely ill-posed problems; see Sect. 5.2 below for discussions.

The preceding observations remain largely valid for general Tikhonov regularization; see Table 2. Since the construction of the approximation $\hat{x}_\alpha$ does not retain the structure of the regularized solution x_α, the error $\tilde{e}_{xz}$ can potentially be much larger than $\tilde{e}_{ij}$, which can indeed be observed. The errors e, e_{xz} and e_{ij} are mostly comparable, except for `deriv2`. For `deriv2`, the approximation $\hat{x}_\alpha$ suffers from grave errors, since the projection of L into $\mathscr{R}(Q)$ is very inaccurate for preserving L. It is expected that the loss occurs whenever general Tikhonov penalty is much more effective than the standard one. This shows the importance of structure preservation. Note that, for a general L, $\tilde{x}_\alpha$ takes only about 1.5 times the computing time of $\hat{x}_\alpha$. This cost can be further reduced since L is highly structured and admits fast inversion. Thus preserving the range structure of x_α in (4) does not incur much overhead.

Last, we present some results on the computing time for `deriv2` versus the problem dimension, and at two truncation levels for RSVD, i.e., $k = 20$ and $k = 30$. The numerical results are given in Fig. 1. The cubic scaling of the standard approach and quadratic scaling of the approach based on RSVD are clearly observed,

Table 2 Numerical results by general Tikhonov regularization (with the first-order derivative penalty) for the examples at two noise levels

example	δ	$\tilde{e}_{xz}$	$\tilde{e}_{ij}$	e	e_{xz}	e_{ij}
baart	1%	3.35e−10	2.87e−10	1.43e−1	1.43e−1	1.43e−1
	5%	3.11e−10	8.24e−12	1.48e−1	1.48e−1	1.48e−1
deriv2	1%	1.36e−1	4.51e−4	1.79e−2	1.48e−1	1.78e−2
	5%	1.57e−1	3.85e−4	2.40e−2	1.77e−1	2.40e−2
foxgood	1%	4.84e−2	2.26e−8	9.98e−1	1.02e0	9.98e−1
	5%	1.90e−2	1.51e−9	2.27e0	2.28e0	2.27e0
gravity	1%	3.92e−2	2.33e−5	1.39e0	1.41e0	1.39e0
	5%	1.96e−2	9.47e−6	3.10e0	3.10e0	3.10e0
heat	1%	5.54e−1	8.74e−1	8.95e−1	1.06e0	1.32e0
	5%	8.90e−1	1.01e0	1.87e0	1.76e0	1.99e0
phillips	1%	3.25e−3	3.98e−4	6.14e−2	6.06e−2	6.14e−2
	5%	5.64e−3	5.82e−4	8.37e−2	8.18e−2	8.34e−2
shaw	1%	3.79e−4	3.70e−8	3.32e0	3.32e0	3.32e0
	5%	9.73e−4	2.17e−8	9.23e0	9.23e0	9.23e0



(a) standard Tikhonov

(b) general Tikhonov

Fig. 1 The computing time t (in seconds) for the example deriv2 at different dimension n ($m = n$). The red, green and blue curves refer to Tikhonov regularization, existing approach [32, 33] and the new approach, respectively, and the sold and dashed curves denote $k = 20$ and $k = 30$, respectively

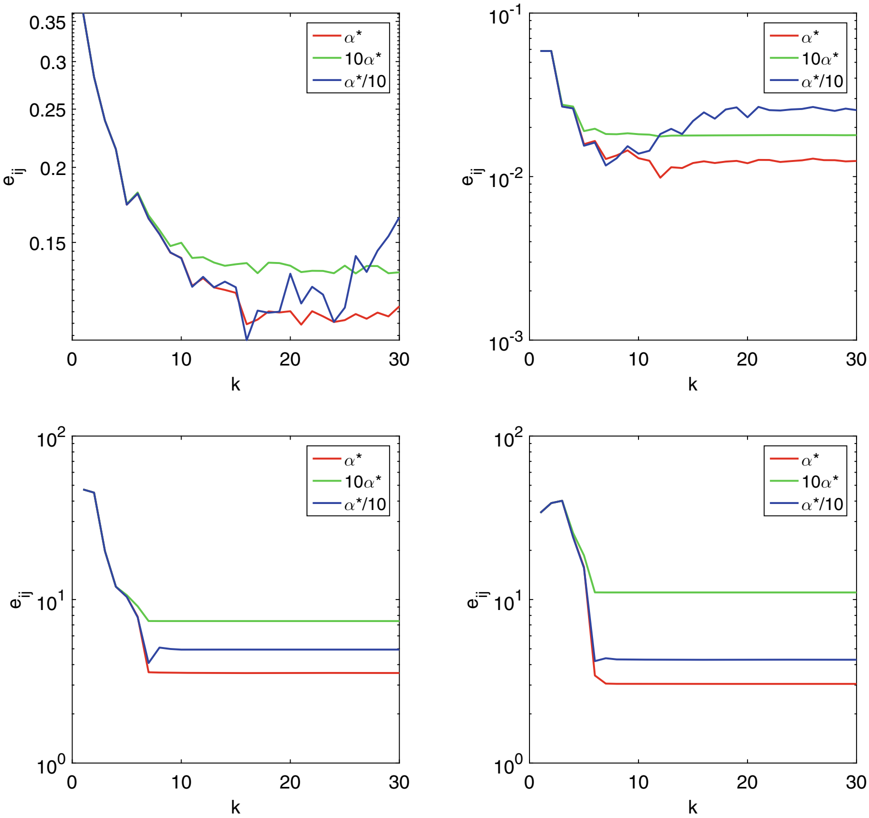
confirming the complexity analysis in Sects. 2 and 3. In both (9) and (16), computing RSVD represents the dominant part of the overall computational efforts, and thus the increase of the rank k from 20 to 30 adds very little overheads (compare the dashed and solid curves in Fig. 1). Further, for Tikhonov regularization, the two randomized variants are equally efficient, and for the general one, the proposed approach is slightly more expensive due to its direct use of L in constructing the approximation

$\tilde{B}_k$ to $B := AL^{\#}$. Although not presented, we note that the results for other examples are very similar.

5.2 Convergence of the Algorithm

There are several factors influencing the quality of $\tilde{x}_\alpha$ the regularization parameter α, the noise level δ and the rank k of the RSVD approximation. The optimal truncation level k should depend on both α and δ. This part presents a study with `deriv2` and `shaw`, which are mildly and severely ill-posed, respectively.

First, we examine the influence of α on the optimal k. The numerical results for three different levels of regularization are given in Fig. 2. In the figure, the notation α^*



(a) standard Tikhonov (b) general Tikhonov

Fig. 2 The convergence of the error e_{ij} with respect to the rank k for `deriv2` (top) and `shaw` (bottom) with $\delta = 1\%$ and different regularization parameters

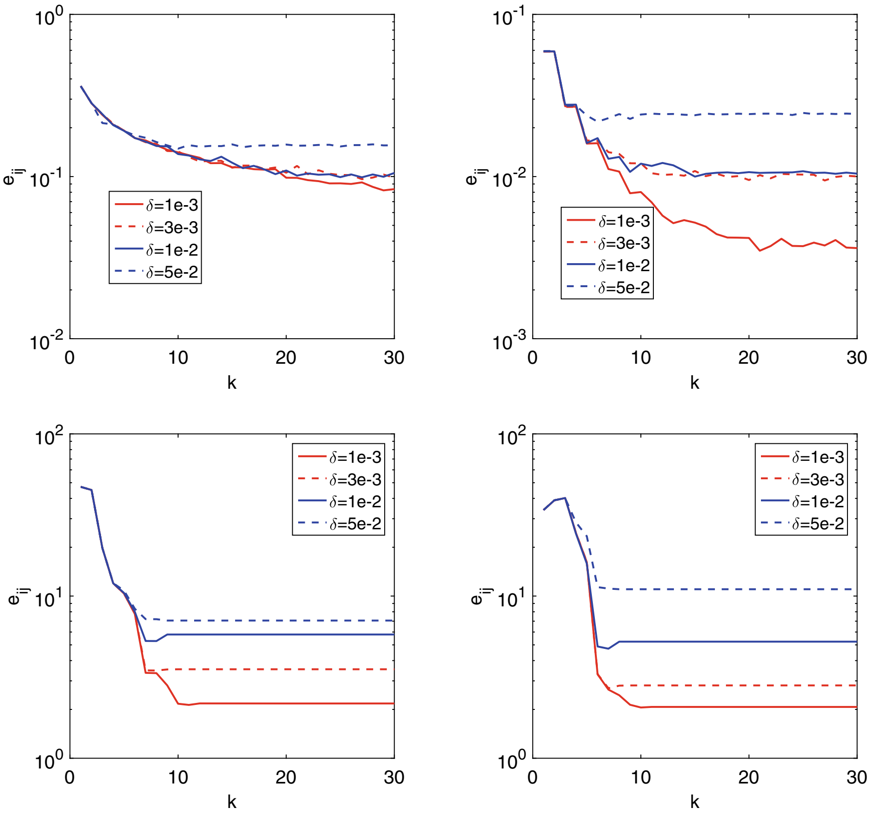
refers to the value attaining the the smallest error for Tikhonov solution x_α, and thus $10\alpha^*$ and $\alpha^*/10$ represent respectively over- and under-regularization. The optimal k value decreases with the increase of α when $\alpha \gg \alpha^*$. This may be explained by the fact that a too large α causes large approximation error and thus can tolerate large errors in the approximation $\tilde{A}_k$ (for a small k). The dependence can be sensitive for mildly ill-posed problems, and also on the penalty. The penalty influences the singular value spectra in the RSVD approximation implicitly by preconditioning: since L is a discrete differential operator, the (weighted) pseudoinverse $L^\#$ (or $L^\dagger$) is a smoothing operator, and thus the singular values of $B = AL^\#$ decay faster than that of A. In all cases, the error e_{ij} is nearly monotonically decreasing in k (and finally levels off at e, as expected). In the under-regularized regime (i.e., $\alpha \ll \alpha^*$), the behavior is slightly different: the error e_{ij} first decreases, and then increases before eventually leveling off at e. This is attributed to the fact that proper low-rank truncation of A induces extra regularization, in a manner similar to TSVD in Sect. 3.1. Thus, an approximation that is only close to x_α (see e.g., [1, 30, 32, 33]) is not necessarily close to $x^\dagger$, when α is not chosen properly.

Next we examine the influence of the noise level δ; see Fig. 3. With the optimal choice of α, the optimal k increases as δ decreases, which is especially pronounced for mildly ill-posed problems. Thus, RSVD is especially efficient for the following two cases: (a) highly noisy data (b) severely ill-posed problem. These observations agree well with Theorem 4: a low-rank approximation $\tilde{A}_k$ whose accuracy is commensurate with δ is sufficient, and in either case, a small rank is sufficient for obtaining an acceptable approximation. For a fixed k, the error e_{ij} almost increases monotonically with the noise level δ.

These empirical observations naturally motivate developing an adaptive strategy for choosing the rank k on the fly so as to effect the optimal complexity. This requires a careful analysis of the balance between k, δ, α, and suitable *a posteriori* estimators. We leave this interesting topic to a future work.

5.3 Electrical Impedance Tomography

Last, we illustrate the approach on 2D electrical impedance tomography (EIT), a diffusive imaging modality of recovering the electrical conductivity from boundary voltage measurement. This is one canonical nonlinear inverse problem. We consider the problem on a unit circle with sixteen electrodes uniformly placed on the boundary, and adopt the complete electrode model [24] as the forward model. It is discretized by the standard Galerkin FEM with conforming piecewise linear basis functions, on a quasi-uniform finite element mesh with 2129 nodes. For the inversion step, we employ ten sinusoidal input currents, unit contact impedance and measure the voltage data (corrupted by $\delta = 0.1\%$ noise). The reconstructions are obtained with an $H^1(\Omega)$-seminorm penalty. We refer to [7, 15] for details on numerical implementation. We test the RSVD algorithm with the linearized model. It can be implemented efficiently without explicitly computing the linearized map. More precisely, let F



(a) standard Tikhonov

(b) general Tikhonov

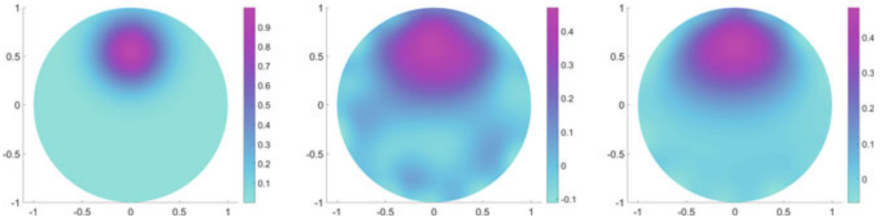
Fig. 3 The convergence of the error e_{ij} with respect to the rank k for `deriv2` (top) and `shaw` (bottom) at different noise levels

be the (nonlinear) forward operator, and σ_0 be the background (fixed at 1). Then the random probing of the range $\mathscr{R}(F'(\sigma_0))$ of the linearized forward operator $F'(\sigma_0)$ (cf. Step 4 of Algorithm 1) can be approximated by

$$F'(\sigma_0)\omega_i \approx F(\sigma_0 + \omega_i) - F(\sigma_0), \quad i = 1, \ldots k + p,$$

and it can be made very accurate by choosing a small variance for the random vector ω_i. Step 6 of Algorithm 1 can be done efficiently via the adjoint technique.

The numerical results are presented in Fig. 4, where linearization refers to the reconstruction by linearizing the nonlinear forward model at the background σ_0. This is one of the most classical reconstruction methods in EIT imaging. The rank k is taken to be $k = 30$ for $\tilde{x}_\alpha$, which is sufficient given the severe ill-posed nature of the EIT inverse problem. Visually, the RSVD reconstruction is indistinguishable



exact linearization RSVD

Fig. 4 Numerical reconstructions for EIT with 0.1% noise

from the conventional approach. Note that contrast loss is often observed for EIT reconstructions obtained by a smoothness penalty. The computing time (in seconds) for RSVD is less than 8, whereas that for the conventional method is about 60. Hence, RSVD can greatly accelerate EIT imaging.

6 Conclusion

In this work, we have provided a unified framework for developing efficient linear inversion techniques via RSVD and classical regularization methods, building on a certain range condition on the regularized solution. The construction is illustrated on three popular linear inversion methods for finding smooth solutions, i.e., truncated singular value decomposition, Tikhonov regularization and general Tikhonov regularization with a smoothness penalty. We have provided a novel interpretation of the approach via convex duality, i.e., it first approximates the dual variable via randomized SVD and then recovers the primal variable via duality relation. Further, we gave rigorous error bounds on the approximation under the canonical sourcewise representation, which provide useful guidelines for constructing a low-rank approximation. We have presented extensive numerical experiments, including nonlinear tomography, to illustrate the efficiency and accuracy of the approach, and demonstrated its competitiveness with existing methods.

Algorithm 3 Iterative refinement of RSVD-Tikhonov solution.

```
Give $A$, $b$ and $J$, and initialize $(x^0, p^0) = (0, 0)$.
Compute RSVD $(\tilde{U}_k, \tilde{\Sigma}_k, \tilde{V}_k)$ to $AL^\dagger$ by Algorithm 1.
for $j = 1, \ldots, J$ do
    Compute the auxiliary variable $z^j$ by (21).
    Update the dual variable $p^{j+1}$ by (22).
    Update the primal variable $x^{j+1}$ by (23).
    Check the stopping criterion.
end for
Output $x^J$ as an approximation to $x_\alpha$.
```

Appendix A: Iterative refinement

Proposition 1 enables iteratively refining the inverse solution when RSVD is not sufficiently accurate. This idea was proposed in [29, 34] for standard Tikhonov regularization, and we describe the procedure in a slightly more general context. Suppose $\mathscr{N}(L) = \{0\}$. Given a current iterate x^j, we define a functional $J_\alpha^j(\delta x)$ for the increment δx by

$$J_\alpha^j(\delta x) := \|A(\delta x + x^j) - b\|^2 + \alpha\|L(\delta x + x^j)\|^2.$$

Thus the optimal correction δx_α satisfies

$$(A^*A + \alpha L^*L)\delta x_\alpha = A^*(b - Ax^j) - \alpha L^*Lx^j,$$

i.e.,

$$(B^*B + \alpha I)L\delta x_\alpha = B^*(b - Ax^j) - \alpha Lx^j, \tag{20}$$

with $B = AL^\dagger$. However, its direct solution is expensive. We employ RSVD for a low-dimensional space $\tilde{V}_k$ (corresponding to B), parameterize the increment $L\delta x$ by $L\delta x = \tilde{V}_k^* z$ and update z only. That is, we minimize the following functional in z

$$J_\alpha^j(z) := \|A(L^\dagger \tilde{V}_k^* z + x^j) - b\|^2 + \alpha\|z + \tilde{V}_k L x^j\|^2.$$

Since $k \ll m$, the problem can be solved efficiently. More precisely, given the current estimate x^j, the optimal z solves

$$(\tilde{V}_k B^* B \tilde{V}_k^* + \alpha I)z = \tilde{V}_k B^*(b - Ax^j) - \alpha \tilde{V}_k L x^j. \tag{21}$$

It is the Galerkin projection of (20) for δx_α onto the subspace $\tilde{V}_k$. Then we update the dual ξ and the primal x by the duality relation in Sect. 6:

$$\xi^{j+1} = b - Ax^j - B\tilde{V}_k^* z^j, \tag{22}$$

$$x^{j+1} = \alpha^{-1}\Gamma A^* \xi^{j+1}. \tag{23}$$

Summarizing the steps gives Algorithm 3. Note that the duality relation (17) enables A and A^* to enter into the play, thereby allowing progressively improving the accuracy. The main extra cost lies in matrix-vector products by A and A^*.

The iterative refinement is a linear fixed-point iteration, with the solution x_α being a fixed point and the iteration matrix being independent of the iterate. Hence, if the first iteration is contractive, i.e., $\|x^1 - x_\alpha\| \leq c\|x^0 - x_\alpha\|$, for some $c \in (0, 1)$, then Algorithm 3 converges linearly to x_α. It can be satisfied if the RSVD approximation $(\tilde{U}_k, \tilde{\Sigma}_k, \tilde{V}_k)$ is reasonably accurate to B.

References

1. Boutsidis, C., Magdon-Ismail, M.: Faster SVD-truncated regularized least-squares. In: 2014 IEEE International Symposium on Information Theory, pp. 1321–1325 (2014). https://doi.org/10.1109/ISIT.2014.6875047
2. Chen, S., Liu, Y., Lyu, M.R., King, I., Zhang, S.: Fast relative-error approximation algorithm for ridge regression. In: Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence, pp. 201–210 (2015)
3. Ekeland, I., Témam, R.: Convex Analysis and Variational Problems. SIAM, Philadelphia, PA (1999). https://doi.org/10.1137/1.9781611971088
4. Eldén, L.: A weighted pseudoinverse, generalized singular values, and constrained least squares problems. BIT **22**(4), 487–502 (1982). https://doi.org/10.1007/BF01934412
5. Engl, H.W., Hanke, M., Neubauer, A.: Regularization of Inverse Problems. Kluwer, Dordrecht (1996). https://doi.org/10.1007/978-94-009-1740-8
6. Frieze, A., Kannan, R., Vempala, S.: Fast Monte-Carlo algorithms for finding low-rank approximations. J. ACM **51**(6), 1025–1041 (2004). https://doi.org/10.1145/1039488.1039494
7. Gehre, M., Jin, B., Lu, X.: An analysis of finite element approximation in electrical impedance tomography. Inverse Prob. **30**(4), 045,013,24 (2014). https://doi.org/10.1088/0266-5611/30/4/045013
8. Golub, G.H., Van Loan, C.F.: Matrix Computations, 3rd edn. Johns Hopkins University Press, Baltimore, MD (1996)
9. Griebel, M., Li, G.: On the decay rate of the singular values of bivariate functions. SIAM J. Numer. Anal. **56**(2), 974–993 (2018). https://doi.org/10.1137/17M1117550
10. Gu, M.: Subspace iteration randomization and singular value problems. SIAM J. Sci. Comput. **37**(3), A1139–A1173 (2015). https://doi.org/10.1137/130938700
11. Halko, N., Martinsson, P.G., Tropp, J.A.: Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. SIAM Rev. **53**(2), 217–288 (2011). https://doi.org/10.1137/090771806
12. Horn, R.A., Johnson, C.R.: Matrix Analysis. Cambridge University Press, Cambridge (1985). https://doi.org/10.1017/CBO9780511810817
13. Ito, K., Jin, B.: Inverse Problems: Tikhonov Theory and Algorithms. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ (2015)
14. Jia, Z., Yang, Y.: Modified truncated randomized singular value decomposition (MTRSVD) algorithms for large scale discrete ill-posed problems with general-form regularization. Inverse Prob. **34**(5), 055,013,28 (2018). https://doi.org/10.1088/1361-6420/aab92d
15. Jin, B., Xu, Y., Zou, J.: A convergent adaptive finite element method for electrical impedance tomography. IMA J. Numer. Anal. **37**(3), 1520–1550 (2017). https://doi.org/10.1093/imanum/drw045
16. Kluth, T., Jin, B.: Enhanced reconstruction in magnetic particle imaging by whitening and randomized SVD approximation. Phys. Med. Biol. **64**(12), 125,026,21 (2019). https://doi.org/10.1088/1361-6560/ab1a4f
17. Kluth, T., Jin, B., Li, G.: On the degree of ill-posedness of multi-dimensional magnetic particle imaging. Inverse Prob. **34**(9), 095,006,26 (2018). https://doi.org/10.1088/1361-6420/aad015
18. Maass, P.: The x-ray transform: singular value decomposition and resolution. Inverse Prob. **3**(4), 729–741 (1987). http://stacks.iop.org/0266-5611/3/729
19. Maass, P., Rieder, A.: Wavelet-accelerated Tikhonov-Phillips regularization with applications. In: Inverse Problems in Medical Imaging and Nondestructive Testing (Oberwolfach, 1996), pp. 134–158. Springer, Vienna (1997)
20. Musco, C., Musco, C.: Randomized block Krylov methods for stronger and faster approximate singular value decomposition. In: NIPS 2015 (2015)
21. Neubauer, A.: An a posteriori parameter choice for Tikhonov regularization in the presence of modeling error. Appl. Numer. Math. **4**(6), 507–519 (1988). https://doi.org/10.1016/0168-9274(88)90013-X

22. Pilanci, M., Wainwright, M.J.: Iterative Hessian sketch: fast and accurate solution approximation for constrained least-squares. J. Mach. Learn. Res. **17**, 38 (2016)
23. Sarlos, T.: Improved approximation algorithms for large matrices via random projections. In: Foundations of Computer Science, 2006. FOCS '06. 47th Annual IEEE Symposium on (2006). https://doi.org/10.1109/FOCS.2006.37
24. Somersalo, E., Cheney, M., Isaacson, D.: Existence and uniqueness for electrode models for electric current computed tomography. SIAM J. Appl. Math. **52**(4), 1023–1040 (1992). https://doi.org/10.1137/0152060
25. Stewart, G.W.: On the perturbation of pseudo-inverses, projections and linear least squares problems. SIAM Rev. **19**(4), 634–662 (1977). https://doi.org/10.1137/1019104
26. Szlam, A., Tulloch, A., Tygert, M.: Accurate low-rank approximations via a few iterations of alternating least squares. SIAM J. Matrix Anal. Appl. **38**(2), 425–433 (2017). https://doi.org/10.1137/16M1064556
27. Tao, T.: Topics in Random Matrix Theory. AMS, Providence, RI (2012). https://doi.org/10.1090/gsm/132
28. Tautenhahn, U.: Regularization of linear ill-posed problems with noisy right hand side and noisy operator. J. Inverse Ill-Posed Probl. **16**(5), 507–523 (2008). https://doi.org/10.1515/JIIP.2008.027
29. Wang, J., Lee, J.D., Mahdavi, M., Kolar, M., Srebro, N.: Sketching meets random projection in the dual: a provable recovery algorithm for big and high-dimensional data. Electron. J. Stat. **11**(2), 4896–4944 (2017). https://doi.org/10.1214/17-EJS1334SI
30. Wei, Y., Xie, P., Zhang, L.: Tikhonov regularization and randomized GSVD. SIAM J. Matrix Anal. Appl. **37**(2), 649–675 (2016). https://doi.org/10.1137/15M1030200
31. Witten, R., Candès, E.: Randomized algorithms for low-rank matrix factorizations: sharp performance bounds. Algorithmica **72**(1), 264–281 (2015). https://doi.org/10.1007/s00453-014-9891-7
32. Xiang, H., Zou, J.: Regularization with randomized SVD for large-scale discrete inverse problems. Inverse Prob. **29**(8), 085,008,23 (2013). https://doi.org/10.1088/0266-5611/29/8/085008
33. Xiang, H., Zou, J.: Randomized algorithms for large-scale inverse problems with general Tikhonov regularizations. Inverse Prob. **31**(8), 085,008,24 (2015). https://doi.org/10.1088/0266-5611/31/8/085008
34. Zhang, L., Mahdavi, M., Jin, R., Yang, T., Zhu, S.: Random projections for classification: a recovery approach. IEEE Trans. Inform. Theory **60**(11), 7300–7316 (2014). https://doi.org/10.1109/TIT.2014.2359204

Parameter Selection in Dynamic Contrast-Enhanced Magnetic Resonance Tomography

Kati Niinimäki, M. Hanhela, and V. Kolehmainen

Abstract In this work we consider the image reconstruction problem of sparsely sampled dynamic contrast-enhanced (DCE) magnetic resonance imaging (MRI). DCE-MRI is a technique for acquiring a series of MR images before, during and after intravenous contrast agent administration, and it is used to study microvascular structure and perfusion. To overcome the ill-posedness of the related spatio-temporal inverse problem, we use regularization. In regularization one of the main problems is how to determine the regularization parameter which controls the balance between data fitting term and regularization term. Most methods for selecting this parameter require the computation of a large number of estimates even in stationary problems. In dynamic imaging, the parameter selection is even more time consuming since separate regularization parameters are needed for the spatial and temporal regularization functionals. In this work, we study the possibility of using the S-curve with DCE-MR data. We select the spatial regularization parameter using the S-curve, leaving the temporal regularization parameter as the only free parameter in the reconstruction problem. In this work, the temporal regularization parameter is selected manually by computing reconstructions with several values of the temporal regularization parameter.

Keywords Sparsely sampled MRI · Dynamic contrast-enhanced MRI · Image reconstruction · S-curve regularization

MSC: 65R20 · 65R32

K. Niinimäki (✉)
IR4M UMR8081, CNRS, Université Paris-Sud, Université Paris-Saclay, SHFJ,
4 place du Général Leclerc, 91401 Orsay, France
e-mail: kati.niinimaki@u-psud.fr; kati.niinimaki@planmeca.com

Xray Division, Planmeca Oy, Asentajankatu 6, 00880 Helsinki, Finland

M. Hanhela · V. Kolehmainen
Department of Applied Physics, University of Eastern Finland, Kuopio, Finland
e-mail: matti.hanhela@uef.fi

V. Kolehmainen
e-mail: ville.kolehmainen@uef.fi

L. Beilina et al. (eds.), *Mathematical and Numerical Approaches for Multi-Wave Inverse Problems*, Springer Proceedings in Mathematics & Statistics 328,
https://doi.org/10.1007/978-3-030-48634-1_6

1 Introduction

Dynamic contrast-enhanced MRI (DCE-MRI) is an imaging method which is used to study microvascular structure and tissue perfusion. The method has many applications including blood-brain-barrier assessment after acute ischemic stroke [20, 30] and treatment monitoring of breast cancer [19, 24] and glioma [25]. The operation principle of DCE-MRI is to inject a bolus of gadolinium based contrast agent into the blood stream, and acquire a time series of MRI data with a suitable T_1-weighting to obtain a time series of 2D (or 3D) images which exhibit contrast changes induced by concentration changes of the contrast agent in the tissues.

The analysis of the contrast agent dynamics during imaging requires high resolution in space and in time; the high temporal resolution is necessary to measure when the contrast is passing through the artery and it is used for determining the subject specific Arterial Input function (AIF), and the high spatial resolution is necessary to adequately capture boundaries of perfused tissues. In many cases, sufficient time resolution can only be obtained by utilizing an imaging protocol during which only partial k-space sampling can be obtained for each image in the time series. However, acquiring less samples than required by the Nyquist criterion makes the image reconstruction problem *ill-posed* (and non-unique), causing artifacts and deterioration of the image quality when conventional reconstruction methods such as inverse Fourier transform or re-gridding are employed.

According to the theory of compressed sensing (CS) [5–7], images that have a sparse representation can be recovered from undersampled measurements of a linear transform, i.e. sampling rate below Nyquist rate, using appropriate nonlinear reconstruction algorithms and appropriate (random) sampling of the data space. The compressibility of MR images and the fact that MR scanner measures samples of a linear transform of the unknown image (the k-space samples can be mathematically considered as Fourier coefficients) suggest that the idea of CS is applicable to MR imaging, offering thus a potentially significant scan time reduction without sacrificing the image quality. Since the seminal work of Candès, Romberg and Tao in 2006, CS has been extensively applied to MRI. In 2007 CS was applied to MRI in [18] and in 2018 it received FDA approval for clinical use.

The undersampling of the k-space for speeding up the dynamic MRI data acquisition can in principle be done in many different ways. However, for many applications, the low frequency features that are present in the center of the k-space are of importance. In [33], for example, it was demonstrated that center weighted random sampling patterns were preferable to purely random sampling of the k-space within the CS approach. Radial sampling has the advantage that the center of the k-space is sampled densely, even when the sampling (i.e., number of radial spokes) is remarkably reduced.

In this paper, we consider image reconstruction problem of DCE-MRI with sparsely sampled golden angle radial data, where the angle of subsequent spokes is $\sim$111.25°. The number of measurement spokes used for reconstructing a single time frame is chosen to be a Fibonacci number, which was shown to be an optimal choice in [32]. This type of sparse sampling between the time frames results in each time frame having different spokes, and eventually the spokes cover the whole k-space.

We also combine the Golden Angle (GA) sampling with concentric squares sampling. This sampling strategy resembles the linogram method [9] developed for computed tomography imaging, but the angles of subsequent spokes were chosen according to the golden angle method as opposed to the angles being equidistant in $\tan\theta$ in linogram sampling. Unlike the conventional radial sampling pattern with spokes of equal length, the concentric squares sampling strategy also covers the corners of the k-space. The sampling pattern therefore also collects information of the high frequencies in the corners of the k-space, which leads to a reduction of artifacts originating from the lack of sampling in the corners.

To overcome the ill-posedness of our image reconstruction problem, we use a variational framework. Thus we solve the following optimization problem

$$\hat{f} = \arg\min D(f, m) + \alpha S(f) + \beta T(f), \tag{1}$$

where $f = \{f_1, f_2, \ldots, f_{N_t}\}$ denotes the image sequence, $m = \{m_1, m_2, \ldots, m_{N_t}\}$ the data sequence, N_t the number of time frames, D the data-fitting term, S the spatial regularization, T the temporal regularization and α and β are the spatial and temporal regularization parameters, respectively. The selection of the regularization parameters is crucial in terms of resulting image quality. There exists several proposed parameter selection methods, but most of them require the computation of a large number of reconstructions with varying parameters. Having to fix two regularization parameters makes the selection even more time consuming.

In this work, we propose to use the S-curve method [12, 15, 22, 23] for automatic selection of the spatial regularization parameter α. The idea in the S-curve method is to select the regularization parameter so that the reconstruction has a priori defined level of sparsity in the chosen transformation domain. In DCE-MRI, a reliable a priori estimate for the sparsity level can be extracted from an anatomical MRI image which is based on full-sampling of the k-space and is always taken as part of the MRI measurement protocol but is usually used only for visualization purposes. For the selection of the spatial regularization parameter α, we employ one time frame of the GA data from the baseline measurement before the contrast agent administration. After fixing the spatial regularization parameter, we compute dynamic reconstructions with several values of parameter β and select a suitable temporal regularization parameter manually. Furthermore we study the performance of three different temporal regularization functionals, namely temporal smoothness, temporal total variation and total generalized variation.

The proposed method is evaluated using simulated GA DCE-MRI data from a rat brain phantom. The results are compared to re-gridding approach, which is the most widely used non-iterative algorithm for reconstructing images from non-Cartesian MRI data. Our re-gridding method was developed in IR4M UMR8081, CNRS, Université Paris-Sud using Matlab®. This re-gridding approach does not need additional density correction and it was first used in [16], see also [11].

2 Image Reconstruction in Radial DCE-MRI

2.1 Forward Problem

The forward problem in 2D MRI can be modelled for most measurement protocols by the Fourier transform

$$m(k_x, k_y) = \int_{\Omega} f(x, y) e^{-i2\pi(k_x x + k_y y)} \mathrm{d}x \mathrm{d}y, \tag{2}$$

where Ω is the image domain $f(x, y)$ is the unknown image, $m(k_x, k_y)$ is the measured data, and k_x, k_y denotes the k-space trajectories. In the discrete framework, the Fourier transform is typically approximated with the multidimensional FFT when using cartesian k-space trajectories and with the non-uniform FFT when using non-cartesian k-space trajectories.

In this work, we consider non-uniform k-space trajectories and approximate the Fourier transform by the non-uniform fast Fourier transform (nuFFT) operator [10]. We discretize our functions as follows; temporal direction is divided into a sequence of N_t (vectorized) images $f = \{f_1, f_2, \ldots, f_{N_t}\}$ and data vectors $m = \{m_1, m_2, \ldots, m_{N_t}\}$, where each $f_t \in \mathbb{C}^{N_p}$ and $m_t \in \mathbb{C}^M$, respectively. The number of data per frame M is equal to the number of GA spokes per frame times the number of samples per spoke. The number of image pixels is $N_p = N \times N$. Thus, using nuFFT we re-write (2) in a discretized form at time t as

$$m_t = A_t f_t + \epsilon_t, \quad t = 1, \ldots, N_t, \tag{3}$$

$A_t = P_t \mathscr{F} S_t$, where P_t is an interpolation matrix between Cartesian k-space and non-cartesian k-space, $\mathscr{F}$ is the 2D FFT operation and S_t is a scaling matrix.

2.2 Inverse Problem of Dynamic Image Reconstruction

The dynamic inverse problem related to the Eq. (3) is: given measurement time series $m = \{m_1, m_2, \ldots, m_{N_t}\}$ and the associated k-space trajectories, solve the unknown images $f = \{f_1, f_2, \ldots, f_{N_t}\}$. To recover f from m, we define the inverse problem as the optimization problem

$$\hat{f} = \arg\min \left\{ \sum_{t=1}^{N_t} \|A_t f_t - m_t\|_2^2 + \alpha S(f) + \beta T(f) \right\}, \tag{4}$$

where $S(f)$ denotes the spatial regularization functional, α the spatial regularization parameter, β the temporal regularization parameter and $T(f)$ the temporal regularization functional.

In this work, we study the applicability of S-curve method for selecting the spatial regularization parameter α. Once α has been fixed, the temporal regularization parameter is then selected manually by computing estimates with different values of β. Our minimization problem is based on L_2-data fidelity term for the measurement model and we use spatial total variation regularization for promoting sparsity of the gradients of each image [27]. Furthermore we study the performance of three different temporal regularization functionals for promoting temporal regularity of the image series. Our spatio-temporal image reconstruction problem thus writes

$$\hat{f} = \arg\min \left\{ \sum_{t=1}^{N_t} \left(\|A_t f_t - m_t\|_2^2 + \alpha \|\nabla f_t\|_{2,1} \right) + \beta T(f) \right\}, \tag{5}$$

where the isotropic 2D spatial total variation norm for complex valued image f_t is defined by

$$\|\cdot\|_{2,1} = \sum_{k=1}^{N} \sqrt{(Re(D_x f_{t,k}))^2 + (Re(D_y f_{t,k}))^2 + (Im(D_x f_{t,k}))^2 + (Im(D_y f_{t,k}))^2}, \tag{6}$$

where Re and Im denoting the real and imaginary parts of f_t respectively and D_x and D_y the discrete forward first differences in horizontal and vertical directions, respectively.

2.2.1 Temporal Regularization 1: Temporal Smoothness (TS)

The temporal smoothness regularization (hereafter referred as TS) is defined as the L_2 norm of forward first differences in time:

$$T(f) = \sum_{t=1}^{N_t - 1} \|f_{t+1} - f_t\|_2^2. \tag{7}$$

This model promotes smooth slowly changing signals, and it has been used in [3] for radial DCE myocardial perfusion imaging. TS regularization was compared with temporal TV regularization in the same application in [1].

2.2.2 Temporal Regularization 2: Temporal Total Variation (TV)

Temporal total variation (hereafter referred as TV) is defined by the L_1 norm of the forward first differences in time:

$$T(f) = \sum_{t=1}^{N_t-1} \|f_{t+1} - f_t\|_1. \tag{8}$$

The temporal total variation model promotes sparsity of the time derivative of the pixel signals, being a highly feasible regularization model for reconstruction of piece-wise regular signals which may exhibit large jumps. The smoothed form of temporal total variation was used in [2] for multislice myocardial perfusion imaging.

2.2.3 Temporal Regularization 3: Total Generalized Variation (TGV)

The total generalized variation model [4] is a total variation model that is generalized to higher order differences. Here we use the second-order total generalized variation, which in the discrete 1-dimensional form is of the form

$$T(f) = \sum_{t=1}^{N_t-1} \min_{v} \|f_{t+1} - f_t - v_t\|_1 + \gamma \|v_{t+1} - v_t\|_1, \tag{9}$$

where v is an auxiliary vector and γ is a parameter, which balances the first and second order terms, and is set to $\sqrt{2}$ here.

This functional balances between minimizing the first-order and second-order differences of the signal. The difference with TV regularization is most clear in smooth regions where piecewise linear solutions are favored over the piecewise constant solutions of TV. From hereafter this temporal regularization is referred as TGV.

TGV was first used in MRI as a spatial prior in [14], and it has also been used in DCE-MRI as a temporal prior in [31], where different temporal priors were compared in cartesian MRI of the breast.

2.3 Regularization Parameter Selection

2.3.1 Spatial Regularization Parameter Selection

The spatial regularization parameter is selected using the S-curve method, originally proposed in [12, 15, 22, 23], but here modified for TV regularization.

Assume that we have an *a priori* estimate $\hat{S}$ for the total variation norm of the unknown function. In practice we can use an anatomical image of the same slice in order to obtain a reliable estimate for $\hat{S}$. Such an anatomical image is practically always acquired as part of the DCE MRI acquisition experiment but usually only used for visualization purposes. However, if such an anatomical image was not acquired, we could, in case of GA acquisition, estimate the expected sparsity level from a conventional reconstruction of a long sequence of baseline data taken before the contrast agent injection. Or the anatomical image could be estimated from the entire data-acquisition similarly as the composite image in [21].

Now, given the estimate $\hat{S}$ we select the regularization parameter α using the S-curve method as follows

(1) Take a sequence of regularization parameters α ranging on the interval $[0, \infty]$ such that

$$0 < \alpha^{(1)} < \alpha^{(2)} < \cdots < \alpha^{(L)} < \infty.$$

(2) Compute the corresponding estimates $\hat{f}_1(\alpha^{(1)}), \ldots, \hat{f}_1(\alpha^{(L)})$.
With DCE-MRI data, reconstructions $\hat{f}_1(\alpha^{(\ell)})$ are computed as follows; we take the data that correspond to the first time frame, i.e. m_1 which has number of elements equal to the number of GA spokes per frame times the number of points per spoke. We reconstruct f_1 for given value $\alpha^{(\ell)}$ by

$$\hat{f}_1(\alpha^{(\ell)}) = \arg\min_{f_1}\{\|A_1 f_1 - m_1\|_2^2 + \alpha^{(\ell)}\|\nabla f_1\|_{2,1}\}.$$

Here it is important that $\alpha^{(1)}$ is taken to be so small that the problem is under regularized and the corresponding reconstruction $\hat{f}_1(\alpha^{(1)})$ results to a very noisy image with a big TV-norm value and $\alpha^{(L)}$ is taken so large that the problem is over regularized and TV norm of reconstruction $\hat{f}_1(\alpha^{(L)})$ is very close to zero.

(3) Compute the TV-norms of the recovered estimates $\hat{f}_1(\alpha^{(\ell)}), \quad \ell = 1, \ldots, L$.

(4) Fit a smooth interpolation curve to the data $\{\alpha^{(\ell)}, S(\alpha^{(\ell)}), \quad \ell = 1, \ldots, L\}$ and use the interpolated sparsity curve to find the value of α for which $S(\alpha) = \hat{S}$. For the interpolation we use Matlab's® *interp* function and we interpolate our original S-curve to a more dense discretization of the regularization parameter α.

2.4 Temporal Regularization Parameter Selection

Once the spatial regularization parameter α has been fixed using the S-curve, the temporal regularization parameter β can be tuned by computing estimates with different values of β and selecting a suitable value manually, for example, by visual assessment of the results.

In this work, we compute the results with three different temporal regularization models using simulated measurement data. Since we consider a simulated test case where a ground truth is available, we select an optimal value of β for each temporal regularization model by selecting the value of β which produces the reconstruction with the smallest root mean square error (RMSE). The RMSE values were calculated separately for three regions; tumor, vascular region and the rest of the image domain. The RMSEs of different ROIs were then used to define a joint RMSE as

$$RMSE_{joint} = \sqrt{RMSE^2_{\Omega_{roi1}} + RMSE^2_{\Omega_{roi2}} + RMSE^2_{\Omega_{roi3}}}, \tag{10}$$

where Ω_{roi1} corresponds to pixels in the vascular region, Ω_{roi2} correspond to pixels in the tumour region and Ω_{roi3} corresponds to pixels in rest of the image domain. The RMSE was calculated this way to weigh the small tumour and vascular regions appropriately. In estimating the pharmacokinetic parameters of tissues, obtaining an accurate arterial input function (AIF) is required [28]. The AIF can be obtained via population averaging, however, usage of patient specific AIF produces more accurate estimates of the kinetic parameters [26]. The AIF is preferably extracted from an artery feeding the tissues of interest, but it can also be estimated from a venous sinus or vein when an artery is not visible [17]. Here the AIF is estimated from the superior sagittal sinus.

3 Materials and Methods

A simulated test case modelling DCE-MRI measurements of a glioma in rat brain was created. The rat brain phantom was based on the rat brain atlas in [29], and scaled to a size of 128 × 128. The rat brain image was divided into three subdomains of different signal behaviour: vascular region (labelled '1' and highlighted with red in Fig. 1), tumour region (labelled '2' and highlighted with blue in Fig. 1) and the rest of the brain tissue. The vascular signal region corresponds to the location of the superior sagittal sinus.

A time series of 2800 ground truth images was simulated by multiplying the signal of each pixel with the template of the corresponding region and adding that to the baseline value of the pixels. The tumour signal templates were based on an experimental DCE-MRI measurement described in [13], where the three different ROIs were identified. Figure 1 shows the signal templates for each of the different tissue regions.

One spoke of GA k-space data was simulated for each of the simulated images, leading to a dynamic experiment with 2800 spokes of k-space data. The time scale of the simulation was set to be similar to the in vivo measurements in [13] where the measurement time between consecutive GA spokes was 38.5 ms. Gaussian complex noise at 5% of the mean of the absolute values of the signal was added to the simulated k-space signal. The simulated test case was carried out using a k-space trajectory which combines the golden angle and the concentric squares sampling strategies.

In [13] it was found that reconstruction of the form (5) performed optimally with segment length[1] of 34 for a similar data set, thus we selected 34 as the segment length for our reconstructions leading to a temporal resolution of ∼1.3 s.

All the regularized reconstructions in this work were computed using the Chambolle-Pock primal-dual algorithm [8]. In the NUFFT implementation of the

[1] Segment length equals the number of radial spokes per image. The number of elements M in the data vector m_t is segment length times number of samples per spoke.

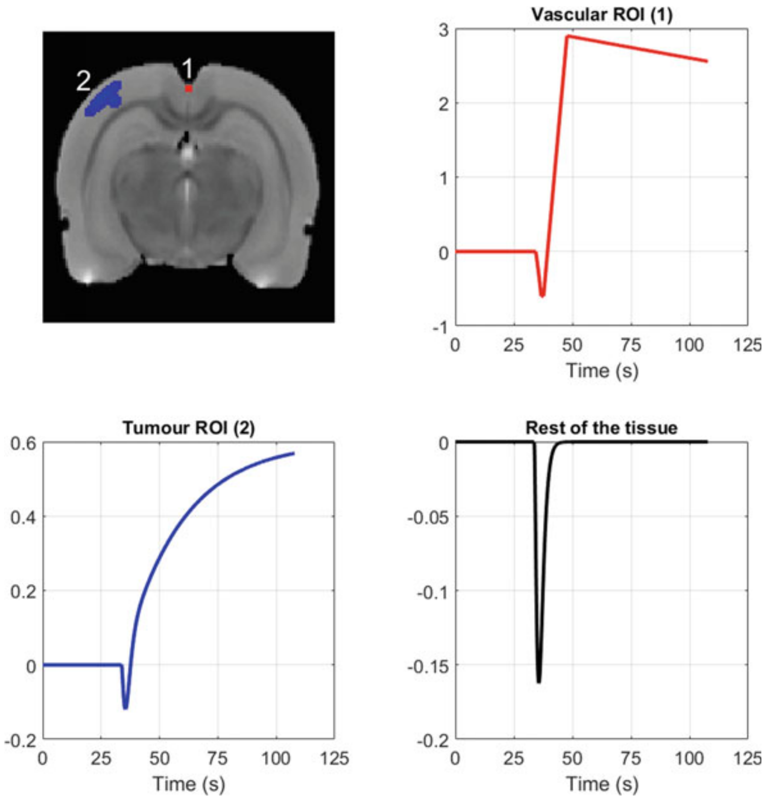


Fig. 1 Template signals of different ROIs. Top left: the simulated image with vascular ROI labelled '1' and marked with red and tumour ROI labelled '2' and marked with blue. Top right: the template signal of the vascular ROI. Bottom left: the template signal of the tumour ROI. Bottom right: the template signal of tissue outside the two ROIs. The vertical axis in the three template signals is the multiplier for the signal added to the base signal

forward model, the measurements were interpolated into a twice oversampled cartesian grid with min-max Kaiser-Bessel interpolation with a neighbourhood size of 4 [10]. The regridding reconstructions were computed using a Matlab code developped at Imagerie par résonance magnétique médicale et multi-modalités (IR4M) UMR8081, Université Paris-Sud, France.

We remark that when computing the RMS error (10), the reconstructed time signals of each pixel were linearly interpolated in the temporal direction to match the temporal resolution of the ground truth phantom.

4 Results

The selection of α was carried out using the first 34 spokes (i.e. the first frame m_1) of the DCE-MRI data and then the selected α was used for all the spatio-temporal reconstructions with different temporal regularizations. The rat brain phantom of Sect. 3 was used to compute the a priori level of sparsity, i.e. in our case we computed the TV norm of the first time frame of the dynamic phantom. This resulted in a sparsity level of $\hat{S} = 0.0259$. The spatial regularization term α was selected using the S-curve as described in Sect. 2.3.1 and resulted in spatial regularization parameter value of $\alpha = 7.3e{-}4$. The TV norms of the reconstructions for the S-curve were computed with 15 values of alpha ranging on interval $[10^{-7}, 10^3]$. These 15 values of TV norm were then interpolated using Matlab's® *interp* function to 405 values. The resulting S-curve for the determination of α is presented in Fig. 2.

In many practical applications, the a priori information, which we use to estimate the value of $\hat{S}$, may come from a different modality or from acquisition with different pulse sequence than the one used in the dynamic measurement. Therefore, in order to compute meaningful estimate of $\hat{S}$ for the TV-regularized case, the reference image has to be scaled such that it is compatible with the measured dynamic data. This normalization of the reference image can be obtained by

$$f_{\text{ref}} = \frac{\|m_t\|}{\|A_t f_{\text{ref}}\|} f_{\text{ref}},$$

where f_{ref} denotes the reference image, m_t the frame of dynamic data that is used in the S-curve and A_t the respective forward model.

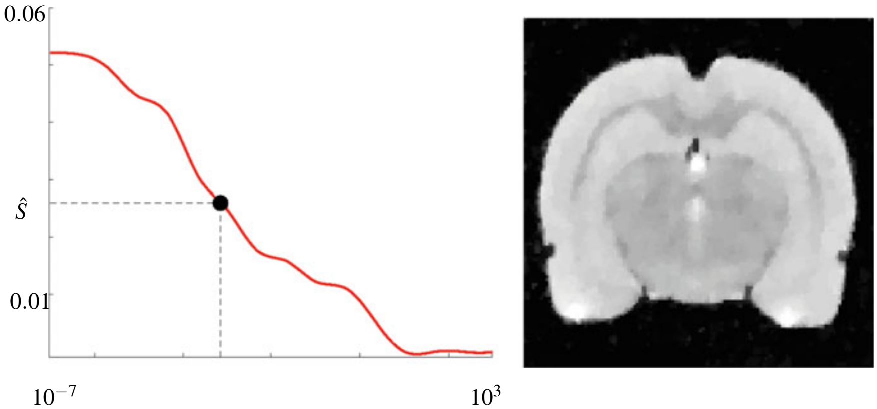
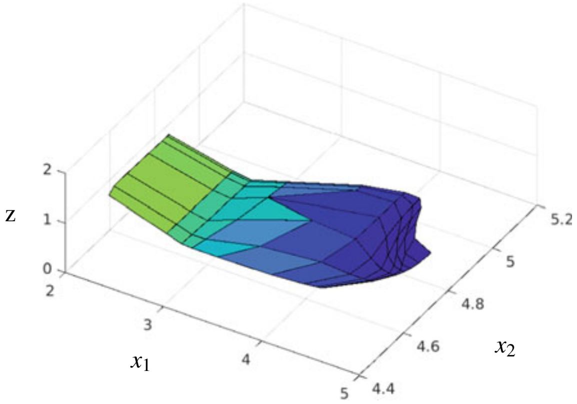


Fig. 2 S-curve for selecting spatial regularization parameter for the simulated data case. Left: plot of the interpolation curve used to determine the value of α such that $S(\alpha) = \hat{S} = 0.0259$. Right: reconstruction (resolution 128×128) of the first time frame ($t = 1$) using the selected value of $\alpha = 7.3e{-}4$

Fig. 3 L-surface for selecting the spatial regularization parameter and the temporal regularization parameter simultaneously, resulting in $\alpha = 3.1623e{-}4$ and $\beta = 3.1623e{-}6$. The axes in the image are: $x_1 = \log(\|f_{t+1} - f_t\|)$, $x_2 = \log(\|\nabla f_t\|_{2,1})$ and $z = \log(\|A_t f_t - m_t\|_2)$



The temporal regularization parameter β was selected by computing reconstructions (5) with different values of β and then selecting the values which had the smallest joint RMSE error, see (10).

This selection resulted in smallest joint RMSE of 0.085 for TS model which corresponded to $\beta = 0.01$, smallest joint RMSE of 0.058 for the TV model corresponding to $\beta = 0.0017$ and smallest joint RMSE of 0.063, corresponding to $\beta = 0.0022$ for the TGV model. As a reference method we selected the regularization parameters α and β using L-surface method for the case where we use spatial total variation and temporal total variation as our regularization model. The resulted L-surface is presented in Fig. 3. Application of L-surface on our data resulted parameters $\alpha = 3.1623e{-}4$ and $\beta = 3.1623e{-}6$.

Figure 4 shows slices of all reconstructions before, during and after contrast agent administration. Figure 5 shows the reconstructed images with different methods for one time frame. The top row of Fig. 5 shows the phantom with a red square denoting a domain that is presented as a closeup in Fig. 6 for all of the reconstructions.

As can be seen from Figs. 4, 5 and 6, the classical regridding method fails on such high time resolution data as employed here, and thus we leave out the regridded reconstruction from the figures of the temporal evolution of the ROI signals. However the L-surface method seems to work nicely on our data, so we include it in the temporal evolution studies. The temporal evolutions in the vascular domain and in the tumor region are averages of Ω_{roi1} and Ω_{roi2}, respectively.

Figure 7 shows the time signals of the reconstructions in the vascular region (Ω_{roi1}) with the different temporal regularization models. Corresponding signals of the reconstructions in the tumor region (i.e. in Ω_{roi2}) are shown in Fig. 8. The tumor region is accurately reconstructed with all the methods, with only small differences between the methods, L-surface method being the noisiest. In the vascular region, the methods have some differences with TGV having the best reconstruction quality and TS having the worst reconstruction quality. The TS method shows smoothing at both the maximum and minimum signal levels whereas TGV reconstructs the fast signal change of the vascular region most reliably. L-surface method is again more

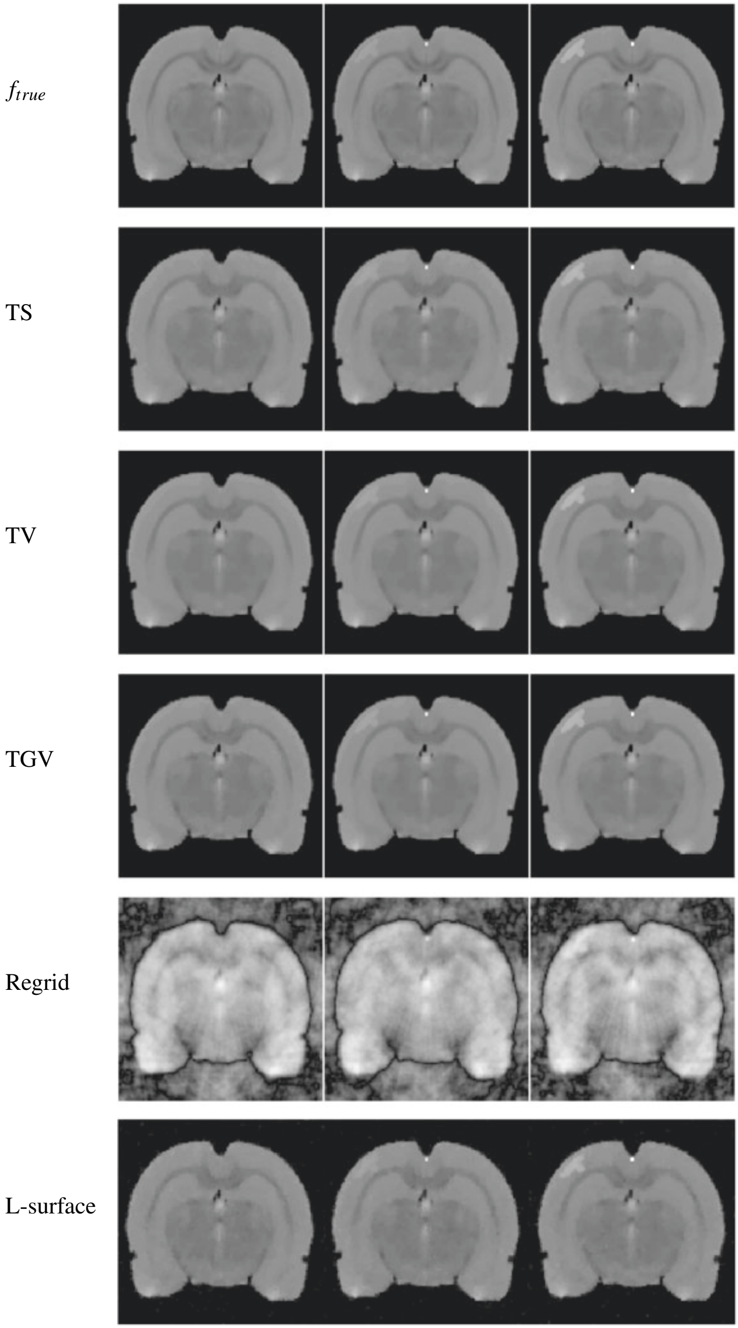


Fig. 4 Reconstructions with different temporal regularizations at different time frames. From top to bottom: true phantom, TS, TV, TGV, regrid and L-surface. From left to right time frames: before, during and after CA administration

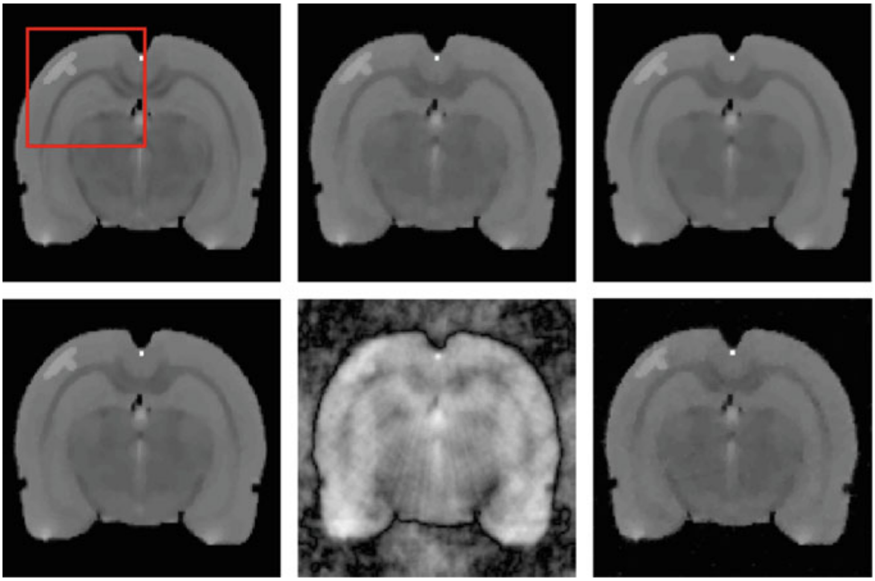


Fig. 5 Reconstructions with different temporal regularizations after CA administration. Top row: true phantom (left), TS reconstruction (middle) and TV reconstruction (right). Bottom row: TGV reconstruction (left), regrid-reconstruction (middle) and L-surface reconstruction (right). The area highlighted in red is presented as closeups in Fig. 6

noisy than the other reconstructions in the vascular region and its performance of reconstructing the vascular signal falls between TS and TV most likely due to the temporal regularization parameter being too small whereas the spatial regularization parameter is on the same order of magnitude as the parameter obtained with S-curve.

5 Conclusions

Variational regularization based solutions for dynamic MRI problems usually include two regularization parameters, one for the spatial and one for the temporal regularization, that the user has to select. Typically, the selection of both of the parameters is carried out manually based on visual assessment of the reconstructed images. In this work we proposed to use the S-curve method for the automatic selection of the spatial regularization parameter, leaving the time regularization parameter the only free parameter. The S-curve method selects the regularization parameter based on the expected sparsity of unknown images in domain of the regularization functional. Furthermore, the method requires computation of the reconstructions with relatively few values of the parameter, making it computationally efficient. The approach was demonstrated to lead to a feasible choice of the spatial regularization parameter in

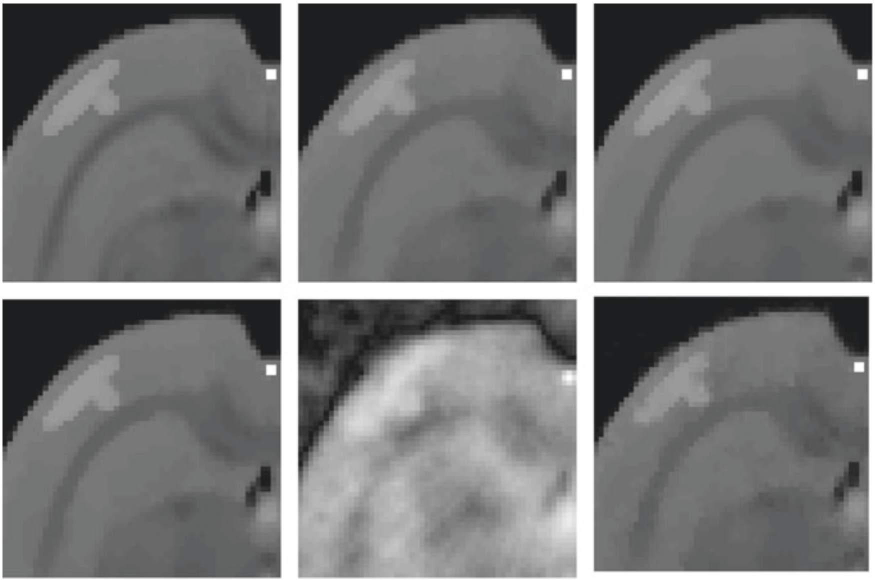


Fig. 6 Closeups of reconstructions with different temporal regularization methods after CA administration. Top row: true phantom (left), TS reconstruction (middle) and TV reconstruction (right). Bottom row: TGV reconstruction (left), regrid-reconstruction (middle) and L-surface reconstruction (right)

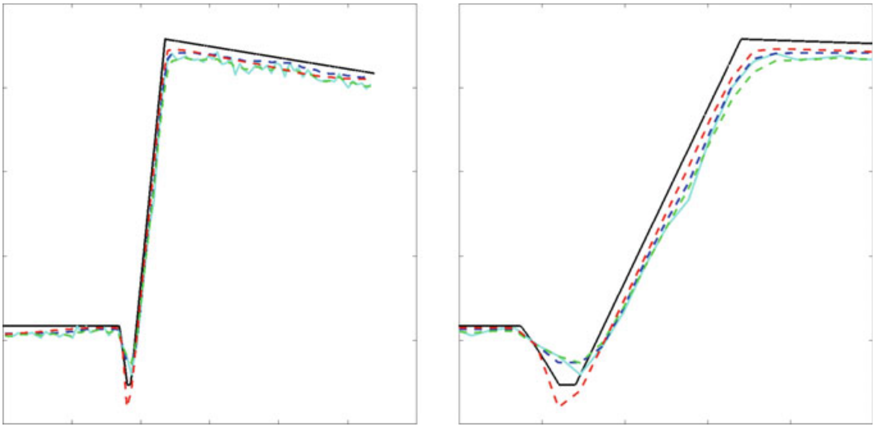


Fig. 7 Reconstructions of vascular region (Ω_{roi1}) with the different methods at their optimal parameters according to the joint RMSE. Black line: true signal, blue line: TV reconstruction, green line: TS reconstruction, red line: TGV reconstruction and light blue line: L-surface reconstruction. Left: temporal evolution during all time frames. Right: closeup image during CA administration

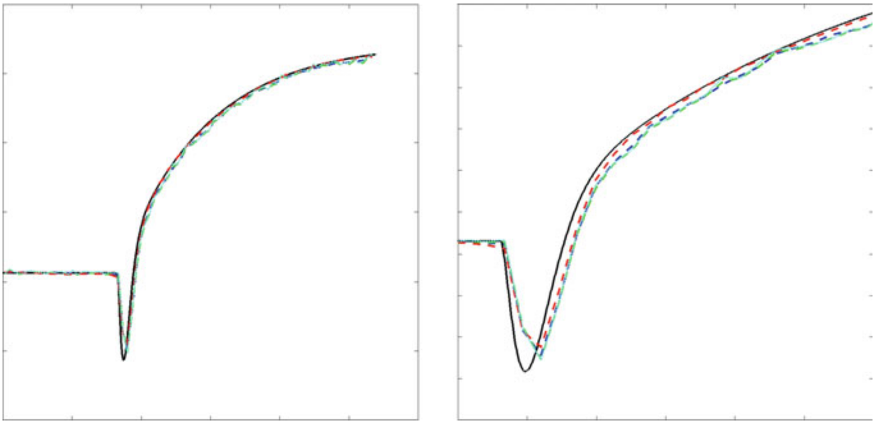


Fig. 8 Reconstructions of tumor region (Ω_{roi2}) with the different methods at their optimal parameters according to the joint RMSE. Black line: true signal, blue line: TV reconstruction, green line: TS reconstruction, red line: TGV reconstruction and light blue line: L-surface reconstruction. Left: temporal evolution during all time frames. Right: closeup image during CA administration

a simulated DCE-MRI experiment of rat brain. The reconstructions were also computed with three different temporal regularization models with the same fixed spatial regularization parameter, demonstrating the robustness of the approach with different time regularization models.

While we proposed automatic selection of the spatial regularization parameter, the temporal regularization parameter was still selected manually. In this work, we selected the temporal regularization parameter by computing RMS errors with respect to a ground truth. If on the other hand the ground truth is unknown, the temporal regularization parameter could be selected based on visual assessment of reconstructed images, leaving β to be the only free parameter. In the future work, we aim to study methods for automatic selection of the time regularization parameter as well. One possibility could be to extend the S-curve to select a parameter which leads to an expected sparsity level in the domain of the temporal regularization. A feasible estimate for the expected level of sparsity in the time direction could potentially be extracted from the changes in the dynamic measurement data.

Acknowledgements This work was supported by Jane and Aatos Erkko foundation and the Academy of Finland, Centre of Excellence in Inverse Modelling and imaging (project 312343).

References

1. Adluru, G., DiBella, E.V.R.: A comparison of L1 and L2 norms as temporal constraints for reconstruction of undersampled dynamic contrast enhanced cardiac scans with respiratory motion. In: Proceedings of International Society for Magnetic Resonance in Medicine, vol. 16, p. 340 (2008)
2. Adluru, G., McGann, C., Speier, P., Kholmovski, E.G., Shaaban, A., Dibella, E.V.R.: Acquisition and reconstruction of undersampled radial data for myocardial perfusion magnetic resonance imaging. J. Magn. Reson. Imaging **29**(2), 466–473 (2009)
3. Adluru, G., Whitaker, R.T., DiBella, E.V.R.: Spatio-temporal constrained reconstruction of sparse dynamic contrast enhanced radial MRI data. In: Proceedings of IEEE International Symposium on Biomedical Imaging, pp. 109–112 (2007)
4. Bredies, K., Kunisch, K., Pock, T.: Total generalized variation. SIAM J. Imaging Sci. **3**(3), 492–526 (2010)
5. Candès, E., Romberg, K.J., Tao, T.: Stable signal recovery from incomplete and inaccurate measurements. Commun. Pure Appl. Math. **59**(8), 1207–1223 (2006)
6. Candès, E.J., Romberg, J., Tao, T.: Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Trans. Inf. Theory **52**, 489–509 (2006)
7. Candès, E.J., Romberg, J.K., Tao, T.: Stable signal recovery from incomplete and inaccurate measurements. Commun. Pure Appl. Math. **59**(8), 1207–1223 (2006)
8. Chambolle, A., Levine, S.E., Lucier, B.J.: An upwind finite-difference method for total variation-based image smoothing. SIAM J. Imaging Sci. **4**(1), 277–299 (2011)
9. Edholm, P.R., Herman, G.T.: Linograms in image reconstruction from projections. IEEE Trans. Med. Imaging **6**(4), 301–307 (1987)
10. Fessler, J.A., Sutton, B.P.: Nonuniform fast fourier transforms using min-max interpolation. IEEE Trans. Signal Process. **51**, 560–574 (2003)
11. Guillot, G., Giovannelli.: Iddn.fr.001.080011.000.s.p.2019.000.31230. International Identifier of Digital Works, 2, 2019. An optional note
12. Hämäläinen, K., Kallonen, A., Kolehmainen, V., Lassas, M., Niinimäki, K., Siltanen, S.: Sparse tomography. SIAM J. Sci. Comput. **35**(3), B644–B665 (2013)
13. Hanhela, M., Kettunen, M., Gröhn, O., Vauhkonen, M., Kolehmainen, V.: Temporal Huber regularization for DCE-MR. J. Math. Imaging Vis. manuscr. (2019)
14. Knoll, F., Bredies, K., Pock, T., Stollberger, R.: Second order total generalized variation (TGV) for MRI. Magn. Reson. Med. **65**(2), 480–491 (2011)
15. Kolehmainen, V., Lassas, M., Niinimäki, K., Siltanen, S.: Sparsity-promoting Bayesian inversion. Inverse Probl. **28**(2), 025005 (2012)
16. Kusmia, S., Eliav, U., Navon, G., Guillot, G.: DQF-MT MRI of connective tissues: application to tendon and muscle. Magn. Reson. Mater. Phys. Biol. Med. **26**, 203–214 (2013)
17. Lavini, C., Verhoeff, J.J.C.: Reproducibility of the gadolinium concentration measurements and of the fitting parameters of the vascular input function in the superior sagittal sinus in a patient population. Magn. Reson. Imaging **28**(10), 1420–1430 (2010)
18. Lustig, M., Donoho, D., Pauly, J.M.: Sparse MRI: the application of compressed sensing for rapid MR imaging. Magn. Reson. Med. **58**, 118–195 (2007)
19. Martincich, L., Montemurro, F., De Rosa, G., Marra, V., Ponzone, R., Cirillo, S., Gatti, M., Biglia, N., Sarotto, I., Sismondi, P., Regge, D., Aglietta, M.: Monitoring response to primary chemotherapy in breast cancer using dynamic contrast-enhanced magnetic resonance imaging. Breast Cancer Res. Treat. **83**(1), 67–76 (2004)
20. Merali, Z., Huang, K., Mikulis, D., Silver, F., Kassner, A.: Evolution of blood-brain-barrier permeability after acute ischemic stroke. PLoS One **12**(2), 1–11 (2017)
21. Mistretta, C.A., Wieben, O., Velikina, J., Block, W., Perry, J., Wu, Y., Johnson, K., Wu, Y.: Highly constrained backprojection for time-resolved MRI. Magn. Reson. Med. **55**(1), 30–40 (2006)
22. Niinimäki, K.: Computational optimization methods for large-scale inverse problems. Ph.D. thesis, University of Eastern Finland (2013)

23. Niinimäki, K., Lassas, M., Hämäläinen, K., Kallonen, A., Kolehmainen, V., Niemi, E., Siltanen, S.: Multi-resolution parameter choice method for total variation regularised tomography (2015). Submitted. http://arxiv.org/abs/1407.2386
24. Pickles, M., Lowry, M., Manton, D., Gibbs, P., Turnbull, L.: Role of dynamic contrast enhanced MRI in monitoring early response of locally advanced breast cancer to neoadjuvant chemotherapy. Breast Cancer Res. Treat. **91**(1), 1–10 (2005)
25. Piludu, F., Marzi, S., Pace, A., Villani, V., Fabi, A., Carapella, C., Terrenato, I., Antenucci, A., Vidiri, A.: Early biomarkers from dynamic contrast-enhanced magnetic resonance imaging to predict the response to antiangiogenic therapy in high-grade gliomas. Neuroradiology **57**(12), 1269–1280 (2015)
26. Port, R.E., Knopp, M.V., Brix, G.: Dynamic contrast-enhanced MRI using Gd-DTPA: interindividual variability of the arterial input function and consequences for the assessment of kinetics in tumors. Magn. Reson. Med. **45**(6), 1030–1038 (2001)
27. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Phys. D Nonlinear Phenom. **60**(1–4), 259–268 (1992)
28. Tofts, P.S., Brix, G., Buckley, D.L., Evelhoch, J.L., Henderson, E., Knopp, M.V., Larsson, H.B.W., Lee, T.-Y., Mayr, N.A., Parker, G.J.M., Port, R.E., Taylor, J., Weisskoff, R.M.: Estimating kinetic parameters from dynamic contrast-enhanced T1-weighted MRI of a diffusable tracer: standardized quantities and symbols. J. Magn. Reson. **10**(3), 223–232 (1999)
29. Valdés-Hernández, P.A., Sumiyoshi, A., Nonaka, H., Haga, R., Aubert-Vásquez, E., Ogawa, T., Iturria-Medina, Y., Riera, J.J., Kawashima, R.: An in vivo MRI template set for morphometry, tissue segmentation, and fMRI localization in rats. Front. Neuroinformatics **5**, 26 (2011)
30. Villringer, K., Sanz Cuesta, B.E., Ostwaldt, A.-C., Grittner, U., Brunecker, P., Khalil, A.A., Schindler, K., Eisenblätter, O., Audebert, H., Fiebach, J.B.: DCE-MRI blood–brain barrier assessment in acute ischemic stroke. Neurology **88**(5), 433–440 (2017)
31. Wang, D., Arlinghaus, L.R., Yankeelov, T.E., Yang, X., Smith, D.S.: Quantitative evaluation of temporal regularizers in compressed sensing dynamic contrast enhanced MRI of the breast. Int. J. Biomed. Imaging 7835749 (2017)
32. Winkelmann, S., Schaeffter, T., Koehler, T., Eggers, H., Doessel, O.: An optimal radial profile order based on the golden ratio for time-resolved MRI. IEEE Trans. Med. Imaging **26**(1), 68–76 (2007)
33. Zong, X., Lee, J., Poplawsky, A.J., Kim, S.-G., Jong, C.Y.: Compressed sensing fMRI using gradient-recalled echo and epi sequences. NeuroImage **92**, 312–321 (2014)

Convergence of Explicit P_1 Finite-Element Solutions to Maxwell's Equations

Larisa Beilina and V. Ruas

Abstract This paper is devoted to the numerical validation of an explicit finite-difference scheme for the integration in time of Maxwell's equations in terms of the sole electric field. The space discretization is performed by the standard P_1 finite element method assorted with the treatment of the time-derivative term by a technique of the mass-lumping type. The rigorous reliability analysis of this numerical model was the subject of authors' another paper [2]. More specifically such a study applies to the particular case where the electric permittivity has a constant value outside a sub-domain, whose closure does not intersect the boundary of the domain where the problem is defined. Our numerical experiments in two-dimension space certify that the convergence results previously derived for this approach are optimal, as long as the underlying CFL condition is satisfied.

Keywords CFL condition · Explicit scheme · Mass-lumping · Maxwell's equations · P_1 finite elements

MSC: 65N12 · 65N15 · 78M10

L. Beilina (✉)
Department of Mathematical Sciences, Chalmers University of Technology
and University of Gothenburg, SE-412 96, Gothenburg, Sweden
e-mail: larisa.beilina@chalmers.se

V. Ruas
Institut Jean Le Rond d'Alembert, UMR 7190 CNRS - Sorbonne Universit é,
75005 Paris, France
e-mail: vitoriano.ruas@upmc.fr

CNPq, Brasilia, Brazil

L. Beilina et al. (eds.), *Mathematical and Numerical Approaches for Multi-Wave Inverse Problems*, Springer Proceedings in Mathematics & Statistics 328,
https://doi.org/10.1007/978-3-030-48634-1_7

1 Introduction

The purpose of this article is to provide a numerical validation of an explicit scheme based on P_1 finite-element space discretizations, to solve hyperbolic Maxwell's equations for the electric field with constant dielectric permittivity in a neighborhood of the boundary of the computational domain. This numerical model was thoroughly studied in [2] from the theoretical point of view.

The standard continuous P_1 FEM is a tempting possibility to solve Maxwell's equations, owing to its simplicity. It is well known however that, for different reasons, this method is not always well suited for this purpose. The first reason is that in general the natural function space for the electric field is not the Sobolev space $\mathbf{H}^1$, but rather the space $\mathbf{H}(curl)$. Another issue difficult to overcome with continuous Lagrange finite elements is the prescription of the zero tangential-component boundary conditions for the electric field, which hold in many important applications. All this motivated the proposal by Nédélec about four decades ago of a family of $\mathbf{H}(curl)$-conforming methods to solve these equations (cf. [23]). These methods are still widely in use, as much as other approaches well adapted to such specific conditions (see e.g. [1, 13, 25]). A comprehensive description of finite element methods for Maxwell's equations can be found in [20].

There are situations however in which the P_1 finite element method does provide an inexpensive and reliable way to solve the Maxwell's equations. In this work we address one of such cases, characterized by the fact that the electric permittivity is constant in a neighborhood of the whole boundary of the domain of interest. This is because, at least in theory, whenever the electric permittivity is constant, the Maxwell's equations simplify into as many wave equations as the space dimension under consideration. More precisely here we show by means of numerical examples that, in such a particular case, a space discretization with conforming linear elements, combined with a straightforward explicit finite-difference scheme of the mass-lumping type for the time integration, gives rise to optimal approximations of the electric field, as long as a classical CFL condition is satisfied.

Actually this work is strongly connected with studies presented in [3, 4] for a combination of a finite difference discretization in a sub-domain with constant permittivity with a finite element discretization in the complementary sub-domain. As pointed out above, the Maxwell's equation reduces to the wave equation in the former case. Since the study of finite-difference methods for this type of equation is well established, only P_1 finite element space discretizations of Maxwell's equations are considered in this paper.

In [3, 4] a stabilized domain-decomposition finite-element/finite-difference approach for the solution of the time-dependent Maxwell's system for the electric field was proposed and numerically verified. In these works [3, 4] different manners to handle the divergence-free condition in a finite-element formulation were considered. The main idea behind the domain decomposition methods in [3, 4] is that a rectangular computational domain is decomposed into two sub-domains, in which two different type of discretizations are employed, namely, the finite-element domain in which a classical P_1 finite element approach is employed, and the finite-difference domain, in which the standard five- or seven-point finite difference scheme is applied, according to the space dimension. The finite-element domain lies strictly inside the finite- difference domain, in such a way that both domains overlap in two layers of structured nodes. First order absorbing boundary conditions [15] are enforced on the boundary of the computational domain, i.e. on the outer boundary of the finite-difference domain. In [3, 4] it was assumed that the dielectric permittivity function is strictly positive and has a constant value in the overlapping nodes as well as in a neighborhood of the boundary of the domain. An explicit time-integration scheme was used both in the finite-element and in the finite-difference domain.

We recall that from the theoretical point of view, for a stable finite-element solution of Maxwell's equation, divergence-free edge elements are the most satisfactory [20, 23]. However the edge elements are less attractive for solving time-dependent problems, since a linear system of equations has to be solved at every time iteration. In contrast, P_1 elements with lumped mass matrix can be efficiently used in connection with an explicit solution scheme [14, 19]. On the other hand it is also well known that the numerical solution of Maxwell's equations with nodal finite elements can result in unstable spurious solutions [21, 24]. Nevertheless a number of techniques are available to remove them, and in this respect we refer for example to [16–18, 22, 24]. In the current work, similar to [3, 4], the spurious solutions are removed from the finite-element solution by adding the divergence-free term to the model equation for the electric field. Numerical tests given in [4] demonstrate that spurious solutions are removable indeed, in case an explicit scheme with P_1 finite elements is employed.

Efficient usage of an explicit scheme combined with P_1 finite-element discretizations for the solution of coefficient inverse problems (CIPs), in the particular context described above was made evident in [5]. In many algorithms aimed at solving electromagnetic CIPs, a qualitative collection of experimental measurements is necessary on the boundary of a computational domain, in order to determine the dielectric permittivity function therein. In this case, in principle the numerical solution of the time-dependent Maxwell's equations is required in the entire space $\mathbb{R}^3$ (see e.g. [5–10], but instead it can be more efficient to consider Maxwell's equations with a constant dielectric permittivity in a neighborhood of the boundary of a computational domain. The explicit scheme with P_1 finite elements considered in this work was numerically tested in the solution of the time-dependent Maxwell's system in both two- and three-dimensional geometry (cf. [4]). It was also combined with a few algorithms to solve different CIPs for determining the dielectric permittivity function in connection with the time-dependent Maxwell's equations, using both simulated and experimentally generated data (see [6–10]). In short, the validation of our formal reliability analysis for such a method conducted in this work, confirms the previously observed adequacy of this numerical approach.

An outline of this paper is as follows: In Sect. 2 we describe the model problem, and give its equivalent variational form. In Sect. 3 we set up the discretizations of the model problem in both space and time, and recall the main results of the reliability analysis conducted in [2] for the underlying numerical model. Section 4 is devoted to the numerical experiments that validate such results. We conclude in Sect. 5 with a few comments.

2 The Model Problem

The particular form of Maxwell's equations for the electric field $\mathbf{e} = (e_1, e_2)$ in a bounded domain Ω of $\Re^2$ with boundary $\partial\Omega$ that we deal with in this work is as follows. First we consider that $\Omega = \bar{\Omega}_{in} \cup \Omega_{out}$, where Ω_{in} is an interior open set whose boundary does not intersect $\partial\Omega$ and Ω_{out} is the complementary set of $\bar{\Omega}_{in}$ with respect to Ω. Now in case $\mathbf{e}$ satisfies (homogeneous) Dirichlet boundary conditions, we are given $\mathbf{e}_0 \in [H^1(\Omega)]^2$ and $\mathbf{e}_1 \in \mathbf{H}(div, \Omega)$ satisfying $\nabla \cdot (\varepsilon \mathbf{e}_0) = \nabla \cdot (\varepsilon \mathbf{e}_1) = 0$ where ε is the electric permittivity. ε is assumed to belong to $W^{2,\infty}(\Omega)$ and to fulfill $\varepsilon \equiv 1$ in Ω_{out} and $\varepsilon \geq 1$ otherwise. Incidentally, throughout this article we denote the standard semi-norm of $C^m(\bar{\Omega})$ by $|\cdot|_{m,\infty}$ for $m > 0$ and the standard norm of $C^0(\bar{\Omega})$ by $\|\cdot\|_{0,\infty}$.

In doing so, the problem to solve is:

$$\begin{array}{ll} \varepsilon \partial_{tt}\mathbf{e} + \nabla \times \nabla \times \mathbf{e} = \mathbf{0} & \text{in } \Omega \times (0, T), \\ \mathbf{e}(\cdot, 0) = \mathbf{e}_0(\cdot), \text{ and } \partial_t \mathbf{e}(\cdot, 0) = \mathbf{e}_1(\cdot) & \text{in } \Omega, \\ \mathbf{e} = \mathbf{0} & \text{on } \partial\Omega \times (0, T), \\ \nabla \cdot (\varepsilon \mathbf{e}) = \mathbf{0} & \text{in } \Omega. \end{array} \tag{1}$$

Remark 1 The analysis carried out in [2] extends in a rather straightforward manner to absorbing conditions $\partial_n \mathbf{e} = -\partial_t \mathbf{e}$ prescribed on the boundary, where $\partial_n \mathbf{e}$ represents the outer normal derivative of $\mathbf{e}$ on $\partial\Omega$. This case is important for it corresponds to practical situations considered in [6–10]. Details on such an extension will be addressed in a forthcoming paper. ■

Next, we set (1) in variational form. With this aim we denote the standard inner product of $[L^2(\Omega)]^2$ by $(\cdot, \cdot)$ and the corresponding norm by $\| \{\cdot\} \|$. Further, for a given non-negative function $\omega \in L^\infty(\Omega)$ we introduce the weighted $L^2(\Omega)$-semi-norm $\|\{\cdot\}\|_\omega := \sqrt{\int_\Omega |\omega||\{\cdot\}|^2 d\mathbf{x}}$, which is actually a norm if $\omega \neq 0$ everywhere in $\bar{\Omega}$. We also introduce, the notation $(\mathbf{a}, \mathbf{b})_\omega := \int_\Omega \omega \mathbf{a} \cdot \mathbf{b} d\mathbf{x}$ for two fields $\mathbf{a}, \mathbf{b}$ which are square integrable in Ω. Notice that if ω is strictly positive this expression defines an inner product associated with the norm $\|\{\cdot\}\|_\omega$.

Then requiring that $\mathbf{e}_{|t=0} = \mathbf{e}_0$ and $\{\partial_t \mathbf{e}\}_{|t=0} = \mathbf{e}_1$ and $\mathbf{e} = 0$ on $\partial\Omega \times [0, T]$, we write for all $\mathbf{v} \in [H_0^1(\Omega)]^2$,

$$(\partial_{tt}\mathbf{e}, \mathbf{v})_\varepsilon + (\nabla \mathbf{e}, \nabla \mathbf{v}) + (\nabla \cdot \varepsilon \mathbf{e}, \nabla \cdot \mathbf{v}) - (\nabla \cdot \mathbf{e}, \nabla \cdot \mathbf{v}) = 0 \; \forall t \in (0, T). \tag{2}$$

3 The Numerical Model

Henceforth we restrict our studies to the case where Ω is a polygonal domain.

3.1 Space Semi-discretization

Let V_h be the usual P_1 FE-space of continuous functions related to a mesh $\mathscr{T}_h$ fitting Ω, consisting of triangles with maximum edge length h, belonging to a quasi-uniform family of meshes (cf. [12]).

Setting $\mathbf{V}_h := [V_h \cap H_0^1(\Omega)]^2$ we define $\mathbf{e}_{0h}$ (resp. $\mathbf{e}_{1h}$) to be the usual $\mathbf{V}_h$-interpolate of $\mathbf{e}_0$ (resp. $\mathbf{e}_1$). Then the semi-discretized problem in space that we wish to solve reads,

Find $\mathbf{e}_h \in \mathbf{V}_h$ *such that* $\forall \mathbf{v} \in \mathbf{V}_h$

$$\begin{aligned} &(\partial_{tt}\mathbf{e}_h, \mathbf{v})_\varepsilon + (\nabla \mathbf{e}_h, \nabla \mathbf{v}) + (\nabla \cdot [\varepsilon \mathbf{e}_h], \nabla \cdot \mathbf{v}) - (\nabla \cdot \mathbf{e}_h, \nabla \cdot \mathbf{v}) = 0, \\ &\mathbf{e}_h(\cdot, 0) = \mathbf{e}_{0h}(\cdot) \text{ and } \partial_t \mathbf{e}_h(\cdot, 0) = \mathbf{e}_{1h}(\cdot) \text{ in } \Omega. \end{aligned} \tag{3}$$

3.2 Full Discretization

To begin with we consider a natural centered time-discretization scheme to solve (3), namely: Given a number N of time steps we define the time increment $\tau := T/N$. Then we approximate $\mathbf{e}_h(k\tau)$ by $\mathbf{e}_h^k \in \mathbf{V}_h$ for $k = 1, 2, \ldots, N$ according to the following scheme for $k = 1, 2, \ldots, N-1$:

$$\begin{aligned} &\left(\frac{\mathbf{e}_h^{k+1} - 2\mathbf{e}_h^k + \mathbf{e}_h^{k-1}}{\tau^2}, \mathbf{v}\right)_\varepsilon + (\nabla \mathbf{e}_h^k, \nabla \mathbf{v}) + (\nabla \cdot \varepsilon \mathbf{e}_h^k, \nabla \cdot \mathbf{v}) - (\nabla \cdot \mathbf{e}_h^k, \nabla \cdot \mathbf{v}) = 0 \;\forall \mathbf{v} \in \mathbf{V}_h, \\ &\mathbf{e}_h^0 = \mathbf{e}_{0h} \text{ and } \mathbf{e}_h^1 = \mathbf{e}_h^0 + \tau \mathbf{e}_{1h} \text{ in } \Omega. \end{aligned} \tag{4}$$

Owing to its coupling with $\mathbf{e}_h^k$ and $\mathbf{e}_h^{k-1}$ on the left hand side of (4), $\mathbf{e}_h^{k+1}$ cannot be determined explicitly by (4) at every time step. In order to enable an explicit solution we resort to the classical mass-lumping technique. We recall that for a constant ε this consists of replacing on the left hand side the inner product $(\mathbf{u}, \mathbf{v})_\varepsilon$ by a discrete inner product $(\mathbf{u}, \mathbf{v})_{\varepsilon.h}$, using the trapezoidal rule to compute the integral of $\int_K \varepsilon \mathbf{u}_{|K} \cdot \mathbf{v}_{|K} d\mathbf{x}$ (resp. $\int_{K \cap \partial\Omega} \mathbf{u}_{|K} \cdot \mathbf{v}_{|K} dS$), for every element K in $\mathscr{T}_h$, where $\mathbf{u}$ stands for $\mathbf{e}_h^{k+1} - 2\mathbf{e}_h^k + \mathbf{e}_h^{k-1}$. It is well-known that in this case the matrix associated with $(\varepsilon \mathbf{e}_h^{k+1}, \mathbf{v})_h$ for $\mathbf{v} \in \mathbf{V}_h$, is a diagonal matrix. In our case ε is not constant, but the same property will hold if we replace in each element K the integral of $\varepsilon \mathbf{u}_{|K} \cdot \mathbf{v}_{|K}$ in a triangle $K \in \mathscr{T}_h$ as follows:

$$\int_K \varepsilon \mathbf{u}_{|K} \cdot \mathbf{v}_{|K} d\mathbf{x} \approx \varepsilon(G_K) area(K) \sum_{i=1}^{3} \frac{\mathbf{u}(S_{K,i}) \cdot \mathbf{v}(S_{K,i})}{3},$$

where $S_{K,i}$ are the vertexes of K, $i = 1, 2, 3$, G_K is the centroid of K.

Before pursuing we define the auxiliary function ε_h whose value in each $K \in \mathscr{T}_h$ is constant equal to $\varepsilon(G_K)$. Then still denoting the approximation of $\mathbf{e}_h(k\tau)$ by $\mathbf{e}_h^k$, for $k = 1, 2, \ldots, N$ we determine $\mathbf{e}_h^{k+1}$ by,

$$\left(\frac{\mathbf{e}_h^{k+1} - 2\mathbf{e}_h^k + \mathbf{e}_h^{k-1}}{\tau^2}, \mathbf{v} \right)_{\varepsilon_h, h} + (\nabla \mathbf{e}_h^k, \nabla \mathbf{v}) + (\nabla \cdot \varepsilon \mathbf{e}_h^k, \nabla \cdot \mathbf{v}) - (\nabla \cdot \mathbf{e}_h^k, \nabla \cdot \mathbf{v}) = 0 \; \forall \mathbf{v} \in \mathbf{V}_h,$$
$$\mathbf{e}_h^0 = \mathbf{e}_{0h} \text{ and } \mathbf{e}_h^1 = \mathbf{e}_h^0 + \tau \mathbf{e}_{1h} \text{ in } \Omega. \tag{5}$$

3.3 Convergence Results

Recalling the assumption that $\varepsilon \in W^{2,\infty}(\Omega)$ we first set

$$\eta := 2 + |\varepsilon|_{1,\infty} + 2|\varepsilon|_{2,\infty}; \tag{6}$$

Next we recall the classical inverse inequality (cf. [12]) together with a result in [11] according to which,

$$\|\nabla v\| \leq Ch^{-1} \|v\|_{\varepsilon_h, h} \text{ for all } v \in V_h, \tag{7}$$

where C is a mesh-independent constant.

Now we assume that τ satisfies the following CFL-condition:

$$\tau \leq h/\nu \text{ with } \nu = C(1 + 3\|\varepsilon - 1\|_\infty)^{1/2}. \tag{8}$$

Further we assume that the solution $\mathbf{e}$ to Eq. (1) belongs to $[H^4\{\Omega \times (0, T)\}]^2$.

Let us define a function $\mathbf{e}_h$ in $\bar{\Omega} \times [0, T]$ whose value at $t = k\tau$ equals $\mathbf{e}_h^k$ for $k = 1, 2, \ldots, N$ and that varies linearly with t in each time interval $([k-1]\tau, k\tau)$, in such a way that $\partial_t \mathbf{e}_h(\mathbf{x}, t) = \dfrac{\mathbf{e}_h^k(\mathbf{x}) - \mathbf{e}_h^{k-1}(\mathbf{x})}{\tau}$ for every $\mathbf{x} \in \bar{\Omega}$ and $t \in ([k-1]\tau, k\tau)$. We also define $\mathbf{a}^{m+1/2}(\cdot)$ for any field $\mathbf{a}(\cdot, t)$ to be $\mathbf{a}(\cdot, [m + 1/2]\tau)$.

Provided the CFL condition (8) is fulfilled and τ also satisfies $\tau \leq 1/[2\eta]$, under the above regularity assumption on $\mathbf{e}$, there exists a constant $\mathscr{C}$ depending only on Ω, ε and T such that,

$$\boxed{\begin{aligned}&\max_{1\le m\le N-1}\left\|[\partial_t(\mathbf{e}_h-\mathbf{e})]^{m+1/2}\right\|+\max_{2\le m\le N}\|\nabla(\mathbf{e}_h^m-\mathbf{e}^m)\|\\ &\le \mathscr{C}(\tau+h+h^2/\tau)\left\{\|\mathbf{e}\|_{H^4[\Omega\times(0,T)]}+|\mathbf{e}_0|_2+|\mathbf{e}_1|_2\right\}.\blacksquare\end{aligned}} \tag{9}$$

(9) means that, as long as τ varies linearly with h, first order convergence of scheme (5) in terms of either τ or h holds in the sense of the norms on the left hand side of (9).

4 Numerical Validation

We perform numerical tests in time $(0, T) = (0, 0.5)$ in the computational domain $\Omega = [0, 1] \times [0, 1]$ for the model problem in two space dimension, namely

$$\begin{aligned}&\varepsilon\partial_{tt}\mathbf{e}-\nabla^2\mathbf{e}-\nabla\nabla\cdot(\varepsilon-1)\mathbf{e}=\mathbf{f} && \text{in } \Omega\times(0,T),\\ &\mathbf{e}(\cdot,0)=\mathbf{0} \text{ and } \partial_t\mathbf{e}(\cdot,0)=\mathbf{0} && \text{in } \Omega,\\ &\mathbf{e}=\mathbf{0} && \text{on } \partial\Omega\times(0,T).\end{aligned} \tag{10}$$

for the electric field $\mathbf{e} = (e_1, e_2)$.

The source data $\mathbf{f}$ (the right hand side) is chosen such that the functions

$$\begin{aligned}e_1&=\frac{1}{\varepsilon}2\pi\sin^2\pi x\cos\pi y\sin\pi y\,\frac{t^2}{2},\\ e_2&=-\frac{1}{\varepsilon}2\pi\sin^2\pi y\cos\pi x\sin\pi x\,\frac{t^2}{2}\end{aligned} \tag{11}$$

are the components of the exact solution to the model problem (10). In (11) the function ε is defined to be,

$$\varepsilon(x,y)=\begin{cases}1+\sin^m\pi(2x-0.5)\cdot\sin^m\pi(2y-0.5) & \text{in } [0.25,0.75]\times[0.25,0.75],\\ 1 & \text{otherwise,}\end{cases} \tag{12}$$

where m is an integer greater than one. In Fig. 1 the function ε is illustrated for different values of m.

The solution given by (11) satisfies homogeneous initial conditions together with homogeneous Dirichlet conditions on the boundary $\partial\Omega$ of the square Ω for every time t. In our computations we used the software package WavES [26] only for the finite element method applied to the solution of the model problem (10). We note that this package was also used in [4] to solve the the same model problem (10) by a domain decomposition FEM/FDM method.

We discretized the computational domain $\Omega \times (0, T)$ denoting by $K_{hl} = \{K\}$ a partition of the spatial domain Ω into triangles K of sizes $h_l = 2^{-l}, l = 1, \ldots, 6$. We let J_{τ_l} be a partition of the time domain $(0, T)$ into time intervals $J = (t_{k-1}, t_k]$

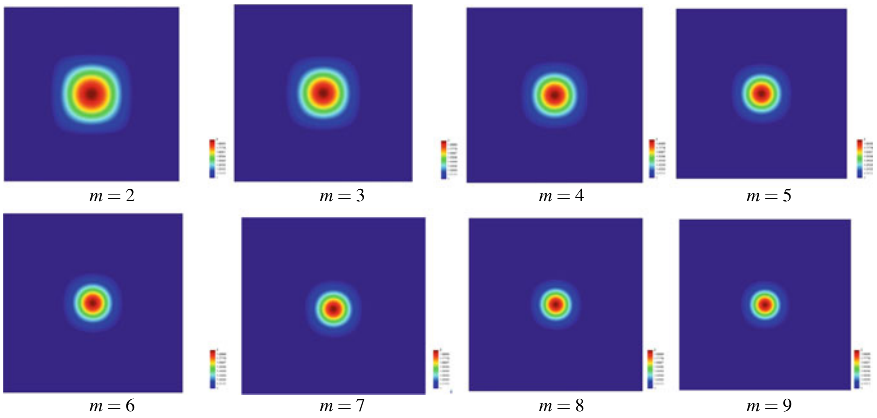


Fig. 1 Function $\varepsilon(x, y)$ in the domain $\Omega = [0, 1] \times [0, 1]$ for different values of m in (12)

of uniform length τ_l for a given number of time intervals N, $l = 1, \ldots, 6$. We choose the time step $\tau_l = 0.025 \times 2^{-l}$, $l = 1, \ldots, 6$, which provides numerical stability for all meshes.

We performed numerical tests taking $m = 2, \ldots, 9$ in (12) and computed the maximum value over the time steps of the relative errors measured in the L_2-norm, the H^1-semi-norm and in the L_2 norm for the time-derivative, respectively:

$$\begin{aligned} e_l^1 &= \frac{\max\limits_{1 \le k \le N} \|\mathbf{e}^k - \mathbf{e}_h^k\|}{\max\limits_{1 \le k \le N} \|\mathbf{e}^k\|}, \\ e_l^2 &= \frac{\max\limits_{1 \le k \le N} \|\nabla(\mathbf{e}^k - \mathbf{e}_h^k)\|}{\max\limits_{1 \le k \le N} \|\nabla \mathbf{e}^k\|}, \\ e_l^3 &= \frac{\max\limits_{1 \le k \le N-1} \|\{\partial_t(\mathbf{e} - \mathbf{e}_h)\}^{k+1/2}\|}{\max\limits_{1 \le k \le N-1} \|\{\partial_t \mathbf{e}\}^{k+1/2}\|}. \end{aligned} \tag{13}$$

Here $\mathbf{e}$ is the exact solution of (10) given by (11) and $\mathbf{e}_h$ is the computed solution, while $N = T/\tau_l$.

In Tables 1 and 2 method's convergence in these three senses is observed for $m = 2, 7$.

Figure 2 shows convergence rates of our numerical scheme based on a P_1 space discretization, taking the function ε defined by (12) with $m = 2$ (on the left) and $m = 7$ (on the right) for $\varepsilon(x)$. Notice that we obtained similar convergence results taking $m = 3, 4, 6, 8, 9$ in (12).

Observation from these tables and figures clearly indicates that our scheme behaves like a first order method in the (semi-)norm of $L^\infty[(0, T); H^1(\Omega)]$ for $\mathbf{e}$

Table 1 Maximum over the time steps of relative errors in the L_2-norm, in the H^1-seminorm and in the L^2-norm of the time derivative for mesh sizes $h_l = 2^{-l}$, $l = 1, \ldots, 6$ taking $m = 2$ in (12)

l	nel	nno	e_l^1	e_{l-1}^1/e_l^1	e_l^2	e_{l-1}^2/e_l^2	e_l^3	e_{l-1}^3/e_l^3
1	8	9	0.054247		0.2767		1.0789	
2	32	25	0.013902	3.902100	0.1216	2.2755	0.4811	2.2426
3	128	81	0.003706	3.751214	0.0532	2.2857	0.2544	1.8911
4	512	289	0.000852	4.349765	0.0234	2.2735	0.1279	1.9891
5	2048	1089	0.000229	3.720524	0.0121	1.9339	0.0641	1.9953
6	8192	4225	0.000059	3.881356	0.0061	1.9836	0.0321	1.9969

Table 2 Maximum over the time steps of relative errors in the L_2-norm, in the H^1-seminorm and in the L^2-norm of the time derivative for mesh sizes $h_l = 2^{-l}$, $l = 1, \ldots, 6$ taking $m = 7$ in (12)

l	nel	nno	e_l^1	e_{l-1}^1/e_l^1	e_l^2	e_{l-1}^2/e_l^2	e_l^3	e_{l-1}^3/e_l^3
1	8	9	0.054224		0.5710		1.1208	
2	32	25	0.012483	4.343828	0.1505	3.7940	0.5024	2.2309
3	128	81	0.002751	4.537623	0.0686	2.1939	0.2688	1.8690
4	512	289	0.000627	4.387559	0.0240	2.8583	0.1339	2.0075
5	2048	1089	0.000158	3.968354	0.0114	2.1053	0.0669	2.0015
6	8192	4225	0.000040	3.949999	0.0057	2	0.0334	2.0030

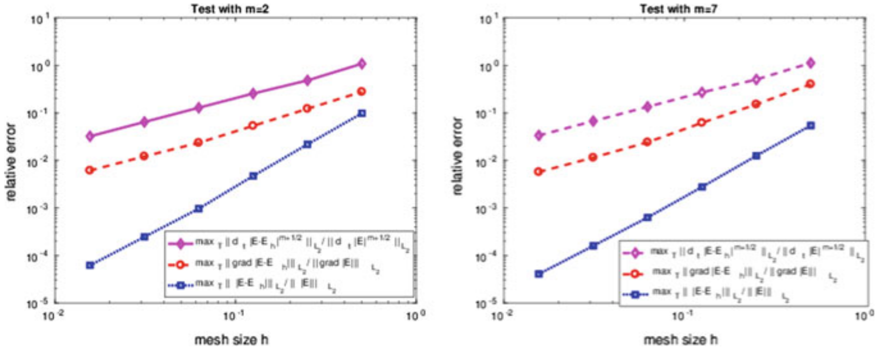
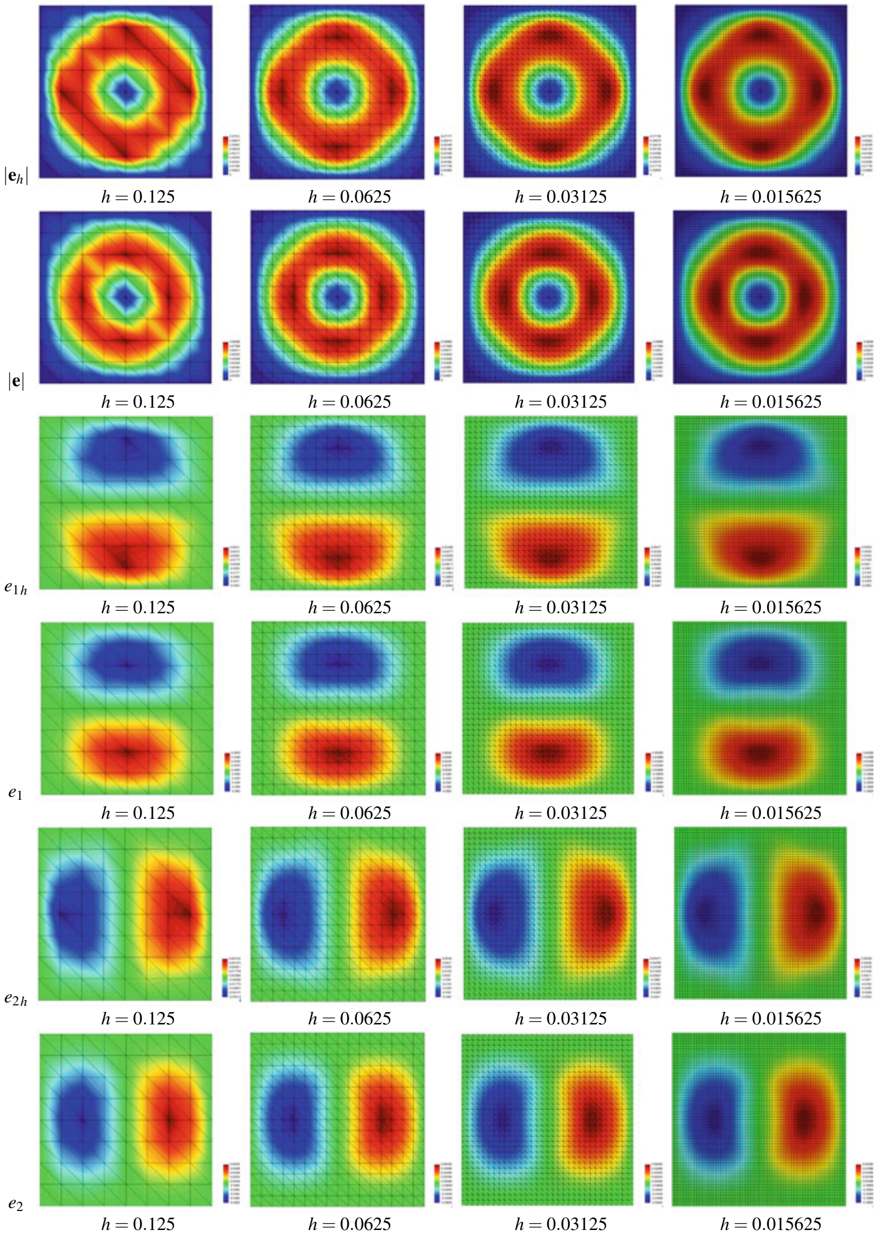


Fig. 2 Maximum in time of relative errors for $m = 2$ (left) and $m = 7$ (right)

and in the norm of $L^\infty[(0, T); L^2(\Omega)]$ for $\partial_t \mathbf{e}$ for the chosen values of m. As far as the value $m = 7$ is concerned this perfectly conforms to the a priori error estimates given in [2] under the assumption that $\mathbf{e} \in \{H^4[\Omega \times (0, T)]\}^2$. On the other hand Table 2 and Fig. 2 also show that the theoretical predictions of [2] extend to the cases not considered therein such as $m = 2$, in which the regularity of the exact solution is lower than assumed. Otherwise stating some of our assumptions seem to be of academic interest only and a lower regularity of the solution such as $H^2[\Omega \times (0, T)]$ should be sufficient to attain optimal first order convergence in both senses.



$h = 0.125$ $h = 0.0625$ $h = 0.03125$ $h = 0.015625$

$h = 0.125$ $h = 0.0625$ $h = 0.03125$ $h = 0.015625$

$h = 0.125$ $h = 0.0625$ $h = 0.03125$ $h = 0.015625$

$h = 0.125$ $h = 0.0625$ $h = 0.03125$ $h = 0.015625$

$h = 0.125$ $h = 0.0625$ $h = 0.03125$ $h = 0.015625$

$h = 0.125$ $h = 0.0625$ $h = 0.03125$ $h = 0.015625$

Fig. 3 Computed versus exact solution at $t = 0.25$ for different meshes taking $m = 2$ in (12)

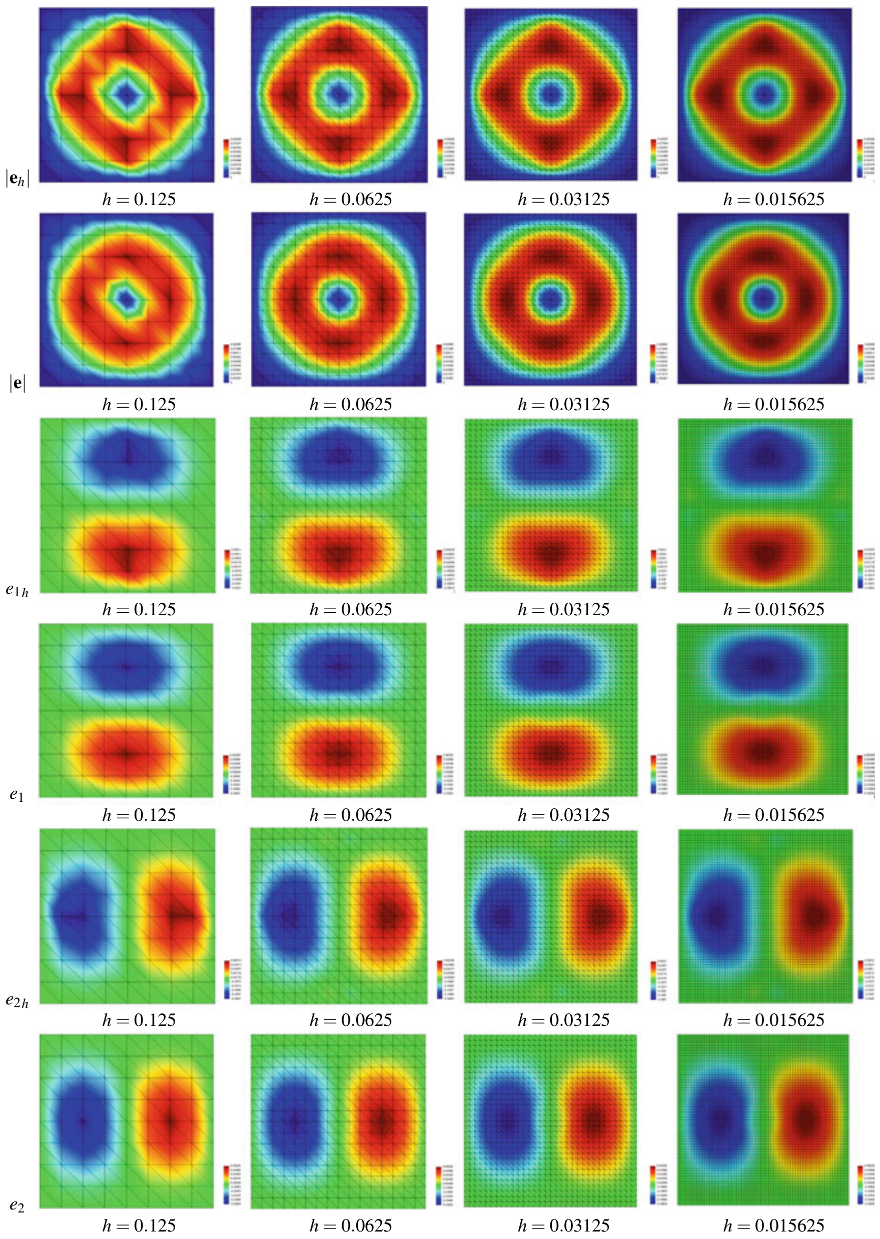


Fig. 4 Computed versus exact solution at $t = 0.25$ for different meshes taking $m = 7$ in (12)

Finally we observe that second-order convergence can be expected from our scheme in the norm $L^{\infty}[(0,T);L^2(\Omega)]$ for $\mathbf{e}$, according to Tables 1, 2 and Fig. 2.

The above representations of the numerical results are enriched by Figs. 3 and 4, in which a graphic comparison between the exact solution and the approximate solutions at time $t = 0.25$ generated by our scheme with different meshes and corresponding time steps are supplied.

5 Conclusions

In this work we validated the reliability analysis conducted in [2] for a numerical scheme to solve Maxwell's equations of electromagnetism, combining an explicit finite difference time discretization with a lumped-mass P_1 finite element space discretization. The scheme is effective in the particular case where the dielectric permittivity is constant in a neighborhood of the boundary of the spatial domain. After presenting the problem under consideration for the electric field we supplied the detailed description of such a scheme and recalled the a priori error estimates that hold for the latter under suitable regularity assumptions specified in [2]. Then we showed by means of numerical experiments performed for a test-problem in two-dimension space with known exact solution, that the convergence results given in [2] are confirmed in practice. Furthermore we presented convincing evidence that such theoretical predictions extend to solutions with much lower regularity than the one assumed in our analysis. Similar optimal second-order convergence is observed in a norm other than those in which convergence was formally established. In short we undoubtedly indicated that Maxwell's equations can be efficiently solved with classical conforming linear finite elements in some relevant particular cases, among those is the model problem stated in (10).

Acknowledgements The research of the first author is supported by the Swedish Research Council grant VR 2018-03661. The second author gratefully acknowledges the financial support provided by CNPq/Brazil through grant 307996/2008-5.

References

1. Assous, F., Degond, P., Heintze, E., Raviart, P.: On a finite-element method for solving the three-dimensional Maxwell equations. J. Comput. Phys. **109**, 222–237 (1993)
2. Beilina, L., Ruas, V.: An explicit P_1 finite element scheme for Maxwell's equations with constant permittivity in a boundary neighborhood. arXiv:1808.10720
3. Beilina, L., Grote, M.: Adaptive hybrid finite element/difference method for Maxwell's equations. TWMS J. Pure Appl. Math. **1**(2), 176–197 (2010)

4. Beilina, L.: Energy estimates and numerical verification of the stabilized domain decomposition finite element/finite difference approach for time-dependent Maxwell's system. Cent. Eur. J. Math. **11**(4), 702–733 (2013). https://doi.org/10.2478/s11533-013-0202-3
5. Beilina, L., Klibanov, M.V.: Approximate Global Convergence and Adaptivity for Coefficient Inverse Problems. Springer, New York (2012)
6. Beilina, L., Cristofol, M., Niinimaki, K.: Optimization approach for the simultaneous reconstruction of the dielectric permittivity and magnetic permeability functions from limited observations. Inverse Probl. Imaging **9**(1), 1–25 (2015)
7. Beilina, L., Thanh, N.T., Klibanov, M.V., Malmberg, J.B.: Globally convergent and adaptive finite element methods in imaging of buried objects from experimental backscattering radar measurements. J. Comp. Appl. Maths. Elsevier (2015). https://doi.org/10.1016/j.cam.2014.11.055
8. Bondestam-Malmberg, J., Beilina, L.: An adaptive finite element method in quantitative reconstruction of small inclusions from limited observations. Appl. Math. Inf. Sci. **12–1**, 1–19 (2018)
9. Bondestam-Malmberg, J., Beilina, L.: Iterative regularization and adaptivity for an electromagnetic coefficient inverse problem. In: AIP Conference Proceedings, vol. 1863, p. 370002 (2017). https://doi.org/10.1063/1.4992549
10. Bondestam-Malmberg, J.: Efficient Adaptive Algorithms for an Electromagnetic Coefficient Inverse Problem, Doctoral thesis. University of Gothenburg, Sweden (2017)
11. Carneiro de Araujo, J.H., Gomes, P.D., Ruas, V.: Study of a finite element method for the time-dependent generalized Stokes system associated with viscoelastic flow. J. Comput. Appl. Math. **234–8**, 2562–2577 (2010)
12. Ciarlet, P.G.: The Finite Element Method for Elliptic Problems. North Holland (1978)
13. Ciarlet Jr., P., Zou, J.: Fully discrete finite element approaches for time-dependent Maxwell's equations. Numerische Mathematik **82**(2), 193–219 (1999)
14. Elmkies, A., Joly, P.: Finite elements and mass lumping for Maxwell's equations: the 2D case. *Numerical Analysis*, C. R. Acad. Sci. Paris, **324**, 1287–1293 (1997)
15. Engquist, B., Majda, A.: Absorbing boundary conditions for the numerical simulation of waves. Math. Comp. **31**, 629–651 (1977)
16. Jiang, B.: The Least-Squares Finite Element Method. Theory and Applications in Computational Fluid Dynamics and Electromagnetics. Springer, Heidelberg (1998)
17. Jiang, B., Wu, J., Povinelli, L.A.: The origin of spurious solutions in computational electromagnetics. J. Comput. Phys. **125**, 104–123 (1996)
18. Jin, J.: The finite element method in electromagnetics. Wiley (1993)
19. Joly, P.: Variational methods for time-dependent wave propagation problems. In: Lecture Notes in Computational Science and Engineering. Springer, Berlin (2003)
20. Monk, P.: Finite Element Methods for Maxwell's Equations. Clarendon Press (2003)
21. Monk, P.B., Parrott, A.K.: A dispersion analysis of finite element methods for Maxwell's equations. SIAM J. Sci. Comput. **15**, 916–937 (1994)
22. Munz, C.D., Omnes, P., Schneider, R., Sonnendrucker, E., Voss, U.: Divergence correction techniques for Maxwell Solvers based on a hyperbolic model. J. Comput. Phys. **161**, 484–511 (2000)
23. Nédélec, J.-C.: Mixed finite elements in R^3. Numerische Mathematik **35**, 315–341 (1980)
24. Paulsen, K.D., Lynch, D.R.: Elimination of vector parasites in finite element Maxwell solutions. IEEE Trans. Microw. Theory Technol. **39**, 395–404 (1991)
25. Ruas, V., Ramos, M.A.S.: A hermite method for Maxwell's equations. Appl. Math. Inf. Sci. **12**(2), 271–283 (2018)
26. WavES, the software package. http://www.waves24.com

Reconstructing the Optical Parameters of a Layered Medium with Optical Coherence Elastography

Peter Elbau, Leonidas Mindrinos, and Leopold Veselka

Abstract In this work we consider the inverse problem of reconstructing the optical properties of a layered medium from an elastography measurement where optical coherence tomography is used as the imaging method. We hereby model the sample as a linear dielectric medium so that the imaging parameter is given by its electric susceptibility, which is a frequency- and depth-dependent parameter. Additionally to the layered structure (assumed to be valid at least in the small illuminated region), we allow for small scatterers which we consider to be randomly distributed, a situation which seems more realistic compared to purely homogeneous layers. We then show that a unique reconstruction of the susceptibility of the medium (after averaging over the small scatterers) can be achieved from optical coherence tomography measurements for different compression states of the medium.

Keywords Optical coherence tomography · Optical coherence elastography · Inverse problem · Parameter identification

MSC: 65J22 · 65M32 · 78A46

1 Introduction

Optical Coherence Tomography is an imaging modality producing high resolution images of biological tissues. It measures the magnitude of the back-scattered light of a focused laser illumination from a sample as a function of depth and provides

P. Elbau (✉) · L. Mindrinos · L. Veselka
Faculty of Mathematics, University of Vienna, Oskar-Morgenstern-Platz 1,
1090 Vienna, Austria
e-mail: peter.elbau@univie.ac.at

L. Mindrinos
e-mail: leonidas.mindrinos@univie.ac.at

L. Veselka
e-mail: leopold.veselka@univie.ac.at

L. Beilina et al. (eds.), *Mathematical and Numerical Approaches for Multi-Wave Inverse Problems*, Springer Proceedings in Mathematics & Statistics 328,
https://doi.org/10.1007/978-3-030-48634-1_8

cross-sectional or volumetric data by performing a series of multiple axial scans at different positions. Initially, it used to operate in time where a movable mirror was giving the depth information. Later on, frequency-domain optical coherence tomography was introduced where the detector is replaced by a spectrometer and no mechanical movement is needed. We refer to [3, 4] for an overview of the physics of the experiment and to [6] for a mathematical description of the problem.

Only lately, the inverse problems arising in optical coherence tomography have attracted the interest from the mathematical community, see, for example, [2, 7, 11, 13]. For many years, the proposed and commonly used reconstruction method was just the inverse Fourier transform. This approach is valid only if the properties of the medium are assumed to be frequency-independent in the spectrum of the light source. However, the less assumptions one takes, the more mathematically interesting but also difficult the problem becomes.

The main assumption, we want to make is that the medium can be (at least locally in the region where the laser beam illuminates the object) well described by a layered structure. Since there are in real measurement images typically multiple small particles visible inside these layers, we will additionally include small, randomly distributed scatterers into the model and calculate the averaged contribution of these particles to the measured fields.

To obtain a reconstruction of the medium, that is, of its electric susceptibility, we consider an elastography setup where optical coherence tomography is used as the imaging system. This so-called optical coherence elastography is done by recording optical coherence tomography data for different compression states of the medium, see [1, 5, 9, 12] for some recent works dealing with this interesting problem.

Under the assumption that the sample can be described as a linear elastic medium, we show that these measurements can be used to achieve a unique reconstruction of the electric susceptibility of the layered medium.

The paper is organised as follows: In Sect. 2 we review the main equations describing mathematically how the data in optical coherence tomography is collected and its relation to the optical properties of the medium. In Sect. 3, we show that the calculation of the back-scattered field can be decomposed into the corresponding subproblems for the single layers, for which we derive the resulting formulæ in Sect. 4. Finally, we present in Sect. 5 that from the measurements at different compression states a unique reconstruction of the susceptibility becomes feasible.

2 Modelling the Optical Coherence Tomography Measurement

We model the sample by a dispersive, isotropic, non-magnetic, linear dielectric medium characterised by its scalar electric susceptibility. To include randomly distributed scatterers in the model, we introduce the susceptibility as a random variable; so let $(\mathscr{X}, \mathscr{A}, P)$ be a probability space and write

$$\chi : \mathscr{X} \times \mathbb{R} \times \mathbb{R}^3 \to \mathbb{R},\ (\sigma, t, x) \mapsto \chi_\sigma(t, x)$$

for the electric susceptibility of the medium in the state σ. (Hereby, to have a causal model, we require an electric susceptibility $\chi : \mathbb{R} \times \mathbb{R}^3 \to \mathbb{R}$ to be a function fulfilling $\chi(t, x) = 0$ for all $t < 0$.)

The object (in a certain realisation state $\sigma \in \mathscr{X}$) is then probed with a laser beam, described by an incident electric field $E^{(0)} : \mathbb{R} \times \mathbb{R}^3 \to \mathbb{R}^3$ characterised by the following properties.

Definition 1 We call $E^{(0)} : \mathbb{R} \times \mathbb{R}^3 \to \mathbb{R}^3$ an incident wave for a given susceptibility $\chi : \mathbb{R} \times \mathbb{R}^3 \to \mathbb{R}$ in a homogeneous background $\chi_0 : \mathbb{R} \to \mathbb{R}$ if it is a solution of Maxwell's equations for χ_0, that is,

$$\varDelta E^{(0)}(t, x) = \frac{1}{c^2} \partial_{tt} D^{(0)}(t, x),$$

where c denotes the speed of light in vacuum and

$$D^{(0)}(t, x) = E^{(0)}(t, x) + \int_{\mathbb{R}} \chi_0(\tau) E^{(0)}(t - \tau, x) \mathrm{d}\tau,$$

and $E^{(0)}$ does not interact with the inhomogeneity for negative times, meaning that

$$E^{(0)}(t, x) = 0 \text{ for all } t \in (-\infty, 0),\ x \in \varOmega \tag{1}$$

with $\varOmega = \{x \in \mathbb{R}^3 \mid \chi(\cdot, x) \neq \chi_0\}$.

We then measure the resulting electric field $E_\sigma : \mathbb{R} \times \mathbb{R}^3 \to \mathbb{R}^3$ induced by the incident field $E^{(0)}$ in the presence of the dielectric medium described by the susceptibility χ_σ.

Definition 2 Let $\chi : \mathbb{R} \times \mathbb{R}^3 \to \mathbb{R}$ be a susceptibility and $E^{(0)} : \mathbb{R} \times \mathbb{R}^3 \to \mathbb{R}^3$ be an incident wave for χ. Then, we call E the electric field induced by $E^{(0)}$ in the presence of χ if E is a solution of the equation system

$$\operatorname{curl} \operatorname{curl} E(t, x) + \frac{1}{c^2} \partial_{tt} D(t, x) = 0 \text{ for all } t \in \mathbb{R},\ x \in \mathbb{R}^3, \tag{2}$$

$$E(t, x) - E^{(0)}(t, x) = 0 \text{ for all } t \in (-\infty, 0),\ x \in \mathbb{R}^3 \tag{3}$$

with c being the speed of light in vacuum and with the electric displacement field $D : \mathbb{R} \times \mathbb{R}^3 \to \mathbb{R}^3$ being related to the electric field via

$$D(t, x) = E(t, x) + \int_{\mathbb{R}} \chi(\tau, x) E(t - \tau, x) \mathrm{d}\tau.$$

Remark 1 The fact that $E^{(0)}$ does not interact with the object before time $t = 0$, see (1), guarantees that $E^{(0)}$ is a solution of (2) and thus the initial condition in (3) is compatible with (2).

Remark 2 We do not want to specify the solution concept for solving (2) here (since we are going for a layered and therefore discontinuous susceptibility, there exists only a weak solution), but will silently assume that the susceptibility and the incident field are such that they induce an electric field with sufficient regularity and the appearing integrals and Fourier transforms are well-defined.

Equation (2) is more conveniently written in Fourier space, where we use the convention

$$\mathscr{F}[f](k) = \frac{1}{(2\pi)^{\frac{n}{2}}} \int_{\mathbb{R}^n} f(x) \mathrm{e}^{-\mathrm{i}\langle k,x\rangle} \mathrm{d}x$$

for the Fourier transform of an integrable function $f : \mathbb{R}^n \to \mathbb{R}$. For convenience, we also use the shorter notation

$$\check{F}(\omega, x) = \sqrt{2\pi}\, \mathscr{F}^{-1}[t \mapsto F(t, x)](\omega) = \int_{\mathbb{R}} F(t, x) \mathrm{e}^{\mathrm{i}\omega t} \mathrm{d}t$$

for this rescaled inverse Fourier transformation of a sufficiently regular function of the form $F : \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}^n$ with respect to the time variable.

Lemma 1 *Let $\chi : \mathbb{R} \times \mathbb{R}^3 \to \mathbb{R}$ be a susceptibility, $E^{(0)} : \mathbb{R} \times \mathbb{R}^3 \to \mathbb{R}^3$ be an incident wave for χ, and E be the induced electric field. Then, $\check{E}$ solves the vector Helmholtz equation*

$$\operatorname{curl}\operatorname{curl} \check{E}(\omega, x) - \frac{\omega^2}{c^2}(1 + \check{\chi}(\omega, x))\check{E}(\omega, x) = 0 \textit{ for all } \omega \in \mathbb{R},\ x \in \mathbb{R}^3, \tag{4}$$

with the constraint

$$\check{E} \in \mathscr{H}(\check{E}^{(0)}), \tag{5}$$

where $\mathscr{H}(\check{E}^{(0)})$ is the space of all functions $F : \mathbb{R} \times \mathbb{R}^3 \to \mathbb{R}^3$ so that the map $\omega \mapsto (F - \check{E}^{(0)})(\omega, x)$ can be holomorphically extended to the space $\mathbb{H} \times \mathbb{R}^3$, where $\mathbb{H} = \{z \in \mathbb{C} \mid \Im\mathrm{m}\, z > 0\}$ denotes the upper half complex plane, and the extension fulfils

$$\sup_{\lambda>0} \int_{\mathbb{R}} |(F - \check{E}^{(0)})(\omega + \mathrm{i}\lambda, x)|^2 \mathrm{d}\omega < \infty$$

for every $x \in \mathbb{R}^3$.

Proof Equation (4) is obtained directly from the application of the Fourier transform to (2). The condition (5) is according to the Paley–Wiener theorem, see, for example, [10, Theorem 9.2], equivalent to the condition (3), which states that $t \mapsto (E - E^{(0)})(t, x)$ has for every $x \in \mathbb{R}^3$ only support in $[0, \infty)$.

In frequency-domain optical coherence tomography, we detect with a spectrometer at a position $x_0 \in \mathbb{R}^3$ outside the medium the intensity of the Fourier components of the superposition of the back-scattered light from the sample and the reference beam, which is the reflection of the incident laser beam from a mirror at some fixed position.

Here, we consider two independent measurements for two different positions of the mirror in order to overcome the problem of phase-less data, see [8]. Thus, we record for some realisation $\sigma \in \mathscr{X}$ and all $\omega \in \mathbb{R}$ the data

$$m_{0,\sigma}(\omega) = |\check{E}_\sigma(\omega, x_0)| \text{ and } m_{i,\sigma}(\omega) = |\check{E}_\sigma(\omega, x_0) + \check{E}_i^{(\mathrm{r})}(\omega, x_0)|,\ i \in \{1, 2\},$$

where $E_1^{(\mathrm{r})} : \mathbb{R} \times \mathbb{R}^3 \to \mathbb{R}^3$ and $E_2^{(\mathrm{r})} : \mathbb{R} \times \mathbb{R}^3 \to \mathbb{R}^3$ denote the two known reference waves, which are solutions of Maxwell's equations in the homogeneous background medium (usually well approximated by the vacuum).

We can uniquely recover from this data the (complex-valued) Fourier transform $\check{E}(\omega, x_0)$ of the electric field for every $\omega \in \mathbb{R}$ by intersecting the three circles

$$\partial B_{m_{0,\sigma}(\omega)}(0) \cap \partial B_{m_{1,\sigma}(\omega)}(-\check{E}_1^{(\mathrm{r})}(\omega, x_0)) \cap \partial B_{m_{2,\sigma}(\omega)}(-\check{E}_2^{(\mathrm{r})}(\omega, x_0))$$

provided that the points 0, $\check{E}_1^{(\mathrm{r})}(\omega, x_0)$, and $\check{E}_2^{(\mathrm{r})}(\omega, x_0)$ in the complex plane do not lie on a single straight line. In the following, we assume that the fields $E_1^{(\mathrm{r})}$ and $E_2^{(\mathrm{r})}$ are chosen such that this condition is satisfied and we can recover the function

$$m_\sigma(\omega) = \check{E}_\sigma(\omega, x_0) \text{ for all } \omega \in \mathbb{R}.$$

However, this information is still not enough for reconstructing the material parameter χ_σ, see, for example, [6]. Thus, we make the a priori assumption that the illuminated region of the medium can be well approximated by a layered medium. Since the layers are typically not completely homogeneous, we also allow for randomly distributed small inclusions in every layer.

Thus, we describe χ to be of the form

$$\chi_\sigma(t, x) = \chi_j(t) + \psi_{j,\sigma_j}(t, x) \tag{6}$$

in the jth layer $\{x \in \mathbb{R}^3 \mid z_{j+1} < x_3 < z_j\}$, $j \in \{1, \dots, J\}$, where we write the measure space as a product $\mathscr{X} = \prod_{j=1}^J \mathscr{X}_j$ with each factor representing the state of one layer. Here, χ_j is the homogeneous background susceptibility of the layer and ψ_j is the random contribution caused by some small particles in the layer. Outside these layers, we set $\chi_\sigma(t, x) = \chi_0(t)$ for some homogeneous background susceptibility χ_0.

To simplify the analysis, we will assume that the scatterers in the jth layer only occur at some distance to the layer boundaries z_j and z_{j+1}, say between Z_j and ζ_j, where $z_{j+1} < Z_j < \zeta_j < z_j$. Moreover, we choose the particles independently, identically, uniformly distributed on the part $U_{j,L_j} = [-\frac{1}{2}L_j, \frac{1}{2}L_j] \times [-\frac{1}{2}L_j, \frac{1}{2}L_j] \times$

$[Z_j, \zeta_j]$ of the layer for some width $L_j > 0$. Concretely, we assume that we have in the jth layer for some number N_j of particles the probability measure P_{j,N_j,L_j} on the probability space $\mathscr{X}_j = (U_{j,L_j})^{N_j}$ given by

$$P_{j,N_j,L_j}(\prod_{\ell=1}^{N_j} A_\ell) = \prod_{\ell=1}^{N_j} \frac{|A_\ell|}{L_j^2(\zeta_j - Z_j)} \tag{7}$$

for all measurable subsets $A_\ell \subset U_{j,L_j}$, where $|A_\ell|$ denotes the three dimensional Lebesgue measure of the set A_ℓ.

The full probability measure $P = P_{N,L}$ is consistently chosen as the direct product $P_{N,L} = \prod_{j=1}^{J} P_{j,N_j,L_j}$ on $\mathscr{X} = \prod_{j=1}^{J} \mathscr{X}_j$.

The particles themselves, we model in each layer as identical balls with a sufficiently small radius R and a homogeneous susceptibility $\chi_j^{(\mathrm{p})}$. Thus, we define for a realisation $\sigma_j \in \mathscr{X}_j$ of the jth layer the contribution of the particles to the susceptibility by

$$\psi_{j,\sigma_j}(t,x) = \sum_{\ell=1}^{N_j} \boldsymbol{\chi}_{B_R(\sigma_{j,\ell})}(x)\,(\chi_j^{(\mathrm{p})}(t) - \chi_j(t)), \tag{8}$$

where we ignore the problem of overlapping particles. Here, we denote by $\boldsymbol{\chi}_A$ the characteristic function of a set A and by $B_r(y)$ the open ball with radius r around a point y.

3 Domain Decomposition of the Solution

The layered structure of the medium allows us to decompose the solution as a series of solution operators for the single layers. To do so, we split the medium at a horizontal stripe where the medium is homogeneous and consider the two subproblems where once the region above and once the region below is replaced by the homogeneous susceptibility $X_0 : \mathbb{R} \to \mathbb{R}$ in the stripe. We write the stripe as the set $\{x \in \mathbb{R}^3 \mid z - \varepsilon < x_3 < z + \varepsilon\}$ for some $z \in \mathbb{R}$ and some height $\varepsilon > 0$ and parametrise the electric susceptibility in the form

$$\chi(t,x) = \begin{cases} X_1(t,x) & \text{if } x \in \Omega_1 = \{y \in \mathbb{R}^3 \mid y_3 > z - \varepsilon\}, \\ X_2(t,x) & \text{if } x \in \Omega_2 = \{y \in \mathbb{R}^3 \mid y_3 < z + \varepsilon\}. \end{cases} \tag{9}$$

with the necessary compatibility condition that X_1 and X_2 coincide in the intersection $\Omega_1 \cap \Omega_2$, where they should both be equal to the homogeneous susceptibility X_0.

Additionally, we have the assumption that the medium is bounded in vertical direction. We can therefore assume that for some $z_- < z_+$, the susceptibilities X_1 and X_2 are homogeneous in $\Omega_+ = \{x \in \mathbb{R}^3 \mid x_3 > z_+\} \subset \Omega_1$ and $\Omega_- = \{x \in \mathbb{R}^3 \mid x_3 < z_-\} \subset \Omega_2$, respectively. We set

$$X_1(t,x) = X_+(t) \text{ for all } x \in \Omega_+ \text{ and } X_2(t,x) = X_-(t) \text{ for all } x \in \Omega_-.$$

Since we are solving Maxwell's equations on the whole space, we extend X_1 and X_2 by the homogeneous susceptibility X_0:

$$X_1(t,x) = X_0(t) \text{ for all } x \in \Omega_2 \text{ and } X_2(t,x) = X_0(t) \text{ for all } x \in \Omega_1,$$

see Picture (a) in Fig. 1 for an illustration of the notation.

The aim is then to reduce the calculation of the electric field in the presence of χ to the subproblems of determining the electric fields in the presence of X_1 and X_2, independently. To do so, we consider the solution in the intersection $\Omega_1 \cap \Omega_2$ and split it there into waves moving in the positive and negative e_3 direction.

Lemma 2 *Let a homogeneous susceptibility $\chi : \mathbb{R} \to \mathbb{R}$ be given on a stripe $\Omega_0 = \{x \in \mathbb{R}^3 \mid x_3 \in (z_0 - \varepsilon, z_0 + \varepsilon)\}$. Then, every solution $\check{E} : \mathbb{R} \times \Omega_0 \to \mathbb{C}^3$ of*

$$\operatorname{curl}\operatorname{curl} \check{E}(\omega,x) - \frac{\omega^2}{c^2}(1 + \check{\chi}(\omega))\check{E}(\omega,x) = 0 \textit{ for all } \omega \in \mathbb{R},\ x \in \Omega_0, \tag{10}$$

admits the form

$$\begin{aligned}\check{E}(\omega,x) = &\int_{\mathbb{R}^2} e_1(k_1,k_2) \mathrm{e}^{-\mathrm{i}x_3\sqrt{\frac{\omega^2}{c^2}(1+\check{\chi}(\omega))-k_1^2-k_2^2}} \mathrm{e}^{\mathrm{i}(k_1x_1+k_2x_2)} \mathrm{d}(k_1,k_2) \\ &+ \int_{\mathbb{R}^2} e_2(k_1,k_2) \mathrm{e}^{\mathrm{i}x_3\sqrt{\frac{\omega^2}{c^2}(1+\check{\chi}(\omega))-k_1^2-k_2^2}} \mathrm{e}^{\mathrm{i}(k_1x_1+k_2x_2)} \mathrm{d}(k_1,k_2)\end{aligned} \tag{11}$$

for all $\omega \in \mathbb{R}$ and $x \in \Omega_0$ with some coefficients $e_1,\ e_2 : \mathbb{R}^2 \to \mathbb{C}^3$.

Proof Taking the divergence of (10), we see that we have $\operatorname{div} \check{E} = 0$ on the stripe Ω_0 with homogeneous susceptibility. Then, Eq. (10) reduces to the three independent Helmholtz equations

$$\Delta \check{E}(\omega,x) + \frac{\omega^2}{c^2}(1 + \check{\chi}(\omega))\check{E}(\omega,x) = 0 \text{ for all } \omega \in \mathbb{R},\ x \in \Omega_0.$$

Applying the Fourier transform with respect to x_1 and x_2 and solving the resulting ordinary differential equation in x_3 gives us (11).

Definition 3 Let $\check{E}$ be a solution of the Eq. (10) on some stripe Ω_0, written in the form (11). We then call $\check{E}$ a downwards moving solution if $e_2 = 0$ and an upwards moving solution if $e_1 = 0$.

Moreover, we define the solution operators $\mathscr{G}_1$ and $\mathscr{G}_2$. To avoid having to define an incident wave on the whole space, we replace the condition (5) by radiation conditions of the form that we specify the upwards moving part on a stripe below the region and the downwards moving part on a stripe above the region.

Definition 4 Let χ be given as in (9) and $\check{E}_0$ be an upwards moving solution in $\Omega_1 \cap \Omega_2$. Then, we define $\mathscr{G}_1 \check{E}_0$ as a solution $\check{E}$ of the equation

$$\operatorname{curl}\operatorname{curl} \check{E}(\omega, x) - \frac{\omega^2}{c^2}(1 + \check{X}_1(\omega, x))\check{E}(\omega, x) = 0$$

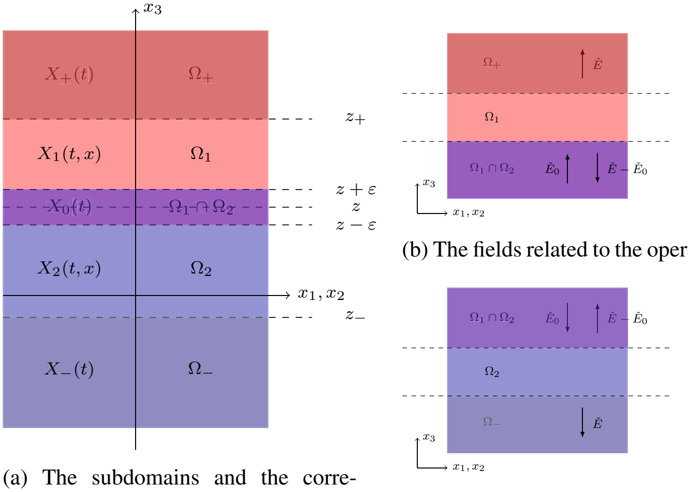
fulfilling the radiation condition that $\check{E} - \check{E}_0$ is a downwards moving solution in $\Omega_1 \cap \Omega_2$ and that $\check{E}$ is an upwards moving solution in Ω_+, see Picture (b) in Fig. 1.

Analogously, we define $\mathscr{G}_2 \check{E}_0$ for a downwards moving solution $\check{E}_0$ in $\Omega_1 \cap \Omega_2$ as a solution $\check{E}$ of the equation

$$\operatorname{curl}\operatorname{curl} \check{E}(\omega, x) - \frac{\omega^2}{c^2}(1 + \check{X}_2(\omega, x))\check{E}(\omega, x) = 0$$

fulfilling the radiation condition that $\check{E} - \check{E}_0$ is an upwards moving solution in $\Omega_1 \cap \Omega_2$ and that $\check{E}$ is a downwards moving solution in Ω_-, see Picture (c) in Fig. 1.

Remark 3 We do not consider the uniqueness of these solutions at this point since we will only need the result for particular, simplified problems where the verification that this gives the desired solution can be done directly.



(a) The subdomains and the corresponding optical parameters

(b) The fields related to the operator $\mathscr{G}_1$

(c) The fields related to the operator $\mathscr{G}_2$

Fig. 1 The geometry and the notation used in this section

Instead we will simply assume that the susceptibilities χ, X_1, and X_2 are such that the only solution $\check{E}$ in the presence of this susceptibility for which $\check{E}$ is upwards moving on Ω_+ and downwards moving on Ω_- is the trivial solution $\check{E} = 0$, meaning that there is only the trivial solution in the absence of an incident wave.

Lemma 3 *Let χ be given by (9) and denote by $\mathscr{G}_1$, $\mathscr{G}_2$ the solution operators as in Definition 4. Let further $E^{(0)}$ be an incident wave on χ which is moving downwards and E_1 be the induced electric fields in the presence of X_1.*

Then, provided the following series converge, we have that the function E defined by

$$\check{E}(\omega,x) = \begin{cases} \check{E}_1(\omega,x) + \sum\limits_{j=0}^{\infty} \mathscr{G}_1(\tilde{\mathscr{G}}_2\tilde{\mathscr{G}}_1)^j\tilde{\mathscr{G}}_2\check{E}_1(\omega,x) & \text{if } x\in\Omega_1, \\ \sum\limits_{j=0}^{\infty} \mathscr{G}_2(\tilde{\mathscr{G}}_1\tilde{\mathscr{G}}_2)^j\check{E}_1(\omega,x) & \text{if } x\in\Omega_2, \end{cases}$$

where we set $\tilde{\mathscr{G}}_i = \mathscr{G}_i - \mathrm{id}$, $i \in \{1,2\}$, is an electric field in the presence of χ fulfilling the radiation conditions that $\check{E} - \check{E}^{(0)}$ is an upwards moving wave in Ω_+ and $\check{E}$ is a downwards moving wave in Ω_-.

Proof First, we remark that the composition of the operators is well defined, since $\check{E}_1 \in \mathscr{H}(\check{E}^{(0)})$ is a downwards moving solution in $\Omega_1\cap\Omega_2$, see Lemma 1, the range of $\tilde{\mathscr{G}}_2$ consists of upwards moving solutions, and the range of $\tilde{\mathscr{G}}_1$ consists of downwards moving solutions.

The field $\check{E}$ is seen to satisfy (4) in Ω_1 by using the definitions of E_1 and the solution operator $\mathscr{G}_1$ on Ω_1. Similarly, using the definition of $\mathscr{G}_2$, we get that the function $\check{E}$ satisfies (4) in Ω_2.

Therefore, it only remains to check that the two formulas coincide in the intersection $\Omega_1\cap\Omega_2$. Using that $\mathscr{G}_i = \tilde{\mathscr{G}}_i + \mathrm{id}$, $i\in\{1,2\}$, we find that

$$\begin{aligned} \check{E}_1 + \sum_{j=0}^{\infty}\mathscr{G}_1(\tilde{\mathscr{G}}_2\tilde{\mathscr{G}}_1)^j\tilde{\mathscr{G}}_2\check{E}_1 &= \check{E}_1 + \sum_{j=0}^{\infty}\tilde{\mathscr{G}}_1(\tilde{\mathscr{G}}_2\tilde{\mathscr{G}}_1)^j\tilde{\mathscr{G}}_2\check{E}_1 + \sum_{j=0}^{\infty}(\tilde{\mathscr{G}}_2\tilde{\mathscr{G}}_1)^j\tilde{\mathscr{G}}_2\check{E}_1 \\ &= \sum_{j=0}^{\infty}(\tilde{\mathscr{G}}_1\tilde{\mathscr{G}}_2)^j\check{E}_1 + \sum_{j=0}^{\infty}\tilde{\mathscr{G}}_2(\tilde{\mathscr{G}}_1\tilde{\mathscr{G}}_2)^j\check{E}_1 = \sum_{j=0}^{\infty}\mathscr{G}_2(\tilde{\mathscr{G}}_1\tilde{\mathscr{G}}_2)^j\check{E}_1. \end{aligned}$$

Moreover, we have that $\check{E} - \check{E}_1$ is by construction an upwards moving wave in Ω_+, and therefore so is $\check{E} - \check{E}^{(0)}$. Similarly, the wave $\check{E}$ is a downwards moving wave in Ω_-.

If we are in a case where our uniqueness assumption mentioned in Remark 3 holds, then Lemma 3 allows us to iteratively reduce the problem of determining the electric field in the presence of the susceptibility χ_σ, defined in (6), to problems of simpler susceptibilities. To this end, we could, for example, successively apply the result to values $z \in (\zeta_j, z_j)$ and $z \in (z_{j+1}, Z_j)$, $j = 1, \dots, J$, where each successive step is only used to further simplify the operator $\mathscr{G}_2$ from the previous step. This then leads to a sort of layer stripping algorithm, see, for example, [8], where a similar argument was presented.

4 Wave Propagation Through a Scattering Layer

Using the above analysis, we can calculate the electric field in the presence of a layered medium of the form (6) as a combination of the solutions of the following two subproblems.

Problem 1 Let $j \in \{0, \dots, J-1\}$. Find the electric field induced by some incident field in the presence of the piecewise homogeneous susceptibility χ given by

$$\chi(t, x) = \begin{cases} \chi_j(t) & \text{if } x_3 > z_{j+1}, \\ \chi_{j+1}(t) & \text{if } x_3 < z_{j+1}. \end{cases} \tag{12}$$

Problem 2 Let $\sigma \in \mathscr{X}$ and $j \in \{1, \dots, J\}$. Find the electric field induced by some incident field in the presence of the susceptibility χ given by

$$\chi(t, x) = \chi_j(t) + \psi_{j,\sigma}(t, x), \tag{13}$$

where the function ψ_j is described by (8).

We thus fix a layer $j \in \{0, \dots, J\}$, and to simplify the calculations, we restrict ourselves in both subproblems to an illumination by a downwards moving plane wave of the form

$$\check{E}^{(0)}(\omega, x) = \check{f}(\omega) \mathrm{e}^{-\mathrm{i}\frac{\omega}{c} n_j(\omega) x_3} \eta \tag{14}$$

for some function $f : \mathbb{R} \to \mathbb{R}$ and a polarisation vector $\eta \in \mathbb{S}^1 \times \{0\}$. Here we define the complex-valued refractive indices for all $j \in \{0, \dots, J\}$ by

$$n_j : \mathbb{R} \to \mathbb{H},\ n_j(\omega) = \sqrt{1 + \check{\chi}_j(\omega)}. \tag{15}$$

Then, the solution of Problem 1 can be explicitly written down.

Lemma 4 *Let $j \in \{0, \dots, J-1\}$ and $E^{(0)}$ be the incident wave given in (14). Then, the electric field E induced by $E^{(0)}$ in the presence of a susceptibility χ of the form (12) is given by*

$$\check{E}(\omega, x) = \check{f}(\omega) \left(\mathrm{e}^{-\mathrm{i}\frac{\omega}{c} n_j(\omega) x_3} - \frac{n_{j+1}(\omega) - n_j(\omega)}{n_{j+1}(\omega) + n_j(\omega)} \mathrm{e}^{-\mathrm{i}\frac{\omega}{c} n_j(\omega) z_{j+1}} \mathrm{e}^{\mathrm{i}\frac{\omega}{c} n_j(\omega)(x_3 - z_{j+1})} \right) \eta$$

for $x_3 > z_{j+1}$*, and by*

$$\check{E}(\omega,x) = \check{f}(\omega)\frac{2n_j(\omega)}{n_{j+1}(\omega)+n_j(\omega)}\mathrm{e}^{-\mathrm{i}\frac{\omega}{c}n_j(\omega)z_{j+1}}\mathrm{e}^{-\mathrm{i}\frac{\omega}{c}n_{j+1}(\omega)(x_3-z_{j+1})}\eta$$

for $x_3 < z_{j+1}$*, where the refractive indices* n_j *and* n_{j+1} *are defined by* (15)*.*

Proof Clearly, $\check{E}$ satisfies the differential equation (4) in both regions $x_3 > z_{j+1}$ and $x_3 < z_{j+1}$. Moreover, $\check{E}^{(0)}$ is the only incoming wave in $\check{E}$. Therefore, it only remains to check that $\check{E}$ has sufficient regularity to be the weak solution along the discontinuity of the susceptibility at $x_3 = z_{j+1}$, meaning that

$$\lim_{x_3\uparrow z_{j+1}}\check{E}(\omega,x) = \lim_{x_3\downarrow z_{j+1}}\check{E}(\omega,x),$$
$$\lim_{x_3\uparrow z_{j+1}}n_{j+1}(\omega)\partial_{x_3}\check{E}(\omega,x) = \lim_{x_3\downarrow z_{j+1}}n_j(\omega)\partial_{x_3}\check{E}(\omega,x).$$

Both identities are readily verified.

For Problem 2, the situation is more complicated and we settle for an approximate solution for the electric field. For that, we assume (using the same notation as in (8)) that the susceptibility $\chi_j^{(\mathrm{p})}$ of the random particles does not differ much from the background χ_j, so that the difference between the induced field and the incident field becomes small, and we do a first order approximation in the difference $\chi_j^{(\mathrm{p})} - \chi_j$. For that purpose, we write the differential Eq. (4) in the form

$$\operatorname{curl}\operatorname{curl}\check{E}(\omega,x) - \frac{\omega^2}{c^2}n_j^2(\omega)(1+\bar{\phi}_{j,\sigma_j}(\omega,x))\check{E}(\omega,x) = 0,$$

where, according to (8),

$$\bar{\phi}_{j,\sigma_j}(\omega,x) = \sum_{\ell=1}^{N_j}\boldsymbol{\chi}_{B_R(\sigma_{j,\ell})}(x)\,\phi_j(\omega),$$

and we abbreviate

$$\phi_j(\omega) = \frac{\check{\chi}_j^{(\mathrm{p})}(\omega)-\check{\chi}_j(\omega)}{1+\check{\chi}_j(\omega)}. \tag{16}$$

In first order in $\bar{\phi}$, we then approximate the field by the solution $\check{E}^{(1)}_{N_j,\sigma_j}$ of the equation

$$\operatorname{curl}\operatorname{curl}\check{E}^{(1)}(\omega,x) - \frac{\omega^2}{c^2}n_j^2(\omega)\check{E}^{(1)}(\omega,x) = \frac{\omega^2}{c^2}n_j^2(\omega)\bar{\phi}_{j,\sigma_j}(\omega,x)\check{E}^{(0)}(\omega,x),$$

the so called Born approximation. Using that the fundamental solution G of the Helmholtz equation, which by definition fulfils

$$\Delta G(\kappa, x) + \kappa^2 G(\kappa, x) = -\delta(x),$$

is given by

$$G(\kappa, x) = \frac{\mathrm{e}^{\mathrm{i}\kappa|x|}}{4\pi|x|},$$

we obtain the expression

$$\begin{aligned} E^{(1)}_{N_j,\sigma_j}(\omega, x) &= E^{(0)}(\omega, x) \\ &+ \left(\frac{\omega^2}{c^2}n_j^2(\omega) + \operatorname{grad}\operatorname{div}\right)\sum_{\ell=1}^{N_j}\int_{B_R(\sigma_{j,\ell})} G(\tfrac{\omega}{c}n_j(\omega), x-y)\phi_j(\omega)E^{(0)}(\omega, y)\mathrm{d}y \end{aligned} \tag{17}$$

for the Born approximation of the induced field, see, for example, [6, Proposition 4].

We now want to determine the expected value of $E^{(1)}_{N_j,\sigma_j}$ in the limit where the number of particles N_j and the width L_j of the region where the particles are horizontally distributed tend to infinity, while keeping the ratio $\rho_j = \frac{N_j}{L_j^2}$ of particles per surface area constant, that is, we want to calculate the expression

$$\bar{E}^{(1)}(\omega, x) = \lim_{N_j\to\infty}\int_{\mathscr{X}_j} \check{E}^{(1)}_{N_j,\sigma_j}(\omega, x)\mathrm{d}P_{j,N_j,L_j(N_j)}(\sigma_j), \tag{18}$$

where $L_j(N_j) = \sqrt{\frac{N_j}{\rho_j}}$ and P denotes the probability measure introduced in (7).

Lemma 5 *Let $j \in \{1, \dots, J\}$ and $\rho_j > 0$ be fixed, $E^{(0)}$ be an incident field of the form (14), and χ be the susceptibility specified in (13).*

Then, the expected value $\bar{E}^{(1)}$ of the Born approximation of the field induced by $E^{(0)}$ in the presence of the susceptibility χ in the limit $N_j \to \infty$ with $L_j^2\rho_j = N_j$, as introduced in (18), is given by

$$\begin{aligned} \bar{E}^{(1)}(\omega, x) &= \check{E}^{(0)}(\omega, x) + (2\pi)^4\rho_j\phi_j(\omega)\check{f}(\omega) \\ &\times h(2R\tfrac{\omega}{c}n_j(\omega))\left(\mathrm{e}^{-\mathrm{i}\frac{\omega}{c}n_j(\omega)Z_j} - \mathrm{e}^{-\mathrm{i}\frac{\omega}{c}n_j(\omega)\zeta_j}\right)\mathrm{e}^{\mathrm{i}\frac{\omega}{c}n_j(\omega)(x_3-\mu_j)}\eta \end{aligned} \tag{19}$$

for $x_3 > \zeta_j + R$ and by

$$\begin{aligned} \bar{E}^{(1)}(\omega, x) &= \check{E}^{(0)}(\omega, x) + \frac{(2\pi)^4}{3}\rho_j\phi_j(\omega)\check{f}(\omega) \\ &\times \left(\mathrm{e}^{-\mathrm{i}\frac{\omega}{c}n_j(\omega)Z_j} - \mathrm{e}^{-\mathrm{i}\frac{\omega}{c}n_j(\omega)\zeta_j}\right)\mathrm{e}^{-\mathrm{i}\frac{\omega}{c}n_j(\omega)(x_3-\mu_j)}\eta \end{aligned} \tag{20}$$

for $x_3 < Z_j - R$*, where* $\mu_j = \frac{1}{2}(\zeta_j + Z_j)$ *and*

$$h(\xi) = \frac{\sin(\xi) - \xi\cos(\xi)}{\xi^3}. \tag{21}$$

Proof Inserting the expression (17) for the Born approximation of the electric field into the formula (18) for the expected value, we obtain the equation

$$\begin{aligned}\bar{E}^{(1)}(\omega, x) &= \check{E}^{(0)}(\omega, x)\\ &\quad + \lim_{N_j\to\infty} N_j\phi_j(\omega)\check{f}(\omega)\left(\frac{\omega^2}{c^2}n_j^2(\omega) + \operatorname{grad}\operatorname{div}\right)K_{L_j(N_j)}(\omega, x)\eta,\end{aligned} \tag{22}$$

where

$$K_L(\omega, x) = \int_{U_{j,L}}\int_{B_R(\sigma_{j,1})} G(\tfrac{\omega}{c}n_j(\omega), x - y)\mathrm{e}^{-\mathrm{i}\frac{\omega}{c}n_j(\omega)x_3}\mathrm{d}y\,\mathrm{d}\sigma_{j,1}.$$

We recall that $U_{j,L} = [-\frac{1}{2}L, \frac{1}{2}L] \times [-\frac{1}{2}L, \frac{1}{2}L] \times [Z_j, \zeta_j]$ is for $L = L_j$ the region in which the particles in the jth layer are lying. To symmetrise the expression, we set

$$\mu_j = \frac{1}{2}(\zeta_j + Z_j) \text{ and } d_j = \frac{1}{2}(\zeta_j - Z_j)$$

and shift $U_{j,L}$ to the origin, by defining $\tilde{U}_{j,L} = U_{j,L} - \mu_j e_3$ with $e_3 = (0, 0, 1)$.

Introducing the probability density

$$p_L(\xi) = \frac{1}{|U_{j,L}|}\boldsymbol{\chi}_{U_{j,L}}(\mu_j e_3 + \xi) = \frac{1}{2L^2 d_j}\boldsymbol{\chi}_{\tilde{U}_{j,L}}(\xi)$$

for the variable $\xi = \sigma_{j,1} - \mu_j e_3$, we rewrite K_L in the form

$$\begin{aligned}K_L(\omega, x) &= \int_{\mathbb{R}^3} p_L(\xi)\mathrm{e}^{-\mathrm{i}\frac{\omega}{c}n_j(\omega)(\mu_j+\xi_3)}\\ &\qquad\times\int_{\mathbb{R}^3}\boldsymbol{\chi}_{B_R(0)}(y)G(\tfrac{\omega}{c}n_j(\omega), x - \mu_j e_3 - \xi - y)\mathrm{e}^{-\mathrm{i}\frac{\omega}{c}n_j(\omega)y_3}\mathrm{d}y\,\mathrm{d}\xi\\ &= (2\pi)^{\frac{3}{2}}\int_{\mathbb{R}^3} p_L(\xi)\mathrm{e}^{-\mathrm{i}\frac{\omega}{c}n_j(\omega)(\mu_j+\xi_3)}\\ &\qquad\times\mathscr{F}[y \mapsto \boldsymbol{\chi}_{B_R(0)}(y)G(\tfrac{\omega}{c}n_j(\omega), x - \mu_j e_3 - \xi - y)](\tfrac{\omega}{c}n_j(\omega)e_3)\mathrm{d}\xi\\ &= (2\pi)^3\int_{\mathbb{R}^3} p_L(\xi)\mathrm{e}^{-\mathrm{i}\frac{\omega}{c}n_j(\omega)(\mu_j+\xi_3)}\\ &\qquad\times\big(\mathscr{F}[\boldsymbol{\chi}_{B_R(0)}] * \mathscr{F}[y \mapsto G(\tfrac{\omega}{c}n_j(\omega), x - \mu_j e_3 - \xi - y)]\big)(\tfrac{\omega}{c}n_j(\omega)e_3)\mathrm{d}\xi.\end{aligned}$$

The shift in the function G now translates in Fourier space to a multiplication by a phase factor; explicitly, we find (using the symmetry $G(\kappa, y) = G(\kappa, -y)$ and the notation $\hat{G}(\kappa, k) = \mathscr{F}[y \mapsto G(\kappa, y)](k)$) that

$$\mathscr{F}[y \mapsto G(\kappa, x - \mu_j e_3 - \xi - y)](k) = \mathrm{e}^{-\mathrm{i}\langle k, x-\mu_j e_3 - \xi\rangle} \hat{G}(\kappa, k).$$

Therefore, we can write K_L (again using the shorter notation $\hat{\boldsymbol{\chi}}_{B_R(0)} = \mathscr{F}[\boldsymbol{\chi}_{B_R(0)}]$ and $\hat{p}_L = \mathscr{F}[p_L]$) as

$$\begin{aligned} K_L(\omega, x) &= (2\pi)^3 \mathrm{e}^{-\mathrm{i}\frac{\omega}{c} n_j(\omega)\mu_j} \int_{\mathbb{R}^3} \hat{\boldsymbol{\chi}}_{B_R(0)}(\tfrac{\omega}{c} n_j(\omega) e_3 - k) \hat{G}(\tfrac{\omega}{c} n_j(\omega), k) \\ &\qquad \times \mathrm{e}^{-\mathrm{i}\langle k, x - \mu_j e_3\rangle} \int_{\mathbb{R}^3} p_L(\xi) \mathrm{e}^{-\mathrm{i}\langle \frac{\omega}{c} n_j(\omega) e_3 - k), \xi\rangle} \mathrm{d}\xi \,\mathrm{d}k \\ &= (2\pi)^{\frac{9}{2}} \mathrm{e}^{-\mathrm{i}\frac{\omega}{c} n_j(\omega)\mu_j} \int_{\mathbb{R}^3} \hat{\boldsymbol{\chi}}_{B_R(0)}(\tfrac{\omega}{c} n_j(\omega) e_3 - k) \\ &\qquad \times \hat{p}_L(\tfrac{\omega}{c} n_j(\omega) e_3 - k) \hat{G}(\tfrac{\omega}{c} n_j(\omega), k) \mathrm{e}^{-\mathrm{i}\langle k, x - \mu_j e_3\rangle} \mathrm{d}k. \end{aligned} \tag{23}$$

Remarking that the Fourier transform of p_L is explicitly given by

$$\begin{aligned} \hat{p}_L(k) &= \frac{1}{(2\pi)^{\frac{3}{2}} L^2} \int_{-\frac{L}{2}}^{\frac{L}{2}} \mathrm{e}^{-\mathrm{i}k_1\xi_1} \mathrm{d}\xi_1 \int_{-\frac{L}{2}}^{\frac{L}{2}} \mathrm{e}^{-\mathrm{i}k_2\xi_2} \mathrm{d}\xi_2 \int_{\mathbb{R}} \boldsymbol{\chi}_{[-d_j, d_j]}(\xi_3) \mathrm{e}^{-\mathrm{i}k_3\xi_3} \mathrm{d}\xi_3 \\ &= \frac{1}{(2\pi)^{\frac{3}{2}} L^2} \frac{2\sin(\frac{1}{2} L k_1)}{k_1} \frac{2\sin(\frac{1}{2} L k_2)}{k_2} \int_{\mathbb{R}} \boldsymbol{\chi}_{[-d_j, d_j]}(\xi_3) \mathrm{e}^{-\mathrm{i}k_3\xi_3} \mathrm{d}\xi_3, \end{aligned}$$

we find with the abbreviation $\hat{\boldsymbol{\chi}}_{[-d_j, d_j]} = \mathscr{F}[\boldsymbol{\chi}_{[-d_j, d_j]}]$ in the limit $N_j \to \infty$ that

$$N_j \hat{p}_{L_j(N_j)}(k) \to 2\pi \rho_j \delta(k_1) \delta(k_2) \hat{\boldsymbol{\chi}}_{[-d_j, d_j]}(k_3) \;(N_j \to \infty). \tag{24}$$

Using (23) in (24), we can calculate the behaviour of K_L in this limit to be

$$\begin{aligned} \lim_{N_j \to \infty} N_j K_{L_j(N_j)}(\omega, x) &= (2\pi)^{\frac{11}{2}} \mathrm{e}^{-\mathrm{i}\frac{\omega}{c} n_j(\omega)\mu_j} \rho_j \\ \times \int_{\mathbb{R}} \hat{\boldsymbol{\chi}}_{B_R(0)}((\tfrac{\omega}{c} n_j(\omega) - k_3) e_3) &\hat{G}(\tfrac{\omega}{c} n_j(\omega), k_3 e_3) \hat{\boldsymbol{\chi}}_{[-d_j, d_j]}(k_3) \mathrm{e}^{-\mathrm{i}k_3(x_3 - \mu_j)} \mathrm{d}k_3. \end{aligned}$$

Using further that $\hat{G}$ can be computed by taking the Fourier transform of the Helmholtz equation, giving us

$$\hat{G}(\kappa, k) = \frac{1}{(2\pi)^{\frac{3}{2}}} \frac{1}{|k|^2 - \kappa^2},$$

and calculating the Fourier transform of the characteristic function of a sphere to be

$$\begin{aligned}\hat{\chi}_{B_R(0)}(k) &= \frac{1}{\sqrt{2\pi}}\int_0^R\int_0^\pi r^2\sin\theta \mathrm{e}^{-\mathrm{i}r|k|\cos\theta}\mathrm{d}\theta\mathrm{d}r = \frac{1}{\sqrt{2\pi}}\int_0^R\frac{r}{\mathrm{i}|k|}(\mathrm{e}^{\mathrm{i}r|k|}-\mathrm{e}^{-\mathrm{i}r|k|})\mathrm{d}r\\ &= \frac{1}{|k|^3}\sqrt{\frac{2}{\pi}}\int_0^{R|k|}\alpha\sin(\alpha)\mathrm{d}\alpha = \frac{1}{|k|^3}\sqrt{\frac{2}{\pi}}(\sin(R|k|)-R|k|\cos(R|k|));\end{aligned}$$

we are left with

$$\begin{aligned}\lim_{N_j\to\infty}N_jK_{L_j(N_j)}(\omega,x) &= (2\pi)^4\mathrm{e}^{-\mathrm{i}\frac{\omega}{c}n_j(\omega)\mu_j}\rho_j\sqrt{\frac{2}{\pi}}\\ &\times\int_{\mathbb{R}}h(R(\tfrac{\omega}{c}n_j(\omega)-k_3))\frac{1}{k_3^2-\frac{\omega^2}{c^2}n_j^2(\omega)}\hat{\chi}_{[-d_j,d_j]}(k_3)\mathrm{e}^{-\mathrm{i}k_3(x_3-\mu_j)}\mathrm{d}k_3, \qquad (25)\end{aligned}$$

where we used the abbreviation h from (21).

Inserting finally

$$\hat{\chi}_{[-d_j,d_j]}(k_3) = \frac{1}{\sqrt{2\pi}}\int_{-d_j}^{d_j}\mathrm{e}^{-\mathrm{i}k_3x_3}\mathrm{d}x_3 = \frac{1}{\sqrt{2\pi}}\frac{1}{\mathrm{i}k_3}\left(\mathrm{e}^{\mathrm{i}k_3d_j}-\mathrm{e}^{-\mathrm{i}k_3d_j}\right),$$

we see that the integrand in (25) can for $x_3-\mu_j>d_j+R$ (that is, for $x_3>\zeta_j+R$) be meromorphically extended to a function of k_3 in the lower half complex plane which decays sufficiently fast at infinity, so that the residue theorem yields

$$\begin{aligned}\lim_{N_j\to\infty}N_jK_{L_j(N_j)}(\omega,x) &= (2\pi)^4\mathrm{e}^{-\mathrm{i}\frac{\omega}{c}n_j(\omega)\mu_j}\rho_j\frac{h(2R\frac{\omega}{c}n_j(\omega))}{\frac{\omega^2}{c^2}n_j^2(\omega)}\\ &\times\left(\mathrm{e}^{\mathrm{i}\frac{\omega}{c}n_j(\omega)d_j}-\mathrm{e}^{-\mathrm{i}\frac{\omega}{c}n_j(\omega)d_j}\right)\mathrm{e}^{\mathrm{i}\frac{\omega}{c}n_j(\omega)(x_3-\mu_j)}.\end{aligned}$$

Putting this into (22), we obtain with $\mu_j+d_j=\zeta_j$ and $\mu_j-d_j=Z_j$ the formula (19).

Similarly, we extend the integrand for $x_3-\mu_j<-d_j-R$ (that is, for $x_3<Z_j-R$) meromorphically to a function of k_3 in the upper half plane and find with the residue theorem that

$$\begin{aligned}\lim_{N_j\to\infty}N_jK_{L_j(N_j)}(\omega,x) &= (2\pi)^4\rho_j\frac{h(0)}{\frac{\omega^2}{c^2}n_j^2(\omega)}\\ &\times\left(\mathrm{e}^{\mathrm{i}\frac{\omega}{c}n_j(\omega)d_j}-\mathrm{e}^{-\mathrm{i}\frac{\omega}{c}n_j(\omega)d_j}\right)\mathrm{e}^{-\mathrm{i}\frac{\omega}{c}n_j(\omega)x_3},\end{aligned}$$

which gives us with (22) and with $h(0)=\frac{1}{3}$ the formula (20).

5 Recovering the Susceptibility with Optical Coherence Elastography

So far, we have presented a way to model the measurements of an optical coherence tomography setup for a layered medium of the form (6). The question we are really interested in, however, is how to reconstruct the properties of the medium from this data.

Let us first consider one of the layer stripping steps for a susceptibility χ of the form (9) with X_1 being either of the form (12) of Problem 1 or of the form (13) of Problem 2. We make the additional assumption that supp $\chi_j \subset [0, T]$ and supp $\chi_j^{(\mathrm{p})} \subset [0, T]$ for a sufficiently small $T > 0$. Then, we see that by choosing a sufficiently short pulse as incident wave, that is, $E^{(0)}(t, x) = f(t + \frac{x_3}{c})\eta$ (assuming for the background medium $\chi_0 = 0$) with f having a sufficiently narrow support (this ability is of course limited by the available frequencies), we can arrange it such that the field E in the presence of χ and the field E_1 in the presence of X_1 (where we are content with the averaged Born approximation of the electric field, see (18), in the case of Problem 2) are such that $E_1(t, x_0) = E(t, x_0)$ for all $t < t_0$ and $E_1(t, x_0) = 0$ for $t \geq t_0$ at the detector $x_0 \in \mathbb{R}^3$ for some time $t_0 \in \mathbb{R}$. This allows us to split the reconstruction of the electric susceptibility by a layer stripping method and reconstruct each layer separately.

We will therefore only describe the inductive steps, in which we independently consider the subproblems described in Sect. 4.

We want to start with measurements from an optical coherence elastography setup, that is, we have optical coherence tomography data for different elastic states of the medium. Concretely, we apply a force proportional to some parameter $\delta \in \mathbb{R}$ perpendicular to the layers of the medium, which causes under the assumption of a linear elastic medium a linear displacement of the position z_j of the layer. Correspondingly, the refractive indices in the medium, defined by (15), will change, which we assume to be linear as well. Thus, each layer at the compression state corresponding to δ will be characterised by a refractive index $\bar{n}_j$ and a vertical position $\bar{z}_j$ of the beginning of the layer of the form

$$\bar{n}_j(\omega, \delta) = n_j(\omega) + \delta n_j'(\omega) \text{ and } \bar{z}_j(\delta) = z_j + \delta z_j'$$

for some functions $n_j' : \mathbb{R} \to \mathbb{C}$ and some slopes $z_j' \in \mathbb{R}$.

In the first reconstruction step, we have that the first layer is the background in which the medium resides, which we assume to be well described by the vacuum $n_0 = 1$ and not to be affected by the compression, that is, $n_0' = 0$. Moreover, the distance between the detector and the medium shall be kept fixed during the compression so that $z_1' = 0$ as well.

According to Lemma 4, the measurements at the detector $x_0 \in \mathbb{R}^3$ with $x_{0,3} > z_1$ then allow us to extract (knowing $\bar{n}_0 = 1$, the incident field $E^{(0)}$, and the vertical position $x_{0,3}$ of the detector explicitly) the information

$$m_0[n_1, n_1', z](\omega, \delta) = \frac{\bar{n}_1(\omega, \delta) - 1}{\bar{n}_1(\omega, \delta) + 1} e^{-2\mathrm{i}\frac{\omega}{c} z_1}. \tag{26}$$

From this data, we can uniquely compute the functions n_1, n_1', and z_1.

Lemma 6 *Let $I \subset \mathbb{R}$ be a set which contains at least two incommensurable points $\omega_1, \omega_2 \in I \setminus \{0\}$ (that is, $\frac{\omega_1}{\omega_2} \in \mathbb{R} \setminus \mathbb{Q}$). Assume that we have (n_1, n_1', z_1) and $(\tilde{n}_1, \tilde{n}_1', \tilde{z}_1)$ with $n_1'(\omega) \neq 0$, $\tilde{n}_1'(\omega) \neq 0$, and*

$$m_0[n_1, n_1', z_1](\omega, \delta) = m_0[\tilde{n}_1, \tilde{n}_1', \tilde{z}_1](\omega, \delta) \text{ for all } \omega \in I,\ \ \delta \in \mathbb{R}. \tag{27}$$

Then, we have

$$n_1(\omega) = \tilde{n}_1(\omega),\ \ n_1'(\omega) = \tilde{n}_1'(\omega),\ \ \text{and } z_1 = \tilde{z}_1 \text{ for all } \omega \in I.$$

Proof Expanding the fractions in (27), the equation reduces to the zeroes of a quadratic polynomial in δ. Comparing the coefficients of second order of δ, we find that

$$n_1'(\omega)\tilde{n}_1'(\omega)\left(e^{-2\mathrm{i}\frac{\omega}{c}z_1} - e^{-2\mathrm{i}\frac{\omega}{c}\tilde{z}_1}\right) = 0.$$

Thus, we get

$$e^{-2\mathrm{i}\frac{\omega}{c}z_1} = e^{-2\mathrm{i}\frac{\omega}{c}\tilde{z}_1} \text{ for all } \omega \in I.$$

Evaluating this at ω_1 and ω_2, we have that there exist two integers $\lambda_1, \lambda_2 \in \mathbb{Z}$ with

$$z_1 - \tilde{z}_1 = \frac{\pi c}{\omega_1}\lambda_1 = \frac{\pi c}{\omega_2}\lambda_2.$$

If $\lambda_2 \neq 0$, then we would get the contradiction $\frac{\lambda_1}{\lambda_2} = \frac{\omega_1}{\omega_2} \in \mathbb{R} \setminus \mathbb{Q}$. Therefore, $\lambda_2 = 0$, which means that $z_1 = \tilde{z}_1$.

With this, (27) evaluated at $\delta = 0$ simplifies to

$$n_1(\omega) = \tilde{n}_1(\omega) \text{ for all } \omega \in I.$$

Finally, looking at the terms of first order in δ in the expanded version of (27), we find that they have been reduced to give the equation

$$n_1'(\omega) = \tilde{n}_1'(\omega).$$

After having recovered the parameters up to the jth layer, $j \in \{1, \ldots, J\}$, we can clean our measurement data from all effects caused by the previous layers and consider the next subproblem, namely the signal originating from the region of the randomly distributed particles. Here, the unknown parameters consist of

- the radius R of the particles, which we will assume to be so small that the approximation $R = 0$ is reasonable and that the particles can also after compression be considered to have a round shape;
- the ratio $\rho_j > 0$ of particles per surface area, which we assume to be invariant under the compression;
- the refractive index $\bar\nu_j$ of the particles, which we assume to deform linearly according to
$$\bar\nu_j(\omega,\delta) = \nu_j(\omega) + \delta\nu_j'(\omega), \text{ where } \nu_j(\omega) = \sqrt{1+\check\chi_j^{(\mathrm{p})}(\omega)},$$
under compression; and
- the vertical positions $\bar\zeta_j$ and $\bar Z_j$ of the beginning and the end of the random medium inside the jth layer, which are also assumed to change linearly according to
$$\bar\zeta_j(\delta) = \zeta_j + \delta\zeta_j' \text{ and } \bar Z_j(\delta) = Z_j + \delta Z_j'.$$

We collect these unknowns in the tuple $S_j = (\rho_j, \nu_j, \nu_j', \zeta_j, \zeta_j', Z_j, Z_j')$. The (corrected) incident wave $E^{(0)}$ and the refractive index n_j and its rate n_j' of change under compression are presumed to be already calculated.

From the measurements of the electric field for this subproblem, provided that it can be well approximated by the expected value of the Born approximation as calculated in Lemma 5, we can extract the data (rewriting the expression (16) for ϕ_j in (19) in terms of the refractive indices)

$$\begin{aligned} M_j[S_j](\omega,\delta) &= \rho_j(\bar\nu_j^2(\omega,\delta) - \bar n_j^2(\omega,\delta)) \\ &\qquad \times \left(\mathrm{e}^{-\mathrm{i}\frac{\omega}{2c}\bar n_j(\omega,\delta)(\bar\zeta_j(\delta)+3\bar Z_j(\delta))} - \mathrm{e}^{-\mathrm{i}\frac{\omega}{2c}\bar n_j(\omega,\delta)(3\bar\zeta_j(\delta)+\bar Z_j(\delta))}\right), \end{aligned}$$

Lemma 7 *Let $j \in \{1,\ldots,J\}$ be fixed, $I \subset \mathbb{R}$ be an arbitrary subset and n_j, n_j' be given such that $n_j(\omega) \neq 0$ for every $\omega \in I$ and that there exists a value $\omega_0 \in I \setminus \{0\}$ with $\Im\mathrm{m}(n_j'(\omega_0)) > 0$. Assume that we have $S_j = (\rho_j, \nu_j, \nu_j', \zeta_j, \zeta_j', Z_j, Z_j')$ and $\tilde S_j = (\tilde\rho_j, \tilde\nu_j, \tilde\nu_j', \tilde\zeta_j, \tilde\zeta_j', \tilde Z_j, \tilde Z_j')$ with*

$$M_j[S_j](\omega,\delta) = M_j[\tilde S_j](\omega,\delta) \text{ for all } \omega \in I,\ \delta \in \mathbb{R}. \tag{28}$$

Additionally, we enforce the ordering $Z_j < \zeta_j$ and $\tilde Z_j < \tilde\zeta_j$ about the beginning and the end of the random layer and make the assumptions $Z_j' > \zeta_j' > 0$ and $\tilde Z_j' > \tilde\zeta_j' > 0$ that the layer shrinks when being compressed.

Moreover, we assume the existence of an element $\omega_1 \in I$ so that

$$\frac{n_j'(\omega_1)}{n_j(\omega_1)} \neq \frac{\nu_j'(\omega_1)}{\nu_j(\omega_1)}. \tag{29}$$

Then, we have

$$S_j = \tilde{S}_j.$$

Proof Considering the different orders of decay in δ in the exponents in (28), we require that all of them match, which yields the equation system

$$\begin{aligned}\delta^2\frac{\omega}{2c}\,\Im\mathrm{m}(n'_j(\omega))(\zeta'_j+3Z'_j) &= \delta^2\frac{\omega}{c}\,\Im\mathrm{m}(n'_j(\omega))(\tilde{\zeta}'_j+3\tilde{Z}'_j) \text{ and}\\ \delta^2\frac{\omega}{2c}\,\Im\mathrm{m}(n'_j(\omega))(3\zeta'_j+Z'_j) &= \delta^2\frac{\omega}{c}\,\Im\mathrm{m}(n'_j(\omega))(3\tilde{\zeta}'_j+\tilde{Z}'_j)\end{aligned}$$

for the exponents quadratic in δ, which implies for the frequency $\omega=\omega_0$ for which $\Im\mathrm{m}(n'_j(\omega_0))>0$ that $\zeta'_j=\tilde{\zeta}'_j$ and $Z'_j=\tilde{Z}'_j$, and, using this result, the equation system

$$\begin{aligned}\delta\frac{\omega}{2c}\,\Im\mathrm{m}(n'_j(\omega))(3\zeta_j+Z_j) &= \delta\frac{\omega}{2c}\,\Im\mathrm{m}(n'_j(\omega))(3\tilde{\zeta}_j+\tilde{Z}_j) \text{ and}\\ \delta\frac{\omega}{2c}\,\Im\mathrm{m}(n'_j(\omega))(\zeta_j+3Z_j) &= \delta\frac{\omega}{2c}\,\Im\mathrm{m}(n'_j(\omega))(\tilde{\zeta}_j+3\tilde{Z}_j)\end{aligned}$$

for the exponents linear in δ, which further implies $\zeta_j=\tilde{\zeta}_j$ and $Z_j=\tilde{Z}_j$.

At this point, (28) is reduced to

$$\rho_j\left((\nu_j+\delta\nu'_j)^2-(n_j+\delta n'_j)^2\right)=\tilde{\rho}_j\left((\tilde{\nu}_j+\delta\tilde{\nu}'_j)^2-(n_j+\delta n'_j)^2\right).$$

Comparing coefficients with respect to δ gives us the equation system

$$\rho_j\left({\nu'_j}^2-{n'_j}^2\right)=\tilde{\rho}_j\left({\tilde{\nu}'_j}^2-{n'_j}^2\right), \tag{30}$$

$$\rho_j\left(\nu_j\nu'_j-n_jn'_j\right)=\tilde{\rho}_j\left(\tilde{\nu}_j\tilde{\nu}'_j-n_jn'_j\right), \tag{31}$$

$$\rho_j\left(\nu_j^2-n_j^2\right)=\tilde{\rho}_j\left(\tilde{\nu}_j^2-n_j^2\right). \tag{32}$$

We use Eqs. (32) in (30) and (31) to eliminate of the variables ρ_j and $\tilde{\rho}_j$, and interpret the result as an equation system for the variables $\tilde{\nu}_j$ and $\tilde{\nu}'_j$. Solving these equations then for $\tilde{\nu}'_j$, gives us

$$\begin{aligned}(\nu_j^2-n_j^2){\tilde{\nu}'_j}^2 &= (\tilde{\nu}_j^2-n_j^2){\nu'_j}^2+(\nu_j^2-\tilde{\nu}_j^2){n'_j}^2,\\ (\nu_j^2-n_j^2)\tilde{\nu}_j\tilde{\nu}'_j &= (\tilde{\nu}_j^2-n_j^2)\nu_j\nu'_j+(\nu_j^2-\tilde{\nu}_j^2)n_jn'_j.\end{aligned}$$

Eliminating further $\tilde{\nu}'_j$ by multiplying the first equation with $\tilde{\nu}_j$ and subtracting the squared second equation, we find after some algebraic manipulations

$$(\tilde{\nu}_j^2-n_j^2)(\nu_j^2-\tilde{\nu}_j^2)(\nu'_jn_j-\nu_jn'_j)^2=0.$$

Evaluating this at the value ω_1, we see that the last factor is by assumption (29) not zero. Thus, there are only two cases.

1. Either we have $\tilde{\nu}_j(\omega_1) = \nu_j(\omega_1) \neq n_j(\omega_1)$ and therefore by (32) that $\tilde{\rho}_j = \rho_j$; then we get with (32) and (30) that $\tilde{\nu}_j = \nu_j$ and $\tilde{\nu}'_j = \nu'_j$ holds on the whole set I, which means that we have shown $\tilde{S}_j = S_j$.
2. Or we have that $\tilde{\nu}_j(\omega_1) = n_j(\omega_1)$. Then, (32) tells us that also $\nu_j(\omega_1) = n_j(\omega_1)$ and thus, by combining (30) and (31), that $\tilde{\nu}'_j(\omega_1) = \nu'_j(\omega_1)$. Furthermore, we know from assumption (29) that in this case $\nu'_j(\omega_1) \neq n'_j(\omega_1)$, and therefore (30) implies $\tilde{\rho}_j = \rho_j$ from which we again conclude that $\tilde{S}_j = S_j$.

As last type of subproblem, we encounter then the interface between the layer j and the layer $j+1$. Similarly to the case of the initial layer, we obtain here from Lemma 4 the data

$$m_j[n_{j+1}, n'_{j+1}, z_{j+1}, z'_{j+1}](\omega, \delta) = \frac{\bar{n}_{j+1}(\omega, \delta) - \bar{n}_j(\omega, \delta)}{\bar{n}_{j+1}(\omega, \delta) + \bar{n}_j(\omega, \delta)} \mathrm{e}^{-2\mathrm{i}\frac{\omega}{c}\bar{n}_j(\omega,\delta)\bar{z}_{j+1}(\delta)}.$$

Again, this data allows us to uniquely obtain the variables n_{j+1}, n'_{j+1}, z_{j+1}, and z'_{j+1} from the already reconstructed values n_j and n'_j.

Lemma 8 *Let $j \in \{1, \ldots, J-1\}$ be fixed, $I \subset \mathbb{R}$ be an arbitrary subset and n_j, n'_j be given such that $n_j(\omega) \neq 0$ for every $\omega \in I$ and that there exists a value $\omega_0 \in I \setminus \{0\}$ with $\Im\mathrm{m}(n'_j(\omega_0)) > 0$. Assume that we have $(n_{j+1}, n'_{j+1}, z_{j+1}, z'_{j+1})$ and $(\tilde{n}_{j+1}, \tilde{n}'_{j+1}, \tilde{z}_{j+1}, \tilde{z}'_{j+1})$ with*

$$m_j[n_{j+1}, n'_{j+1}, z_{j+1}, z'_{j+1}](\omega, \delta) = m_j[\tilde{n}_{j+1}, \tilde{n}'_{j+1}, \tilde{z}_{j+1}, \tilde{z}'_{j+1}](\omega, \delta) \tag{33}$$

for all $\omega \in I$ and $\delta \in \mathbb{R}$.

Then, we have

$$n_{j+1}(\omega) = \tilde{n}_{j+1}(\omega),\ \ n'_{j+1}(\omega) = \tilde{n}'_{j+1}(\omega),\ \ z_{j+1} = \tilde{z}_{j+1},\ \text{ and } z'_{j+1} = \tilde{z}'_{j+1}$$

for all $\omega \in I$.

Proof Comparing again the different orders of decay in δ in the exponents in (33), we require that the coefficients on both sides coincide:

$$2\delta^2 \frac{\omega}{c} \Im\mathrm{m}(n'_j(\omega))(z'_{j+1} - \tilde{z}'_{j+1}) = 0 \text{ and}$$

$$4\delta \frac{\omega}{c} \left(\Im\mathrm{m}(n_j(\omega))(z'_{j+1} - \tilde{z}'_{j+1}) + \Im\mathrm{m}(n'_j(\omega))(z_{j+1} - \tilde{z}_{j+1})\right) = 0.$$

Because of the assumption that $\Im\mathrm{m}(n'_j(\omega_0)) > 0$, this is equivalent to

$$z'_{j+1} = \tilde{z}'_{j+1} \text{ and } z_{j+1} = \tilde{z}_{j+1}.$$

As in the proof of Lemma 6, Eq. (33) for $\delta = 0$ then gives us

$$2n_j(\omega)(n_{j+1}(\omega) - \tilde{n}_{j+1}(\omega)) = 0,$$

resulting in $n_{j+1}(\omega) = \tilde{n}_{j+1}(\omega)$.

Finally, dividing both sides of (33) by the exponential factors (which we already know to be the same), we get a quadratic equation for δ and equating the first order terms in δ, we obtain

$$2n_j(\omega)(n'_{j+1}(\omega) - \tilde{n}'_{j+1}(\omega)) = 0,$$

which yields $n'_{j+1}(\omega) = \tilde{n}'_{j+1}(\omega)$.

Conclusion

We have thus shown that by analysing a layered medium endued with independently uniformly distributed scatterers in each layer with optical coherence tomography, we can reduce the inverse problem of reconstructing the electric susceptibility of the medium to subproblems for each layer separately by a layer stripping argument, provided the homogeneous parts between the different regions are not too small.

Then by combining this imaging method with an elastography setup by recording measurements for different compression states (normal to the layered structure), we find out that this allows for the reconstruction of the optical parameters and leads to a unique reconstructability of all the optical parameters: the electric susceptibilities and positions of the layers, the electric susceptibilities of the randomly distributed particles, their particle density, and the locations of the regions of these particles (at every compression state). Of course, the recovered shifts of the layer boundaries for the different compression states could then be used in a next step to determine elastic parameters of the medium.

Acknowledgements This work was made possible by the greatly appreciated support of the Austrian Science Fund (FWF) via the special research programme SFB F68 “Tomography Across the Scales”: Peter Elbau and Leopold Veselka have been supported via the subproject F6804-N36 “Quantitative Coupled Physics Imaging”, and Leonidas Mindrinos acknowledges support from the subproject F6801-N36.

References

1. Ammari, H., Bretin, E., Millien, P., Seppecher, L., Seo, J.K.: Mathematical modeling in full-field optical coherence elastography. SIAM J. Appl. Math. **75**(3), 1015–1030 (2015)
2. Ammari, H., Romero, F., Shi, C.: A signal separation technique for sub-cellular imaging using dynamic optical coherence tomography. Multiscale Model. Simul. **15**(3), 1155–1175 (2017)

3. Brezinski, M.E.: Optical Coherence Tomography Principles and Applications. Academic Press, New York (2006)
4. Drexler, W., Fujimoto, J.G.: Optical Coherence Tomography: Technology and Applications, 2nd edn. Springer International Publishing, Switzerland (2015)
5. Drexler, W., Hubmer, S., Krainz, L., Neubauer, A., Scherzer, O., Schmid, J., Sherina, E.: Lamé parameter estimation from static displacement field measurements. In: Burger, M., Hahn, B., Quinto, E.T. (eds) Oberwolfach Conference: Tomographic Inverse Problems: Theory and Applications, Oberwolfach reports, pp. 74–76. EMS Publishing House (2019)
6. Elbau, P., Mindrinos, L., Scherzer, O.: Mathematical methods of optical coherence tomography. In: Scherzer, O. (ed.) Handbook of Mathematical Methods in Imaging, pp. 1169–1204. Springer, New York (2015)
7. Elbau, P., Mindrinos, L., Scherzer, O.: The inverse scattering problem for orthotropic media in polarization-sensitive optical coherence tomography. GEM. Int. J. Geomath. **9**(1), 145–165 (2018)
8. Elbau, P., Mindrinos, L., Veselka, L.: Quantitative OCT reconstructions for dispersive media. In: Kaltenbacher, B., Wald, A., Schuster, T. (eds) Time-dependent problems in imaging and parameter identification, to appear. Springer, Heidelberg (2020)
9. Nahas, A., Bauer, M., Roux, S., Boccara, A.C.: 3D static elastography at the micrometer scale using full field oct. Biomed. Opt. Express **4**(10), 2138–2149 (2013)
10. Rudin, W.: Real and Complex Analysis, 3rd edn. McGraw-Hill, New York (1987)
11. Santos, M., Araüjo, A., Barbeiro, S., Cramelo, F., Correia, A., Marques, M.I., Morgado, M., Pinto, L., Serranho, P., Bernardes, A.: Maxwell's equations based 3d model of light scattering in the retina. In: 4th Portuguese Meeting on Bioengineering (ENBENG), p. 5. IEEE (2015)
12. Sun, C., Standish, B.A., Yang, V.: Optical coherence elastography: current status and future applications. J. Biomed. Opt. **16**(4), 1–13 (2011)
13. Tricoli, U., Carminati, R.: Modeling of full-field optical coherence tomography in scattering media. J. Opt. Soc. Am. A **36**(11), C122–C129 (2019)

The Finite Element Method and Balancing Principle for Magnetic Resonance Imaging

Larisa Beilina, Geneviève Guillot, and Kati Niinimäki

Abstract This work considers a finite element method in combination with balancing principle for a posteriori choice of the regularization parameter for image reconstruction problem appearing in magnetic resonance imaging (MRI). The fixed point iterative algorithm is formulated and it's performance is demonstrated on the image reconstruction from experimental MR data.

Keywords MRI · Fredholm integral equation of the first kind · Finite element method · Regularization · Balancing principle

MSC: 65R20 · 65R32

1 Introduction

In this work is studied MRI problem described by a Fredholm integral equation of the first kind, via applying finite element method (FEM) to its solution. Fredholm integral equation of the first kind is an ill-posed problem and to approach it, a minimization of the Tikhonov functional [15–17] is usually used.

L. Beilina (✉)
Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, 412 96 Gothenburg, Sweden
e-mail: larisa.beilina@chalmers.se

G. Guillot
IR4M UMR8081, CNRS, Université Paris-Sud, Université Paris-Saclay, bâtiment 220, 4 place du Général Leclerc, 91401 Orsay, France
e-mail: genevieve.guilloti@u-psud.fr

K. Niinimäki
IR4M UMR8081, CNRS, Université Paris-Sud, Université Paris-Saclay, SHFJ, 4 place du Général Leclerc, 91401 Orsay, France
e-mail: kati.niinimaki@u-psud.fr; kati.niinimaki@planmeca.com

Xray Division, Planmeca Oy, Asentajankatu 6, 00880 Helsinki, Finland

L. Beilina et al. (eds.), *Mathematical and Numerical Approaches for Multi-Wave Inverse Problems*, Springer Proceedings in Mathematics & Statistics 328,
https://doi.org/10.1007/978-3-030-48634-1_9

The paper is focused on the efficiency of applying of FEM for image reconstruction in magnetic resonance imaging. A FEM for integral equation of the first kind was elaborated in [5] and an adaptive FEM with a posteriori error estimates for Tikhonov functional and the regularized solution of this functional was developed in [10]. The present work employs the finite element method for minimization of the regularized Tikhonov functional where the regularized term is performed in the H^1 norm. Further, we derive the Fréchet derivative of the regularized Tikhonov functional and formulate the finite element method to find optimal solution of this functional. Balancing principle [6] is then used for a posteriori choice of the regularization parameter. Finally, a fixed point iterative algorithm, which combined the finite element method and balancing principle, is formulated and applied to the reconstruction of images from experimental MR data acquired at IR4M laboratory in Paris-Sud University, France.

The outline of this paper is as follows. The statements of forward and inverse MRI problems are presented in Sect. 2. In Sect. 3, the finite element method for minimization of the Tikhonov functional is formulated. In Sect. 4, the balancing principle for choosing of the regularization parameter in the Tikhonov functional is briefly presented and in Sect. 4.1, the fixed point iterative algorithm for solution of MRI problem is formulated. Section 4.2 presents convergence analysis of the algorithm of Sect. 4.1. Finally, Sect. 5 presents numerical results of reconstruction from experimental MR data using proposed finite element method.

2 Statement of the Forward and Inverse MRI Problem

Throught the paper, by $H^k(\Omega)$ denotes the Hilbert space of all L_2-functions $\omega(x)$ defined in the domain Ω which are k times continuously differentiable in Ω and with all partial derivatives of the order $|\alpha| \leq k : D^\alpha w \in L_2(\Omega)$. The inner product in $H^k\,(\Omega)$ is defined as

$$(w, v)_{H^k(\Omega)} = \sum_{|\alpha|\leq k} \int\limits_{\Omega} D^\alpha w\; D^\alpha v\; dx.$$

We denote the domain of image reconstruction by $\Omega \subset \mathbb{C}^2$ with the boundary $\partial\Omega$, and the domain where MR data are collected, by $\Omega_\kappa \subset \mathbb{C}^2$ with the boundary $\partial\Omega_\kappa$, and call them as image-space and k-space, respectively.

The goal of this work is to solve a two-dimensional Fredholm integral equation of the first kind

$$u(k_x, k_y) = \int_\Omega f(x, y) G(x, y, k_x, k_y)\, \mathrm{d}x\mathrm{d}y, \tag{1}$$

where $f(x, y) \in H(\Omega)$ denotes the unknown image function which should be reconstructed from the experimental MR data $u(k_x, k_y) \in L_2(\Omega_\kappa)$. In (1) the kernel function is defined by

$$G(x, y, k_x, k_y) = e^{-2\pi \mathbf{i}(k_x x + k_y y)} \in C^k, \quad k > 0 \tag{2}$$

where $\mathbf{i}$ is an imaginary unit. In Eqs. (1)–(2) (k_x, k_y) denote the k-space trajectories which correspond to the coordinates of measured data $u \in \Omega_\kappa$ and are defined in time t as

$$k_x(t) = \frac{\gamma}{2\pi} \int_0^t G_x(\tau) d\tau = \frac{\gamma}{2\pi} G_x t, \tag{3}$$

$$k_y(t) = \frac{\gamma}{2\pi} \int_0^t G_y(\tau) d\tau = \frac{\gamma}{2\pi} G_y t, \tag{4}$$

where G_x, G_y are the known magnetic field gradients which prescribe how the k-space data in Ω_κ is acquired. For more details about the statement of MRI problem we refer to [4].

The Eq. (1) can be written as an operator equation

$$Af = u, \tag{5}$$

with a bounded linear operator $A : H^1(\Omega) \to L_2(\Omega_\kappa)$ defined as

$$Af := \int_\Omega f(x, y) e^{-2\pi \mathbf{i}(k_x x + k_y y)} \, \mathrm{d}x \mathrm{d}y. \tag{6}$$

Further we will consider the following ill-posed problem

Ill-posed problem (IP)
Find $f(x, y)$ in (1) when the measured MR data $u(k_x, k_y) \in \Omega_\kappa$, the k-space coordinates (k_x, k_y) and the kernel $G(x, y, k_x, k_y)$ are known.

IP needs regularization [1, 2, 9, 13, 15–17]. Thus, to find a solution for **IP** we construct the Tikhonov regularization functional

$$M_\alpha(f) = \frac{1}{2} \|Af - u\|^2_{L_2(\Omega_\kappa)} + \frac{\alpha}{2} \|f\|^2_{H^1(\Omega)}, \tag{7}$$

$$M_\alpha(f) : H^1(\Omega) \to \mathbb{C}, \quad u \in L_2(\Omega_\kappa).$$

In (7) $\alpha = \alpha(\delta) > 0$ is a regularization parameter depending on the noise δ in data such that

$$\|u - u^*\|_{L_2(\Omega_\kappa)} \le \delta,$$

where u^* denote perfect noiseless data corresponding to the exact solution f^* of (5) such that

$$Af^* = u^*. \tag{8}$$

Our goal is to find minimum of (7)

$$\inf_{f \in H^1(\Omega)} M_\alpha(f) = \inf_{f \in H^1(\Omega)} \left\{ \frac{1}{2} \left\| Af - u \right\|^2_{L_2(\Omega_\kappa)} + \frac{\alpha}{2} \left\| f \right\|^2_{H^1(\Omega)} \right\} \tag{9}$$

such that for all $b \in H^1(\Omega)$

$$M'_\alpha(f)(b) = 0, \tag{10}$$

where $M'_\alpha(f)$ denotes the Fréchet derivative of the functional (7).

The following lemma is well-known for the operator $A : L_2(\Omega) \to L_2(\Omega_\kappa)$, see [2].

Lemma 1 *Let $A : L_2(\Omega) \to L_2(\Omega_\kappa)$ be a bounded linear operator. Then the Fréchet derivative of the functional (7) is*

$$M'_\alpha(f)(b) = (A^*Af - A^*u, b) + \alpha(f, b), \ \forall b \in L_2(\Omega). \tag{11}$$

In the case when the operator $A : H^1(\Omega) \to L_2(\Omega_\kappa)$ the derivation of the Fréchet derivative is more complicated because of the presence a H^1 norm in the regularization term. Below we formulate a lemma concerning the Fréchet derivative of the operator $A : H^1(\Omega) \to L_2(\Omega_\kappa)$.

Lemma 2 *Let $A : H^1(\Omega) \to L_2(\Omega_\kappa)$ be a bounded linear operator. Then the Fréchet derivative of the functional*

$$M_\alpha(f) = \frac{1}{2} \left\| Af - u \right\|^2_{L_2(\Omega_\kappa)} + \frac{\alpha}{2} \left\| |\nabla f| \right\|^2_{L^2(\Omega)}, \tag{12}$$

is

$$M'_\alpha(f)(b) = (A^*Af - A^*u, b) + \alpha(|\nabla f|, |\nabla b|), \ \forall b \in H^1(\Omega), \tag{13}$$

with a convex growth factor b, i.e., $|\nabla b| < b$.

Proof We have

$$\begin{aligned} M_\alpha(f) &= \frac{1}{2} \left\| Af - u \right\|^2_{L_2(\Omega_\kappa)} + \frac{\alpha}{2} \left\| |\nabla f| \right\|^2_{L^2(\Omega)} \\ &= \frac{1}{2} \int_{\Omega_k} \Big[\int_\Omega f(x, y) G(x, y, k_x, k_y) dx dy - u(k_x, k_y) \Big]^2 dk_x dk_y \\ &+ \frac{\alpha}{2} \int_\Omega |\nabla f|^2 dx dy. \end{aligned} \tag{14}$$

To find the Fréchet derivative (13) of the functional (12) we consider $M_\alpha(f + b) - M_\alpha(f) \ \forall b \in H^1(\Omega)$:

$$\begin{aligned} M_\alpha(f+b) - M_\alpha(f) = \frac{1}{2}\int_{\Omega_k}\Big[\int_\Omega (f+b)Gdxdy - u(k_x,k_y)\Big]^2 dk_x dk_y \\ + \frac{\alpha}{2}\int_\Omega |\nabla(f+b)|^2 dxdy \\ - \frac{1}{2}\int_{\Omega_k}\Big[\int_\Omega fGdxdy - u(k_x,k_y)\Big]^2 dk_x dk_y \\ - \frac{\alpha}{2}\int_\Omega |\nabla f|^2 dxdy = I_1 + I_2, \end{aligned} \tag{15}$$

where

$$\begin{aligned} I_1 &:= \frac{1}{2}\int_{\Omega_k}\Big[\int_\Omega (f+b)Gdxdy - u(k_x,k_y)\Big]^2 dk_x dk_y \\ &\quad - \frac{1}{2}\int_{\Omega_k}\Big[\int_\Omega fGdxdy - u(k_x,k_y)\Big]^2 dk_x dk_y, \\ I_2 &:= \frac{\alpha}{2}\int_\Omega |\nabla(f+b)|^2 dxdy - \frac{\alpha}{2}\int_\Omega |\nabla f|^2 dxdy. \end{aligned} \tag{16}$$

We consider separately terms I_1 and I_2. As for the term I_1

$$\begin{aligned} I_1 &= \frac{1}{2}\int_{\Omega_k}\Big[\int_\Omega fGdxdy - u + \int_\Omega bGdxdy\Big]^2 dk_x dk_y - \frac{1}{2}\int_{\Omega_k}\Big[\int_\Omega fGdxdy - u\Big]^2 dk_x dk_y \\ &= \frac{1}{2}\int_{\Omega_k}\Big\{(\int_\Omega fGdxdy - u)^2 + 2(\int_\Omega fGdxdy - u)\cdot\int_\Omega bGdxdy + (\int_\Omega bGdxdy)^2 \\ &- \int_\Omega (fGdxdy - u)^2\Big\} dk_x dk_y = \int_{\Omega_k}\Big[\int_\Omega fGdxdy - u)\cdot\int_\Omega bGdxdy\Big] dk_x dk_y \\ &+ \frac{1}{2}\int_{\Omega_k}\Big[\int_\Omega bGdxdy\Big]^2 dk_x dk_y. \end{aligned} \tag{17}$$

Similarly we rewrite the term I_2 as:

$$I_2 = \frac{\alpha}{2}\int_\Omega (|\nabla f + \nabla b|^2 - |\nabla f|^2)dxdy \le \frac{\alpha}{2}\int_\Omega (2|\nabla f||\nabla b| + |\nabla b|^2)dxdy. \tag{18}$$

Taking limits in I_1 and I_2 in the definition of Fréchet derivative we get

$$\begin{aligned} 0 &= \lim_{\|b\|\to 0}\frac{I_1}{\|b\|_2} \\ &= \lim_{\|b\|\to 0}\frac{\int_{\Omega_k}\Big[\int_\Omega fGdxdy - u)\cdot\int_\Omega bGdxdy\Big] dk_x dk_y + \frac{1}{2}\int_{\Omega_k}\Big[\int_\Omega bGdxdy\Big]^2 dk_x dk_y}{||b||_2}, \\ 0 &= \lim_{\|b\|\to 0}\frac{I_2}{\|b\|_2} \le \lim_{\|b\|\to 0}\frac{\alpha[\int_\Omega |\nabla f||\nabla b|dxdy + \frac{1}{2}\int_\Omega |\nabla b|^2 dxdy]}{\|b\|_2}. \end{aligned} \tag{19}$$

The second terms in limits for I_1 and I_2 in (19) are satisfying

$$\begin{aligned}
&\lim_{\|b\|\to 0} \frac{\frac{1}{2}\int_{\Omega_k}\left[\int_\Omega bGdxdy\right]^2 dk_x dk_y}{||b||_2} \to 0,\\
&\lim_{\|b\|\to 0} \frac{\frac{\alpha}{2}\int_\Omega |\nabla b|^2 dxdy}{\|b\|_2} \le \frac{\alpha}{2}\lim_{\|b\|\to 0} C(b)\frac{||b||_2^2}{||b||_2} \to 0, \quad C(b) = const.
\end{aligned} \tag{20}$$

Thus, the factors in the first terms for I_1 and I_2 of (19) should be also zero, from which (13) follows. □

Similarly can be proven the following Lemma.

Lemma 3 *Let $A : H^1(\Omega) \to L_2(\Omega_\kappa)$ be a bounded linear operator. Then the Fréchet derivative of the functional*

$$M_\alpha(f) = \frac{1}{2}\|Af-u\|^2_{L_2(\Omega_\kappa)} + \frac{\alpha}{2}\|f+|\nabla f|\|^2_{L^2(\Omega)} = \frac{1}{2}\|Af-u\|^2_{L_2(\Omega_\kappa)} + \frac{\alpha}{2}\|f\|^2_{H^1(\Omega)}, \tag{21}$$

is given by

$$M'_\alpha(f)(b) = (A^*Af - A^*u, b) + \alpha[(f,b) + (f,|\nabla b|) + (|\nabla f|, b) + (|\nabla f|, |\nabla b|)],\ \forall b \in H^1(\Omega), \tag{22}$$

with a convex growth factor b, i.e., $|\nabla b| < b$.

Proof Once again, we consider $M_\alpha(f+b) - M_\alpha(f)\ \forall b \in H^1(\Omega)$:

$$\begin{aligned}
M_\alpha(f+b) - M_\alpha(f) = &\frac{1}{2}\int_{\Omega_k}\left[\int_\Omega (f+b)Gdxdy - u(k_x,k_y)\right]^2 dk_x dk_y\\
&+ \frac{\alpha}{2}\int_\Omega (f+b+|\nabla(f+b)|)^2 dxdy\\
&- \frac{1}{2}\int_{\Omega_k}\left[\int_\Omega fGdxdy - u(k_x,k_y)\right]^2 dk_x dk_y\\
&- \frac{\alpha}{2}\int_\Omega (f+|\nabla f|)^2 dxdy = I_1 + I_2,
\end{aligned} \tag{23}$$

where I_1 is given in (16) and

$$I_2 := \frac{\alpha}{2}\int_\Omega (f+|\nabla f + \nabla b| + b)^2 dxdy - \frac{\alpha}{2}\int_\Omega (f+|\nabla f|)^2 dxdy. \tag{24}$$

The term I_1 is estimated in Lemma 2. Thus, it remains to consider only the term I_2.

$$
\begin{aligned}
I_2 &= \frac{\alpha}{2}\int_{\Omega}(f + |\nabla f + \nabla b| + b)^2 dxdy - \frac{\alpha}{2}\int_{\Omega}(f + |\nabla f|)^2 dxdy \\
&\leq \frac{\alpha}{2}\int_{\Omega}\Big(\Big[(f + |\nabla f|) + (b + |\nabla b|)\Big]^2 - (f + |\nabla f|)^2\Big)dxdy \\
&= \frac{\alpha}{2}\int_{\Omega}\Big(2(f + |\nabla f|)(b + |\nabla b|) + (b + |\nabla b|)^2\Big)dxdy.
\end{aligned}
\tag{25}
$$

Taking limit in I_2 in the definition of Fréchet derivative we get:

$$
0 = \lim_{\|b\|\to 0}\frac{I_2}{\|b\|_2} \leq \lim_{\|b\|\to 0}\frac{\alpha\Big[\int_{\Omega}(f + |\nabla f|)(b + |\nabla b|)dxdy + \frac{1}{2}\int_{\Omega}(b + |\nabla b|)^2 dxdy\Big]}{\|b\|_2}.
\tag{26}
$$

The second term in limit for I_2 in (26) can be written as

$$
\lim_{\|b\|\to 0}\frac{\frac{\alpha}{2}\int_{\Omega}(b + |\nabla b|)^2 dxdy}{\|b\|_2} \leq \frac{\alpha}{2}\lim_{\|b\|\to 0} D(b)\frac{||b||_2^2}{||b||_2} \to 0, \quad D(b) = const.
\tag{27}
$$

which approaches zero when b goes to zero. Thus, the factors in the first terms for I_1 in (19) and I_2 in (26) should be also zero, from which (22) follows.

3 The Finite Element Method for Minimization of the Tikhonov Functional

To formulate the finite element method for (13) we discretize the domains $\Omega \subset \mathbb{R}^2$, $\Omega_\kappa \subset \mathbb{R}^2$ by the meshes K_h, K_{h_κ}, respectively, consisting of non-overlapping triangles K such that

$$
\begin{aligned}
\Omega &= \cup_{K\in K_h} K = K_1 \cup K_2 ... \cup K_s, \quad K_h = \{K_1, ..., K_s\}, \\
\Omega_\kappa &= \cup_{K\in K_{h_\kappa}} K = K_{k_1} \cup K_{k_2} ... \cup K_{k_s}, \quad K_{h_\kappa} = \{K_{k_1}, ..., K_{k_s}\}.
\end{aligned}
$$

with the standard mesh regularity assumption [7]. We note that the number of elements s in both meshes is the same.

We define the finite element space $V_h \subset V$ as

$$
V_h = \big\{v \in L_2(\Omega) : v \in C(\Omega), v|_{\partial\Omega} = 0,\ v|_K \in P_1(K)\ \forall K \in K_h\big\}.
\tag{28}
$$

The finite element method for (13) reads: find $f_h \in V_h$ such that for all $v \in V_h$

$$
M'_\alpha(f_h)(v) = (A^*Af_h - A^*u, v) + \alpha(\nabla f_h, \nabla v) = 0.
\tag{29}
$$

The function f is approximated by $f_h \in V_h$, such that

$$f_h = \sum_{i=1}^{N} f_i \varphi_i, \tag{30}$$

where $\{\varphi_i\}_{i=1}^N$ are the standard continuous piecewise linear functions and f_i denote the unknown discrete function-values at the mesh point $x_i \in K_h$.

Substituting (30) into (29) with $v = \varphi_j$ and taking discrete function values of u_i at the mesh point $x_i \in K_{h_k}$ we get the discrete system of equations

$$\sum_{i,j=1}^{N} (A\varphi_i, A\varphi_j) f_i - \sum_{i,j=1}^{N} (u_i, A\varphi_j) + \alpha \sum_{i,j=1}^{N} (\nabla\varphi_i, \nabla\varphi_j) f_i = 0. \tag{31}$$

This system can be rewritten as

$$\sum_{i,j=1}^{N} (A\varphi_i, A\varphi_j) f_i + \alpha \sum_{i,j=1}^{N} (\nabla\varphi_i, \nabla\varphi_j) f_i = \sum_{i,j=1}^{N} (u_i, A\varphi_j), \tag{32}$$

which is equivalent to the following linear system of equations

$$(C + \alpha K)\mathbf{f} = \mathbf{b}. \tag{33}$$

In system (33), matrices C, K are the finally assembled block matrices, corresponding to the first two terms in the left hand side of (32), $\mathbf{f}$ denotes the vector of nodal values of finite element approximation f_h, $\mathbf{b}$ is the finally assembled right hand side of (32), see details in [3].

4 Balancing Principle

In this section, we briefly describe the balancing principle for finding the regularization parameter α in the functional (7) according to [6]. For this purpose the functional (7) is rewritten here as

$$M_\alpha(f) = \frac{1}{2}\,\|Af - u\|^2_{L_2(\Omega_k)} + \alpha\frac{1}{2}\|f\|^2_{H^1(\Omega)} = \varphi(f) + \alpha\psi(f). \tag{34}$$

For the functional (34) the value function $F(\alpha) : \mathbb{C} \to \mathbb{C}$ is defined according to [14] as

$$F(\alpha) = \inf_f M_\alpha(f). \tag{35}$$

If there exists derivative $F'(\alpha)$ at $\alpha > 0$ then from (34) and (35) follows that

$$F(\alpha) = \inf_f M_\alpha(f) = \underbrace{\varphi'(f)}_{\bar{\varphi}(\alpha)} + \alpha \underbrace{\psi'(f)}_{\bar{\psi}(\alpha)}. \tag{36}$$

Since $F'_\alpha(\alpha) = \psi'(f) = \bar{\psi}(\alpha)$ then from (36) we get

$$\bar{\psi}(\alpha) = F'(\alpha), \quad \bar{\varphi}(\alpha) = F(\alpha) - \alpha F'(\alpha). \tag{37}$$

For the functional (34) balancing principle (or Lepskii, see [11, 12]) finds $\alpha > 0$ such that the following expression is fulfilled

$$\bar{\varphi}(\alpha) = \gamma \alpha \bar{\psi}(\alpha), \tag{38}$$

where γ is determined by the statistical a priori knowledge from shape parameters in Gamma distributions. When $\gamma = 1$ the method is called zero crossing method, see details in [8].

Let us show that the balancing rule (38) finds optimal $\alpha > 0$ minimizing the balancing function

$$\varPhi_\gamma(\alpha) = \frac{F^{1+\gamma}(\alpha)}{\alpha}. \tag{39}$$

From conditions (37) it follows that

$$0 = \bar{\varphi}(\alpha) - \gamma \alpha \bar{\psi}(\alpha) = F(\alpha) - \alpha F'(\alpha) - \gamma \alpha F'(\alpha) = F(\alpha) - \alpha F'(\alpha)(1+\gamma),$$

which can be rewritten as

$$F(\alpha) = \alpha F'(\alpha)(1+\gamma). \tag{40}$$

Dividing both sides of (40) by $\alpha F(\alpha)$ we get

$$\frac{1}{\alpha} = \frac{F'(\alpha)}{F(\alpha)}(1+\gamma) = \frac{dF/d\alpha}{F(\alpha)}(1+\gamma)$$

or

$$\frac{d\alpha}{\alpha} = \frac{dF}{F(\alpha)}(1+\gamma).$$

Integrating with respect to α both sides of the above equation we obtain

$$\ln \alpha + C_1 = (1+\gamma) \ln F(\alpha) + C_2.$$

Now choosing $C_1 = C_2 = const.$ the above equation is rewritten as

$$\alpha = \exp^{(1+\gamma)\ln F(\alpha)} = F(\alpha)^{1+\gamma}$$

which in turn can be rewritten as the balancing function (39) to be minimized in the balancing principle.

We can check that the minimum of $\varPhi_\gamma(\alpha)$ is achieved at

$$0 = (\Phi_\gamma(\alpha))'_\alpha = \frac{(1+\gamma)F'(\alpha)F^\gamma(\alpha)\alpha - F^{1+\gamma}(\alpha)}{\alpha^2}.$$

From the above equation we get

$$(1+\gamma)F'(\alpha)F^\gamma(\alpha)\alpha = F^{1+\gamma}(\alpha) \to (1+\gamma)F'(\alpha)\alpha = F(\alpha).$$

This equation is the same as the Eq. (40) which gives the balancing principle

$$\bar{\varphi}(\alpha) = \gamma\alpha\bar{\psi}(\alpha). \qquad (41)$$

Thus, the balancing principle computes optimal value of α where $(\Phi_\gamma(\alpha))'_\alpha = 0$.

4.1 Fixed Point Algorithm for Finding Optimal α

For the Tikhonov functional (34) the following fixed point algorithm for computing α is proposed.

- Step 1. Start with the initial approximation of α_0, for example, choose $\alpha_0 = \delta^\mu$, $\mu \in (0, 2)$, and compute the sequence of α_k in the following steps.
- Step 2. Compute the value function $F(\alpha_k) = \inf_f M_{\alpha_k}(f)$ and get f_{α_k} via solving (33)
- Step 3. Update the regularization parameter $\alpha := \alpha_{k+1}$ as

$$\alpha_{k+1} = \frac{1}{\gamma}\frac{\varphi(f_{\alpha_k})}{\psi(f_{\alpha_k})}$$

- Step 4. For the tolerance $0 < \theta < 1$ chosen by the user, stop computing regularization parameters α_k if computed α_k are stabilized, or $|\alpha_k - \alpha_{k-1}| \leq \theta$. Otherwise, set $k := k + 1$ and go to Step 2.

4.2 Study of Convergence of Fixed Point Algorithm

The local convergence of the fixed point algorithm is developed under the following assumptions for the functional (34). Let the interval for finding optimal α be defined as $[\alpha_l, \alpha_r]$ and it is such that

- 1. $\bar{\psi}(\alpha_r) > 0 \to \bar{\psi}(\alpha) > 0$ for $\forall\alpha \in [0, \alpha_r]$.
- 2. Then $\exists\alpha_b \in [\alpha_l, \alpha_r] : D^\pm\Phi_\gamma(\alpha) < 0$ for $\alpha \in [\alpha_l, \alpha_b]$ and $D^\pm\Phi_\gamma(\alpha) > 0$ for $\alpha \in [\alpha_b, \alpha_r]$, where

$$D^+F(\alpha) = \lim_{h\to 0-} \frac{F(\alpha) - F(\alpha - h)}{h},$$
$$D^-F(\alpha) = \lim_{h\to 0+} \frac{F(\alpha + h) - F(\alpha)}{h}.$$

The first assumption guarantees well-posedness of the algorithm which is valid for a broad class of ill-posed problems with different regularization terms like $L^2 - l^1$ and L^2-TV. The second assumption guarantees that there exists only one local minimizer α_b for Φ_γ on $[\alpha_l, \alpha_r]$.

Let us define the residual

$$r(\alpha) = \bar{\varphi}(\alpha) - \gamma\alpha\bar{\psi}(\alpha). \tag{42}$$

The following Lemma will be used in the convergence theorem.

Lemma [6]

Under the above assumptions and with $\alpha_0 = [\alpha_l, \alpha_r]$ the sequence $\{\alpha_k\}$ generated by the fixed point algorithm is such that

- It is either finite or infinite and strictly monotone, and increasing if $r(\alpha) > 0$ and decreasing if $r(\alpha) < 0$.
- If $r(\alpha) > 0$, then the sequence $\{\alpha_k\} \in [\alpha_l, \alpha_b]$
- if $r(\alpha) < 0$, then the sequence $\{\alpha_k\} \in [\alpha_b, \alpha_r]$.

Theorem [6]

Under above assumptions with $\alpha_0 = [\alpha_l, \alpha_r]$ the sequence $\{\alpha_k\}$ generated by the fixed point algorithm is such that

- The sequence $\{\Phi_\gamma(\alpha_k)\}$ generated by the function

$$\Phi_\gamma(\alpha) = \frac{F^{1+\gamma}(\alpha)}{\alpha}$$

  is monotonically decreasing.
- The sequence $\{\alpha_k\}$ converges to the local minimizer α_b.

Sketch of the proof

Let us consider the case $r(\alpha_0) > 0$, then the sequence $\{\alpha_k\}$ is increasing and we have $\alpha_k < \alpha_{k+1}$. The function F is concave, thus Lipschitz continuous. Thus $\Phi_\gamma(\alpha)$ is locally Lipschitz continuous and there exists $\Phi'_\gamma(\alpha)$ such that

$$\begin{aligned}\Phi'_\gamma(\alpha) &= \frac{(1+\gamma)F^\gamma(\alpha)F'(\alpha)\alpha - F^{1+\gamma}(\alpha)}{\alpha^2},\\ &= \frac{F^\gamma(\alpha)}{\alpha^2}((1+\gamma)F'(\alpha)\alpha - F(\alpha)) = \frac{F^\gamma(\alpha)}{\alpha^2}(-r(\alpha)) < 0.\end{aligned} \tag{43}$$

In (43) we have used the fact that

$$-r(\alpha) = (1+\gamma)F'(\alpha)\alpha - F(\alpha). \tag{44}$$

Let us check (44). Since

$$\bar{\psi}(\alpha) = F'(\alpha), \quad \bar{\varphi}(\alpha) = F(\alpha) - \alpha F'(\alpha), \tag{45}$$

then using the balancing principle we get

$$r(\alpha) = \bar{\varphi}(\alpha) - \gamma\alpha\bar{\psi}(\alpha) = F(\alpha) - \alpha F'(\alpha) - \gamma\alpha F'(\alpha) = F(\alpha) - \alpha F'(\alpha)(1+\gamma),$$

and thus, (44) holds.

Next, the function $\Phi'_\gamma(\alpha)$ is locally integrable and

$$\Phi_\gamma(\alpha_{k+1}) = \Phi_\gamma(\alpha_k) + \int_{\alpha_k}^{\alpha_{k+1}} \Phi'_\gamma(\alpha)\, d\alpha. \tag{46}$$

Since by (43) we have $\Phi'_\gamma(\alpha) < 0$ then from (46) it follows that $\Phi_\gamma(\alpha_{k+1}) < \Phi_\gamma(\alpha_k)$. Thus, the sequence $\{\Phi_\gamma(\alpha_k)\}$ is monotonically decreasing.

By Lemma [6] there exists a limit $\alpha^* \in [\alpha_l, \alpha_r]$ of the sequence $\{\alpha_k\}$. If $\alpha_k < \alpha_{k+1}$, $\Phi_\gamma(\alpha_{k+1}) < \Phi_\gamma(\alpha_k)$ we have for the finite sequence $\{\alpha_k\}_{k=1}^{k_0}$

$$\lim_{k\to k_0} D^+\Phi_\gamma(\alpha_k) \le \lim_{k\to k_0} \frac{F^\gamma(\alpha_k)}{\alpha_k^2}(-r(\alpha_k)) \le \lim_{k\to k_0} D^-\Phi_\gamma(\alpha_k), \tag{47}$$

then $D^\pm\Phi_\gamma(\alpha_{k_0}) = 0$ since $-r(\alpha_{k_0}) = 0$. By our assumption, this local minimizer $\alpha_{k_0} = \alpha_b$. Now from iterations in the fixed point algorithm we have

$$\frac{1}{\gamma}\frac{F(\alpha_k) - \alpha_k D^- F(\alpha_k)}{D^- F(\alpha_k)} \le \alpha_{k+1} = \frac{1}{\gamma}\frac{\bar{\varphi}(\alpha_k)}{\bar{\psi}(\alpha_k)} \le \frac{1}{\gamma}\frac{F(\alpha_k) - \alpha_k D^+ F(\alpha_k)}{D^+ F(\alpha_k)}. \tag{48}$$

Since $\lim_{k\to\infty} D^\pm F(\alpha_k) = D^- F(\alpha^*)$ then the local minimizer is $\alpha_b = \alpha^*$ given by

$$\alpha^* = \frac{1}{\gamma}\frac{F(\alpha^*) - \alpha^* D^- F(\alpha^*)}{D^- F(\alpha^*)}. \tag{49}$$

□

5 Numerical Experiment

In this section, the performance of the algorithm in Sect. 4.1 on the reconstruction of phantoms from experimentally measured MR data is presented. The MR data acquisition is described in details in the recent paper by authors [3] where the performance

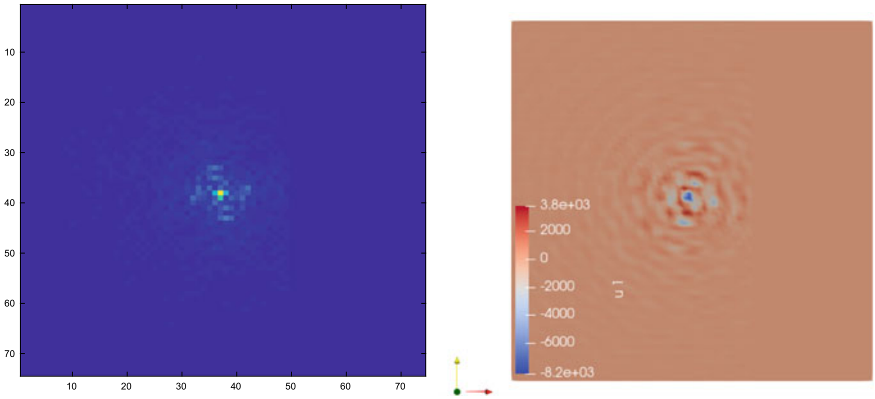


Fig. 1 Two-dimensional slice of data $|u|$ used in tests 1 and 2. Left figure: visualized data $|u|$ on rectangular mesh in matlab. Right figure: visualized data $|u|$ on the finite element mesh K_{h_k}

of applying different interpolation techniques for obtaining the reconstructed images was discussed. For the MR data, a cylindrical plastic phantom of 10 mm diameter is used. This cylindrical phantom contained 8 small cylinders of hardened polymeric bone cement (Osteopal®). The diameters of the 8 small cylinders were all equal to 2 mm. The plastic cylinder was filled with 0.4 mMol/L MnCl$_2$ water solution. Using this cylindrical phantom, a three-dimensional experimental dataset was acquired in Ω_κ. Next, 2D slice of this dataset was selected for for the computations of the function u in (12) in Ω_κ, see Fig. 1.

The 2D computational domain Ω is set of discrete elements $n_x \times n_y$ defined by

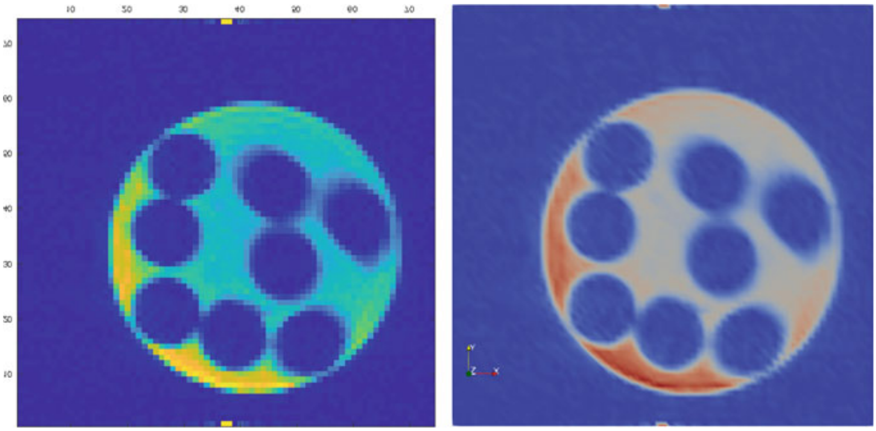


Fig. 2 Reconstructions $|f_h|$ obtained without using regularization parameter in the functional (12): via inverse Fourier transform (IDFT), left figure, and using the finite element method, right figure

$$\Omega = \{(x, y) \in (-37, 36) \times (-37, 36)\},$$

and the domain Ω_k is a set of discrete elements $n_{k_x} \times n_{k_y}$ defined by

$$\Omega_k = \left\{(k_x, k_y) \in (-0.5, 0.5) \times (-0.5, 0.5)\right\}.$$

We choose the mesh size $h = 1$ in Ω and $h = 1/73$ in Ω_k. In both meshes the number of points in x and y direction is the same, $n_x = n_y = n_{k_x} = n_{k_y} = n = 74$.

Results of reconstructions without using the regularization term ($\alpha = 0$) are presented in the Fig. 2. Using this figure, we observe that the reconstruction obtained via inverse discrete Fourier transform (IDFT) (Matlab®'s FFT routines), see left Fig. 2, is less sharp than the result obtained using the finite element method.

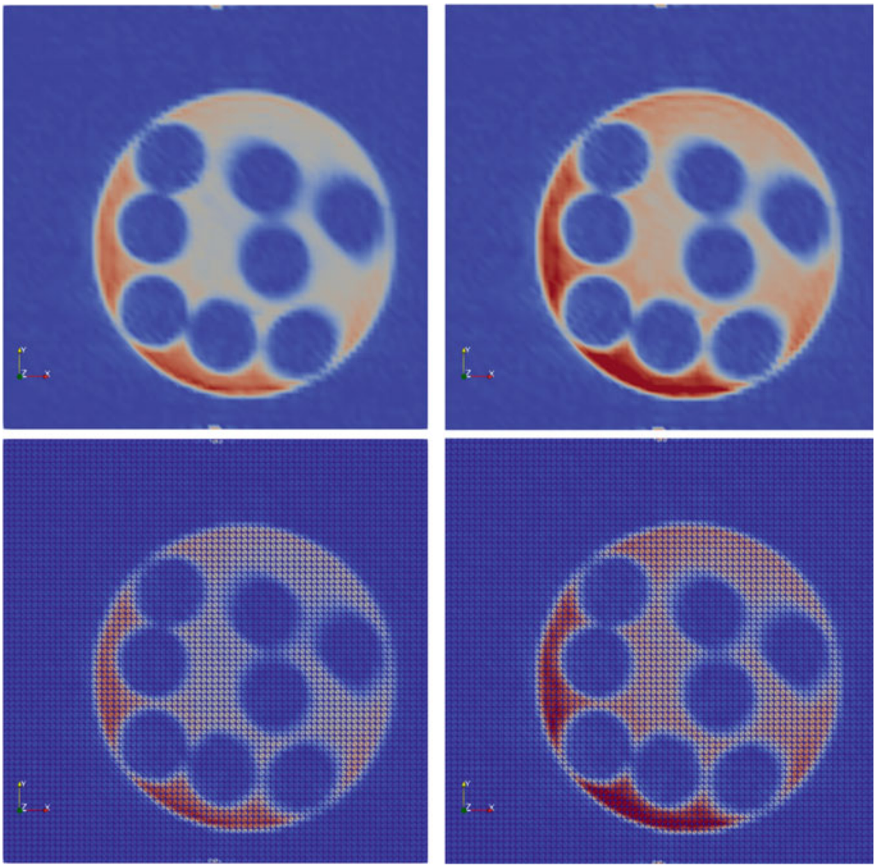


Fig. 3 Reconstructions $|f_h|$ obtained using the finite element method for minimization of functional (12) (left) and functional (21) (right)

Table 1 The optimal values of the computed regularization parameter $\alpha_b = \alpha^*$ using the fixed point algorithm of Sect. 4.1, and corresponding computed residuals $\||Af_h - u|\|_2$ for different regularization functions $\psi(f_h)$

	$\|\|\nabla f_h\|\|_{L_1}$	$\|\|\nabla f_h\|\|_{L_2}^2$	$\|\|f_h +$ $\|\nabla f_h\|\|_{L_1}$	$\|\|f_h +$ $\|\nabla f_h\|\|_{L_2}^2$	No reg.
α^*	0.59	4.73	0.41	3.52	0
$\|\|Af_h - u\|\|_2$	6.44	6.52	6.43	6.48	6.43

Results of reconstruction using finite element method for minimizing of functional (12) (Test 1) and (21) (Test 2), are presented in Fig. 3 and Table 1. For choosing the regularization parameter α, the fixed point algorithm of Sect. 4.1 on the interval $[\alpha_l, \alpha_r] = [0.01, 1]$ with tolerance $\Theta = 10^{-7}$ was used. The optimal values of the computed regularization parameter $\alpha_b = \alpha^*$ are presented in Table 1. Convergence in fixed point algorithm was achieved after 9 and 11 iterations in Tests 1 and 2, respectively.

6 Conclusions

The finite element method for image reconstruction problem in magnetic resonance imaging (MRI) was developed. The balancing principle for a posteriori choice of the regularization parameter was also presented and analyzed. The fixed point iterative algorithm was formulated and tested on the image reconstruction from experimental MR data.

Numerical results compare reconstruction obtained via usual inverse discrete Fourier transform (IDFT) (Matlab®'s FFT routines) and via the finite element method with and without using the regularization terms in the functional to be minimized. Numerical results show effectiveness of using the finite element method to get qualitative MR image reconstruction compared with standard techniques.

Acknowledgements The research of LB was supported by the Swedish Research Council grant VR 2018-03661 and by the French Institute in Sweden, FRÖ program. The research of KN was supported by the French Institute in Sweden, TOR program.

References

1. Basistov, Y.A., Goncharsky, A.V., Lekht, E.E., Cherepashchuk, A.M., Yagola, A.G.: Application of the regularization method for increasing of the radiotelescope resolution power. Astron. zh. **56**(2), 443–449 (1979) (in Russian)
2. Bakushinsky A.B., Kokurin, M.Y., Smirnova, A.: Iterative Methods for Ill-Posed Problems. Walter de Gruyter GmbH&Co. (2011)

3. Beilina, L., Guillot, G., Niinimäki, K.: On finite element method for magnetic resonance imaging. Springer Proc. Mathemat. Statist. **243**, 119–132 (2018)
4. Brown, R.W., Cheng, Y.-C.N., Haacke, E.M., Thompson, M.R., Venkatesan, R.: Magnetic Resonance Imaging: Physical Principles and Sequence Design, 2nd edn. Wiley, Inc. (1999)
5. Hsiao, G.C., Wendland, W.L.: A finite element method for some integral equations of the first kind. J. Mathemat. Anal. Appl. Elsevier **58**, 449–481 (1977)
6. Ito, K., Jin, B.: Inverse Problems: Tikhonov Theory and Algorithms. Series on Applied Mathematics, vol. 22. World Scientific (2015)
7. Johnson, C.: Numerical Solution of Partial Differential Equations by the Finite Element Method. Dover Books on Mathematics (2009)
8. Johnston, P.R., Gulrajani, R.M.: A new method for regularization parameter determination in the inverse problem of electrocardiography. IEEE Trans. Biomed. Eng. **44**(1), 19–39 (1997)
9. Kaipio, J., Somersalo, E.: Statistical and Computational Inverse Problems. Springers Applied Mathematical Sciences, vol. 160. New York (2005)
10. Koshev, N., Beilina, L.: An adaptive finite element method for Fredholm integral equations of the first kind and its verification on experimental data. Central Euro. J. Mathemat. **11**(8), 1489–1509 (2013)
11. Lazarov, R.D., Lu, S., Pereverzev, S.V.: On the balancing principle for some problems of numerical analysis. Numer. Math. **106**(4), 659–689
12. Mathé, P.: The Lepskii principle revised. Inver. Prob. **22**(3), L11–L15 (2006)
13. Mueller, J.L., Siltanen, S.: Linear and Nonlinear Inverse Problems with Practical Applications. SIAM Computational Science & Engineering (2012)
14. Tikhonov, A.N., Arsenin, V.Y.: Solutions of Ill-Posed Problems. Wiley, New-York (1977)
15. Tikhonov, A.N., Leonov, A.S., Yagola, A.G.: Nonlinear Ill-Posed Problems. Chapman & Hall (1998)
16. Tikhonov, A.N., Goncharsky, A.V., Stepanov, V.V., Yagola, A.G.: Numerical Methods for the Solution of Ill-Posed Problems. Kluwer, London (1995)
17. Tikhonov, A.N., Goncharskiy, A.V., Stepanov, V.V., Kochikov, I.V.: Ill-Posed problem in image processing. DAN USSR, Moscow **294**(4), 832–837 (1987) (in Russian)