Springer Proceedings in Mathematics & Statistics

Larisa Beilina
Yury V. Shestopalov *Editors*

Inverse Problems and Large-Scale Computations

Springer

Springer Proceedings in Mathematics & Statistics

Volume 52

For further volumes:
http://www.springer.com/series/10533

Springer Proceedings in Mathematics & Statistics

This book series features volumes composed of select contributions from workshops and conferences in all areas of current research in mathematics and statistics, including OR 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.

Larisa Beilina • Yury V. Shestopalov
Editors

Inverse Problems and Large-Scale Computations

Springer

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

Yury V. Shestopalov
Karlstad University
Karlstad, Sweden

ISSN 2194-1009 ISSN 2194-1017 (electronic)
ISBN 978-3-319-00659-8 ISBN 978-3-319-00660-4 (eBook)
DOI 10.1007/978-3-319-00660-4
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2013945312

Mathematics Subject Classification (2010): 35J05, 35J25, 35J65, 35J66, 58J10, 58J20, 35J20, 45A05, 49N30, 49N45, 65N06, 65N12, 65N20, 65N21, 65N30, 78M22, 78M10

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface for Volume II: Inverse Problems and Large-Scale Computations

In this volume we collected some of the articles presented on the Second Annual Workshop on Inverse Problems and works presented at the Workshop on Large-Scale Modeling. Both workshops were supported by the Swedish Institute, Visby program and co-organized by the Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg and Karlstad University and took place from May 1 to 6, 2012, in Sunne, Sweden.

All papers in this volume highlight the most recent findings in new solution techniques for the inverse problems, analysis of wave propagation in nonlinear media, and some other research areas. The numerical and mathematical methods are developed for powerful supercomputers employing parallel computations and applicable for solving large-scale problems and very large equation systems. The book is maybe the first attempt to unite rigorous mathematical statements and methods and the most advanced to date numerical techniques and algorithms aimed at solution to large-scale problems with an account of uncertain data. The first three papers of this book reflect topics presented at the Second Annual Workshop on Inverse Problems. Construction of new reliable methods for solution of coefficient inverse problems is a very challenging task. An adaptive approximately globally convergent method for a hyperbolic coefficient inverse problem in the case of backscattering data is studied in the paper by M. Asadzadeh and L. Beilina. In this paper authors present also numerical examples for reconstruction of land mines from backscattered data using an adaptive approximate globally convergent algorithm. A mathematical formulation of a coefficient inverse problem and a procedure on how to find the distribution of electrical conductivity and magnetic permeability in the isotropic geological medium from the frequency domain measurements is reported in the paper by V. Gubatenko. A new approximate globally convergent method for the reconstruction of an unknown conductivity function from backscattered electric field measured at the boundary of geological medium under assumptions that dielectric permittivity and magnetic permeability are known functions is developed in the paper by J.B. Malmberg and L. Beilina. Authors presented their method for the typical case of a coefficient inverse problem arising in electrical prospecting.

The remaining papers of this volume reflects variety of subjects discussed on the Workshop on Large-Scale Modeling. One of them is the development of models based on self-consistent statement of propagation, resonance scattering, and generation of waves in nonlinear layered dielectrics, including elaboration and comparison of different numerical algorithms for simulating the field effects at multiple frequencies on the wave scattering and generation. These results are important for creating nonlinear dielectrics with controllable permittivity and various applications in device technology and electronics. The studies are performed by L. Angermann, V. Yatsyk, and D. Valovik. Implementation of the algorithms and techniques developed for the analysis of nonlinear problems using high-performance multi-core and multiprocessor computers is reported in the paper by V. Trofimov, O. Matusevich, I. Shirokov, and M. Fedotov. A. Smirnov with coauthors developed the methods and algorithms that can be used for a wide class of forward problems of electromagnetic field theory when numerical solution by conventional FDTD methods met substantial difficulties due to complex geometries or computational requirements. The solver created on the basis of the proposed approach employs algorithms of parallel computations and is implemented on supercomputers of last generation for solving large-scale problems with characteristic matrix dimensions achieving 1012. Modern analytical and numerical approaches in optical waveguide theory employing spectral theory of operator-valued functions and the integral equation method are considered in the paper by A. Frolov and E. Kartchevskiy. Linear with respect to observations, optimal estimates of solutions and right-hand sides of Maxwell equations with uncertain data (called minimax or guaranteed estimates) are studied by Y. Podlipenko and Y. Shestopalov. The methods for finding these estimates are proposed, estimation errors expressed in terms of solutions to special variational equations are obtained, and the convergence of Galerkin approximations is proved. Development of efficient analytical and numerical solution techniques for the inverse problems occupies a special place in the book. In fact, the determination of electromagnetic parameters of dielectric bodies of complicated structure is an urgent problem because these parameters cannot be directly measured (due to composite character of the material and small size of samples), which leads to the necessity of applying methods of mathematical modeling and numerical solution of the corresponding forward and inverse electromagnetic problems. It is especially important to develop the solution techniques when the inverse problem for bodies of complicated shape is considered in the resonance frequency range. In the paper by Y. Smirnov, Y. Shestopalov, and E. Derevyanchuk a method is developed for the solution to the inverse problem of reconstructing (complex) permittivity of layered dielectrics in a waveguide from the transmission coefficients measured at different frequencies. The method enables in particular solutions in a closed form for one- and multi-sectional diaphragms. Numerical results of calculating (complex) permittivity of the layers are presented and the case of metamaterials is also considered. The results can be applied in nanotechnology, optics, and design of microwave devices. The paper by A. Samokhina and E. Trahtengerts is of special value as far as identification and analysis of large-scale problems with uncertain data in different areas of science and technology are concerned. The authors consider

the algorithms of facilitating the decision-making process when simultaneous or almost simultaneous emergencies take place. In the presence of huge volumes of incoming information effective algorithms of emergency identification are proposed and developed for the analysis and solution of corresponding large-scale problems. The issues related to dynamic computer support are also examined in the paper. Among the book features it should be noted that in many articles a reader finds the whole description of the approach, from the accurate problem statement to numerical results obtained using most powerful to date computer resources and facilities. The intended audience of the book is: university students (knowledge of mathematics: bachelor level and higher), PhD students (specializing in applied mathematics, mathematics, electrical engineering, physics), Dr Sci, researchers, university teachers, RD engineers, and electrical engineers with deeper knowledge and interest in mathematics.

Gothenburg, Sweden — Larisa Beilina
Karlstad, Sweden — Yury V. Shestopalov

Contents

Adaptive Approximate Globally Convergent Algorithm with Backscattered Data 1
Mohammad Asadzadeh and Larisa Beilina

On the Formulation of Inverse Problem in Electrical Prospecting 21
V.P. Gubatenko

Approximate Globally Convergent Algorithm with Applications in Electrical Prospecting 29
John Bondestam Malmberg and Larisa Beilina

Preset Field Approximation and Self-consistent Analysis of the Scattering and Generation of Oscillations by a Layered Structure 41
Lutz Angermann, Vasyl V. Yatsyk, and Mykola V. Yatsyk

***A Posteriori* Estimates for Errors of Functionals on Finite Volume Approximations to Solutions of Elliptic Boundary-Value Problems** 57
Lutz Angermann

Electromagnetic Wave Propagation in Nonlinear Layered Waveguide Structures: Computational Approach to Determine Propagation Constants 69
Dmitry V. Valovik

Performance of Multi-cores and Multiprocessor Computers for Some 3D Problems of Nonlinear Optics and Gaseous Dynamics 91
Vyacheslav A. Trofimov, Olga V. Matusevich, Ivan A. Shirokov, and Mikhail V. Fedotov

Modeling of Electromagnetic Wave Propagation in Guides with Inhomogeneous Dielectric Inclusions 113
Alexander Smirnov, Alexey Semenov, and Yury Shestopalov

Integral Equation Methods in Optical Waveguide Theory 119
Alexander Frolov and Evgeny Kartchevskiy

Guaranteed Estimates of Functionals from Solutions and Data of Interior Maxwell Problems Under Uncertainties 135
Yury Podlipenko and Yury Shestopalov

Permittivity Reconstruction of Layered Dielectrics in a Rectangular Waveguide from the Transmission Coefficients at Different Frequencies 169
Yu. G. Smirnov, Yu. V. Shestopalov, and E. D. Derevyanchuk

Computer Algorithms for Processing Large Information Volumes to Make Decision on Countermeasures for Multiple Emergencies Occurring Simultaneously 183
A.S. Samokhina and E.A. Trahtengerts

System of Nonlinear Boundary-Value Problems and Self-Consistent Analysis of Resonance Scattering and Generation of Oscillations by a Cubically Polarisable Layered Structure 199
Vasyl V. Yatsyk

Adaptive Approximate Globally Convergent Algorithm with Backscattered Data

Mohammad Asadzadeh and Larisa Beilina

Abstract We construct, analyze and implement an approximately globally convergent finite element scheme for a hyperbolic coefficient inverse problem in the case of backscattering data. This extends the computational aspects introduced in Asadzadeh and Beilina (Inv. Probl. 26, 115007, 2010), where using Laplace transformation, the continuous problem is reduced to a nonlinear elliptic equation with a gradient dependent nonlinearity. We investigate the behavior of the nonlinear term and discuss the stability issues as well as optimal a posteriori error bounds, based on an adaptive procedure and due to the maximal available regularity of the exact solution. Numerical implementations justify the efficiency of adaptive a posteriori approach in the globally convergent setting.

1 Introduction

The inverse algorithms have a wide spectrum of application areas raging from mining, detecting oil reservoirs, earth layers, explosives in airports to medical optical imaging, etc. Efficiency of this problem, through *approximate globally convergent approximation (AGCA)* [10], was recently verified on blind imaging of the experimental data that was measured in picoseconds scale regime. In [1] we performed adaptive finite element technique directly inside the AGCA and derived optimal a posteriori error estimates for a finite element approximation of a nonlinear elliptic integro-differential equation. To further improving this efficiency we invoke an adaptivity procedure inside the AGCA algorithm, introduced in [1] for the numerical study of the hyperbolic coefficient inverse problem in two dimensions in the case of the full data collection.

M. Asadzadeh (✉) • L. Beilina
Department of Mathematics, Chalmers University of Technology and the University of Gothenburg, SE-412 96, Gothenburg, Sweden
e-mail: mohammad@chalmers.se; larisa.beilina@chalmers.se

L. Beilina and Y.V. Shestopalov (eds.), *Inverse Problems and Large-Scale Computations*, Springer Proceedings in Mathematics & Statistics 52, DOI 10.1007/978-3-319-00660-4_1,

A direct numerical approach to solve coefficient inverse problems (CIP) is through a minimization procedure for the least square residual functional. This, however, may lead to multiple local minima for the functionals. To avoid such an obstacle, in [9] a convexification algorithm was introduced for solution of the one-dimensional CIP in imaging electromagnetic frequency. This algorithm was further extended in [8] to higher dimensions with applications in diffusive optical mammography. Convexification is the origin of the AGCA methods. Some modified approaches to the AGCA algorithms were introduced in [2–5] and summarized in [6], where a layer-stripping procedure was performed with respect to the pseudo-frequency rather than the spatial variable which is the case in the convexification. The Carleman weight function in [2–6] depends on the pseudo-frequency and not on the spatial variable, as in [8, 9]. These new approaches contribute to improved stability in the globally convergent reconstruction algorithm.

An alternative approach to solve CIP is a synthesis of an AGCA method and a strongly converging, however, local scheme such as the adaptive finite element method. In [3, 5] it was shown that the AGCA method provides a good initial guess for the locally convergent adaptive method. A first application of these results for the acoustic wave equation shows a good performance [3–5]. To compare with [3–5], the present work introduces extensive implementation results for a new such combination. Here adaptivity is performed directly inside the AGCA algorithm in the case when we have only backscattered data at the observation boundary.

A concise description of the theoretical procedure is as follows: A Laplace transformation in time converts the model problem to a convection-diffusion-type equation. The finite elements perform more accurately for elliptic and parabolic equations than the hyperbolic ones. Hence, the study of the CIP through combining a time transformation followed by a finite element procedure not only reduces the dimension of the underlying problem but also shifts the equation to a more desirable one from the finite element point of view. To our knowledge, the combination of the AGCA method, for a nonlinear elliptic problem and a posteriori procedure, using adaptive algorithm, is not considered elsewhere.

The paper is organized as follows: In Sect. 2 we formulate both forward and inverse problems and transfer the inverse problem to a Dirichlet boundary value problem for a nonlinear integro-differential equation with a removed unknown coefficient. In Sect. 3 we introduce the layer-stripping procedure with respect to $s > 0$, the parameter of the Laplace transform in the original hyperbolic PDE. We point out that here we do not use the inverse Laplace transform, since approximations for the unknown coefficient are obtained in the "Laplace's domain". In Sect. 4 we describe a finite element method, state bounds for coefficients (derived in [1]), and formulate a corresponding dual problem. Section 5 is devoted to derivation of bounds for the nonlinear operator and a priori error estimates. In Sect. 6 we develop reliable and efficient a posteriori error estimates, for the full problem. In Sect. 7 we introduce a new adaptive globally convergent algorithm based on a posteriori error estimate of Sect. 6. Finally, in our concluding Sect. 8 we present the results of reconstruction of the function in two dimensions based on adaptive AGCA algorithm.

2 The Forward and Inverse Problems

Consider the Cauchy problem for the hyperbolic equation

$$c(x)u_{tt} = \Delta u, \text{ in } \mathbf{R}^n \times (0,\infty), n = 2,3, u(x,0) = 0, u_t(x,0) = \delta(x - x_0). \quad (1)$$

Equation (1) describes, e.g., propagation of acoustic and electromagnetic waves.

Let $\Omega \subset \mathbb{R}^n, n = 2,3$ be a convex bounded domain with the boundary $\partial\Omega \in C^n, n = 2,3$. We shall assume that $c(x)$ satisfies the following conditions:

$$\begin{cases} c(x) \in C^2(\mathbf{R}^n), \quad 2d_1 \le c(x) \le 2d_2, \quad d_1 > 0, d_2 > 0, \\ c(x) = 2d_1, \quad \text{for } x \in \mathbf{R}^n \setminus \Omega, \quad \Omega \subset \mathbf{R}^n, \quad n = 2,3, \end{cases} \quad (2)$$

where, d_1 and d_2 are given bounds for the function $c(x)$,

In this work we consider the case of the *backscattered* data, or such data which are given only at a part of the boundary of the computational domain. Let us define our computational domain Ω with the backscattered boundary Γ:

$$\Omega \subset \{x = (x_1, x_2, x_3) : x_3 > 0\},$$
$$\Gamma = \partial\Omega \cap \{x_3 = 0\} \neq \varnothing.$$

In our computations we will consider the case when the wave field is initialized by the incident plane wave propagating along the positive direction of the x_3-axis in the half space $\{x_3 < 0\}$ and "falling" on the half space $\{x_3 > 0\}$. Numerical tests in Sect. 8 are performed for the given function g_0 and g_1, where $u(x,t) = g_1(x,t)$ at Γ and $u(x,t) = g_0(x,t)$ at $\partial\Omega \setminus \Gamma$, with $u(x,t)$ satisfying the Cauchy problem

$$u_{tt} - \Delta u = 0, \quad \text{in } \Omega \times (0,\infty), \quad (3)$$
$$u(x,0) = 0, \ u_t(x,0) = f(x), \text{ in } \Omega.$$

Hence, in these tests we set

$$u(x,t) := g_2(x,t) = \begin{cases} g_1(x,t), (x,t) \in \Gamma \times (0,\infty), \\ g_0, (x,t) \in (\partial\Omega \setminus \Gamma) \times (0,\infty) \end{cases} \quad (4)$$

and consider the following inverse problem:

Inverse Problem with Backscattered Data (IPB). *Suppose that the coefficient $c(x)$ satisfies conditions (2) and it is unknown in the domain Ω. Determine the function $c(x)$ for $x \in \Omega$, assuming that the function $g_2(x,t)$ in Eq. (4) is known for a single direction of the incident plane wave propagating along the positive direction of x_3-axis in the half space $\{x_3 < 0\}$ and falling on the half space $\{x_3 > 0\}$*

We note that our formulation of IPB is for the case of a plane wave. In the case of problem (1), with a Dirac delta function as initial data, the formulation of inverse problem IPB is similar. In this case we should replace the wording "for a

single direction of the incident plane wave propagating along the positive direction of x_3-axis in the half space $\{x_3 < 0\}$ and falling on the half space $\{x_3 > 0\}$", by the expression "for a single source position $x_0 \in \{x_3 < 0\}$".

Next, we use the Laplace transform

$$U(x,s) = \int_0^\infty u(x,t)e^{-st}dt, \qquad \text{for } s > \underline{s} > 0, \tag{5}$$

where $\underline{s}$ is the *pseudo-frequency* constant. Recall that it suffices to choose $\underline{s}$ such that the integral (5) and its first partial derivatives in x and t converge. Then U satisfies

$$\begin{cases} \Delta U - s^2 c(x) U = -\delta(x - x_0)\, c(x_0), & \forall s \geq \underline{s} > 0, \\ \lim_{|x|\to\infty} U(x,s) = 0, & \forall s \geq \underline{s} > 0. \end{cases} \tag{6}$$

For every $s \geq \underline{s}$, the Eq. (6) possesses a positive, unique solution U.

2.1 The Nonlinear Integro-Differential Equation with Eliminated Unknown Coefficient

Introducing the function $v = \ln U$, since $x_0 \notin \overline{\Omega}$, then Eq. (6) yields

$$\Delta v + |\nabla v|^2 = s^2 c(x), \qquad \text{in } \Omega, \tag{7}$$

$$v(x,s) = \ln G(x,s), \ \forall (x,s) \in \partial\Omega \times [\underline{s}, \overline{s}], \tag{8}$$

where $G(x,s)$ is the Laplace transform of the data function $g(x,t)$. To single out the unknown coefficient $c(x)$ in Eq. (7), we introduce a new function

$$H(x,s) = \frac{v}{s^2}. \tag{9}$$

Assuming certain regularity conditions [2], it follows that H satisfies

$$\Delta H + s^2 |\nabla H|^2 = c(x). \tag{10}$$

Next let

$$q(x,s) = \partial_s H(x,s), \tag{11}$$

then using Eq. (11)

$$H(x,s) = -\int_s^\infty q(x,\tau)\, d\tau := -\int_s^{\overline{s}} q(x,\tau)\, d\tau + W(x,\overline{s}), \tag{12}$$

where $\bar{s} > s_0$ is a large number and

$$W(x,\bar{s}) \approx H(x,\bar{s}) = \frac{\ln U(x,\bar{s})}{\bar{s}^2}. \tag{13}$$

$W(x,\bar{s})$ is known as the *tail function*. To determine W we need to choose the parameter $\bar{s}$ numerically. We include W either on the right hand side in iteration steps as data, or study it as an unknown in a coupled system of equations.

Differentiating Eq. (10) with respect to s, from Eqs. (12) and (13), we obtain the following nonlinear integro-differential equation for $q = q(x,s)$,

$$\Delta q - 2s^2 \nabla q \cdot \int_s^{\bar{s}} \nabla q(x,\tau)\, d\tau + 2s \left[\int_s^{\bar{s}} \nabla q(x,\tau)\, d\tau \right]^2 \tag{14}$$
$$+ 2s^2 \nabla q \nabla W - 2s \nabla W \cdot \int_s^{\bar{s}} \nabla q(x,\tau)\, d\tau + 2s(\nabla W)^2 = 0.$$

By Eqs. (8), (9) and (11) we may impose the following Dirichlet boundary condition

$$q(x,s) = \psi(x,s), \ \forall (x,s) \in \partial\Omega \times [\underline{s},\bar{s}], \tag{15}$$

where ψ satisfies

$$\psi\Big(x,s\Big) = \frac{G_s}{Gs^2} - \frac{2\ln G}{s^3}. \tag{16}$$

Suppose that $D_x^\alpha q$, $|\alpha| \leq 2$ are already approximated. Then the coefficient $c(x)$ can be, approximately, determined using Eq. (10), where H is given by Eq. (12), which requires an initial guess for W as well.

3 A Sequence of Elliptic Dirichlet Boundary Value Problems

We approximate $q(x,s)$ with a piecewise constant function with respect to s. Assume a partition $\underline{s} = s_N < s_{N-1} < \ldots < s_1 < s_0 = \bar{s}$, $s_{n-1} - s_n = k$ of $[\underline{s},\bar{s}]$ with a sufficiently small and uniform step size k such that $q(x,s) = q_n(x)$ for $s \in (s_n, s_{n-1})$. Hence,

$$\int_s^{\bar{s}} \nabla q(x,\tau) d\tau = (s_{n-1} - s)\nabla q_n(x) + k \sum_{j=1}^{n-1} \nabla q_j(x), \quad s \in (s_n, s_{n-1}). \tag{17}$$

We approximate the boundary condition (15) as being piecewise constant on s,

$$q_n(x) = \overline{q}_n(x), \quad x \in \partial\Omega, \quad j = 1,\ldots,n, \tag{18}$$

where

$$\overline{f}_n(x) = \frac{1}{k}\int_{s_n}^{s_{n-1}} f(x,s)\,ds. \tag{19}$$

On each subinterval $(s_n, s_{n-1}], n \geq 1$, we assume that $q_j(x)$, $j = 1,\ldots,n-1$ are known. In this way, for each n, $n = 1,\ldots,N$, we obtain an approximate equation for $q_n(x)$. Now we insert Eq. (17) in Eq. (14) and multiply the resulting equation by the Carleman weight function (CWF):

$$C_{n,\lambda}(s) = e^{\lambda(s-s_{n-1})}, \qquad s \in (s_n, s_{n-1}], \qquad \lambda >> 1, \tag{20}$$

and integrate over $s \in (s_n, s_{n-1}]$ (see Theorem 6.1 [2]). We obtain for $n = 1,\ldots,N$,

$$\begin{aligned}\mathscr{L}_n\left(q_n, W_n\right) - \varepsilon q_n =: \Delta q_n - A_{1,n}\left(k\sum_{i=1}^{n-1}\nabla q_i\right)\nabla q_n + A_{1n}\nabla q_n \nabla W_n - \varepsilon q_n \\ \approx 2\frac{I_{1,n}}{I_0}\left(\nabla q_n\right)^2 - A_{2,n}k^2\Big(\sum_{i=1}^{n-1}\nabla q_i(x)\Big)^2 + 2A_{2,n}\nabla W_n\Big(k\sum_{i=1}^{n-1}\nabla q_i\Big) - A_{2,n}\Big(\nabla W_n\Big)^2.\end{aligned} \tag{21}$$

The term $-\varepsilon q_n$ is added for regularizing purpose. The coefficients are computed as:

$$\begin{aligned} I_0 : \quad &= \int_{s_n}^{s_{n-1}} C_{n,\lambda}(s)\,ds, \qquad I_{1,n} := \int_{s_n}^{s_{n-1}} s(s_{n-1}-s)[s-(s_{n-1}-s)] \\ A_{1,n} : &= \frac{2}{I_0}\int_{s_n}^{s_{n-1}} s[s-2(s_{n-1}-s)]C_{n,\lambda}(s)\,ds, \qquad A_{2,n} := \frac{2}{I_0}\int_{s_n}^{s_{n-1}} sC_{n,\lambda}(s)\,ds. \end{aligned}$$

Thus we have the Dirichlet boundary value problem (21), with the boundary data (18). In this system the tail function W is also unknown. Observe that

$$\frac{|I_{1,n}(\lambda,k)|}{I_0(\lambda,k)} \leq \frac{4\bar{s}^2}{\lambda}, \qquad \text{for } \min(\lambda k, \bar{s}) \geq 1. \tag{22}$$

Therefore taking $\lambda >> 1$ we mitigate the influence of the nonlinear term with $(\nabla q_n)^2$ in Eq. (21), which enables us to solve a linear problem on each iterative step.

4 A Finite Element Discretization

We approximate the solution for Eq. (21) by a finite element method with continuous piecewise linear functions on *a partially structured mesh* in space and implement resulting scheme using a hybrid code. More specifically, we decompose the computational domain G into $\Omega \subset G$ and $\Omega^c = G \setminus \Omega$ and discretize Ω by an unstructured mesh and Ω^c by a quasi-uniform mesh. In Ω, for each n, we use a

partition $\mathscr{T}_{n,h} = \{K\}$. Here $h = h(x)$ denotes a piecewise constant mesh function $h = h(x)$ representing the diameter of the element K containing x, and $(\cdot,\cdot)$ and $\|\cdot\|$ denote the L_2-inner product and norm, respectively.

Choosing $c(x) = 1$ for $x \in \Omega^c$, given $g(x,t) = u\,|_{\partial\Omega}$, we can uniquely determine the function $u(x,t)$ as the solution of the boundary value problem for Eq. (1) with boundary conditions on both boundaries ∂G and $\partial\Omega$. Next, using Laplace transform of $u(x,t)$, Eqs. (9) and (11) one can uniquely determine $\tilde{q}(x)$,

$$\tilde{q}(x) =: \frac{\partial q}{\partial \mathbf{n}}\Big|_{\partial\Omega}, \tag{23}$$

here $\mathbf{n}$ is the outward unit normal to the boundary $\partial\Omega$ at the point $x \in \partial\Omega$. In our computations the functions $p(x,t), \tilde{q}(x)$, and $g(x,t)$ are calculated from the solution of the forward problem (21) with the exact value of the coefficient $c(x)$. A variational formulation for Eq. (21) is for $n = 1,\ldots,N$; find $V_n,\ q_n \in H^1(\Omega)$ such that

$$\begin{aligned}\mathscr{F}(q_n,V_n;\varphi) =: {} & (\nabla q_n,\nabla\varphi) + (A_{1,n}(k\sum_{i=1}^{n-1}\nabla q_i)\nabla q_n,\varphi) - (A_{1n}\nabla q_n\nabla W_n,\varphi) + (\varepsilon q_n,\varphi)\\ & +(2\tfrac{I_{1,n}}{I_0}(\nabla q_n)^2,\varphi) - (A_{2,n}k^2(\sum_{i=1}^{n-1}\nabla q_i(x))^2,\varphi) + (2A_{2,n}\nabla W_n(k\sum_{i=1}^{n-1}\nabla q_i),\varphi)\\ & -(A_{2,n}(\nabla W_n)^2,\varphi) \approx (\tilde{q}_n,\varphi)_{\partial\Omega}, \quad \forall\varphi \in H^1(\Omega).\end{aligned} \tag{24}$$

To formulate a finite element method for Eq. (21), we introduce the trial space $V^q_{n,h}$,

$$V^q_{n,h} := \{v_n \in H^1(\Omega) : v_n|_K \in P_1(K),\ \partial_{\mathbf{n}} v_n|_{\partial\Omega} = \tilde{q}_{n,h},\ \forall K \in \mathscr{T}_{n,h}\},$$

where $n = 1,\ldots,N$, $P_1(K)$ denotes the set of linear functions on K, and $\tilde{q}_{n,h}$ is an approximation for $\tilde{q}(x)$. We also introduce the test function space $V_{n,h}$ defined as

$$V_{n,h} := \{v_n : v_n \text{ is continuous on } \Omega, \text{ and } w_n|_K \in P_1(K), \quad \forall K \in \mathscr{T}_{n,h}\}.$$

$V_{n,h}$ and $V^q_{n,h} \subset H^1(\Omega)$. The finite element for Eq. (21) is formulated as for $n = 1,\ldots,N$, find $q_{n,h}$ and $W_{n,h} \in V^q_{n,h}$, approximations of q_n and W_n, respectively, such that

$$\mathscr{F}(q_{n,h},W_{n,h};\varphi) \approx (\tilde{q}_{n,h},\varphi)_{\partial\Omega}, \quad \forall\varphi \in V_{n,h}. \tag{25}$$

Subtracting Eq. (25) from Eq. (24) we get the classical *Galerkin orthogonality:*

$$\mathscr{F}(q_n,W_n;\varphi) - \mathscr{F}(q_{n,h},W_{n,h};\varphi) \approx 0, \quad \forall\varphi \in V_{n,h}. \tag{26}$$

Now, we introduce the residual, $\mathscr{R}_n := \mathscr{R}_n(q_{nh},W_{n,h})$, for a discrete solution for Eq. (21) as follows: for $n = 1,\ldots,N$; find $q_{nh}, W_{n,h} \in V^q_{n,h}$ such that

$$
\begin{aligned}
&-\Delta_h q_{nh} + A_{1,n}\left(k\sum_{i=1}^{n-1}\nabla q_{ih}\right)\nabla q_{nh} - A_{1n}\nabla q_{nh}\nabla W_{nh} + \varepsilon q_{nh} + 2\tfrac{I_{1,n}}{I_0}(q_{nh})^2\\
&-A_{2,n}k^2\left(\sum_{i=1}^{n-1}\nabla q_{ih}(x)\right)^2 + 2A_{2,n}\nabla W_{nh}\left(k\sum_{i=1}^{n-1}\nabla q_{ih}\right) - A_{2,n}(\nabla W_{nh})^2 := \mathscr{R}_n,\\
&\quad q_{nh}|_{\partial\Omega} = \tilde{q},
\end{aligned}
\tag{27}
$$

where $\Delta_h q_{nh}$ denotes the discrete Laplacian defined by

$$
(\Delta_h q_{nh}, \eta) = (\nabla q_{nh}, \nabla\eta), \qquad \forall \eta \in W_{n,h}. \tag{28}
$$

Let now $e_{n,h} = q_n - q_{n,h}$, $n = 1,\ldots,N$; then a modified form of the Galerkin orthogonality Eq. (26) yields the strong error representation formula:

$$
\begin{aligned}
&-\Delta_h e_{n,h} + I_1\nabla e_{n,h} + \varepsilon e_{n,h} + 2\frac{I_{1,n}}{I_0}\left[(\nabla q_n)^2 - (\nabla q_{n,h})^2\right]\\
&\quad + I_2\cdot(k\sum_{i=1}^{n-1}\nabla e_{i,h}) + I_3\cdot\nabla\Theta_n = -\mathscr{R}_n.
\end{aligned}
\tag{29}
$$

For each interval $[s_n, s_{n-1})$, we rewrite Eq. (29) (we suppress n) and consider the equation

$$
\begin{aligned}
\Gamma e &:= -\Delta e + C_1\nabla e + \varepsilon e + \delta\Lambda e = -C_2\left(k\sum_{i=1}^{n-1}\nabla e_i\right) - \mathscr{R} - C_3\nabla\Theta\\
e|_{\partial\Omega} &= 0,
\end{aligned}
\tag{30}
$$

where C_j, $j = 1,2,3$ are corresponding to the spatially continuous versions of I_j:s, $\delta := I_{1,n}/I_0$ and Λ, the nonlinear term, is defined by

$$
\Lambda e := |\nabla q|^2 - |\nabla q_h|^2. \tag{31}
$$

In Eq. (30) the error in W is included in the Θ-term and the residual term $\mathscr{R}$ satisfies

$$
(\mathscr{R}, \varphi) \approx 0, \qquad \forall \varphi \in V_{n,h}. \tag{32}
$$

5 Bounds for the Nonlinear Operator Λ and A Priori Estimates

Below we derive a bound for Λ, using $f(q) = |\nabla q|^2$, $0 < \theta < 1$, and

$$
\begin{aligned}
\mathscr{D}f(\theta q + (1-\theta)q_h) &= \mathscr{D}\left(|\nabla(\theta q + (1-\theta)q_h)|^2\right)\\
&= 2\left(|\nabla(\theta q + (1-\theta)q_h)|\right)\cdot\left(\mathscr{D}|\nabla(\theta q + (1-\theta)q_h)|\right),
\end{aligned}
\tag{33}
$$

where $\mathscr{D}f$ is given in the Taylor expansion of $f(q_h)$ about q, viz.,

$$f(q_h) = f(q) + (q_h - q)\mathscr{D}f(\theta q + (1-\theta)q_h). \tag{34}$$

We may write Λe in a compact form as

$$\Lambda e = 2e\Big(|\theta\nabla e + \nabla q_h|\Big) \cdot \Big(\mathscr{D}|\nabla(\theta q + (1-\theta)q_h)|\Big). \tag{35}$$

5.1 The Dual Problem for a Linearized Approach

Here, we sketch a framework for the dual approach for a *linear/linearized version* of Eq. (30). To begin with, we assume that Λ is a linear operator and let

$$\Gamma^\star\varphi := -\Delta\varphi - C_1\nabla\varphi + \varepsilon\varphi + \delta\Lambda^\star\varphi = e, \qquad n = 1,\dots,N, \qquad \varphi|_{\partial\Omega} = 0, \tag{36}$$

with $\Gamma^\star$ and $\Lambda^\star$ being the adjoints of Γ and Λ, respectively. By Eq. (30) we have that

$$\|e\|^2_{L_2(\Omega)} = (e, \Gamma^\star\varphi) = (\Gamma e, \varphi) = -(\tilde{\mathscr{R}}, \varphi). \tag{37}$$

The identity (37) is known as *the error representation formula*. Using the identities

$$-(\chi, \varphi - P_h\varphi) = -(\chi - P_h\chi, \varphi - P_h\varphi), \tag{38}$$

for $\chi = \mathscr{R}$, $\chi = C_2\sum_{i=1}^{n-1}\nabla e_i$, or $\chi = C_3\nabla\Theta$, where $P_h : L_2(\Omega) \to W_{n,h}$ is the $L_2(\Omega)$-projection, and we have used the orthogonality $\mathscr{R} \perp W_{n,h}$, and the strong stability estimates for the dual problem, we get from Eq. (37) (see [1] for details) that

$$\|e\|_{L_2(\Omega)} \le C_s C_i \left\|h^2(\tilde{\mathscr{R}} - P_h\tilde{\mathscr{R}})\right\| \le CC_sC_i \left\|h^2(\mathscr{R} - P_h\mathscr{R})\right\|, \tag{39}$$

where C_i and C_s are interpolation and stability constants, respectively. Recalling Eq. (35)

$$(\Lambda^\star\varphi, e) = (\varphi, \Lambda e) = 2\Big(\varphi,\ \Big[|\theta\nabla e + \nabla q_n|\Big] \cdot \Big[\mathscr{D}|\nabla(\theta q + (1-\theta)q_h)|\Big]e\Big). \tag{40}$$

For piecewise linear approximation, successive use of Hölder inequality yields

$$|(\Lambda^\star\varphi, e)| \le C\|\varphi\|\,\|e\|\,\|q\|_{W^2_\infty}\Big(\|q_h\|_{W^1_\infty} + \|e\|_{W^1_\infty}\Big). \tag{41}$$

Thus we get the following estimate for the nonlinear operator Λ:

$$\|\Lambda\| \le \|q\|_{W^2_\infty}\Big(\|q_h\|_{W^1_\infty} + \|e\|_{W^1_\infty}\Big).$$

Theorem 1 (An a priori error bound). *Let $q_n \in W_2^2(\Omega)$ and q_{nh}, be the solutions for Eqs. (24) and (25), respectively. Then for a piecewise linear finite element approximation error $e_n = q_n - q_{n,h}$ we have (see [1]) that*

$$\|e_n\| \le Ch\|q_n\|_{W_2^2} = \mathscr{O}(h). \tag{42}$$

6 A Posteriori Error Estimation

The a posteriori error analysis is based on representing the error in terms of the solution φ of the dual problem, related to Eq. (21). We recall the problem (30) and write the dual problem for all $[s_n, s_{n-1})$, $n = 1, \dots, N$, as

$$-\Delta\varphi - C_1\nabla\varphi + \varepsilon\varphi + \delta\Lambda^\star\varphi + \delta|\nabla\varphi_h|^2 + \tilde{C}_{\varphi,\Theta} = \psi, \qquad \varphi|_{\partial\Omega} = 0, \tag{43}$$

where $\tilde{C}_{\varphi,\Theta} := C_2 k\sum_{i=1}^{n-1}\nabla\varphi_i + C_3\nabla\Theta$ is assumed to be known from the previous iteration steps, and $\Theta = \Theta_n = W_h - W_{n,h}$. We assume that $\Theta \in H^1_{loc}$ and $\varphi_h \in W^{1,4}_{loc}$. Thus, we wish to control the quantity (e,ψ) with $e = q - q_h$ in Ω, where $\psi \in [L^2(\Omega)]^3$ is given. For approximations of spectral order > 1, (for linear approximation the J_5-term below will vanish) we may write

$$\begin{aligned}(\psi,e) \approx &-(\Delta\varphi,e) - (C_1\nabla\varphi,e) + (\varepsilon\varphi,e) - \delta(|\nabla q_h|^2\mathscr{D}(\varphi),e)\\ &+\delta(\mathscr{D}(|\nabla q_h|^2\varphi),e) + \delta(|\nabla\varphi_h|^2,e) + (\tilde{C}_{\varphi,\Theta},e) =: \textstyle\sum_{k=1}^7 J_k.\end{aligned} \tag{44}$$

Due to the limited regularity of the approximate solution $q_{n,h}$, the scalar products $I_j,\ j = 1,\dots,7$, involving $e = q_n - q_{n,h}$, should be performed elementwise: $(f,g) := \sum_K (f,g)_K$. This will introduce accumulative sum of the normal derivatives over enter-element boundaries. Taking into account these boundary terms, by repeated use of Green's formula, we can recompute each $J_j,\ j = 1,\dots,7$, separately. In this way, finally we obtain the following error representation inequality:

Lemma 1. *Let φ be the solution of the dual problem (43), q that of Eq. (24), and q_h the FEM solution of Eq. (25). Then the following error representation inequality holds true:*

$$\left|(\psi,e)\right| \le (|\tilde{\mathscr{R}}_1|,|\sigma|) + (|\tilde{\mathscr{R}}_2|,|\sigma|) + C_3(|\nabla\Theta|,|e|) + \delta(|\nabla\varphi_h|^2,|e|), \tag{45}$$

where the residuals are defined as

$$\tilde{\mathscr{R}}_1 =: \Delta_h e - C_1\nabla e - \varepsilon e - \delta\Lambda e - C_2 k\sum_{i=1}^{n-1}\nabla e_i, \quad \tilde{\mathscr{R}}_2 = \max_{S\subset\partial K} h_K^{-1}\left|\left[\partial_s q_h\right]\right|, \tag{46}$$

and interpolation error is

$$\sigma = h_K \left[\partial_{\mathbf{n}} \varphi_h\right]. \tag{47}$$

Now we use, elementwise, Hölder inequality and Let $\psi = e$ to obtain the following a posteriori error estimate:

Theorem 2. *Let φ be the solution of the dual problem (43), q the solution of Eq. (24), and q_h the FEM solution of Eq. (25). Then there is a constant C, independent of Ω and h, such that for $\psi - \delta |\nabla \varphi_h|^2 = e$ the following a posteriori error estimate holds:*

$$\|e\|^2 \leq Ch \left[\left(\|\mathscr{R}_1\|_{L_2(\Omega)} + \|\mathscr{R}_2\|_{L_2(\Omega)} \right) \|\tilde{\sigma}\|_{L_2(\Omega)} + h|C_3|^2 \right], \tag{48}$$

where $h = \max_K(h_K)$, $\mathscr{R}_1 = \tilde{\mathscr{R}}_1(q_h) = \Delta_h q_h + C_1 \nabla q_h - \varepsilon q_h - \delta \Lambda q_h - C_2 k \sum_{i=1}^{n-1} \nabla q_{h,i}$, $\mathscr{R}_2 = \tilde{\mathscr{R}}_2$ is given in Eq. (46), $\tilde{\sigma} = \left[\partial_{\mathbf{n}} \varphi_h\right]$, and $\mathscr{R}_3\Big|_K := |\nabla \Theta|\Big|_K$ can be estimated as $\|\mathscr{R}_3\|^2_{L_2(\Omega)} \approx C_\Omega \xi^2 \sim Ch^2$, whereas choosing $\psi := e + \delta |\nabla \varphi_h|^2 + C_3 |\nabla \Theta|$ yields

$$\|e\|^2 \leq Ch \left(\|\mathscr{R}_1\|_{L_2(\Omega)} + \|\mathscr{R}_2\|_{L_2(\Omega)} \right) \|\tilde{\sigma}\|_{L_2(\Omega)}. \tag{49}$$

7 The Adaptive Approximate Globally Convergent Algorithm

In this section we present our adaptive globally convergent algorithm, where we use Theorem 2 which states that the error, between the exact and approximate solution for the functions q_n of the Eq. (21), depends on the residuals given by Eq. (46). However, in the case of using continuous piecewise linear finite element approximation of functions q_n, only the first residual $\tilde{\mathscr{R}}_1$ will appear. To calculate it we should find an approximate solution q_n of the Eq. (21) on every mesh. We get q_n as $q_n = \lim_{k \to \infty} q_n^k$, where k is the number of iterations with respect to the tail function $W_n(x, \bar{s})$.

To solve Eq. (21) on a new refined mesh, we first linearly interpolate the function $\bar{\psi}_n$, given by Eq. (15), for each pseudo-frequency interval $[s_n, s_{n-1})$. Then, on every mesh we compute approximations c_n of $c(x)$ using variational formulation of the Eq. (7); see [6] for full details. Thus, we can explicitly compute the function c_n on every frequency interval (s_n, s_{n-1}) through the finite element formulation.

We denote the stopping number k (on which these iterations are stopped) by m_n.

7.1 An Approximate Globally Convergent Algorithm

Below, we briefly describe a globally convergent algorithm of [2,5,6] which we use in our computations and in the adaptive globally convergent algorithm.

Step 0. $n_1, n \geq 1$. Stage 1: iterate with respect to the nonlinear term. Assume that the functions $q_1, \ldots, q_{n-1}, q_{n,1}^0 (:= q_{n-1}) \in C^{2+\alpha}(\overline{\Omega})$ and the tail function $V_{n,0}(x,\overline{s}) \in C^{2+\alpha}(\overline{\Omega})$ are already constructed. Then, we solve, iteratively, the following Dirichlet boundary value problems: For $k = 1, 2, \ldots$, find $q_{n,1}$ such that

$$\Delta q_{n,1}^k - A_{1n}\left(h\sum_{j=1}^{n-1}\nabla q_j\right)\cdot\nabla q_{n,1}^k - \varepsilon q_{n,1}^k + A_{1n}\nabla q_{n,1}^k\cdot\nabla W_{n,0} \tag{50}$$

$$= 2\frac{I_{1n}}{I_0}\left(\nabla q_{n,1}^{k-1}\right)^2 - A_{2n}h^2\left(\sum_{j=1}^{n-1}\nabla q_j(x)\right)^2 \tag{51}$$

$$+2A_{2n}\nabla W_{n,0}\cdot\left(h\sum_{j=1}^{n-1}\nabla q_j(x)\right) - A_{2n}\left(\nabla W_{n,0}\right)^2, \tag{52}$$

$$q_{n,1}^k = \overline{\psi}_n(x), \qquad x \in \partial\Omega. \tag{53}$$

As a result, we obtain the function $q_{n,1} := \lim_{k\to\infty} q_{n,1}^k$ in the $C^{2+\alpha}(\overline{\Omega})$.

Step 1. Compute $c_{n,1}$ via backwards calculations using finite element formulation of Eq. (7); see Chap. 3 of [6] for details.

Step 2. Solve the hyperbolic forward problem with $c_n(x) := c_{n,1}(x)$; calculate the Laplace transform and the function $U_{n,1}(x,\overline{s})$.

Step 3. Find a new approximation for the tail function

$$W_{n,1}(x) = \frac{\ln U_{n,1}(x,\overline{s})}{\overline{s}^2}. \tag{54}$$

Step 4. n_i, $i \geq 2$. We now iterate with respect to the tails Eq. (54). Suppose that functions $q_{n,i-1}, W_{n,i-1}(x,\overline{s}) \in C^{2+\alpha}\left(\overline{\Omega}\right)$ are already constructed.

Step 5. Solve the boundary value probleme

$$\Delta q_{n,i} - A_{1n}\left(h\sum_{j=1}^{n-1}\nabla q_j\right)\cdot\nabla q_{n,i} - \kappa q_{n,i} + A_{1n}\nabla q_{n,i}\cdot\nabla W_{n,i-1} \tag{55}$$

$$= 2\frac{I_{1n}}{I_0}\left(\nabla q_{n,i-1}\right)^2 - A_{2n}h^2\left(\sum_{j=1}^{n-1}\nabla q_j(x)\right)^2 \tag{56}$$

$$+2A_{2n}\nabla W_{n,i-1}\cdot\left(h\sum_{j=1}^{n-1}\nabla q_j(x)\right) - A_{2n}\left(\nabla W_{n,i-1}\right)^2, \tag{57}$$

$$q_{n,i}(x) = \overline{\psi}_n(x), \qquad x \in \partial\Omega. \tag{58}$$

Step 6. Compute $c_{n,i}$ by backwards calculations using finite element formulation of Eq. (7); see Chap. 3 of [6].

Step 7. Solve the hyperbolic forward problem (1) with $c_n(x) := c_{n,i}$, compute the Laplace transform and obtain the function $W_{n,1}(x,\bar{s})$.

Step 8. Find a new approximation for the tail function

$$W_{n,i}(x) = \frac{\ln U_{n,i}(x,\bar{s})}{\bar{s}^2}. \tag{59}$$

Step 9. Iterate with respect to i and stop iterations at $i = m_n$ such that $q_{n,m_n} := \lim_{i\to\infty} q^k_{n,i}$. Stopping criterion for computing functions $q^k_{n,i}$ is

$$\text{either } F^k_n \geq F^{k-1}_n \text{ or } F^k_n \leq \eta, \tag{60}$$

where η is a chosen tolerance and F^k_n are defined as

$$F^k_n = \frac{||q^k_{n,i} - q^{k-1}_{n,i}||_{L_2(\Gamma)}}{||q^{k-1}_{n,i}||_{L_2(\Gamma)}}$$

Step 10. Set

$$q_n := q_{n,m_n}, \qquad c_n(x) := c_{n,m_n}(x), \qquad W_{n+1,0}(x) := \frac{\ln W_{n,m_n}(x,\bar{s})}{\bar{s}^2}.$$

Step 11. We stop computing functions $c^k_{n,i}$ when

$$\text{either} \quad N_n \geq N_{n-1} \text{ or } \quad N_n \leq \eta, \tag{61}$$

where

$$N_n = \frac{||c^k_{n,i} - c^{k-1}_{n,i}||_{L_2(\Omega)}}{||c^{k-1}_{n,i}||_{L_2(\Omega)}}. \tag{62}$$

7.2 *Adaptive Approximate Globally Convergent Algorithm*

In computations of Sect. 8 we use the following adaptive approximate globally convergent algorithm:

Step 0. Choose an initial mesh K_h in Ω and an initial time partition J_0 of the time interval $(0,T)$. Compute an initial approximation c^0_{n,m_n} using an approximate globally convergent algorithm described above on the initial mesh; see [6] for the details. Compute the sequence of functions c^j_{n,m_m}, where $j > 0$ is the number of mesh refinements, on adaptively refined meshes via following steps:

Step 1. Compute the initial approximation for the tail function $W_n(x,\bar{s})$ on a new mesh K_h using the computed solution of the hyperbolic problem (3).
Step 2. Compute the finite element solutions $q_n^j(x,s)$ of Eq. (21) on a refined mesh K_h on the pseudo-frequency interval (s_n, s_{n-1}) using Algorithm of Sect. 7.1.
Step 3. Update the coefficient c_n^j on K_h using the finite element formulation for Eq. (7).
Step 4. Stop computing c_n^j and obtain the function c_{n,m_n}^j using the criterion Eq. (61).
Step 5. Refine the mesh at all the points where

$$c_{n,m_n}^j(x) \geq \beta_1 \max_{\Omega} c_{n,m_n}^j. \tag{63}$$

The tolerance number $\beta_1 \in (0,1)$ is chosen by the user.
Step 6. Construct a new refined mesh K_h in Ω and a new time partition J_τ of the time interval $(0,T)$ satisfying the CFL condition, and return to step 1 and perform all of the above steps on the new mesh.
Step 7. Stop mesh refinements and obtain the function c_{n,m_n}^j if norms defined in the criterion Eq. (61) are fulfilled.

8 Imaging of Land Mines Using an Adaptive Approximate Globally Convergent Algorithm

In this section we present numerical implementation of an adaptive approximate globally convergent method with backscattered data in two dimensions. Our goal is reconstruction of land mines from backscattered data using an adaptive approximate globally convergent algorithm of Sect. 7.2.

Let the ground be $\{\mathbf{x} = (x,z) : z > 0\} \subset \mathbb{R}^2$. Suppose that a polarized electric field is generated by a plane wave, which is initialized at the line $\{z = z^0 < 0, x \in \mathbb{R}\}$ at the moment of time $t = 0$.

In our model we use the well-known fact that the maximal depth of an antipersonnel land mine does not exceed approximately 10 centimeters (cm) = 0.1 meter (m), and we model these mines as small rectangles with length of side 0.2 m and width of side 0.1 m. In our computations we are interested in imaging of land mines when one mine is located very close to the other one. This is an important case in the real-life military applications.

We have modelled such a problem on a domain Ω (see Fig. 1), viz., We set

$$\tilde{\Omega}_{FEM} = \{\mathbf{x} = (x,z) \in (-0.3, 0.3)\,\mathrm{m} \times (0.05, 0.45)\,\mathrm{m}\},$$

and introduce a dimensionless spatial variables $\mathbf{x}' = \mathbf{x}/(0.1\,\mathrm{m})$, so that the domain $\tilde{\Omega}_{FEM}$ is transferred into a dimensionless computational domain

$$\Omega_{FEM} = (-3.0, 3.0) \times (0.5, 4.5).$$

We choose values of function $c(x)$ using tables of dielectric constants [11] and use the fact that in the dry sand $c = 5$ and in the trinitrotoluene (TNT) $c = 22$. Thus, the relation of mine/background contrast is $22/5 \approx 4$; hence, we consider new parameters

$$c' = \frac{c}{5},$$

to get

$$c(\text{dry sand}) = 1,\ c(\text{TNT}) \approx 4. \tag{64}$$

For simulation of backscattered data for the inverse problem IPB, we solve the forward problem using the software package WavES [12]. The dimensionless size of our computational domain is $\Omega = [-4.0, 4.0] \times [0, 5.0]$. This domain is split into a dimensionless finite element domain $\Omega_{FEM} = [-3.0, 3.0] \times [0.5, 4.5]$ and a surrounding domain Ω_{FDM} with a structured mesh, $\Omega = \Omega_{FEM} \cup \Omega_{FDM}$, see Fig. 1. The spatial mesh in Ω_{FEM} and in Ω_{FDM} consists of triangles and squares, respectively. The mesh size is $\tilde{h} = 0.125$ in the overlapping regions. The boundary of the domain Ω is $\partial\Omega = \partial\Omega_1 \cup \partial\Omega_2 \cup \partial\Omega_3$. Here, $\partial\Omega_1$ and $\partial\Omega_2$ are respectively top and bottom sides of the domain Ω, see Fig. 1, and $\partial\Omega_3$ is the union of left and right sides of this domain. We define the boundary of the domain Ω_{FEM} as $\Gamma = \Gamma_1 \cup \Gamma_2 \cup \Gamma_3$. Here, Γ_1 and Γ_2 are respectively top and bottom sides of the domain Ω_{FEM}, see Fig. 1, and Γ_3 is the union of left and right sides of this domain.

We use the hybrid method of [7]. Since in our applications we know value of the coefficient $c(\mathbf{x})$ outside of the domain of interest Ω_{FEM} such that

$$c(\mathbf{x}) = 1 \text{ in } \Omega_{FDM}, \tag{65}$$

hence, we need to determine $c(\mathbf{x})$ only in Ω_{FEM}.

Fig. 1 (**a**) Geometry of the hybrid mesh. This is a combination of the quadrilateral mesh in the subdomain Ω_{FDM} (**b**), where we apply FDM, and the finite element mesh in the inner domain Ω_{FEM} (**c**), where we use FEM. The solution of the inverse problem is computed in Ω_{FEM}. The trace of the solution of the forward problem (66) is recorded at the top boundary Γ_1 of the finite element domain Ω_{FEM}

The forward problem in our computational test is

$$\begin{aligned} c(\mathbf{x})\, u_{tt} - \Delta u = 0, \quad & \text{in } \Omega \times (0,T), \\ u(\mathbf{x},0) = 0,\ u_t(\mathbf{x},0) = 0, \ & \text{in } \Omega, \\ \partial_n u = f(t), \ & \text{on } \partial\Omega_1 \times (0,t_1], \\ \partial_n u = -\partial_t u, \ & \text{on } \partial\Omega_1 \times (t_1,T), \\ \partial_n u = -\partial_t u, \ & \text{on } \partial\Omega_2 \times (0,T), \\ \partial_n u = 0, \ & \text{on } \partial\Omega_3 \times (0,T), \end{aligned} \tag{66}$$

where $f(t)$ is the amplitude of the initialized plane wave,

$$f(t) = \frac{(\sin(\omega t - \pi/2) + 1)}{10}, \quad 0 \le t \le t_1 := \frac{2\pi}{\omega}. \tag{67}$$

To compute the data for the inverse problem we solve the forward problem (66) with $\omega = 7.0$ in Eq. (67) and in the time $T = (0,6)$ with the time step $\tau = 0.01$ which is satisfied the CFL condition and save the solution of this problem at the top boundary Γ_1 of the finite element domain Ω_{FEM}. Figure 2 shows isosurfaces of the computed solution of the problem (66) in the computational domain Ω.

In our test we also define the set of admissible coefficients for the function $c(\mathbf{x})$ in Ω_{FEM} as

$$M_c = \{c(\mathbf{x}) : c(\mathbf{x}) \in [1,8],\ c(\mathbf{x}) = 1\ \forall \mathbf{x} \in \mathbb{R}^2 \setminus \Omega, c(\mathbf{x}) \in C^2\left(\mathbb{R}^2\right).\}$$

8.1 Numerical Results

We have performed two set of tests. In the first test we solve IPB using approximate globally convergent algorithm of Sect. 7.1, and in the second test we solve IPB using adaptive approximate globally convergent algorithm of Sect. 7.2. The goal of both tests was to reconstruct structure given on Fig. 1a.

The backscattered data at the boundary Γ_1 in both tests were computationally simulated using the software package WavES [12] via solving the hyperbolic problem (66) with known values of the coefficient $c = 4$ inside two inclusions of Fig. 1a and with 5 % additive noise in simulated data.

Figure 3 displays sensitivity of the simulated function $q(\mathbf{x},s), \mathbf{x} \in \Gamma_1$ for $s = 2$ and $s = 3$. We observed that all values of the function $|q(\mathbf{x})|$ for $s > 5$ are very noisy and does not show sensitivity to the inclusions. Because of that we decided to take pseudo-frequency interval $s = [2,3]$, where the computed function $q(\mathbf{x},s), \mathbf{x} \in \Gamma_1$ is most sensitive to the presence of two inclusions. We run both tests with the step in the pseudo-frequency $h = 0.05$.

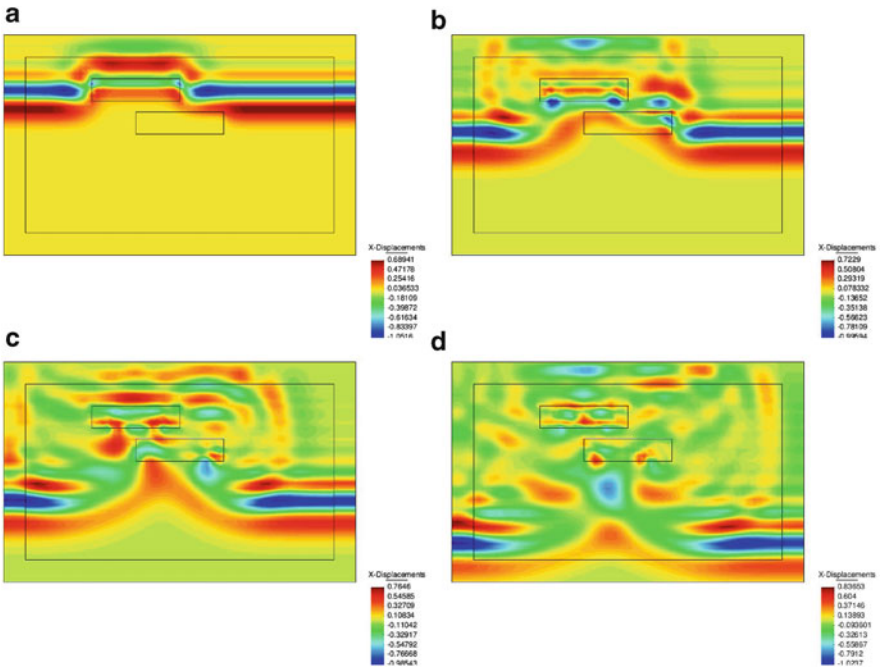

Fig. 2 Isosurfaces of the computed exact solution for the forward problem (66) at different times with a plane wave initialized at the *top boundary* (**a**) t= 3.0, (**b**) t=4.0, (**c**) t=5.0, and (**d**) t=6.0

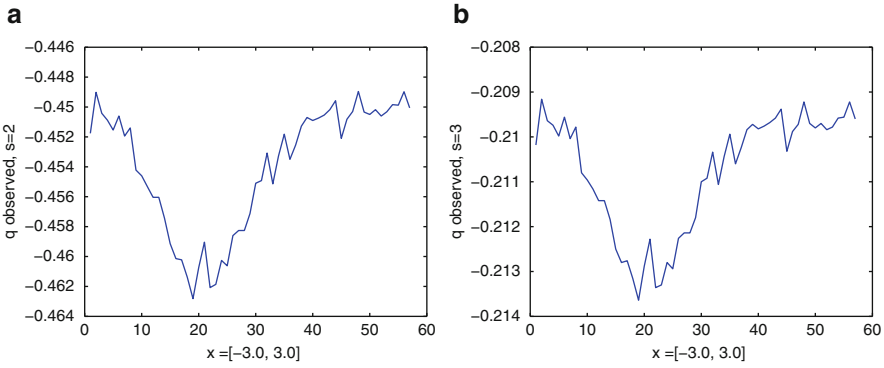

Fig. 3 Backscattered data for the function q at the *top boundary* Γ_1 of the computational domain Ω_{FEM} computed for the different values of the pseudo-frequency s (**a**) $s = 2$, (**b**) $s = 3$

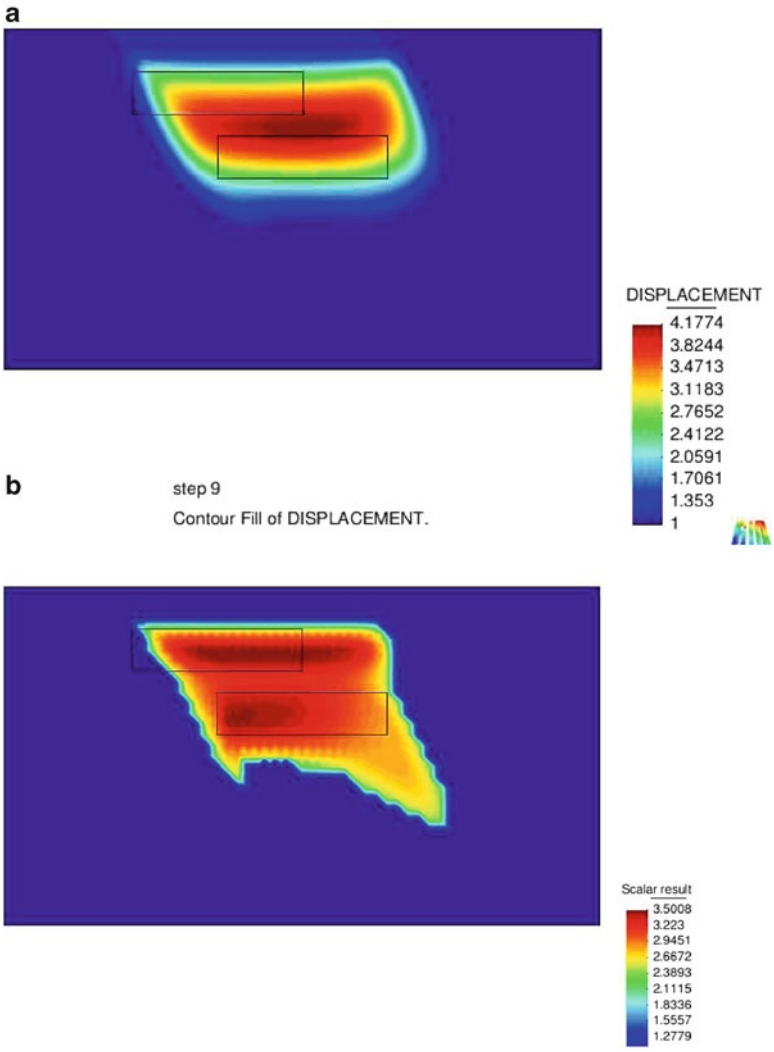

Fig. 4 Computed images using backscattered data obtained from the geometry presented on Fig. 1a. (**a**) Test1: location and contrast of inclusions are accurately imaged ($c_{7,9} \approx 4.17$). (**b**) Test2: location, contrast, and shape of inclusions are accurately imaged. The computed function $c = 1$ outside of imaged inclusions $c^1_{5,4} \approx 3.5$

8.2 Test 1

In this test we solve IPB using globally convergent algorithm of Sect. 7.1. The boundary conditions for the integral-differential equation (50) were replaced with the following Dirichlet boundary conditions:

$$q_n|_{\Gamma_1} = \psi_{1n}(\mathbf{x}),\ q_n|_{\Gamma_2 \cup \Gamma_3} = \psi_{2n}(\mathbf{x}),$$

where function $\psi_{1n}(\mathbf{x})$ and $\psi_{2n}(\mathbf{x})$ are generated by functions $g_1(\mathbf{x},t)$ and $g_0(\mathbf{x},t)$, respectively, defined in Eq. (4). In this test we simulated the function $g_0(\mathbf{x},t)$ at $\Gamma_2 \cup \Gamma_3$ by solution of the forward problem (66) with $c(\mathbf{x}) = 1$ at every point of the computational domain Ω. The Dirichlet boundary condition at $\Gamma_2 \cup \Gamma_3$ is also approximated and it is necessary to solve the integral-differential equation (50).

An approximate globally convergent algorithm of Sect. 7.1 was used to calculate the image of Fig. 4a. We observe that the location of both mine-like targets is reconstructed accurately, and the contrast $\max[c_{comp}(\mathbf{x})] = 4.17$ is also accurately imaged (exact $c(x) = 4$ in both inclusions).

However, in this test we were not able to separate images for both mines. We could only image them as one big inclusion. In the next test we try to improve the result of the reconstruction of the Test 1 using an adaptivity technique inside approximately globally convergent method.

8.3 Test 2

In this test we solve IPB using an adaptive globally convergent algorithm of Sect. 7.2. This algorithm was used to calculate the image of Fig. 4b which was obtained on the one time refined mesh. We observe that not only location and contrast of both mine-like targets are reconstructed accurately, but also the shape of mines is imaged more accurately than in the Test 1: in the Test 2 we are able to separate these two mines. Thus, we conclude that an adaptive approximate globally convergent algorithm of Sect. 7.2 allows better reconstruction of shape of inclusions than an usual approximate globally convergent method of Sect. 7.1 even for the case of backscattered data.

Acknowledgements The research of the authors was supported by the Swedish Research Council, the Swedish Foundation for Strategic Research (SSF) through the Gothenburg Mathematical Modelling Centre (GMMC), and by the Swedish Institute, Visby Program.

References

1. Asadzadeh, M., Beilina, L.: A posteriori error analysis in a globally convergent numerical method for a hyperbolic coefficient inverse. Inv. Probl. **26**, (2010).
2. Beilina, L., Klibanov, M.V.: A globally convergent numerical method for a coefficient inverse problem. SIAM J. Sci. Comp. **31**(1), 478–509 (2008)
3. Beilina, L., Klibanov, M.V.: A posteriori error estimates for the adaptivity technique for the Tikhonov functional and global convergence for a coefficient inverse problem. Inv. Probl. **26**, 045012 (2010)
4. Beilina, L., Klibanov, M.V.: Reconstruction of dielectrics from experimental data via a hybrid globally convergent/adaptive inverse algorithm. Inv. Probl. **26**, 125009 (2010)
5. Beilina, L., Klibanov, M.V.: Synthesis of global convergence and adaptivity for a hyperbolic coefficient inverse problem in 3D. J. Inv. Ill-Posed Probl. **18**(1), 85–132 (2010)
6. Beilina, L., Klibanov, M.V.: Approximate Global Convergence and Adaptivity for Coefficient Inverse Problems. Springer, New-York (2012)
7. Beilina, L., Samuelsson, K., Åhlander, K.: Efficiency of a hybrid method for the wave equation. In: International Conference on Finite Element Methods, Gakuto International Series Mathematical Sciences and Applications, Gakkotosho CO., LTD, 2001
8. Klibanov, M., Timonov, A.A.: A unified framework for constructing of globally convergent numerical algorithms for multidimensional coefficient inverse problems. Appl. Anal. **83**, 933–955 (2004)
9. Klibanov, M.V., Timonov, A.: Carleman Estimates for Coefficient Inverse Problems and Numerical Applications. VSP, Utrecht (2004)
10. Klibanov, M.V., Fiddy, M.A., Beilina, L., Pantong, N., Schenk, J.: Picosecond scale experimental verification of a globally convergent numerical method for a coefficient onverse problem. Inv. Prob. **26**(3), 1–30 (2010)
11. Tables of dielectric constants at http://www.asiinstr.com/technical/DielectricConstants.htm
12. Software package WavES at http://www.waves24.com/

On the Formulation of Inverse Problem in Electrical Prospecting

V.P. Gubatenko

Abstract The following inverse problem can be formulated for the isotropic geological medium with applications in electrical prospecting: *The electromagnetic field is measured on the surface of the ground. Find the distribution of electrical conductivity σ and magnetic permeability μ of the geological medium.* We consider a simplified mathematical formulation of this problem in the frequency domain, assuming that the parameters of the geological medium σ and μ possess the frequency dispersion.

1 The First Inverse Problem

Assume, x, y, and z are the Cartesian coordinates in Euclidean space. Our goal is to find the coefficients σ and μ of Maxwell equations

$$\operatorname{rot} \mathbf{H} = \sigma \mathbf{E}, \tag{1}$$

$$\operatorname{rot} \mathbf{E} = i\omega\mu \mathbf{H} \tag{2}$$

in region $V = \left\{M(x,y,z) \in R^3 \,|\, z > 0\right\}$ (in the ground). Here, $E = E(M,i\omega) = E(x,y,z,i\omega) = (E_x, E_y, E_z)$ and $H = H(M,i\omega) = H(x,y,z,i\omega) = (H_x, H_y, H_z)$ are the complex amplitudes of electric and magnetic fields in the ground, respectively, i is the imaginary unit, and ω is the angular frequency.

Let the unknown parameters $\sigma = \sigma(x,y,z,i\omega)$ and $\mu = \mu(x,y,z,i\omega)$ of the medium satisfy the conditions

V.P. Gubatenko (✉)
Saratov State University, 410012, Saratov, Russia
e-mail: gubatenkovp@gmail.com

L. Beilina and Y.V. Shestopalov (eds.), *Inverse Problems and Large-Scale Computations*, Springer Proceedings in Mathematics & Statistics 52, DOI 10.1007/978-3-319-00660-4_2,

$$\sigma(M,i\omega)\neq 0, \qquad \mu(M,i\omega)\neq 0, \tag{3}$$

$$\operatorname{Re}\sigma(M,i\omega)\geq 0, \qquad \operatorname{Im}\mu(M,i\omega)\leq 0, \tag{4}$$

$$\sigma(M,i\omega)\in C^{k-1}(V),\ \mu(M,i\omega)\in C^{k-1}(V), \quad k\geq 3. \tag{5}$$

Here, the restrictions (3) and (4) indicate the feasibility of the physical parameters of the medium, and the Eq. (5) is the condition of smoothness.

Since the parameters of the medium make available measurements near the ground, we assume that their distributions on the surface $z=+0$ are known:

$$\sigma=\sigma^0(x,y,+0,i\omega), \quad \mu=\mu^0(x,y,+0,i\omega). \tag{6}$$

Suppose also that vector fields $\mathbf{E}$ and $\mathbf{H}$ are known on the surface $z=+0$:

$$\mathbf{E}=\mathbf{E^0}(x,y,+0,i\omega), \quad \mathbf{H}=\mathbf{H^0}(x,y,+0,i\omega), \tag{7}$$

where $\mathbf{E}=\mathbf{E^0}(x,y,+0,i\omega)=\left(E_x^0,E_y^0,E_z^0\right)$, $\mathbf{H}=\mathbf{H^0}(x,y,+0,i\omega)=\left(H_x^0,H_y^0,H_z^0\right)$. Then Eqs. (1) and (2) on the surface $z=+0$ can be written as

$$\begin{aligned}
&\frac{\partial H_z^0}{\partial y}-\left.\frac{\partial H_y}{\partial z}\right|_{z=+0}=\sigma^0E_x^0,\\
&\left.\frac{\partial H_x}{\partial z}\right|_{z=+0}-\frac{\partial H_z^0}{\partial x}=\sigma^0E_y^0,\\
&\frac{\partial H_y^0}{\partial x}-\frac{\partial H_x^0}{\partial y}=\sigma^0E_z^0,
\end{aligned} \tag{8}$$

$$\begin{aligned}
&\frac{\partial E_z^0}{\partial y}-\left.\frac{\partial E_y}{\partial z}\right|_{z=+0}=i\omega\mu^0H_x^0,\\
&\left.\frac{\partial E_x}{\partial z}\right|_{z=+0}-\frac{\partial E_z^0}{\partial x}=i\omega\mu^0H_y^0,\\
&\frac{\partial E_y^0}{\partial x}-\frac{\partial E_x^0}{\partial y}=i\omega\mu^0H_z^0.
\end{aligned} \tag{9}$$

Here, $\left.\frac{\partial E_x}{\partial z}\right|_{z=+0}$, $\left.\frac{\partial E_y}{\partial z}\right|_{z=+0}$, $\left.\frac{\partial H_x}{\partial z}\right|_{z=+0}$ and $\left.\frac{\partial H_y}{\partial z}\right|_{z=+0}$ are the partial derivatives of the electromagnetic field components along the coordinate z on the surface $z=+0$.

Relations (6)–(9) can be taken as the boundary conditions of the inverse problem. However, comparing the expression (7)–(9), we see that the functions appearing in them are dependent. For example, if $E_z^0\neq 0$, then as independent

functions can be selected $\mu=\mu^0(x,y,+0,i\omega), E_x^0, E_y^0, E_z^0, \left.\frac{\partial E_x}{\partial z}\right|_{z=+0}$, and $\left.\frac{\partial E_y}{\partial z}\right|_{z=+0}$. In this case, the function H_x^0, H_y^0, H_z^0 is defined by conditions (9). Then, $\sigma = \sigma^0(x,y,+0,i\omega)$ is determined from the last equality of Eq. (8), and functions $\left.\frac{\partial H_x}{\partial z}\right|_{z=+0}$, $\left.\frac{\partial H_y}{\partial z}\right|_{z=+0}$ are determined from the first and second equations of the same conditions.

Thus, the boundary conditions for the inverse problem can be written as

$$\mu=\mu^0(x,y,+0,i\omega), \quad E_x=E_x^0(x,y,+0,i\omega), \quad E_y=E_y^0(x,y,+0,i\omega), \tag{10}$$

$$E_z=E_z^0(x,y,+0,i\omega), \left.\frac{\partial E_x}{\partial z}\right|_{z=+0}=\varphi(x,y,+0,i\omega), \left.\frac{\partial E_y}{\partial z}\right|_{z=+0}=\psi(x,y,+0,i\omega),$$

where the functions in the right-hand sides of equalities are known. In the case of $E_z^0=0$ to the boundary conditions (10) we add

$$\sigma=\sigma^0(x,y,+0,i\omega).$$

Note, that from Eqs. (1), (2) and conditions (4), (5) we obtain

$$\mathbf{E}(M,i\omega)\in C^k(V), \mathbf{H}(M,i\omega)\in C^k(V), \tag{11}$$

$$\lim_{z\to+\infty}\mathbf{E}(M,i\omega)=0, \lim_{z\to+\infty}\mathbf{H}(M,i\omega)=0. \tag{12}$$

We can assume that the vector fields $\mathbf{E}$ and $\mathbf{H}$ are not identically zero in the region V, and it follows from conditions (3). The same is true for $\mathrm{rot}\,\mathbf{E}$ and $\mathrm{rot}\,\mathbf{H}$.

As follows from the formulation of the inverse problem, the solution to this problem exists, but not its uniqueness is obvious. Clearly, if we could found any solution $\sigma=\tilde{\sigma}(E,y,z,i\omega)$ and $\mu=\tilde{\mu}(E,y,z,i\omega)$ of this problem, then for these parameters there exists a unique solution of Maxwell equations (1) and (2) with respect to the vector fields $\mathbf{E}$, $\mathbf{H}$. The following question arises: can we reduce the first inverse problem to the problem of finding the vector field $\mathbf{E}$? To answer this question, we formulate the next inverse problem.

2 The Second Inverse Problem

Let the scalar functions σ, μ and vector fields $\mathbf{E}$, $\mathbf{H}$ still satisfy the conditions (3), (5) and (11), (12), and at the same time the vector fields are not identically equal to zero in the region V. Let us formulate the following inverse problem:

Suppose, in region V is given a vector field $\mathbf{E}$. *Find in the region V the field scalar functions* σ, μ *and vector* $\mathbf{H}$, *turning the relationships (1) and (2) to identity.*

A similar problem can be considered for the *given* vector $\mathbf{H}$ and the unknown functions σ, μ, $\mathbf{E}$. However, we will not discuss this problem separately, taking into account the symmetry of the Eqs. (1) and (2) with respect to the formal replacement $\mathbf{E}\leftrightarrow\mathbf{H}$, $i\omega\mu\leftrightarrow\sigma$, called the principle of duality commutes [6].

Lemma 1. *For the existence of the second inverse problem is necessary and sufficient that the unknown scalar function μ is a solution of the differential equation*

$$\mathbf{E}\times\operatorname{rot}\left(\frac{1}{\mu}\operatorname{rot}\mathbf{E}\right)=0 \tag{13}$$

except for solutions μ of the equation

$$\operatorname{rot}\left(\frac{1}{\mu}\operatorname{rot}\mathbf{E}\right)=0\,. \tag{14}$$

Remark to Lemma 1. If the function μ is a solution of Eq. (14), the unknown vector $\mathbf{H}$ is identically zero.

Lemma 2. *If for a given vector E there exists a solution of the second inverse problem, then at any point $M\in V$ or*

$$(\mathbf{E}(M),\operatorname{rot}\mathbf{E}(M))=0,\quad (\operatorname{rot}\mathbf{E}(M),\operatorname{rot}\operatorname{rot}\mathbf{E}(M))=0,$$

or

$$(\mathbf{E}(M),\operatorname{rot}\mathbf{E}(M))\neq 0,\quad (\operatorname{rot}\mathbf{E}(M),\operatorname{rot}\operatorname{rot}\mathbf{E}(M))\neq 0.$$

Remark to Lemma 2. If $(\mathbf{E},\operatorname{rot}\mathbf{E})\neq 0$, $(\operatorname{rot}\mathbf{E},\operatorname{rot}\operatorname{rot}\mathbf{E})\neq 0$ at some point $M\in V$ then since the scalar functions are continuous, there exists a neighborhood of this point at which these inequalities are true. Therefore, when setting the vector $\mathbf{E}$ in the second inverse problem, we consider two cases: or $(\mathbf{E},\operatorname{rot}\mathbf{E})\equiv 0$, $(\operatorname{rot}\mathbf{E},\operatorname{rot}\operatorname{rot}\mathbf{E})\equiv 0$ in the area, or $(\mathbf{E},\operatorname{rot}\mathbf{E})$, $(\operatorname{rot}\mathbf{E},\operatorname{rot}\operatorname{rot}\mathbf{E})$ are not identically zero in any subregion V. The first case corresponds to the orthogonal vectors $\mathbf{E}$ and $\mathbf{H}$, but the second case is not orthogonal. The first case includes, for example, three-component flat and axisymmetric electromagnetic fields, and the second- five-component transverse electric and transverse magnetic fields [8].

Theorem 1. *If the vector field $\mathbf{E}$ is a solution of the nonlinear equations*

$$(\mathbf{E},\operatorname{rot}\mathbf{E})=0,\quad (\operatorname{rot}\mathbf{E},\operatorname{rot}\operatorname{rot}\mathbf{E})=0, \tag{15}$$

then for a given vector $\mathbf{E}$, the solution of the second inverse problem exists and is not unique.

The proof of this theorem is based on the theory of the linear partial differential equations of the first order and common solutions to these equations by means of

characteristic systems [5]. After finding the solutions of Eq. (13) and after removing solutions of Eq. (14) from solutions of Eq. (13), we can determine functions σ and $\mathbf{H}$ as follows:

$$\sigma = \frac{1}{i\omega \mathbf{E}^2}\left(\mathbf{E}, \operatorname{rot}\left(\frac{1}{\mu}\operatorname{rot}\mathbf{E}\right)\right), \quad \mathbf{H} = \frac{1}{i\omega\mu}\operatorname{rot}\mathbf{E}. \tag{16}$$

We note here that the first equality of Eq. (16) is determined from

$$\operatorname{rot}\left(\frac{1}{\mu}\operatorname{rot}\mathbf{E}\right) = i\omega\sigma\mathbf{E}. \tag{17}$$

Theorem 2. *Let the vector* $\mathbf{E}$ *is given in the region* V *and* $(\mathbf{E}, \operatorname{rot}\mathbf{E}) \neq 0$, $(\operatorname{rot}\mathbf{E}, \operatorname{rot}\operatorname{rot}\mathbf{E}) \neq 0$ *in this region. For the existence solution of the inverse problem, it is necessary and sufficient that the vector* E *is the solution of the nonlinear equation*

$$\operatorname{rot}\mathbf{F}^E = 0, \tag{18}$$

where

$$\mathbf{F}^E = \frac{1}{(\operatorname{rot}\mathbf{E}, \operatorname{rot}\operatorname{rot}\mathbf{E})}\operatorname{div}\left[\frac{(\operatorname{rot}\mathbf{E}, \operatorname{rot}\operatorname{rot}\mathbf{E})}{(\mathbf{E}, \operatorname{rot}\mathbf{E})}\mathbf{E}\right]\operatorname{rot}\mathbf{E} + \frac{1}{(\mathbf{E}, \operatorname{rot}\mathbf{E})}(\mathbf{E} \times \operatorname{rot}\operatorname{rot}\mathbf{E}).$$

The general solution μ of the second inverse problem has the form

$$\mu = \mu_0(i\omega)\exp\left(\int_{M_0}^{M} F_x^E dx + F_y^E dy + F_z^E dz\right) \tag{19}$$

where $\mathbf{F}^E = \left(F_x^E, F_y^E, F_z^E\right)$; $M(x,y,z) \in V$, $M_0(x_0,y_0,z_0) \in V$; the function $\mu_0(i\omega)$ is arbitrary and does not depend on the coordinates. Electrical conductivity and magnetic field are determined by formulas (16). Let us consider the following inverse problem.

3 The Third Inverse Problem

As follows from Theorems 1 and 2, the vector field $\mathbf{E}$ is accompanied by a family of functions $\{\mathbf{E}, \mathbf{H}, \mu, \sigma\}$, which becomes an identity equation (1) and (2) then and only then, when the vector field $\mathbf{E}$ satisfies to the Eqs. (15) or (18). Of course, not every vector field $\mathbf{E}$ under these conditions uniquely determines the scalar functions σ, μ, which obey the conditions of physics Eq. (4). For a single determination of the parameters of the medium in accordance with Theorems 1 and 2, we require a priori information on the distribution of the permeability function μ in the region V.

Suppose, for example, here and below, $\mu = \mu_0 = 4\pi \cdot 10^{-7}$ H/m, which corresponds to the sedimentary rocks studied in the structural electrical prospecting. In this case, as for the orthogonal fields **E**, **H**, and also for the non-orthogonal fields, vector **E** must be a solution of the equation

$$\mathbf{E} \times \operatorname{rot} \operatorname{rot} \mathbf{E} = 0\,. \tag{20}$$

Then the first inverse problem is reduced to the following inverse problem:

The Inverse Problem 3.1. *Find the solution of Eq. (20) satisfying to the boundary conditions*

$$E_x = E_x^0(x,y,+0,i\omega)\,, \qquad E_y = E_y^0(x,y,+0,i\omega)\,, \tag{21}$$

$$E_z = E_z^0(x,y,+0,i\omega) \neq 0,\ \left.\frac{\partial E_x}{\partial z}\right|_{z=+0} = \varphi(x,y,+0,i\omega)\,,\ \left.\frac{\partial E_y}{\partial z}\right|_{z=+0} = \psi(x,y,+0,i\omega)$$

and in the case when $E_z = 0$

$$\sigma = \sigma^0(x,y,+0,i\omega)\,, \qquad E_x = E_x^0(x,y,+0,i\omega)\,, \qquad E_y = E_y^0(x,y,+0,i\omega)\,, \tag{22}$$

$$\left.\frac{\partial E_x}{\partial z}\right|_{z=+0} = \varphi(x,y,+0,i\omega),\ \left.\frac{\partial E_y}{\partial z}\right|_{z=+0} = \psi(x,y,+0,i\omega)\,.$$

If the first inverse problem with boundary conditions (10) has a unique solution, then the electric field intensity **E** for the inverse problem 3.1 is also unique. Having determined the field **E** of the inverse problem 3.1, we easily find from (16) functions σ and H.

Let us show on a simple example of the classical model of the magnetotelluric sounding that the solution as of the first inverse problem such that for the inverse problem 3.1 is not unique in the case of the frequency dispersion of the electrical conductivity.

Suppose that in the half-space $z > 0$ is situated nonmagnetic medium with electrical conductivity $\sigma = \sigma(z,i\omega)$, and let us initialize an electromagnetic field $\mathbf{E} = (E_x(z,i\omega),0,0)$, $\mathbf{H} = (0,H_y(z,i\omega),0)$ with orthogonal vectors **E** and **H**. Then in the half-space $z > 0$, Eq. (20) [or Eq. (17)] has the form

$$\frac{d^2E_x}{dz^2} + i\omega\mu_0\sigma E_x = 0\,. \tag{23}$$

Assume that the surface boundary conditions (22) have the form

$$\sigma^0(0,i\omega) = \sigma_0,\ E_x = E^0(i\omega)\,,\ \left.\frac{dE_x}{dz}\right|_{z=+0} = \varphi(i\omega) = -\sqrt{-i\omega\mu_0\sigma_0}E^0(i\omega)\,, \tag{24}$$

where $\sigma_0 = \text{const} > 0$; $\text{Re}\sqrt{-i\omega\mu_0} > 0$; $E^0(i\omega)$ is an arbitrary complex function of angular frequency ω. It is easy to see that there are at least two solutions of the first inverse problem and the inverse problem 3.1. These solutions are

$$\sigma = \sigma_0, \quad E_x = E^0(i\omega)\exp(-k_0 z) \tag{25}$$

and

$$\sigma = -\frac{1}{i\omega\mu_0}\frac{k_0^2 + k_1(k_1 - k_0)(2k_0 - k_1)z + \frac{k_1(k_1-k_0)^2}{2}z^2}{1 + (k_1 - k_0)z + \frac{(k_1-k_0)^2}{2}z^2}, \tag{26}$$

$$E_x = E^0(i\omega)\left[1 + (k_1 - k_0)z + \frac{(k_1 - k_0)^2}{2}z^2\right]\exp(-k_1 z),$$

where $k_0 = \sqrt{-i\omega\mu_0\sigma_0}$; $k_1 = \sqrt{-i\omega\mu_0\sigma_1}$, $\sigma_1 = const$, $\sigma_0 < \sigma_1 < 4\sigma_0$.

Solution (25) corresponds to a quasi-stationary field in the homogeneous half-space $z > 0$ with conductivity $\sigma = \sigma_0$ and the solution (26)—the gradient frequency-dispersive medium. It is easy to show that $\text{Re}\sigma > 0$, $\text{Im}\sigma < 0$ for this medium, and all $\omega > 0$, such that this model is the frequency-dispersed medium and this medium is physically realizable. This result, however, does not contradict to the uniqueness theorem [9], since this theorem is proved under the assumption of the absence of the frequency dispersion of electrical conductivity σ.

This example shows that in order to find the unique solution of the inverse problem it is necessary to know the additional information about the nature of the frequency dispersion of conductivity. If, for example, we know that the unknown scalar function does not depend on the angular frequency ω, we use the method developed by Klibanov and Beilina for hyperbolic coefficient inverse problems [1–4, 7].

Indeed, since the conductivity is determined by the formula (16) and does not depend on the angular frequency, then

$$\frac{\partial}{\partial\omega}\left[\frac{1}{\omega\mathbf{E}^2}(\mathbf{E}, \text{rot rot}\,\mathbf{E})\right] = 0, \tag{27}$$

and we can formulate the next inverse problem:

The Inverse Problem 3.2. Find the solution $\mathbf{E}$ of Eqs. (20) and (27) satisfying to (21) under the condition $E_z^0(x, y, +0, i\omega) \neq 0$, or (22) in the case of $E_z = 0$.

The solution to this inverse problem exists and is unique, at least for the one-dimensional inverse problem of magnetotelluric sounding. After finding a solution to this problem, it is easy to find the electrical conductivity σ of relationship (27).

Acknowledgements The author is grateful to the Russian Foundation for Basic Research, grant nr. 10-05-00 753-a, and to the Swedish Institute, Visby Program.

References

1. Beilina, L., Klibanov, M.V.: A globally convergent numerical method for a coefficient inverse problem. SIAM J. Sci. Comp. **31**, 478–509 (2008)
2. Beilina, L., Klibanov, M.V.: Reconstruction of dielectrics from experimental data via a hybrid globally convergent/adaptive inverse algorithm. Inv. Probl. **26**, 125009 (2010)
3. Beilina, L., Klibanov, M.V.: Synthesis of global convergence and adaptivity for a hyperbolic coefficient inverse problem in 3D. J. Inv. Ill-Posed Probl. **18**, 85–132 (2010)
4. Beilina, L., Klibanov, M.V.: Approximate Global Convergence and Adaptivity for Coefficient Inverse Problems. Springer, New-York (2012)
5. Elsholz, L.E.: Differential Equations and Variational Calculus (in russian). Nauka, Moscow (1969)
6. Goldstein, L.D., Zernov, N.V.: Electromagnetic Fields and Waves (in russian). Soviet Radio, Moscow (1971)
7. Klibanov, M.V., Fiddy, M.A., Beilina, L., Pantong, N., Schenk, J.: Picosecond scale experimental verification of a globally convergent numerical method for a coefficient inverse problem. Inv. Probl. **26**, 045003 (2010)
8. Svetov, B.S., Gubatenko, V.P.: Analytic Solutions of the Electrodynamic Problems (in russian). Nauka, Moscow (1988)
9. Tikhonov, A.N.: About mathematical foundation of the theory of electromagnetic exploration (in russian). J. Comput. Math. Math. Phys. 5, No. 3, 545–547 (1965)

Approximate Globally Convergent Algorithm with Applications in Electrical Prospecting

John Bondestam Malmberg and Larisa Beilina

Abstract In this paper we present at the first time an approximate globally convergent method for the reconstruction of an unknown conductivity function from backscattered electric field measured at the boundary of geological medium under assumptions that dielectric permittivity and magnetic permeability functions are known. This is the typical case of an coefficient inverse problem in electrical prospecting. We consider a simplified mathematical model assuming that geological medium is isotropic and non-dispersive.

1 Introduction

In this work we consider a coefficient inverse problem (CIP) for Maxwell equations in time domain and derive an approximate globally convergent method for reconstruction of an unknown conductivity function in space with data resulted from a single measurement. This means that our boundary data are generated by a single source location or a single direction of the propagation of an incident plane wave. We assume that we are working in isotropic medium with known values of electric permeability and magnetic permittivity functions. This is the typical case of electrical prospecting [4] and is of great interest in the geological community.

The first generation of globally convergent algorithms developed in [6, 7, 10] is called convexification algorithms. In this paper we use similar technique as in [3] to derive an approximate globally convergent method of second generation for finding the conductivity function. This method, first appearing in [1], uses a layer stripping procedure with respect to the pseudo-frequency.

J.B. Malmberg (✉) • L. Beilina
Department of Mathematical Sciences, Chalmers University of Technology and Gothenburg University, SE-412 96 Gothenburg, Sweden
e-mail: john.bondestam.malmberg@chalmers.se; larisa.beilina@chalmers.se

L. Beilina and Y.V. Shestopalov (eds.), *Inverse Problems and Large-Scale Computations*, Springer Proceedings in Mathematics & Statistics 52, DOI 10.1007/978-3-319-00660-4_3,

The main difficulty for solution of CIPs is ill-posedness and nonlinearity of these problems. Approximate globally convergent method of [3] gives an answer to the question: How to obtain an unknown coefficient function inside the domain of interest in a small neighborhood of the exact solution without a priori knowledge of any information about this solution? The approximate globally convergent method of [3] is experimentally verified in recent works [2, 8]. Using our recent numerical experience [2, 3, 8] we can conclude that approximate globally convergent method is reliable tool for solution of CIPs using a single measurement data.

In the current work we derive an approximate globally convergent method for explicit computation of the conductivity function using iterative layer stripping procedure in pseudo-frequency. We also present an approximate mathematical model for computation of the so-called "tail" function which is crucial for reliable reconstruction of an unknown conductivity function. Finally, we formulate an approximate globally convergent algorithm which can be used in real computations for reconstruction of an unknown function from backscattered data collected at the boundary.

2 The Maxwell Equations in Time Domain

We consider the Maxwell equations in an isotropic, non-dispersive medium (see, for instance, [5])

$$\frac{\partial B(x,t)}{\partial t} = -\nabla \times E(x,t) - M(x,t) \qquad \text{for } (x,t) \in \mathbb{R}^n \times (0,T), \tag{1}$$

$$\frac{\partial D(x,t)}{\partial t} = \nabla \times H(x,t) - J(x,t) \qquad \text{for } (x,t) \in \mathbb{R}^n \times (0,T). \tag{2}$$

Here, $n = 2, 3$, $T > 0$, $B = \mu H$ is the magnetic flux density; $D = \varepsilon E$ is the electric flux density; $H = (H_1, H_2, H_3)$ is the magnetic field; $E = (E_1, E_2, E_3)$ is the electric field; μ and ε are the magnetic permeability and the dielectric permittivity of the medium, respectively; and J and M are electric and magnetic current densities, respectively. The electric and magnetic fields satisfy the relations

$$\nabla \cdot (\varepsilon E) = \rho, \ \nabla \cdot (\mu H) = 0 \qquad \text{in } \mathbb{R}^n \times (0,T), \tag{3}$$

where $\rho(x,t)$ is a given charge density.

As it suffices for our purposes we consider the case when μ and ε are constants, $M(x,t) \equiv 0$, $\rho(x,t) \equiv 0$, and J is generated by the electric field such that $J = \sigma E$, where σ is the conductivity of the medium. We assume that $\sigma = \sigma(x)$ is dependent only on the spatial variable x.

Under these assumptions (1) and (2) are reduced to

$$\mu \frac{\partial H(x,t)}{\partial t} = -\nabla \times E(x,t) \qquad \text{for } (x,t) \in \mathbb{R}^n \times (0,T), \tag{4}$$

$$\varepsilon \frac{\partial E(x,t)}{\partial t} = \nabla \times H(x,t) - \sigma(x) E(x,t) \qquad \text{for } (x,t) \in \mathbb{R}^n \times (0,T), \tag{5}$$

and since μ and ε are positive constants Gauss' law (3) is reduced to

$$\nabla \cdot E = 0,\ \nabla \cdot H = 0 \qquad \text{in } \mathbb{R}^n \times (0,T). \tag{6}$$

In addition to Eqs. (4) and (5) we impose the following initial conditions on the magnetic and electric fields:

$$H(x,0) = 0, \tag{7}$$

$$E(x,0) = (E_{0,1}, E_{0,2}, E_{0,3})\delta(x - x_0) =: E_0 \delta(x - x_0), \tag{8}$$

where $E_{0,k}$, $k = 1, 2, 3$ are constants, δ is the three dimensional Dirac delta, and x_0 is some specific point in $\mathbb{R}^n$. This corresponds to the initialization of an electric pulse at the point x_0 at time $t = 0$.

The problem described by Eqs. (4), (5), (7), and (8) is similar to those considered in [4]. It describes the electric and magnetic fields generated in response to an electric pulse initiated at $x_0 \in \mathbb{R}^n$ and propagating through "the ground."

Further we will consider the inverse problem when $\sigma(x)$ is included in the equation for the electric field. Hence we eliminate the dependence on the magnetic field from the Cauchy problem described in Eqs. (4), (5), (7), and (8).

Applying the curl operator to Eq. (4) yields

$$\nabla \times \frac{\partial H(x,t)}{\partial t} = -\frac{1}{\mu} \nabla \times \nabla \times E(x,t) \qquad \text{for } (x,t) \in \mathbb{R}^n \times (0,T).$$

Using above equation and differentiating Eq. (5) with respect to t gives

$$\begin{aligned} \varepsilon \frac{\partial^2 E(x,t)}{\partial t^2} &= \nabla \times \frac{\partial H(x,t)}{\partial t} - \sigma(x) \frac{\partial E(x,t)}{\partial t} \\ &= -\frac{1}{\mu} \nabla \times \nabla \times E(x,t) - \sigma(x) \frac{\partial E(x,t)}{\partial t} \qquad \text{for } (x,t) \in \mathbb{R}^n \times (0,T). \end{aligned}$$

Hence, after some rearrangement of the terms, and by applying the identity $\nabla \times \nabla \times E = \nabla(\nabla \cdot E) - \Delta E$ together with Gauss' law (6), we get

$$\mu\varepsilon \frac{\partial^2 E(x,t)}{\partial t^2} + \mu\sigma(x) \frac{\partial E(x,t)}{\partial t} - \Delta E(x,t) = 0 \qquad \text{for } (x,t) \in \mathbb{R}^n \times (0,T). \tag{9}$$

Letting t go to zero in Eq. (5) and using Eqs. (7) and (8), we get

$$\frac{\partial E(x,0)}{\partial t} = -\frac{1}{\varepsilon}\sigma(x)E_0\delta(x-x_0). \tag{10}$$

Noting that the equations for each component of the electric field in Eqs. (8), (9), and (10) are decoupled, we may write the following componentwise Cauchy problems, for $k = 1, 2, 3$:

$$\begin{aligned}
&\mu\varepsilon\frac{\partial^2 E_k(x,t)}{\partial t^2} + \mu\sigma(x)\frac{\partial E_k(x,t)}{\partial t} - \Delta E_k(x,t) = 0, && \text{for } (x,t) \in \mathbb{R}^n \times (0,T),\\
&E_k(x,0) = E_{0,k}\delta(x-x_0), && \text{for } x \in \mathbb{R}^n,\\
&\frac{\partial E_k(x,0)}{\partial t} = -\frac{1}{\varepsilon}\sigma(x)E_{0,k}\delta(x-x_0), && \text{for } x \in \mathbb{R}^n.
\end{aligned} \tag{11}$$

Further we will assume that only the first component E_1 of the electric field $E = (E_1, E_2, E_2)$ is initialized by the function $E_0 = (E_{0,1}, 0, 0)$ and thus, by Eq. (6), the other two components E_2 and E_3 are zero. This yields that the problem (11) is reduced to the solution of the following Cauchy problem:

$$\begin{aligned}
&\mu\varepsilon\frac{\partial^2 E_1(x,t)}{\partial t^2} + \mu\sigma(x)\frac{\partial E_1(x,t)}{\partial t} - \Delta E_1(x,t) = 0, && \text{for } (x,t) \in \mathbb{R}^n \times (0,T),\\
&E_1(x,0) = E_{0,1}\delta(x-x_0), && \text{for } x \in \mathbb{R}^n,\\
&\frac{\partial E_1(x,0)}{\partial t} = -\frac{1}{\varepsilon}\sigma(x)E_{0,1}\delta(x-x_0), && \text{for } x \in \mathbb{R}^n.
\end{aligned} \tag{12}$$

To reduce notations we will in the following drop the index on E_1, writing $E(x,t) = E_1(x,t)$.

A Coefficient Inverse Problem

Let $\Omega \subset \mathbb{R}^n$ be a bounded domain with a piecewise smooth boundary Γ such that $x_0 \notin \overline{\Omega}$. Let $L^2(\Omega)$ be the space of square integrable functions on Ω and define $\Omega_T := \Omega \times (0,T)$ and $\Gamma_T := \Gamma \times (0,T)$. Suppose that $\sigma(x)$ satisfies the Cauchy problem (12), restricted to Ω_T, for known coefficients μ and ε and that $\sigma(x) \in C_{d,\Omega}$, where

$$C_{d,\Omega} := \{u \in C^2(\mathbb{R}^n) : 1 \le u(x) \le d, x \in \mathbb{R}^n, u \equiv 1 \text{ in } \mathbb{R}^n \setminus \Omega\} \tag{13}$$

for some given $d > 1$. We then seek to determine $\sigma(x)$, $x \in \Omega$, under assumption that the function

$$g(x,t) = E(x,t)\Big|_{\Gamma_T} \tag{14}$$

is known. In other words, σ and E satisfy the following initial boundary value problem:

$$\begin{aligned} &\mu\varepsilon\frac{\partial^2 E(x,t)}{\partial t^2}+\mu\sigma(x)\frac{\partial E(x,t)}{\partial t}-\Delta E(x,t)=0, && \text{for } (x,t)\in\Omega_T,\\ &E(x,t)=g(x,t) && \text{for } (x,t)\in\Gamma_T \qquad (15)\\ &E(x,0)=\frac{\partial E(x,0)}{\partial t}=0, && \text{for } x\in\Omega. \end{aligned}$$

3 Approximately Globally Convergent Method

In this section we develop an approximately globally convergent method for the **CIP** to reconstruct the conductivity function. The method uses the Laplace transform of the problem (15), and hence we start by deriving some properties thereof.

3.1 Laplace Transformation of the Initial Boundary Value Problem

Let

$$\mathbf{E}(x,s):=\mathscr{L}\Big[E(x,\cdot)\Big](s):=\int_0^\infty E(x,t)\mathrm{e}^{-st}\,\mathrm{d}t,\quad s\geq\underline{s}$$

for some fixed $\underline{s}>0$ be the Laplace transform of the electric field $E(x,t)$ given by Eq. (15). Applying the Laplace transform to the differential equation in Eq. (15) and using the two well-known properties $\mathscr{L}[f'](s)=s\mathscr{L}[f](s)-f(0)$ and $\mathscr{L}[f''](s)=s^2\mathscr{L}[f](s)-sf(0)-f'(0)$ of the Laplace transform we get

$$\Delta\mathbf{E}(x,s)-(\mu\varepsilon s^2+\mu s\sigma(x))\mathbf{E}(x,s)=-\mu\varepsilon sE_0\delta(x-x_0). \qquad (16)$$

Similarly with Theorems 2.7.1 and 2.7.2 of [3], it can be proved that $\mathbf{E}(x,s)\to 0$ as $|x|\to\infty$ and that $\mathbf{E}(x,s)>0$; hence, the following function

$$v(x,s):=\frac{\ln(\mathbf{E}(x,s))}{s},\qquad x\in\Omega,\ s\in[\underline{s},\overline{s}] \qquad (17)$$

for some $\overline{s}>\underline{s}>0$ is well defined.

Next, we assume that the following asymptotic behavior for the function $\mathbf{E}(x,s)$ holds

$$D_x^\alpha\frac{\partial^n}{\partial s^n}\mathbf{E}(x,s)=D_x^\alpha\frac{\partial^n}{\partial s^n}\left(\frac{\mathrm{e}^{-sl(x,x_0)}}{f(x,x_0)}\left(1+O\left(\frac{1}{s}\right)\right)\right),\quad s\to\infty \qquad (18)$$

where $|\alpha| \leq 3$, $n = 0, 1$, $l(x, x_0)$ is the length of a geodesic line, generated by the eikonal equation corresponding to the function σ, connecting points x and $x_0, x \neq x_0$, and $f(x, x_0)$ is a certain function, nonzero for $x \in \overline{\Omega}$. This Lemma follows from Theorem 4.1 of [9]. We note that asymptotic behavior (18) is fulfilled for the general hyperbolic equation of the second order under the assumption that the geodesic lines are regular; see Remarks 2.3.1 of [3].

Using Eq. (18) we get the following asymptotic behavior for the function $v(x, s)$ of Eq. (17):

$$\left\| \frac{\partial^n v(\cdot, s)}{\partial s^n} \right\|_{C^{2+\alpha}(\overline{\Omega})} = O\left(\frac{1}{s^n}\right), \quad s \to \infty, \ n = 0, 1, \tag{19}$$

where $C^{2+\alpha}(\overline{\Omega})$ is the Hölder space of order $2 + \alpha$, $0 \leq \alpha < 1$.

3.2 The Transformation Procedure

In this section we will show how to reduce the inverse problem **CIP** to the solution of a nonlinear integral differential equation. First, we write $\mathbf{E}(x, s) = \mathrm{e}^{sv(x,s)}$ with v defined in Eq. (17). Substituting this into Eq. (16) and noting that $x_0 \notin \overline{\Omega}$, we get the equation

$$\Delta v(x, s) + s(\nabla v(x, s))^2 = \mu \varepsilon s + \mu \sigma(x). \tag{20}$$

Given knowledge of the functions v, as well as the coefficients μ and ε, we calculate $\sigma(x)$ explicitly from Eq. (20):

$$\sigma(x) = \frac{1}{\mu}(\Delta v(x, s) + s(\nabla v(x, s))^2) - \varepsilon s \tag{21}$$

for any $s \geq \underline{s}$. However, v is at this point unknown. Next, denoting

$$q(x, s) = \frac{\partial v(x, s)}{\partial s} \tag{22}$$

and differentiating equation (20) with respect to s yields

$$\Delta q(x, s) + (\nabla v(x, s))^2 + 2s \nabla v(x, s) \cdot \nabla q(x, s) = \mu \varepsilon. \tag{23}$$

Using asymptotic behavior (19) in Eq. (22) we get

$$v(x, s) = -\int_s^{\infty} q(x, s) ds. \tag{24}$$

Next, we define the so-called tail function $V(x,\bar{s})$ as

$$V(x,\bar{s}) := \int_{\bar{s}}^{\infty} q(x,\tau)\,\mathrm{d}\tau = v(x,s) + \int_{s}^{\bar{s}} q(x,\tau)\,\mathrm{d}\tau, \tag{25}$$

allowing us to rewrite Eq. (23) on the form

$$\begin{aligned} A(q)(x,s) :=& \Delta q(x,s) + (\nabla V(x,\bar{s}))^2 + \left(\int_s^{\bar{s}} \nabla q(x,\tau)\,\mathrm{d}\tau\right)^2 \\ &- 2\nabla V(x,\bar{s}) \cdot \int_s^{\bar{s}} \nabla q(x,\tau)\,\mathrm{d}\tau + 2s\nabla V(x,\bar{s}) \cdot \nabla q(x,s) \\ &- 2s\int_s^{\bar{s}} \nabla q(x,\tau)\,\mathrm{d}\tau \cdot \nabla q(x,s) = \mu\varepsilon. \end{aligned} \tag{26}$$

In view of Eq. (14), we may write

$$q(x,s)\Big|_{\Gamma} = \frac{\partial}{\partial s}\frac{\ln\left(\mathscr{L}\left[g(x,\cdot)\right](s)\right)}{s} =: \varphi(x,s), \tag{27}$$

which, together with Eq. (26), constitutes a nonlinear problem for the unknown function q, given knowledge of the tail function $V(x,\bar{s})$. Under assumption that $V(x,\bar{s})$ or some approximation thereof is known we now derive a frequency discretized analogue of the problem (26) and (27).

Define a partition $\underline{s} = s_N < s_{N-1} < \ldots < s_1 = \bar{s}$ with $s_n - s_{n+1} = h$ for $n = 1,\ldots,N-1$. We assume that q is a constant function of s on each interval $(s_{n+1}, s_n]$ and require that it satisfies Eqs. (26) and (27) in weighted average on each such interval. That is, $q(x,s) = q_n(x)$, $s \in (s_{n+1}, s_n]$,

$$\int_{s_{n+1}}^{s_n} w_{1,n}(s)A(q)(x,s)\,\mathrm{d}s = \mu\varepsilon \int_{s_{n+1}}^{s_n} w_{1,n}(s)\,\mathrm{d}s, \qquad n = 1\ldots,N-1, \tag{28}$$

and

$$\int_{s_{n+1}}^{s_n} w_{2,n}q_n(x)\,\mathrm{d}s = \int_{s_{n+1}}^{s_n} w_{2,n}\varphi(x,s)\,\mathrm{d}s, \qquad n = 1,\ldots,N-1, \tag{29}$$

where $w_{1,n}$ and $w_{2,n}$ are some weight functions.

Similarly with [3] we define so-called Carleman weight functions in pseudo-frequency s, $w_{1,n}(s) = \mathrm{e}^{\lambda(s-s_n)}$, for some parameter $\lambda \gg 1$. This will "reduce" the non-linearity of the Eq. (26). We take $w_{2,n}(s) \equiv 1$ for simplicity.

With these weight functions, and noting that

$$\int_s^{\bar{s}} \nabla q(x,s)\,\mathrm{d}s = (s_n - s)\nabla q_n(x) + \sum_{j=0}^{n-1} h\nabla q_j(x) \text{ for } s \in (s_{n+1}, s_n],$$

where we set $q_0 \equiv 0$, we can now use Eqs. (26) and (28) to get

$$\Delta q_n(x) + B_n(\lambda,h)\left(\nabla V(x,\bar{s}) - \sum_{j=0}^{n-1} h\nabla q_j(x)\right)\nabla q_n(x) \tag{30}$$
$$= \mu\varepsilon - C_n(\lambda,h)(\nabla q_n(x))^2 - \left(\nabla V(x,\bar{s}) - \sum_{j=0}^{n-1} h\nabla q_j(x)\right)^2,$$

where

$$B_n(\lambda,h) = 4\frac{I_1(\lambda,h)}{I_0(\lambda,h)} + 2s_n, \tag{31}$$

$$C_n(\lambda,h) = 3\frac{I_2(\lambda,h)}{I_0(\lambda,h)} + 2s_n\frac{I_1(\lambda,h)}{I_0(\lambda,h)}, \tag{32}$$

$$I_k(\lambda,h) = \int_{-h}^{0} \tau^k e^{\lambda\tau}\,d\tau = (-1)^k \frac{k! - e^{-\lambda h}\sum_{j=0}^{k}\frac{k!}{j!}(\lambda h)^j}{\lambda^{k+1}}. \tag{33}$$

It should be noted that as $\lambda \to \infty$

$$\frac{I_k(\lambda,h)}{I_l(\lambda,h)} = O(\lambda^{l-k}),$$

so that in particular the coefficient $C_n(\lambda,h)$ becomes small, $O(\lambda^{-1})$, for large λ. Thus, for sufficiently large values of λ, we may neglect the first, nonlinear, term of the right hand side of Eq. (30).

Similarly, from Eq. (29) with $w_{n,2}(s) \equiv 1$, we get

$$q_n(x)\Big|_{\Gamma} = \frac{1}{h}\int_{s_n}^{s_{n-1}} \varphi(x,s)\,ds =: \overline{\varphi}_n(x). \tag{34}$$

If $V(x,\bar{s})$ or some approximation thereof is known, we can use the boundary value problem (30), (34) to successively compute q_n for $n = 1, 2, \ldots, N$.

3.3 *Modeling of the Tail Function* $V(x,\bar{s})$

Let the function $\sigma^*(x)$ be the exact solution of our **CIP** for the exact data g^* in Eq. (14) with the known exact functions μ and ε, and let $\mathbf{E}^*(x,s)$ be the Laplace transform of the corresponding solution to Eq. (15). We define the exact tail function

$$V^*(x,\bar{s}) = \frac{\ln(\mathbf{E}^*(x,\bar{s}))}{\bar{s}}. \tag{35}$$

Let $q^*(x,s)$ and $\varphi^*(x,s)$ be the exact functions corresponding to $q(x,s)$ and $\varphi(x,s)$ in Eq. (26), respectively, defined from the following nonlinear integral differential equation

$$\begin{aligned} A(q^*)(x,s) :=& \Delta q^*(x,s) + (\nabla V^*(x,\bar{s}))^2 + \left(\int_s^{\bar{s}} \nabla q^*(x,\tau)\,\mathrm{d}\tau\right)^2 \\ & - 2\nabla V^*(x,\bar{s}) \cdot \int_s^{\bar{s}} \nabla q^*(x,\tau)\,\mathrm{d}\tau + 2s\nabla V^*(x,\bar{s}) \cdot \nabla q^*(x,s) \\ & - 2s\int_s^{\bar{s}} \nabla q^*(x,\tau)\,\mathrm{d}\tau \cdot \nabla q^*(x,s) = \mu\varepsilon \end{aligned} \tag{36}$$

with

$$q^*(x,s)\Big|_{\Gamma} = \varphi^*(x,s). \tag{37}$$

Using Eq. (19) assume that the functions V^* and q^* have the following asymptotic behavior:

$$\begin{aligned} V^*(x,\bar{s}) &= p^*(x) + \frac{f^*(x)}{\bar{s}} + O\left(\frac{1}{\bar{s}^2}\right) \approx p^*(x) + \frac{f^*(x)}{\bar{s}},\ \bar{s} \to \infty, \\ q^*(x,\bar{s}) &= \partial_{\bar{s}} V^*(x,\bar{s}) = -\frac{f^*(x)}{\bar{s}^2} + O\left(\frac{1}{\bar{s}^3}\right) \approx -\frac{f^*(x)}{\bar{s}^2},\ \bar{s} \to \infty. \end{aligned} \tag{38}$$

We take $s = \bar{s}$ in Eqs. (36) and (37) to get

$$\begin{aligned} \Delta q^* + 2\bar{s}\nabla q^* \nabla V^* + (\nabla V^*)^2 &= \mu\varepsilon && \text{in } \Omega, \\ q^*(x,\bar{s}) &= \psi^*(x,\bar{s}) && \text{for } x \in \Gamma. \end{aligned} \tag{39}$$

Then we use the first two terms in the asymptotic behavior (38) for the exact tail $V^*(x,\bar{s}) = p^*(x) + \frac{f^*(x)}{\bar{s}}$ and for the exact function $q^*(x,\bar{s}) = -\frac{f^*(x)}{\bar{s}^2}$ to obtain

$$\begin{aligned} -\frac{\Delta f^*}{\bar{s}^2} - 2\bar{s}\left(\nabla p^* + \frac{\nabla f^*}{\bar{s}}\right) \cdot \frac{\nabla f^*}{\bar{s}^2} + \left(\nabla p^* + \frac{(\nabla f^*)}{\bar{s}}\right)^2 &= \mu\varepsilon && \text{in } \Omega, \\ -\frac{f^*(x)}{\bar{s}^2} &= \psi^*(x,\bar{s}) && \text{for } x \in \Gamma. \end{aligned}$$

Multiplying the above equation by $-\bar{s}^2$, we obtain the following *approximate* Dirichlet boundary value problem for the functions $p^*, f^* \in C^{2+\alpha}$:

$$\Delta f^* + (\nabla f^*)^2 - \bar{s}^2(\nabla p^*)^2 = -\bar{s}^2\mu\varepsilon \qquad \text{in } \Omega, \tag{40}$$

$$f^*(x) = -\bar{s}^2\psi^*(x,\bar{s}) \qquad \text{for } x \in \Gamma. \tag{41}$$

The function $p^*(x)$ in Eq. (40) can be determined by taking only the first term in the asymptotic behavior in Eq. (38) assuming that

$$V^*(x,\bar{s}) = p^*(x), \qquad q^*(x,\bar{s}) = 0. \tag{42}$$

Then substituting Eq. (42) in Eq. (39), we get the following equation for the function $p^*(x)$:

$$(\nabla p^*)^2 = \mu\varepsilon \quad \text{in } \Omega, \qquad p^* = 0 \quad \text{on } \Gamma, \tag{43}$$

where the boundary condition is obtained from the asymptotics for the function $q^*(x,\bar{s}) = 0$.

3.4 New Approximate Mathematical Model

In this subsection we will present the new approximate mathematical model for the solution of our **CIP** using a new representation of the tail function $V(x,\bar{s})$. Let conditions (13) and (14) hold. Then there exists functions $p^*(x), f^*(x) \in C^{2+\alpha}\left(\overline{\Omega}\right)$ such that the exact tail function $V^*(x)$ has the form

$$V^*(x,s) := p^*(x) + \frac{f^*(x)}{s} \tag{44}$$

for $s \geq \bar{s}$. Here we used assumption that

$$V^*(x,\bar{s}) = p^*(x) + \frac{f^*(x)}{\bar{s}} = \frac{\ln(\mathbf{E}^*(x,\bar{s}))}{\bar{s}^2}. \tag{45}$$

Using definition $q^*(x,s) = \partial_s V^*(x,s)$ for $s \geq \bar{s}$, we get from Eq. (44)

$$q^*(x,\bar{s}) = -\frac{f^*(x)}{\bar{s}^2}. \tag{46}$$

Then we can obtain the following explicit formula for reconstruction of the coefficient $\sigma^*(x)$

$$\sigma^*(x) = \frac{1}{\mu}\left(\Delta v^*(x,s) + s(\nabla v^*(x,s))^2\right) - \varepsilon s, \tag{47}$$

where

$$v^* = -\int_s^{\bar{s}} q^*(x,\tau)\,d\tau + p^*(x) + \frac{f^*(x)}{\bar{s}}.$$

Using the new mathematical model above we can obtain the first guess for the tail function $V(x,\bar{s})$ in Eq. (26) as

$$V_{0,0}(x) := p(x) + \frac{f(x)}{\bar{s}}. \tag{48}$$

Here, the function $p(x)$ is determined by solution of the problem (43), and the function $f(x)$ is the solution of the problem (40), (41) with the computed function $p(x)$.

3.5 The Algorithm

We are now ready to present an approximately globally convergent algorithm for **CIP**.

Step 0. Construct the initial approximation $V_{0,0}$ of the tail function $V(x,\bar{s})$. This can be done by first solving Eq. (15) with $\sigma \equiv 1$ or applying Eq. (48) using the new mathematical model of Sect. 3.4. Set $q_0 \equiv 0$, and set counters n and i to 1, and i_0 and m to 0.

Step 1. Calculate an approximation $q_{n,i}^m$ of q_n from Eqs. (30), (34) with $V = V_{n,i-1}$ if $i > 1$ or $V = V_{n-1,i_{n-1}}$ if $i = 1$, and $(\nabla q_n)^2 = (\nabla q_{n,i}^{m-1})^2$ if $m > 0$ or $(\nabla q_n)^2 = 0$ if $m = 0$.

Step 2. If $m = 0$, set $m = 1$ and return to Step 1. Otherwise, calculate

$$d_{n,i}^m = \frac{\|q_{n,i}^m - q_{n,i}^{m-1}\|_{L^2(\Omega)}}{\|q_{n,i}^{m-1}\|_{L^2(\Omega)}}.$$

If either $d_{n,i}^m < \eta_1$ for some predefined tolerance η_1, or $d_{n,i}^m > d_{n,i}^{m-1}$, set $q_{n,i} = q_{n,i}^m$ and $m = 0$, then proceed to Step 3. Otherwise, increase m by 1 and return to Step 1.

Step 3. Calculate $v_{n,i} = -hq_{n,i} - h\sum_{j=0}^{n-1} q_j$ and then $\sigma_{n,i}$ using Eq. (21) with $v = v_{n,i}$ and $s = s_n$ and extend $\sigma_{n,i}$ to all of $\mathbb{R}^n$ so that $\sigma_{n,i} \in C_{d,\Omega}$. Compute $E_{n,i}$ by solving Eq. (15) with $\sigma = \sigma_{n,i}$ and then $\mathbf{E}_{n,i}$ by applying the Laplace transform to $E_{n,i}$ for $s = \bar{s}$. Update the approximation of the tail function V by setting

$$V_{n,i} = \frac{\ln(\mathbf{E}_{n,i})}{\bar{s}}.$$

Step 4. If $i = 1$ set $i = 2$ and return to Step 1. Otherwise, calculate

$$e_{n,i} = \frac{\|\sigma_{n,i} - \sigma_{n,i-1}\|_{L^2(\Omega)}}{\|\sigma_{n,i-1}\|_{L^2(\Omega)}}.$$

If either $e_{n,i} < \eta_2$ for some predefined tolerance η_2, or $e_{n,i} > e_{n,i-1}$, set $q_n = q_{n,i}$, $V_{n+1,0} = V_{n,i}$, $\sigma_n = \sigma_{n,i}$, $i_n = i$, then set $i = 0$ and proceed to Step 5. Otherwise, increase i by 1 and return to Step 1.

Step 5. If $n = 1$, return to Step 1. Otherwise, compute

$$f_n = \frac{\|\sigma_n - \sigma_{n-1}\|_{L^2(\Omega)}}{\|\sigma_{n-1}\|_{L^2(\Omega)}}.$$

If either $f_n < \eta_3$ for some predefined tolerance η_3, $f_n > f_{n-1}$, or $n = N - 1$, we accept $\sigma = \sigma_n$ as an approximate solution of **CIP** and stop the calculations. Otherwise, we increase n by 1 and return to Step 1.

Acknowledgements This research was supported by the Swedish Research Council, the Swedish Foundation for Strategic Research (SSF) through the Gothenburg Mathematical Modelling Centre (GMMC), and the Swedish Institute, Visby Program.

References

1. Beilina, L., Klibanov, M.V.: A globally convergent numerical method for a coefficient inverse problem. SIAM J. Sci. Comp. **31**, 478–509 (2008)
2. Beilina, L., Klibanov, M.V.: Reconstruction of dielectrics from experimental data via a hybrid globally convergent/adaptive inverse algorithm. Inv. Probl. **26**, 125009 (2010)
3. Beilina, L., Klibanov, M.V.: Approximate Global Convergence and Adaptivity for Coefficient Inverse Problems. Springer, New York (2012)
4. Gubatenko, V.P.: On the formulation of inverse problem in electrical prospecting. Inverse Problems and Large-Scale Computations, Springer Proceedings in Mathematics & Statistics, **52**, (2013)
5. Hammond, P., Sykulski, J.K.: Engineering Electromagnetism: Physical Processes and Computation. Oxford University Press, Oxford (1994)
6. Klibanov, M.V., Timonov, A.: A unified framework for constructing the globally convergent algorithms for multidimensional coefficient inverse problems. Appl. Anal. **83**, 933–955 (2004)
7. Klibanov, M.V., Timonov, A.: Carleman Estimates for Coefficient Inverse Problems and Numerical Applications. VSP, Utrecht (2004)
8. Klibanov, M.V., Fiddy, M.A., Beilina, L., Pantong, N., Schenk, J.: Picosecond scale experimental verification of a globally convergent numerical method for a coefficient inverse problem. Inv. Probl. **26**, 045003 (2010)
9. Romanov, V.G.: Inverse Problems of Mathematical Physics. VNU, Utrecht (1986)
10. Xin, J., Klibanov, M.V.: Numerical solution of an inverse problem of imaging of antipersonnel land mines by the globally convergent convexification algorithm. SIAM J. Sci. Comp. **30**, 3170–3196 (2008)

Preset Field Approximation and Self-consistent Analysis of the Scattering and Generation of Oscillations by a Layered Structure

Lutz Angermann, Vasyl V. Yatsyk, and Mykola V. Yatsyk

Abstract Nonlinear dielectrics with controllable permittivity are subject of intense studies and begin to find broad applications, for instance in device technology and electronics. Based on a model of resonance scattering and generation of waves on an isotropic nonmagnetic nonlinear layered dielectric structure which is excited by a packet of plane waves, we compare two numerical algorithms for simulating various effects of the fields at multiple frequencies.

1 Introduction

In this paper we compare two numerical algorithms for simulating the effects of fields at multiple frequencies on the scattering and generation of oscillations by an isotropic, nonmagnetic, linearly polarised (E-polarisation), layered, cubically polarisable, dielectric structure. The analysis is performed in the domain of resonance frequencies. We consider wave packets consisting of both strong electromagnetic fields at the excitation frequency of the nonlinear structure, leading to the generation of waves, and of weak fields at the multiple frequencies, which do not lead to the

L. Angermann (✉)
Technische Universität Clausthal, Institut für Mathematik, Erzstr. 1, 38678, Clausthal-Zellerfeld, Federal Republic of Germany
e-mail: lutz.angermann@tu-clausthal.de

V.V. Yatsyk
O.Ya. Usikov Institute for Radiophysics and Electronics of the National Academy of Sciences of Ukraine, 12 Ac. Proskura Str., Kharkov, 61085, Ukraine
e-mail: yatsyk@vk.kharkov.ua; vasyl.yatsyk@rambler.ru

M.V. Yatsyk
Kharkov National University of Radioelectronics, 14 Lenina Str., Kharkov, 61166, Ukraine
e-mail: kolya.yatsyk@rambler.ru

L. Beilina and Y.V. Shestopalov (eds.), *Inverse Problems and Large-Scale Computations*, Springer Proceedings in Mathematics & Statistics 52, DOI 10.1007/978-3-319-00660-4_4,

generation of harmonics but influence on the process of scattering and generation of waves by the nonlinear structure. As has been shown by the authors in previous papers [1, 2] a self-consistent formulation of a finite nonlinear system of boundary-value problems with respect to the components of the scattered and generated fields forms the appropriate mathematical model. It is also known that this nonlinear system is equivalent to a nonlinear system of one-dimensional nonlinear integral equations of the second kind. The solution of the system of nonlinear integral equations is approximated numerically by the help of the quadrature method. The numerical self-consistent algorithms under consideration are based on block-iterative procedures which require the solution of linear systems in each step. In this way the approximate solution of the self-consistent nonlinear problems is described by means of solutions of linear problems with an induced nonlinear permittivity. The first step of this algorithm is called preset field method. It consists of finding the Kerr approximation of the fields caused by the sources of the incident field. Further, these fields are used as the preset fields. They induce the permittivity and the sources of energy generation caused by the nonlinear terms of the coefficients of the cubic susceptibility of the medium. In this work, results of calculations of characteristics of the scattered and generated fields of plane waves are presented, taking into account the influence of weak fields at multiple frequencies on the cubically polarisable layer. We restrict ourselves to the investigation of the third harmonic generated by layers with both negative and positive values of the cubic susceptibility of the medium. The preset field method allows to obtain a preliminary, approximate solution of the problem and to estimate some of the qualitative characteristics of the scattered and generated oscillations without significant computational costs. The disadvantages of the method include the absence of energy balance conditions. The preset field method is significantly different from the self-consistent approach. Within the framework of the self-consistent system, we show the following. The variation of the imaginary parts of the permittivities of the layer at the multiple frequencies can take both positive and negative values along the height of the nonlinear layer. It is induced by the nonlinear part of the permittivities and is caused by the loss of energy in the nonlinear medium which is spent for the generation of the electromagnetic fields. The magnitudes of these variations are determined by the amplitude and phase characteristics of the fields which are scattered and generated by the nonlinear layer. We present and discuss results of calculations of the scattered field taking into account the third harmonic generated by nonlinear layer. We show that the portion of total energy generated in the third harmonic may reach up to 36 % which exceeds significantly the known results [4].

2 Technique

We consider a layered nonlinear medium which is located in the region $\{\mathbf{r} = (x,y,z)^\top \in \mathbb{R}^3 : \ |z| \le 2\pi\delta\}$; see Fig. 1.

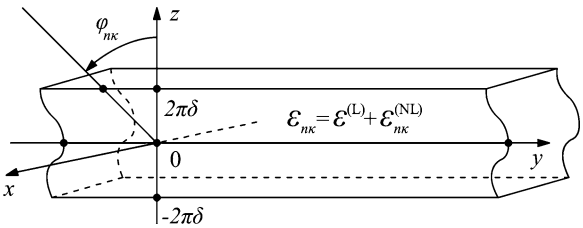

Fig. 1 The nonlinear dielectric layered structure of thickness $4\pi\delta$ with dielectric permittivity $\varepsilon_{n\kappa}$, which is excited by plane waves at the frequencies $n\kappa$ under the incident angles $\varphi_{n\kappa}$

It is assumed that the vector of the polarisation moment $\mathbf{P}$ can be expanded as follows:

$$\mathbf{P} = \chi^{(1)}\mathbf{E} + (\chi^{(2)}\mathbf{E})\mathbf{E} + ((\chi^{(3)}\mathbf{E})\mathbf{E})\mathbf{E} + \text{ h.o.t.},$$

where $\chi^{(1)}$, $\chi^{(2)}$, $\chi^{(3)}$ are the media susceptibility tensors of rank two, three and four. In the case of isotropic media, the quadratic term disappears. It is convenient to split $\mathbf{P}$ into its linear and nonlinear parts $\mathbf{P} = \mathbf{P}^{(L)} + \mathbf{P}^{(NL)} := \chi^{(1)}\mathbf{E} + \mathbf{P}^{(NL)}$. Similarly, with $\varepsilon := \mathbf{I} + 4\pi\chi^{(1)}$ and $\mathbf{D}^{(L)} := \varepsilon\mathbf{E}$, the electric displacement field can be decomposed as

$$\mathbf{D} = \mathbf{D}^{(L)} + 4\pi\mathbf{P}^{(NL)}. \tag{1}$$

Furthermore, if the media under consideration are nonmagnetic, isotropic and transversely inhomogeneous w.r.t. z, i.e., $\varepsilon = \varepsilon^{(L)}\mathbf{I}$ with a scalar, possibly complex-valued function $\varepsilon^{(L)} = \varepsilon^{(L)}(z)$, if the wave is linearly E-polarised, i.e.,

$$\mathbf{E} = (E_1, 0, 0)^\top, \quad \mathbf{H} = (0, H_2, H_3)^\top, \tag{2}$$

and if the electric field $\mathbf{E}$ is homogeneous w.r.t. the coordinate x, i.e., $\mathbf{E}(\mathbf{r},t) = \mathbf{E}(y,z,t) = (E_1(y,z,t),0,0)^\top$, then the Maxwell's equations together with (1) reduce to

$$\Delta\mathbf{E} - \frac{\varepsilon^{(L)}}{c^2}\frac{\partial^2}{\partial t^2}\mathbf{E} - \frac{4\pi}{c^2}\frac{\partial^2}{\partial t^2}\mathbf{P}^{(NL)} = 0, \tag{3}$$

where Δ reduces to the Laplacian w.r.t. y and z, i.e., $\Delta := \partial^2/\partial y^2 + \partial^2/\partial z^2$.

A stationary electromagnetic wave (with the oscillation frequency $\omega > 0$) propagating in a nonlinear dielectric structure gives rise to a field containing all frequency harmonics. Therefore, representing $\mathbf{E}$, $\mathbf{P}^{(NL)}$ via Fourier series

$$\mathbf{E}(\mathbf{r},t) = \frac{1}{2}\sum_{s\in\mathbb{Z}} \mathbf{E}(\mathbf{r},s\omega)e^{-is\omega t}, \quad \mathbf{P}^{(NL)}(\mathbf{r},t) = \frac{1}{2}\sum_{s\in\mathbb{Z}} \mathbf{P}^{(NL)}(\mathbf{r},s\omega)e^{-is\omega t},$$

we obtain from (3) an infinite system of coupled equations w.r.t. the Fourier amplitudes. In the case of a three-component E-polarised electromagnetic field [cf. (2)], this system reduces to a system of scalar equations w.r.t. E_1:

$$\Delta E_1(\mathbf{r},s\omega)+\frac{\varepsilon^{(L)}(s\omega)^2}{c^2}E_1(\mathbf{r},s\omega)+\frac{4\pi(s\omega)^2}{c^2}P_1^{(NL)}(\mathbf{r},s\omega)=0,\quad s\in\mathbb{N} \tag{4}$$

($E_1(\mathbf{r};j\omega)=\overline{E}_1(\mathbf{r};-j\omega)$ and $E_1(\mathbf{r},s\omega)|_{s=0}=0$).

We assume that the main contribution to the nonlinearity is introduced by the term $\mathbf{P}^{(NL)}(\mathbf{r},s\omega)$ (cf. [1,2,6]), and we take only the lowest-order terms in the Taylor series expansion of the nonlinear part $\mathbf{P}^{(NL)}(\mathbf{r},s\omega)=(P_1^{(NL)}(\mathbf{r},s\omega),0,0)^\top$ of the polarisation vector in the vicinity of the zero value of the electric field intensity. In this case, the only non-trivial component of the polarisation vector is determined by the susceptibility tensor $\chi^{(3)}$, and we have that

$$\begin{aligned}&P_1^{(NL)}(\mathbf{r},s\omega)\\ &\dot{=}\frac{3}{4}\sum_{j\in\mathbb{N}}\chi^{(3)}_{1111}(s\omega;j\omega,-j\omega,s\omega)|E_1(\mathbf{r},j\omega)|^2E_1(\mathbf{r},s\omega)\\ &\quad+\frac{1}{4}\sum_{\left\{\begin{smallmatrix}n,m,p\in\mathbb{Z}\setminus\{0\}\\ n\neq -m,\ p=s\\ m\neq -p,\ n=s\\ n\neq -p,\ m=s\\ n+m+p=s\end{smallmatrix}\right.}\chi^{(3)}_{1111}(s\omega;n\omega,m\omega,p\omega)E_1(\mathbf{r},n\omega)E_1(\mathbf{r},m\omega)E_1(\mathbf{r},p\omega),\end{aligned} \tag{5}$$

where the symbol $\dot{=}$ means that higher-order terms are neglected.

If we study nonlinear effects involving the waves at the first three frequency components of E_1 only, it is possible to restrict the system (4), (5) to three equations. Using Kleinman's rule (i.e. the equality of all coefficients $\chi^{(3)}_{1111}$ at multiple frequencies, [3,4]), we obtain the system

$$\begin{aligned}&\Delta E_1(\mathbf{r},n\kappa)+(n\kappa)^2\varepsilon_{n\kappa}(z,\alpha(z),E_1(\mathbf{r},\kappa),E_1(\mathbf{r},2\kappa),E_1(\mathbf{r},3\kappa))\\ &\quad=-\delta_{n1}\kappa^2\alpha(z)E_1^2(\mathbf{r},2\kappa)\overline{E}_1(\mathbf{r},3\kappa)\\ &\qquad-\delta_{n3}(3\kappa)^2\alpha(z)\left\{\frac{1}{3}E_1^3(\mathbf{r},\kappa)+E_1^2(\mathbf{r},2\kappa)\overline{E}_1(\mathbf{r},\kappa)\right\},\ n=1,2,3,\end{aligned} \tag{6}$$

where $\kappa:=\frac{\omega}{c}=\frac{2\pi}{\lambda}$, $\varepsilon_{n\kappa}:=\begin{cases}1, & |z|>2\pi\delta,\\ \varepsilon^{(L)}+\varepsilon^{(NL)}_{n\kappa}, & |z|\le 2\pi\delta,\end{cases}$ and $\varepsilon^{(L)}:=1+4\pi\chi^{(1)}_{11}$,

$$\begin{aligned}\varepsilon^{(NL)}_{n\kappa}:=\alpha(z)\Big[&\sum_{j=1}^{3}|E_1(\mathbf{r},j\kappa)|^2+\delta_{n1}\frac{\left[\overline{E}_1(\mathbf{r},\kappa)\right]^2}{E_1(\mathbf{r},\kappa)}E_1(\mathbf{r},3\kappa)\\ &+\delta_{n2}\frac{\overline{E}_1(\mathbf{r},2\kappa)}{E_1(\mathbf{r},2\kappa)}E_1(\mathbf{r},\kappa)E_1(\mathbf{r},3\kappa)\Big]\end{aligned} \tag{7}$$

with $\alpha(z):=3\pi\chi^{(3)}_{1111}(z)$, δ_{nj} ... Kronecker's symbol. In addition, the following conditions are met:

(C1) $E_1(n\kappa;y,z) = U(n\kappa;z)\exp(i\phi_{n\kappa}y), \quad n=1,2,3$
(the quasi-homogeneity w.r.t. y),
(C2) $\phi_{n\kappa} = n\phi_\kappa, \quad n=1,2,3$
(the condition of phase synchronism of waves),
(C3) $\mathbf{E}_{\mathrm{tg}}(n\kappa;y,z)$ and $\mathbf{H}_{\mathrm{tg}}(n\kappa;y,z)$ (i.e. $E_1(n\kappa;y,z)$ and $H_2(n\kappa;y,z)$)
are continuous across the interfaces,
(C4) $E_1^{\mathrm{scat}}(n\kappa;y,z) = \left\{ \begin{matrix} a_{n\kappa}^{\mathrm{scat}} \\ b_{n\kappa}^{\mathrm{scat}} \end{matrix} \right\} \exp\left(i\left(\phi_{n\kappa}y \pm \Gamma_{n\kappa}(z \mp 2\pi\delta)\right)\right), \quad z \gtrless \pm 2\pi\delta$
(the radiation condition),

where $\phi_{n\kappa} := n\kappa\sin\varphi_{n\kappa}$, $\varphi_{n\kappa}$, is the incident angle of the exciting wave of frequency $n\kappa$, $|\varphi_{n\kappa}| < \pi/2$, and $\Gamma_{n\kappa} := \sqrt{(n\kappa)^2 - \phi_{n\kappa}^2}$ with $\mathfrak{Re}\,\Gamma_{n\kappa} > 0$, $\mathfrak{Im}\,\Gamma_{n\kappa} = 0$.

A detailed discussion of the condition (C2) is given in [1, Sect. 3]. In particular, this condition implies that $\varphi_{n\kappa} = \varphi_\kappa$. The condition (C4) provides a physically consistent behaviour of the energy characteristics of scattering and generation. It guarantees the absence of waves coming from infinity (i.e. $z = \pm\infty$); see [5].

The investigation of quasi-homogeneous fields $E_1(n\kappa;y,z)$ [cf. condition (C1)] in a nonlinear transversely inhomogeneous dielectric layer shows that under the condition of the phase synchronism (C2) the components of the nonlinear polarisation $P_1^{(G)}(\mathbf{r},n\kappa)$ [playing the role of the sources generating radiation in the right-hand sides of the system (6)] satisfy the quasi-homogeneity condition, too. Furthermore, thanks to condition (C2) the nonlinear transversely inhomogeneous dielectric structure does not depend on the longitudinal coordinate y, i.e., it remains as a nonlinear layered dielectric structure [1]. These facts allow to write the desired solution in the form

$$\begin{aligned} E_1(n\kappa;y,z) &= U(n\kappa;z)\exp(i\phi_{n\kappa}y) \\ &= \begin{cases} a_{n\kappa}^{\mathrm{inc}}\exp(i(\phi_{n\kappa}y - \Gamma_{n\kappa}(z-2\pi\delta))) \\ \quad + a_{n\kappa}^{\mathrm{scat}}\exp(i(\phi_{n\kappa}y + \Gamma_{n\kappa}(z-2\pi\delta))), & z > 2\pi\delta, \\ U(n\kappa;z)\exp(i\phi_{n\kappa}y), & |z| \le 2\pi\delta, \\ b_{n\kappa}^{\mathrm{scat}}\exp(i(\phi_{n\kappa}y - \Gamma_{n\kappa}(z+2\pi\delta))), & z < -2\pi\delta, \end{cases} \\ &\quad n=1,2,3. \end{aligned} \tag{8}$$

The substitution of the ansatz (8) into the PDE system (6) results in a system of semilinear boundary-value problems of Sturm–Liouville type:

$$\begin{aligned} &\frac{d^2}{dz^2}U(n\kappa;z) + \left\{\Gamma_{n\kappa}^2 - (n\kappa)^2\left[1 - \varepsilon_{n\kappa}(z,\alpha(z),U(\kappa;z),U(2\kappa;z),U(3\kappa;z))\right]\right\}U(n\kappa;z) \\ &\quad = -(n\kappa)^2\alpha(z)\left(\delta_{n1}U^2(2\kappa;z)\overline{U}(3\kappa;z) + \delta_{n3}\left\{\frac{1}{3}U^3(\kappa;z) + U^2(2\kappa;z)\overline{U}(\kappa;z)\right\}\right), \\ &|z| \le 2\pi\delta, \quad n=1,2,3. \end{aligned} \tag{9}$$

By elementary calculations, from (C3) we obtain the boundary conditions for (9):

$$
\begin{aligned}
&i\Gamma_{n\kappa}U(n\kappa;-2\pi\delta)+\frac{d}{dz}U(n\kappa;-2\pi\delta)=0,\\
&i\Gamma_{n\kappa}U(n\kappa;2\pi\delta)-\frac{d}{dz}U(n\kappa;2\pi\delta)=2i\Gamma_{n\kappa}a_{n\kappa}^{\mathrm{inc}},\quad n=1,2,3.
\end{aligned}
\tag{10}
$$

Problem (6), (C1)–(C4) can also be reduced to finding solutions of one-dimensional nonlinear integral equations (cf. [1, 2]) w.r.t. the unknown functions $U(n\kappa;\cdot)\in L_2(-2\pi\delta,2\pi\delta)$:

$$
\begin{aligned}
&U(n\kappa;z)+\frac{i(n\kappa)^2}{2\Gamma_{n\kappa}}\int_{-2\pi\delta}^{2\pi\delta}\exp(i\Gamma_{n\kappa}|z-z_0|)\\
&\quad\times[1-\varepsilon_{n\kappa}(z_0,\alpha(z_0),U(\kappa;z_0),U(2\kappa;z_0),U(3\kappa;z_0))]U(n\kappa;z_0)dz_0\\
&=\delta_{n1}\frac{i(n\kappa)^2}{2\Gamma_{n\kappa}}\int_{-2\pi\delta}^{2\pi\delta}\exp(i\Gamma_{n\kappa}|z-z_0|)\alpha(z_0)U^2(2\kappa;z_0)\overline{U}(3\kappa;z_0)dz_0\\
&\quad+\delta_{n3}\frac{i(n\kappa)^2}{2\Gamma_{n\kappa}}\int_{-2\pi\delta}^{2\pi\delta}\exp(i\Gamma_{n\kappa}|z-z_0|)\alpha(z_0)\Big\{\frac{1}{3}U^3(\kappa;z_0)\\
&\quad+U^2(2\kappa;z_0)\overline{U}(\kappa;z_0)\Big\}dz_0+U^{\mathrm{inc}}(n\kappa;z),\\
&|z|\le 2\pi\delta,\quad n=1,2,3,
\end{aligned}
\tag{11}
$$

where $U^{\mathrm{inc}}(n\kappa;z)=a_{n\kappa}^{\mathrm{inc}}\exp[-i\Gamma_{n\kappa}(z-2\pi\delta)]$. The system of nonlinear integral equation (11) can be transformed into the system of nonlinear Sturm–Liouville boundary-value problems (9) and (10). This indicates the equivalence of the system (11) with the problems (9), (10) and (6), (C1)–(C4).

The application of a suitable quadrature rule to (11) leads to a system of complex-valued nonlinear algebraic equations:

$$
(\mathbf{I}-\mathbf{B}_{n\kappa}(\mathbf{U}_\kappa,\mathbf{U}_{2\kappa},\mathbf{U}_{3\kappa}))\mathbf{U}_{n\kappa}=\delta_{n1}\mathbf{C}_\kappa(\mathbf{U}_{2\kappa},\mathbf{U}_{3\kappa})+\delta_{n3}\mathbf{C}_{n\kappa}(\mathbf{U}_\kappa,\mathbf{U}_{2\kappa})+\mathbf{U}_{n\kappa}^{\mathrm{inc}},
\tag{12}
$$

$n=1,2,3$, where we use a discrete set $\{z_l\}_{l=1}^N$ of nodes such that $-2\pi\delta=:z_1<z_2<\cdots<z_l<\cdots<z_N=:2\pi\delta$.

$\mathbf{U}_{n\kappa}:=\{U_l(n\kappa)\}_{l=1}^N\approx\{U(n\kappa;z_l)\}_{l=1}^N$, $\mathbf{U}_{n\kappa}^{\mathrm{inc}}:=\{a_{n\kappa}^{\mathrm{inc}}\exp[-i\Gamma_{n\kappa}(z_l-2\pi\delta)]\}_{l=1}^N$, $\mathbf{I}:=\{\delta_{lj}\}_{l,j=1}^N$ denotes the identity matrix, $\mathbf{B}_{n\kappa}(\mathbf{U}_\kappa,\mathbf{U}_{2\kappa},\mathbf{U}_{3\kappa}))$; $\mathbf{C}_\kappa(\mathbf{U}_{2\kappa},\mathbf{U}_{3\kappa})$, $\mathbf{C}_{n\kappa}(\mathbf{U}_\kappa,\mathbf{U}_{2\kappa})$ are the matrices and the right-hand side, resp., generated by the quadrature method.

The solution of (12) can be found iteratively, where at each step a system of linearised algebraic problems is solved.

The analytic continuation of these linearised nonlinear problems into the region of complex values of the frequency parameter allows us to switch to the analysis of spectral problems; see [2]. Then we obtain in a similar manner a set of independent systems of linear algebraic equations depending nonlinearly on the spectral parameter:

$$(\mathbf{I}-\mathbf{B}_{n\kappa}(\kappa_n))\mathbf{U}_{\kappa_n}=\mathbf{0},$$

where $\kappa_n \in \Omega_{n\kappa} \subset \mathsf{H}_{n\kappa}$, at $\kappa = \kappa^{\text{inc}}$, $n = 1,2,3$, $\Omega_{n\kappa}$ are the discrete sets of eigenfrequencies and $\mathsf{H}_{n\kappa}$ denote two-sheeted Riemann surfaces. The matrices $\mathbf{B}_{n\kappa}(\kappa_n) := \mathbf{B}_{n\kappa}(\kappa_n;\mathbf{U}_\kappa,\mathbf{U}_{2\kappa},\mathbf{U}_{3\kappa})$ are defined for *given* $\mathbf{U}_{n\kappa}$. The spectral problem of finding the eigenfrequencies κ_n and the corresponding eigenfields $\mathbf{U}_{\kappa_n}$ (i.e. the non-trivial solutions of the linearised homogeneous integral equations) reduces to the following equations:

$$\begin{cases} f_{n\kappa}(\kappa_n) := \det(\mathbf{I}-\mathbf{B}_{n\kappa}(\kappa_n)) = 0, \\ (\mathbf{I}-\mathbf{B}_{n\kappa}(\kappa_n))\mathbf{U}_{\kappa_n} = \mathbf{0}, \\ \kappa := \kappa^{\text{inc}}, \quad \kappa_n \in \Omega_{n\kappa} \subset \mathsf{H}_{n\kappa}, \quad n = 1,2,3. \end{cases} \tag{13}$$

3 Results

Consider the excitation of the nonlinear structure by a sufficiently strong electromagnetic field at the basic frequency κ only, i.e.,

$$a_\kappa^{\text{inc}} \neq 0, \quad a_{2\kappa}^{\text{inc}} = a_{3\kappa}^{\text{inc}} = 0. \tag{14}$$

In this case, the number of equations in the system (6) can be reduced to two by deleting the second equation and setting $E_1(\mathbf{r},2\kappa) = 0$. The dielectric permittivity (7) of the nonlinear layer simplifies as follows:

$$\begin{aligned} &\varepsilon_{n\kappa}(z,\alpha(z),E_1(\mathbf{r},\kappa),0,E_1(\mathbf{r},3\kappa)) \\ &\quad = \varepsilon_{n\kappa}(z,\alpha(z),U(\kappa;z),U(3\kappa;z)) \\ &\quad =: \varepsilon^{(L)}(z) + \varepsilon_{n\kappa}^{(NL)}(\alpha(z),U(\kappa;z),U(3\kappa;z)), \quad n = 1,3. \end{aligned} \tag{15}$$

Thus, we investigate the problem (6), (7) for $n = 1,3$ and $U(2\kappa;\cdot) = 0$ as in [1, 2]. The algebraic system (12) reduces to

$$\begin{cases} (\mathbf{I}-\mathbf{B}_\kappa(\mathbf{U}_\kappa,\mathbf{U}_{3\kappa}))\mathbf{U}_\kappa = \mathbf{U}_\kappa^{\text{inc}}, \\ (\mathbf{I}-\mathbf{B}_{3\kappa}(\mathbf{U}_\kappa,\mathbf{U}_{3\kappa}))\mathbf{U}_{3\kappa} = \mathbf{C}_{3\kappa}(\mathbf{U}_\kappa). \end{cases} \tag{16}$$

The self-consistent numerical approach (Method SC) consists of the following algorithm. Given a relative error tolerance $\xi > 0$, the solution of (16) is approximated by means of the following iterative method:

$$\left\{ \begin{array}{l} \left\{ \left[\mathbf{I}-\mathbf{B}_\kappa\left(\mathbf{U}_\kappa^{(s-1)},\mathbf{U}_{3\kappa}^{(S_3(q))}\right)\right]\mathbf{U}_\kappa^{(s)} = \mathbf{U}_\kappa^{\text{inc}} \right\}_{s=1}^{S_1(q):\,\eta_1(S_1(q))<\xi} \\ \left\{ \left[\mathbf{I}-\mathbf{B}_{3\kappa}\left(\mathbf{U}_\kappa^{(S_1(q))},\mathbf{U}_{3\kappa}^{(s-1)}\right)\right]\mathbf{U}_{3\kappa}^{(s)} = \mathbf{C}_{3\kappa}\left(\mathbf{U}_\kappa^{(S_1(q))}\right) \right\}_{s=1}^{S_3(q):\,\eta_3(S_3(q))<\xi} \end{array} \right\}_{q=1}^{Q} \tag{17}$$

with $\eta_n(s) := \|\mathbf{U}_{n\kappa}^{(s)} - \mathbf{U}_{n\kappa}^{(s-1)}\| / \|\mathbf{U}_{n\kappa}^{(s)}$, $n = 1,3$, where the terminating index $Q \in \mathbb{N}$ is defined by the requirement $\max\{\eta_1(Q), \eta_3(Q)\} < \xi$. The initial values are $\mathbf{U}_{\kappa}^{(0)} := \mathbf{U}_{3\kappa}^{(0)} := \mathbf{0}$ and $S_3(1) := 0$ for the first outer iteration step in the first system w.r.t. $\mathbf{U}_{\kappa}^{(s)}$.

The first step of this algorithm is called *preset field method.* It consists of finding the solution of the linear problem at the excitation frequency (i.e. the determination of the preset field) and, after this, of solving the linear problem at the generation frequency for an approximation of the field using the already found preset field. This algorithm is called the *zeroth-order preset field method of linear approximation* (Method 0):

$$\begin{aligned} \left[\mathbf{I} - \mathbf{B}_{\kappa}(\mathbf{0},\mathbf{0})\right]\mathbf{U}_{\kappa}^{(1)} &= \mathbf{U}_{\kappa}^{\mathrm{inc}}, \\ \left[\mathbf{I} - \mathbf{B}_{3\kappa}\left(\mathbf{U}_{\kappa}^{(1)},\mathbf{0}\right)\right]\mathbf{U}_{3\kappa}^{(1)} &= \mathbf{C}_{3\kappa}\left(\mathbf{U}_{\kappa}^{(1)}\right). \end{aligned} \tag{18}$$

The first step of the algorithm (17) can also be regarded as a preset field method. Here nonlinear problems are solved. We are looking for an approximation of the Kerr field at the excitation frequency (i.e. a preset field) and for the Kerr approximation at the generation frequency using the already found preset field. We call this algorithm the *first-order preset field method of nonlinear approximation* (Method 1):

$$\begin{aligned} &\left\{\left[\mathbf{I} - \mathbf{B}_{\kappa}\left(\mathbf{U}_{\kappa}^{(s-1)},\mathbf{0}\right)\right]\mathbf{U}_{\kappa}^{(s)} = \mathbf{U}_{\kappa}^{\mathrm{inc}}\right\}_{s=1}^{S_1:\,\eta_1(S_1)<\xi}, \\ &\left\{\left[\mathbf{I} - \mathbf{B}_{3\kappa}\left(\mathbf{U}_{\kappa}^{(S_1)},\mathbf{U}_{3\kappa}^{(s-1)}\right)\right]\mathbf{U}_{3\kappa}^{(s)} = \mathbf{C}_{3\kappa}\left(\mathbf{U}_{\kappa}^{(S_1)}\right)\right\}_{s=1}^{S_3:\,\eta_3(S_3)<\xi}. \end{aligned} \tag{19}$$

Here we have considered linear, noniterative (18) and nonlinear, iterative (19) algorithms for solving the nonlinear problem of scattering and generation of oscillations by the help of a preset field at the excitation frequency. These methods have in common the fact that, in the solution of both the linear (18) and the nonlinear (19) problems, an imaginary part of the nonlinear component of the permittivity is not induced, i.e., $\mathfrak{Im}(\varepsilon_{n\kappa}^{(NL)}) \equiv \mathbf{0}$.

Further, the preset fields [see problems (18) and (19)] can be used as the initial fields in the iteration (17) of the self-consistent approach. They induce the complex permittivities with $\mathfrak{Im}(\varepsilon_{n\kappa}^{(NL)}) \not\equiv \mathbf{0}$ and the sources of energy generation caused by the nonlinear terms of the coefficients of the cubic susceptibility of the medium. Such problems can be solved by means of the self-consistent approach using the algorithm (17), i.e., the Method SC.

Schematically the methods are illustrated in Table 1.

Table 1 The preset field method for the linear (0) and the first nonlinear (1) approximation, SC … self-consistent analysis (known values from the first equation (methods 0,1) or from the previous iteration step (method SC)—cyan, sought values—red)

Method	Algorithm	Permittivity
0	$(\mathbf{I}-\mathbf{B}_\kappa(\mathbf{0},\mathbf{0}))\mathbf{U}_\kappa=\mathbf{U}^{\mathrm{inc}}_\kappa$	$\varepsilon_\kappa\equiv\varepsilon^{(L)}+\varepsilon^{(NL)}_\kappa(\alpha,\mathbf{0},\mathbf{0}))$, $\varepsilon^{(NL)}_\kappa\equiv\mathbf{0}$
	$(\mathbf{I}-\mathbf{B}_{3\kappa}(\mathbf{U}_\kappa,\mathbf{0}))\mathbf{U}_{3\kappa}=\mathbf{C}_{3\kappa}(\mathbf{U}_\kappa)$	$\varepsilon_{3\kappa}\equiv\varepsilon^{(L)}+\varepsilon^{(NL)}_{3\kappa}(\alpha,\mathbf{U}_\kappa,\mathbf{0}))$, $\Im\mathfrak{m}(\varepsilon^{(NL)}_{3\kappa})\equiv\mathbf{0}$
1	$(\mathbf{I}-\mathbf{B}_\kappa(\mathbf{U}_\kappa,\mathbf{0}))\mathbf{U}_\kappa=\mathbf{U}^{\mathrm{inc}}_\kappa$	$\varepsilon_\kappa\equiv\varepsilon^{(L)}+\varepsilon^{(NL)}_\kappa(\alpha,\mathbf{U}_\kappa,\mathbf{0}))$, $\Im\mathfrak{m}(\varepsilon^{(NL)}_\kappa)\equiv\mathbf{0}$
	$(\mathbf{I}-\mathbf{B}_{3\kappa}(\mathbf{U}_\kappa,\mathbf{U}_{3\kappa}))\mathbf{U}_{3\kappa}=\mathbf{C}_{3\kappa}(\mathbf{U}_\kappa)$	$\varepsilon_{3\kappa}\equiv\varepsilon^{(L)}+\varepsilon^{(NL)}_{3\kappa}(\alpha,\mathbf{U}_\kappa,\mathbf{U}_{3\kappa}))$, $\Im\mathfrak{m}(\varepsilon^{(NL)}_{3\kappa})\equiv\mathbf{0}$
SC	$(\mathbf{I}-\mathbf{B}_\kappa(\mathbf{U}_\kappa,\mathbf{U}_{3\kappa}))\mathbf{U}_\kappa=\mathbf{U}^{\mathrm{inc}}_\kappa$	$\varepsilon_\kappa\equiv\varepsilon^{(L)}+\varepsilon^{(NL)}_\kappa(\alpha,\mathbf{U}_\kappa,\mathbf{U}_{3\kappa}))$, $\Im\mathfrak{m}(\varepsilon^{(NL)}_\kappa)\not\equiv\mathbf{0}$
	$(\mathbf{I}-\mathbf{B}_{3\kappa}(\mathbf{U}_\kappa,\mathbf{U}_{3\kappa}))\mathbf{U}_{3\kappa}=\mathbf{C}_{3\kappa}(\mathbf{U}_\kappa)$	$\varepsilon_{3\kappa}\equiv\varepsilon^{(L)}+\varepsilon^{(NL)}_{3\kappa}(\alpha,\mathbf{U}_\kappa,\mathbf{U}_{3\kappa}))$, $\Im\mathfrak{m}(\varepsilon^{(NL)}_{3\kappa})\equiv\mathbf{0}$

3.1 Example 1: Three-Layered Structure

In the first experiment we consider a three-layered structure with a dielectric permittivity of the form (15), where

$$\varepsilon^{(L)}(z):=\begin{cases}16,\ z\in[-2\pi\delta,-2\pi\delta/3],\\ 64,\ z\in[-2\pi\delta/3,2\pi\delta/3],\\ 16,\ z\in[-2\pi\delta/3,2\pi\delta],\end{cases}\quad \alpha(z):=\begin{cases}0.01,\ z\in[-2\pi\delta,-2\pi\delta/3],\\ -0.01,\ z\in[-2\pi\delta/3,2\pi\delta/3],\\ 0.01,\ z\in[-2\pi\delta/3,2\pi\delta],\end{cases}$$

$\delta:=0.5$, $\kappa^{\mathrm{inc}}:=\kappa:=0.25$, and $\varphi_\kappa\in[0^\circ,90^\circ)$.

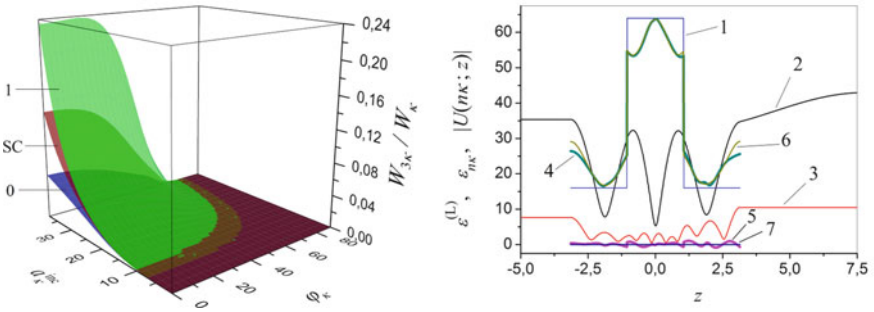

Fig. 2 Portion of energy generated in the third harmonic (*left*): 0… linear approximation in the preset field method, 1… nonlinear approximation in the preset field method, SC… self-consistent approach, properties of the nonlinear layer in the self-consistent approach for $a^{\mathrm{inc}}_\kappa=38$ and $\varphi_\kappa=0^\circ$ (*right*): #1 … $\varepsilon^{(L)}$, #2 … $|U(\kappa;z)|$, #3 … $|U(3\kappa;z)|$, #4 … $\Re\mathfrak{e}(\varepsilon_\kappa)$, #5 … $\Im\mathfrak{m}(\varepsilon_\kappa)$, #6 … $\Re\mathfrak{e}(\varepsilon_{3\kappa})$, #7 … $\Im\mathfrak{m}(\varepsilon_{3\kappa})\equiv0$

The preset field method allows to obtain a preliminary, approximate solution of the problem and to estimate some of the qualitative characteristics of the scattered and generated oscillations without significant computational costs; see Figs. 2–4.

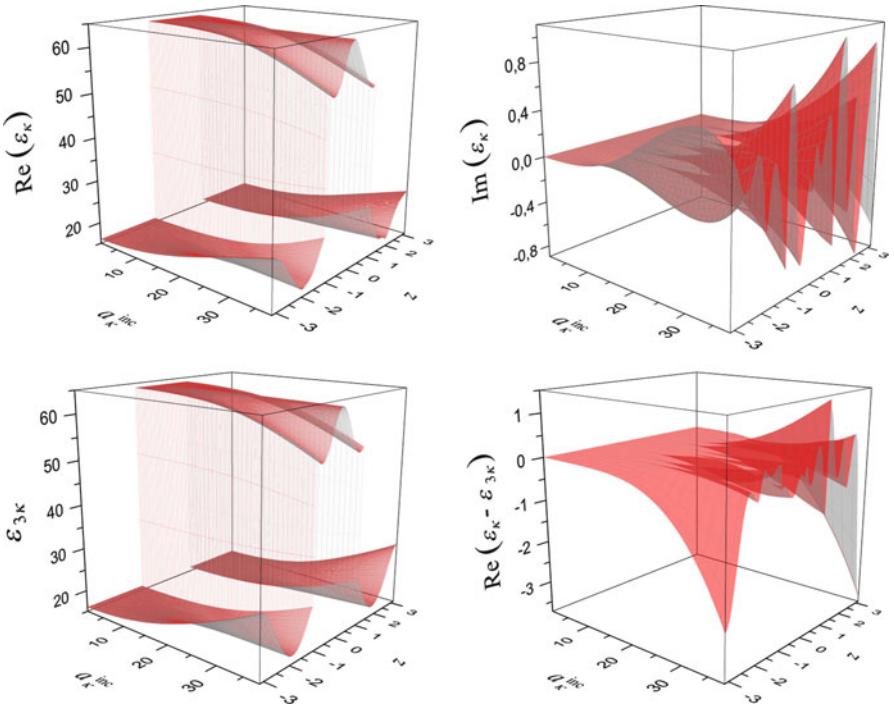

Fig. 3 The nonlinear dielectric permittivity in the self-consistent approach for the excitation frequency $\kappa = 0.25$ and for $\varphi_\kappa = 0°$: $\mathfrak{Re}\left(\varepsilon_\kappa\left[a_\kappa^{\mathrm{inc}}, z\right]\right)$ (*top left*), $\mathfrak{Im}\left(\varepsilon_\kappa\left[a_\kappa^{\mathrm{inc}}, z\right]\right)$ (*top right*), $\varepsilon_{3\kappa}\left[a_\kappa^{\mathrm{inc}}, z\right]$ (*bottom left*), $\mathfrak{Re}\left(\varepsilon_\kappa\left[a_\kappa^{\mathrm{inc}}, z\right] - \varepsilon_{3\kappa}\left[a_\kappa^{\mathrm{inc}}, z\right]\right)$ (*bottom right*)

The scattering and generation properties of the nonlinear layer in the case (14) can be described by means of the *reflection/generation* and *transmission/generation* coefficients

$$R_{n\kappa} := |a_{n\kappa}^{\mathrm{scat}}|^2 / |a_\kappa^{\mathrm{inc}}|^2 \quad \text{and} \quad T_{n\kappa} := |b_{n\kappa}^{\mathrm{scat}}|^2 / |a_\kappa^{\mathrm{inc}}|^2$$

(relative intensities).

In the considered case of a single incident field (14) and for nonabsorbing media with $\mathfrak{Im}(\varepsilon^{(L)}) = 0$, the energy balance equation

$$R_\kappa + T_\kappa + R_{3\kappa} + T_{3\kappa} = 1 \tag{20}$$

is satisfied in the self-consistent formulation, cf. Method SC in Table 1.

If we define by

$$W_{n\kappa} := |a_{n\kappa}^{\mathrm{scat}}|^2 + |b_{n\kappa}^{\mathrm{scat}}|^2 \tag{21}$$

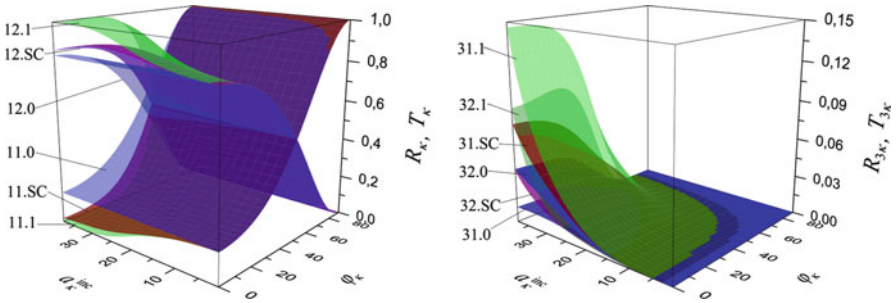

Fig. 4 The scattering properties (11.0, 11.1, 11.SC ... R_κ, 12.0, 12.1, 12.SC ... T_κ) at $\kappa = 0.25$ (*left*) and the generation properties (31.0, 31.1, 31.SC ... $R_{3\kappa}$, 32.0, 32.1, 32.SC ... $T_{3\kappa}$) for $3\kappa = 0.75$ of the nonlinear structure (*right*). The graphs 11.0, 12.0, 31.0, 32.0 depict the results of the linear approximation in the preset field method, the graphs 11.1, 12.1, 31.1, 32.1—the results of the nonlinear approximation in the preset field method, the graphs 11.SC, 12.SC, 31.SC, 32.SC—the results of the self-consistent approach

the total energy of the scattered and generated fields at the frequencies $n\kappa$, $n = 1, 3$, then the energy balance equation (20) can be rewritten as

$$W_\kappa + W_{3\kappa} = |a_\kappa^{\text{inc}}|^2 .$$

In the numerical experiments, the quantities $W_{3\kappa}/W_\kappa$ (which characterises the portion of energy generated in the third harmonic in comparison to the energy scattered in the nonlinear layer) and the function

$$W^{(\text{Error})} := 1 - \left[R_\kappa + T_\kappa + R_{3\kappa} + T_{3\kappa}\right] \tag{22}$$

(which characterises the numerical violation of the energy balance) are of particular interest. In the investigated range of parameters, the dimension of the systems of algebraic equations (17)–(19), resulting from the application of Simpson's quadrature rule, was $N = 501$, and the relative error of calculations did not exceed $\xi = 10^{-7}$. Here we emphasise that in the numerical simulation of scattering and generation processes without any weak fields, i.e., $a_{2\kappa}^{\text{inc}} = a_{3\kappa}^{\text{inc}} = 0$, the residual (22) of the energy balance equation (20) did not exceed the value $|W^{(\text{Error})}| < 10^{-8}$. The property $\Im\mathfrak{m}\left(\varepsilon_\kappa^{(NL)}\left[a_\kappa^{\text{inc}}, z\right]\right) \not\equiv 0$ is typical in the self-consistent formulation, cf. Fig. 2 (right), Fig. 3 (top right) and the Method SC in Table 1. This value is responsible for the loss of energy which is caused by the generation of the electromagnetic field of the third harmonic.

The disadvantages of the preset field method include the absence of the energy balance condition (20). The solution $\mathbf{U}_\kappa$ of the first system of equations plays the role of the preset field and does not depend on the solution $\mathbf{U}_{3\kappa}$ of the second system. For this approximation the property $\Im\mathfrak{m}(\varepsilon_{3\kappa}^{(NL)}) \equiv \mathbf{0}$ is typical; see Methods 0 and 1 in Table 1. Here both systems separately satisfy a law of conservation of energy. For

a nonabsorbing medium, i.e., $\mathfrak{Im}(\varepsilon_\kappa^{(L)}) = 0$, the law of conservation of energy at the frequency of excitation of the nonlinear structure can be written in the form

$$R_\kappa + T_\kappa = 1\,. \tag{23}$$

Using (23) it is easy to obtain the relations characterising the portion of energy generated in the third harmonic for the preset field approximation, cf. (18), (19) and the Methods 0 and 1 in Table 1. Namely, from (21) to (23) we find that $W_{3\kappa}/W_\kappa = R_{3\kappa} + T_{3\kappa}$ and $W^{(\mathrm{Error})} := -\big[R_{3\kappa} + T_{3\kappa}\big]$, therefore $W_{3\kappa}/W_\kappa = W^{(\mathrm{Error})}$. We see that neglecting the portion of energy which is spent for the generation of the third harmonic leads to an error in the energy balance relation. This error is determined by the relative portion of the generated energy.

3.2 Example 2: Single-Layered Structure with a Positive Value of the Cubic Susceptibility

In the second experiment we consider a single-layered structure with a dielectric permittivity of the form (15), where $\varepsilon^{(L)}(z) := 16$ and $\alpha(z) := 0.01$ for $z \in [-2\pi\delta, 2\pi\delta]$, $\delta := 0.5$, $\kappa^{\mathrm{inc}} := \kappa := 0.375$, and $\varphi_\kappa \in [0^\circ, 90^\circ)$.

Since $W_{n\kappa} = |a_{n\kappa}^{\mathrm{scat}}|^2 + |b_{n\kappa}^{\mathrm{scat}}|^2$, the ratio $W_{3\kappa}/W_\kappa$ characterises the relative portion of energy generated in the third harmonic at the value a_κ^{inc}; see Fig. 5 (top and bottom left).

Here we observe that $W_{3\kappa}/W_\kappa = 0.3558$ for $a_\kappa^{\mathrm{inc}} = 14$ and $\varphi_\kappa = 66^\circ$, i.e., the total energy $W_{3\kappa}$ generated in the third harmonic constitutes 35.58 % of the energy W_κ; see Fig. 5 (bottom left) and [2]. The function $\mathfrak{Im}(\varepsilon_\kappa)$ in Fig. 5 (bottom right, curve #5) and Fig. 6 (top right) characterises the loss of energy in the nonlinear layer (w.r.t. excitation frequency κ) spent on the generation of electromagnetic field of the third harmonic (at the frequency 3κ); $\mathfrak{Re}(\varepsilon_\kappa)$ is shown in Fig. 6 (top left). $\mathfrak{Im}\left(\varepsilon_\kappa^{(NL)}\left[a_\kappa^{\mathrm{inc}}, z\right]\right) \not\equiv 0$ is characteristic for the self-consistent formulation. Note that at the frequency 3κ the permittivity $\varepsilon_{3\kappa}$ is real; see Fig. 5 (bottom right, curves #6, #7).

The *reflection/generation* and *transmission/generation* coefficients $R_{n\kappa}$, $T_{n\kappa}$ are depicted in Figs. 7 and 8 (top). We present results corresponding to the case of energy canalisation [2], cf. the curves #1 in Fig. 8 (top left) where $R_\kappa \approx 0$ at $\varphi_\kappa = 60^\circ$ and in Fig. 8 (top right) where $R_\kappa \approx 0$ at $a_\kappa^{\mathrm{inc}} = 14$. The results of calculations are validated numerically by the help of the energy balance equation (20). In the investigated range of parameters, the dimension of the resulting systems of algebraic equations was $N = 301$, and the relative error of calculations did not exceed $\xi = 10^{-7}$. The residual (22) of the energy balance equation (20) did not exceed the value $|W^{(\mathrm{Error})}| < 10^{-8}$. The solutions of (13) are shown in Fig. 8 (second from top and bottom). Comparing Fig. 8 (top left and second from top left) we see that a local maximum in the generated energy at the tripled frequency (curves #3 ... $R_{3\kappa}$,

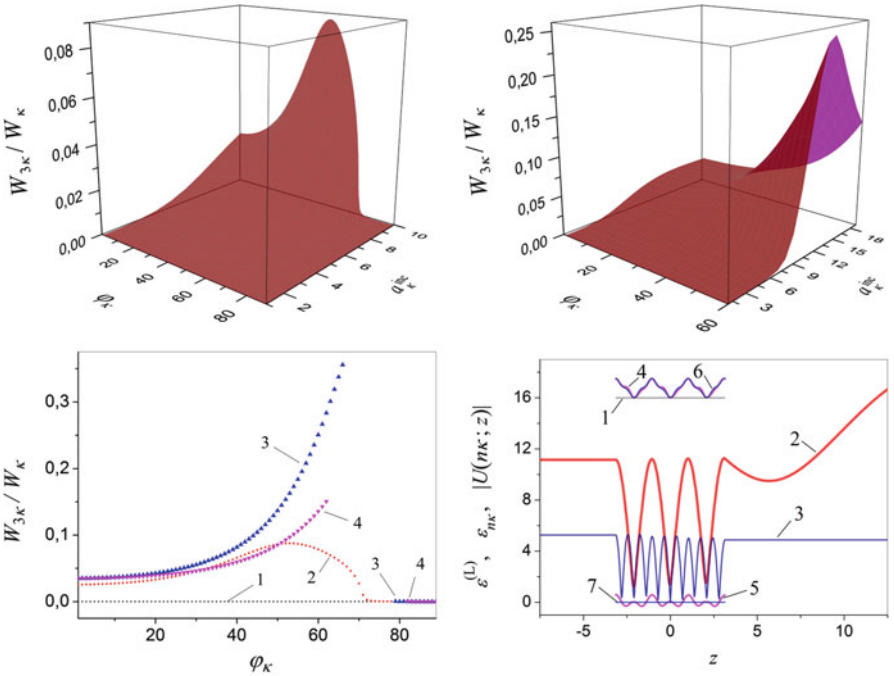

Fig. 5 The portion of energy generated in the third harmonic (*top left/right* and *bottom left*): #1 … $a_\kappa^{\mathrm{inc}} = 1$, #2 … $a_\kappa^{\mathrm{inc}} = 9.93$, #3 … $a_\kappa^{\mathrm{inc}} = 14$, #4 … $a_\kappa^{\mathrm{inc}} = 19$, and some graphs describing the properties of the nonlinear layer for $a_\kappa^{\mathrm{inc}} = 14$ and $\varphi_\kappa = 66^\circ$ (*bottom right*): #1 … $\varepsilon^{(L)}$, #2 … $|U(\kappa;z)|$, #3 … $|U(3\kappa;z)|$, #4 … $\mathfrak{Re}(\varepsilon_\kappa)$, #5 … $\mathfrak{Im}(\varepsilon_\kappa)$, #6 … $\mathfrak{Re}(\varepsilon_{3\kappa})$, #7 … $\mathfrak{Im}(\varepsilon_{3\kappa}) \equiv 0$

#4…$T_{3\kappa}$, #5…$W_{3\kappa}/W_\kappa$) corresponds to a characteristic behaviour of the curve #5.2…$\mathfrak{Im}(\kappa_1^{(NL)})$ in a vicinity of its local minimum. Analogously, the comparison of the computational results depicted in Fig. 8 (top right and second from top right) shows that a local maximum in the generated energy at the tripled frequency (curves #3 …$R_{3\kappa}$, #4…$T_{3\kappa}$, #5…$W_{3\kappa}/W_\kappa$) corresponds to a characteristic behaviour of the curve #5.2…$\mathfrak{Im}(\kappa_1^{(NL)})$ in a region where the absolute value of $\partial\mathfrak{Im}(\kappa_1^{(NL)})/\partial\varphi_\kappa$ is small.

Figure 8 (bottom) presents the characteristic distribution of the eigenfields of the problem (13), where the left figure depicts the eigenfields of the linear problem ($\alpha = 0$) and the right figure depicts the eigenfields of the linearised nonlinear problem ($\alpha = +0.01$).

4 Conclusion

We have investigated the scattering and generation properties of cubically polarisable layered structures which are excited by a sufficiently strong electromagnetic field.

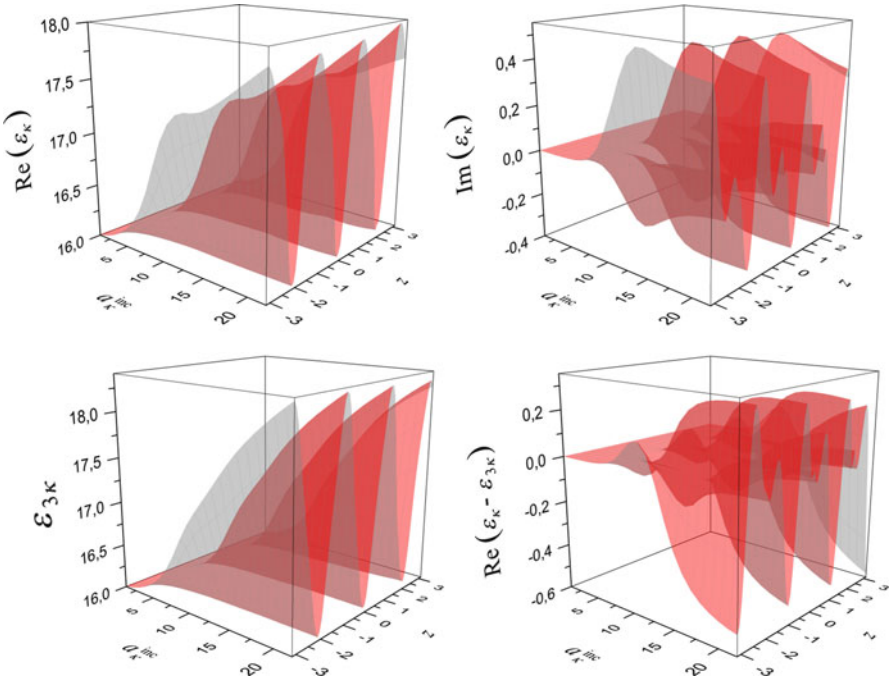

Fig. 6 The nonlinear dielectric permittivity induced by the scattered and generated fields for $\varphi_\kappa = 60°$: $\mathfrak{Re}\left(\varepsilon_\kappa\left[a_\kappa^{\mathrm{inc}}, z\right]\right)$ (*top left*), $\mathfrak{Im}\left(\varepsilon_\kappa\left[a_\kappa^{\mathrm{inc}}, z\right]\right)$ (*top right*), $\varepsilon_{3\kappa}\left[a_\kappa^{\mathrm{inc}}, z\right]$ (*bottom left*), $\mathfrak{Re}\left(\varepsilon_\kappa\left[a_\kappa^{\mathrm{inc}}, z\right] - \varepsilon_{3\kappa}\left[a_\kappa^{\mathrm{inc}}, z\right]\right)$ (*bottom right*)

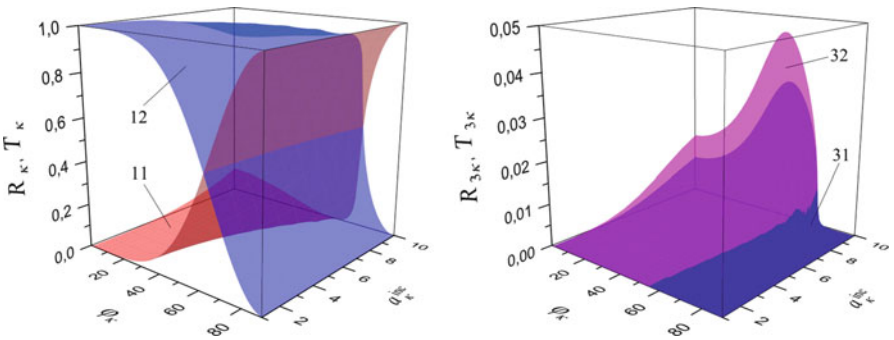

Fig. 7 The scattering and generation properties of the nonlinear structure: #11 ... $R_\kappa\left[\varphi_\kappa, a_\kappa^{\mathrm{inc}}\right]$, #12 ... $T_\kappa\left[\varphi_\kappa, a_\kappa^{\mathrm{inc}}\right]$ (*left*), #31 ... $R_{3\kappa}\left[\varphi_\kappa, a_\kappa^{\mathrm{inc}}\right]$, #32 ... $T_{3\kappa}\left[\varphi_\kappa, a_\kappa^{\mathrm{inc}}\right]$ (*right*)

In particular, we considered a three-layered structure with sign-alternating cubic susceptibility of the medium and a single-layered structure with positive cubic susceptibility of the medium.

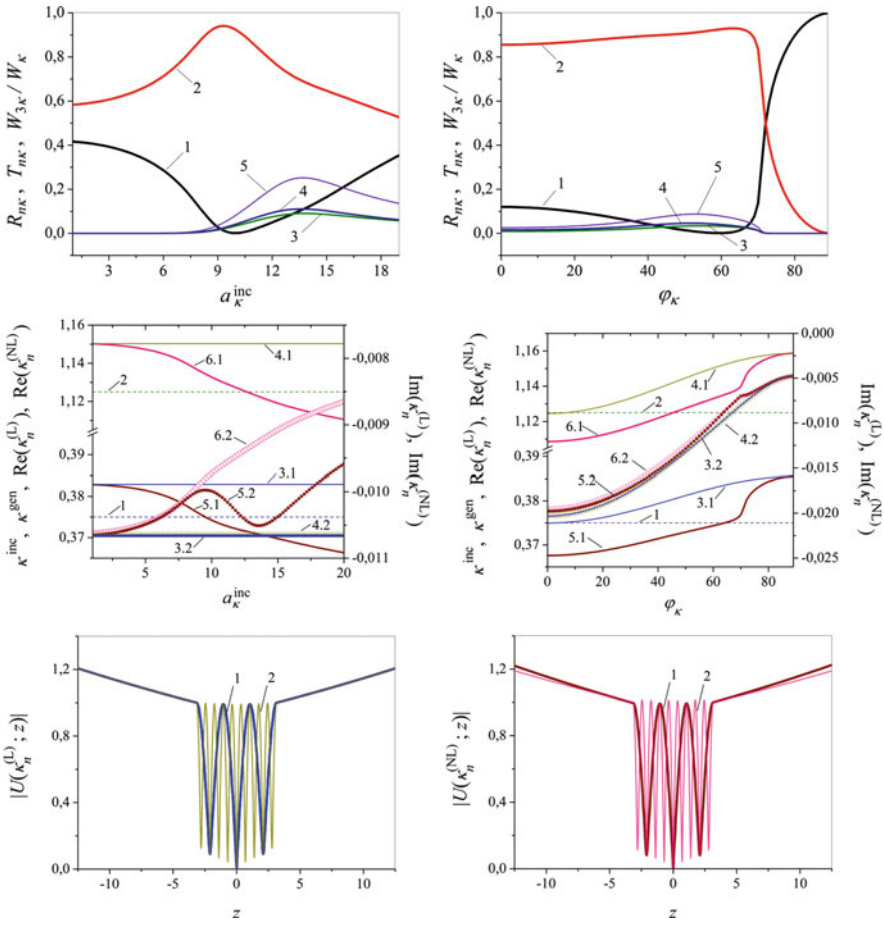

Fig. 8 The curves R_κ (#1), T_κ (#2), $R_{2\kappa}$ (#3), $T_{2\kappa}$ (#4), $R_{3\kappa}$ (#5), $T_{3\kappa}$ (#6), $W_{3\kappa}/W_\kappa$ (#7) for $\varphi_\kappa = 60^\circ$ (*top left*) and $a_\kappa^{\rm inc} = 9.93$ (*top right*), the curves $\kappa := \kappa^{\rm inc} := 0.375$ (#1), $3\kappa = \kappa^{\rm gen} = 3\kappa^{\rm inc} = 1.125$ (#2), the complex eigenfrequencies $\mathfrak{Re}(\kappa_1^{(L)})$ (#3.1), $\mathfrak{Im}(\kappa_1^{(L)})$ (#3.2), $\mathfrak{Re}(\kappa_3^{(L)})$ (#4.1), $\mathfrak{Im}(\kappa_3^{(L)})$ (#4.2) of the linear problem ($\alpha = 0$) and $\mathfrak{Re}(\kappa_1^{(NL)})$ (#5.1), $\mathfrak{Im}(\kappa_1^{(NL)})$ (#5.2), $\mathfrak{Re}(\kappa_3^{(NL)})$ (#6.1), $\mathfrak{Im}(\kappa_3^{(NL)})$ (#6.2) of the linearised nonlinear problem ($\alpha = +0.01$) for $\varphi_\kappa = 60^\circ$ (*second from top left*) and $a_\kappa^{\rm inc} = 9.93$ (*second from top right*), and the graphs of the eigenfields of the layer for $\varphi_\kappa = 60^\circ$, $a_\kappa^{\rm inc} = 14$. The linear problem ($\alpha = 0$, *bottom left*): $|U(\kappa_1^{(L)}; z)|$ with $\kappa_1^{(L)} = 0.3829155 - i\,0.01066148$ (#1), $|U(\kappa_3^{(L)}; z)|$ with $\kappa_3^{(L)} = 1.150293 - i\,0.01062912$ (#2), the linearised nonlinear problem ($\alpha = +0.01$, *bottom right*): $|U(\kappa_1^{(NL)}; z)|$ with $\kappa_1^{(NL)} = 0.3705110 - i\,0.01049613$ (#1), $|U(\kappa_3^{(NL)}; z)|$ with $\kappa_3^{(NL)} = 1.121473 - i\,0.009194824$ (#2)

The mathematical model consists of a system of boundary-value problems of Sturm–Liouville type and of an equivalent system of one-dimensional nonlinear integral equations of the second kind. Various effects caused by the nonlinearity of the structure were investigated using analytical and numerical techniques.

The main focus was on the investigation of the practical behaviour of different numerical algorithms for the solution of the system of nonlinear algebraic equations resulting from the discretisation of the integral equations by means of appropriate quadrature rules. It could be observed that only the self-consistent approach ensures the physically important law of the balance of energy. The results principally indicate how to control the generated field by means of the intensity of the exciting field. In particular, they offer the possibility of designing a frequency multiplier and nonlinear dielectrics with controllable permittivity.

Acknowledgements This work was partially supported by the Visby Program of the Swedish Institute and by the joint Russian–Ukrainian RFBR-NASU grant no. 12.02.90425-2012.

References

1. Angermann, L., Yatsyk, V.V.: Generation and resonance scattering of waves on cubically polarisable layered structures. In: Angermann, L. (ed.) Numerical Simulations – Applications, Examples and Theory, pp. 175–212. InTech, Rijeka (2011)
2. Angermann, L., Yatsyk, V.V.: Resonance properties of scattering and generation of waves on cubically polarisable dielectric layers. In: Zhurbenko, V. (ed.) Electromagnetic Waves, pp. 299–340. InTech, Rijeka (2011)
3. Kleinman, D.A.: Nonlinear dielectric polarization in optical media. Phys. Rev. **126**(6), 1977–1979 (1962)
4. Miloslavsky, V.K.: Nonlinear Optics. V.N. Karazin Kharkov National University, Kharkov (2008)
5. Shestopalov, V.P., Sirenko, Y.K.: Dynamical Theory of Gratings. Naukova Dumka, Kiev (1989)
6. Shestopalov, Y.V., Yatsyk, V.V.: Diffraction of electromagnetic waves by a layer filled with a Kerr-type nonlinear medium. J. Nonlinear Math. Phys. **17**(3), 311–335 (2010)

A Posteriori Estimates for Errors of Functionals on Finite Volume Approximations to Solutions of Elliptic Boundary-Value Problems

Lutz Angermann

Abstract This paper describes the extension of recent methods for a posteriori error estimation such as dual-weighted residual methods to node-centered finite volume discretizations of second-order elliptic boundary-value problems including upwind discretizations. It is shown how different sources of errors, in particular modeling errors and discretization errors, can be estimated with respect to a user-defined output functional.

1 Introduction

Adaptive finite element approaches are in use since several decades. For instance, in the so-called *h*-adaptive methods the computational meshes are refined locally so that the mesh captures the variation of the solution while remaining coarse elsewhere. It has been shown that such approaches are computationally much more efficient than uniform meshes. In recent years, there has been considerable progress in applying these techniques to more involved questions such as the a posteriori error estimation of values of nonlinear functionals of interest (*goal-oriented* estimation, see, e.g., [9, 10]) or the (additional) a posteriori estimation of modeling errors (see, e.g., [13, 18, 20]). The present paper describes the extension of recent techniques for obtaining a posteriori error estimates for modeling and discretization errors to nonlinear second-order elliptic PDEs which are discretized by means of node-centered finite volume schemes including stabilization mechanisms of upwind type. Finite volume methods are attractive methods in selected areas of application, and therefore it is a natural step to develop analogous methods of error control for FVM. However, since finite volume methods suffer, in general, from the so-called

L. Angermann (✉)
Technische Universität Clausthal, Institut für Mathematik, Erzstr. 1, 38678, Clausthal-Zellerfeld, Federal Republic of Germany
e-mail: lutz.angermann@tu-clausthal.de

L. Beilina and Y.V. Shestopalov (eds.), *Inverse Problems and Large-Scale Computations*, Springer Proceedings in Mathematics & Statistics 52, DOI 10.1007/978-3-319-00660-4_5,

property of Galerkin orthogonality, special attention is to be paid to the treatment of the resulting defect term. It is shown that the extension of the dual-weighted a posteriori error estimates to finite volume discretizations is possible in a reasonable way. Furthermore, the latter approach is interesting because of the fact that different sources of errors (i.e. not only discretization errors but, for example, also modeling errors) can be estimated with respect to a rather arbitrary user-defined output functional. For instance, in the field of inverse problems, the Tikhonov functionals can serve as typical output functionals (see, e.g., [11]).

Here we will mainly deal with Voronoi and Donald finite volume partitions on simplicial primary partitions of the computational domain; however, the ideas can be extended to more general primary partitions, in particular quadrilateral or hexahedral partitions (cf., e.g., [4, Sect. 4.2]). A more detailed version of the present paper was published as an e-print (see [6]).

We consider the following boundary value problem with respect to the unknown function $u: \Omega \to \mathbb{R}$:

$$\begin{cases} -\nabla \cdot (\mathbf{A}(\cdot,u)\nabla u) + \mathbf{b}(\cdot,u) \cdot \nabla u + c(\cdot,u)u = f & \text{in } \Omega, \\ u = 0 & \text{on } \Gamma, \end{cases} \tag{1}$$

where $\Omega \subset \mathbb{R}^d$, $d \in \{2,3\}$, is a bounded polygonal or polyhedral domain with a Lipschitzian boundary Γ, and the data in (1) are sufficiently smooth:

$$\mathbf{A}: \Omega \times \mathbb{R} \to \mathbb{R}^{d,d}, \quad \mathbf{b}: \Omega \times \mathbb{R} \to \mathbb{R}^d, \quad c: \Omega \times \mathbb{R} \to \mathbb{R}, \quad f: \Omega \to \mathbb{R}.$$

Using the formal notation

$$\begin{aligned} (w,v) &:= \int_\Omega wv\,dx, \quad (\nabla w, \nabla v) := \int_\Omega \nabla w \cdot \nabla v\,dx, \\ a(w;v) &:= (\mathbf{A}(\cdot,w)\nabla w, \nabla v) + (\mathbf{b}(\cdot,w) \cdot \nabla w, v) + (c(\cdot,w)w, v), \\ \langle f, v\rangle &:= (f,v), \end{aligned}$$

the variational formulation of problem (1) in the space $V := H_0^1(\Omega)$ reads as follows:

Find $u \in V$ such that

$$\forall v \in V: \quad a(u;v) = \langle f, v\rangle. \tag{2}$$

Regarding results for the existence, uniqueness and regularity of solutions of (1) or (2), there is a wide literature both of relatively general nature (see, e.g., [12, Chap. 2] for a short survey) as well as for more specialized equations (see, e.g., [8]).

2 The Finite Volume Scheme

Finite volume methods are attractive discretization methods for partial differential equations of first or second order in conservative form since they adequately transfer the conservation law, which is expressed by the differential equation, to the discrete level. At the same time, due to their proximity to finite difference methods, they are relatively easy to implement even in the nonlinear situation. However, a drawback of many finite volume methods is that there is no p-hierarchy as in finite element methods; therefore, the order of accuracy (related to the grid size) is relatively low. Nevertheless, finite volume methods find wide applications in the computational practice.

In this section we concentrate on node-centered finite volume methods for the discretization of problem (1). Let us first consider a family of Voronoi diagrams such that their straight-line duals are Delaunay triangulations $\mathscr{T}$ of Ω consisting of self-centered simplices. Here a simplex T is called *self-centered* if its circumcenter lies in the interior of T or on the boundary ∂T.

Denote by $\overline{\Lambda} \subset \mathbb{N}$ the index set of all vertices x_i of a particular triangulation $\mathscr{T}$ and by $\Lambda \subset \overline{\Lambda}$ the index set of all vertices lying in Ω. In more detail, let

$$
\begin{aligned}
\Omega_i := \Omega_i^V &:= \{x \in \Omega : \|x - x_i\| < \|x - x_j\| \ \forall j \in \overline{\Lambda} \setminus \{i\}\},\ i \in \overline{\Lambda},\\
&\quad \text{where } \|\cdot\| \text{ denotes the Euclidean norm in } \mathbb{R}^d,\\
m_i &:= \operatorname{meas}_d(\Omega_i),\\
&\quad \text{where } \operatorname{meas}_d(\cdot) \text{ denotes the } d\text{-dimensional volume},\\
\Gamma_{ij} &:= \partial\Omega_i \cap \partial\Omega_j,\ \Gamma_{ij}^T := \Gamma_{ij} \cap T,\ i \in \Lambda, j \in \overline{\Lambda} \setminus \{i\},\ T \in \mathscr{T},\\
m_{ij} &:= \operatorname{meas}_{d-1}(\Gamma_{ij}),\ m_{ij}^T := \operatorname{meas}_{d-1}\left(\Gamma_{ij}^T\right),\\
d_{ij} &:= \|x_i - x_j\|,\\
\Lambda_i &:= \{j \in \overline{\Lambda} \setminus \{i\} : \ m_{ij} \neq 0\},\\
\Lambda_T &:= \{i \in \overline{\Lambda} : \ x_i \in \partial T\},\\
h &:= \max_{T \in \mathscr{T}} h_T, \quad \text{where } h_T := \operatorname{diam} T.
\end{aligned}
$$

The finite volume solution will be interpolated in the discrete space

$$
V_{\mathscr{T}} := \left\{v \in V : \ \left(\forall T \in \mathscr{T} : \ v\big|_T \in \mathscr{P}_1(T)\right)\right\},
$$

where $\mathscr{P}_1(T)$ is the set of all first degree polynomials on T. We introduce a so called *lumping operator*

$$
L_{\mathscr{T}} : \ C(\overline{\Omega}) \to L_\infty(\Omega) \quad \text{acting as} \quad L_{\mathscr{T}} v := \sum_{i \in \overline{\Lambda}} v(x_i) \chi_{\Omega_i},
$$

where χ_{Ω_i} denotes the indicator function of the set Ω_i.

Due to stability reasons, especially for the case of dominating convection, the class of finite volume methods under consideration is characterized by an additional stabilization technique called *upwinding*. For that purpose we introduce a scaling function $K:\ \mathbb{R}\to[0,\infty)$ which is defined by the help of a weighting function $r:\ \mathbb{R}\to[0,1]$ as $K(z):=1-[1-r(z)]z$. A typical example is

$$r(z):=1-\frac{1}{z}\Big(1-\frac{z}{e^z-1}\Big), \tag{3}$$

leading to $K(z)=z/(e^z-1)$, the Bernoulli function.

The discrete problem for the case of a scalar diffusion coefficient, i.e., where $\mathbf{A}$ is of the form $A\mathbf{I}$ with $A:\ \Omega\times\mathbb{R}\to\mathbb{R}$ and $\mathbf{I}$ being the identity in $\mathbb{R}^d$, is formulated as follows:

Find $u_{\mathcal{T}}\in V_{\mathcal{T}}$ such that

$$\forall v_{\mathcal{T}}\in V_{\mathcal{T}}:\quad a_{\mathcal{T}}(u_{\mathcal{T}};v_{\mathcal{T}})=\langle f_{\mathcal{T}},v_{\mathcal{T}}\rangle, \tag{4}$$

where

$$a_{\mathcal{T}}(w_{\mathcal{T}};v_{\mathcal{T}}):=\sum_{i\in\Lambda}v_{\mathcal{T}i}\left\{\sum_{j\in\Lambda_i}\frac{\mu_{ij}}{d_{ij}}K\left(\frac{\gamma_{ij}d_{ij}}{\mu_{ij}}\right)(w_{\mathcal{T}i}-w_{\mathcal{T}j})m_{ij}+c_iw_{\mathcal{T}i}m_i\right\},$$

$$\langle f_{\mathcal{T}},v_{\mathcal{T}}\rangle:=\sum_{i\in\Lambda}f_iv_{\mathcal{T}i}m_i,$$

and

$$\mu_{ij}=\mu_{ij}(w_{\mathcal{T}i},w_{\mathcal{T}j}):=A\left(\frac{x_i+x_j}{2},\frac{w_{\mathcal{T}i}+w_{\mathcal{T}j}}{2}\right),$$

$$\gamma_{ij}=\gamma_{ij}(w_{\mathcal{T}i},w_{\mathcal{T}j}):=v_{ij}\cdot\mathbf{b}\left(\frac{x_i+x_j}{2},\frac{w_{\mathcal{T}i}+w_{\mathcal{T}j}}{2}\right),$$

$$c_i=c_i(w_{\mathcal{T}i}):=c(x_i,w_{\mathcal{T}i}),\quad f_i:=f(x_i).$$

The scheme (4) with the weighting function (3) is often called *exponentially upwinded*. It can be defined for other control functions $r:\mathbb{R}\to[0,1]$, too (see, e.g., [6]).

Finally we mention an equivalent representation of the form $a_{\mathcal{T}}$. By the definition of K, we can write that

$$a_{\mathcal{T}}(w_{\mathcal{T}};v_{\mathcal{T}})=a^0_{\mathcal{T}}(w_{\mathcal{T}};v_{\mathcal{T}})+b_{\mathcal{T}}(w_{\mathcal{T}};v_{\mathcal{T}})+d_{\mathcal{T}}(w_{\mathcal{T}};v_{\mathcal{T}}), \tag{5}$$

where, with $r_{ij} := r\Big(\dfrac{\gamma_{ij} d_{ij}}{\mu_{ij}}\Big)$,

$$a^0_{\mathscr{T}}(w_{\mathscr{T}}; v_{\mathscr{T}}) := \sum_{i\in\Lambda} v_{\mathscr{T}i} \sum_{j\in\Lambda_i} \mu_{ij}(w_{\mathscr{T}i} - w_{\mathscr{T}j}) \frac{m_{ij}}{d_{ij}},$$

$$b_{\mathscr{T}}(w_{\mathscr{T}}; v_{\mathscr{T}}) := \sum_{i\in\Lambda} v_{\mathscr{T}i} \sum_{j\in\Lambda_i} \left[(1-r_{ij}) w_{\mathscr{T}j} - \left(\frac{1}{2} - r_{ij}\right) w_{\mathscr{T}i}\right] \gamma_{ij} m_{ij}, \tag{6}$$

$$d_{\mathscr{T}}(w_{\mathscr{T}}; v_{\mathscr{T}}) := \sum_{i\in\Lambda} \left\{ c_i m_i - \frac{1}{2} \sum_{j\in\Lambda_i} \gamma_{ij} m_{ij} \right\} w_{\mathscr{T}i} v_{\mathscr{T}i}. \tag{7}$$

In the case of a matrix-valued diffusion coefficient $\mathbf{A}: \Omega \times \mathbb{R} \to \mathbb{R}^{d,d}$, in the forms $a^0_{\mathscr{T}}$ and $b_{\mathscr{T}}$ the corresponding values of μ_{ij} are replaced according to the formula

$$\mu_{ij} := \begin{cases} \dfrac{d_{ij}}{m_{ij}} \displaystyle\int_{\partial\Omega_i} (\mathbf{A}\nabla\psi_j)\cdot\nu\, ds, & m_{ij} > 0, \\ 0, & m_{ij} = 0, \end{cases} \tag{8}$$

where $\{\psi_j\}_{j\in\Lambda}$ is the standard nodal basis of $V_{\mathscr{T}}$.

The design of the finite volume method for Donald diagrams starts from an admissible (in the sense of FEM) triangulation $\mathscr{T}$ of Ω. Then, for any $T \in \mathscr{T}$ with local vertices $z_j \equiv x_{i_j}$, $i_j \in \Lambda_T$, $j \in [1, d+1]_{\mathbb{N}}$, we define

$$\Omega^D_{i_j,T} := \left\{ x \in T : (\forall k \in [1,d+1]_{\mathbb{N}} \setminus \{j\} : \lambda_k(x) < \lambda_j(x)) \right\},$$

where $\lambda_j(x)$ is the jth barycentric coordinate of x w.r.t. T. Define for $i \in \overline{\Lambda}$ the sets

$$\Omega^D_i := \operatorname{int}\Big(\bigcup_{T:\partial T \ni x_i} \overline{\Omega^D_{i,T}} \Big).$$

In this way, we get a family of Donald diagrams.

Although it is possible to introduce a discretization like (5), we use the following version:

$$a_{\mathscr{T}}(w_{\mathscr{T}}; v_{\mathscr{T}}) = (\mathbf{A}(\cdot, w_{\mathscr{T}})\nabla w_{\mathscr{T}}, \nabla v_{\mathscr{T}}) + b_{\mathscr{T}}(w_{\mathscr{T}}; v_{\mathscr{T}}) + d_{\mathscr{T}}(w_{\mathscr{T}}; v_{\mathscr{T}}), \tag{9}$$

where the forms $b_{\mathscr{T}}$, $d_{\mathscr{T}}$ are defined analogously to (6), (7). In particular, $\gamma_{ij} \in \mathbb{R}$ is an approximation to $(\nu\cdot\mathbf{b})(\cdot, w_{\mathscr{T}})|_{\Gamma ij}$.

In the case of a matrix-valued diffusion coefficient, we define μ_{ij} analogously to (8) but use it only in $b_{\mathscr{T}}$ to ensure a certain stabilization. The form $a^0_{\mathscr{T}}$ remains as it is, i.e.,

$$a^0_{\mathscr{T}}(w_{\mathscr{T}}; v_{\mathscr{T}}) := (\mathbf{A}(\cdot, w_{\mathscr{T}})\nabla w_{\mathscr{T}}, \nabla v_{\mathscr{T}}).$$

Regarding stability and a priori error estimates, so in the case of a linear equation with a scalar diffusion coefficient, there exists a comparatively well-developed theory since a long time (see, e.g., [3], or the overviews in [17, Chap. 6], [6, Sect. 3]). However, due to the possible structural diversity of the nonlinearities in (1), in the nonlinear situation there is not such a relatively canonical theory. We mention here only a few papers which are concerned with the investigation of node-centered finite volume methods for nonlinear elliptic (or parabolic) equations and refer to the literature cited therein: [14–16].

3 A Posteriori Error Estimates for Nonlinear Problems

In this section we present the general approach that does not depend on the particular discretization. The nonlinear primal problem we are interested in is given by

$$u \in V: \; a(u;v) + a_\delta(u;v) = \langle f, v\rangle \quad \forall v \in V. \tag{10}$$

It represents the weak formulation of the originally given (accurate) boundary-value problem for a partial differential equation in a real Hilbert space V, where f is a linear functional on V and $\langle f, v\rangle$ denotes the value of f at the element $v \in V$. The forms $a: V \times V \to \mathbb{R}$ and $a_\delta: V \times V \to \mathbb{R}$ are linear in the second argument but may be nonlinear in the first one. In the context of the boundary-value problem (2), the left-hand side of (2) is written in (10) as the sum $a + a_\delta$, where a stands for a certain simplified problem and a_δ represents a part of the equation which is to be neglected in the practical computations. That is, the discretization applies only to a in (10). The goal is to estimate the influence of both neglecting a_δ and discretizing a and f with respect to a given output functional $j: V \to \mathbb{R}$. The directional derivatives of $a(u;\cdot)$ and $a_\delta(u;\cdot)$ in u will be denoted by $a'(u;\cdot,\cdot)$ and $a'_\delta(u;\cdot,\cdot)$, respectively. The form

$$a'(u;w,v) := \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \left[a(u + \varepsilon w; v) - a(u;v)\right]$$

is linear in w and v. The second and third directional derivatives are denoted by $a''(u;\cdot,\cdot,\cdot)$ and $a'''(u;\cdot,\cdot,\cdot,\cdot)$, respectively. The dual problem we will use in the analysis is the following:

$$z \in V: \; a'(u;w,z) + a'_\delta(u;w,z) = j'(u;w) \quad \forall w \in V. \tag{11}$$

The primal solution $u_m \in V$ and the dual solution $z_m \in V$ of the reduced problems are given by

$$u_m \in V: \; a(u_m;v) = \langle f, v\rangle \quad \forall v \in V, \tag{12}$$

$$z_m \in V: \; a'(u_m;w,z_m) = j'(u_m;w) \quad \forall w \in V. \tag{13}$$

These variational problems will be formulated in terms of optimization problems. The primal and dual solutions will be expressed by the variables $x := (u,z) \in X := V \times V$ and $x_m := (u_m, z_m) \in X$. In the variational space X, we consider the functionals

$$\begin{aligned} L(x) &:= L_m(x) + L_\delta(x), \\ L_m(x) &:= j(u) + \langle f, z\rangle - a(u;z), \\ L_\delta(x) &:= -a_\delta(u;z). \end{aligned} \tag{14}$$

Obviously, the original primal and dual problems (10) and (11) and the reduced primal and dual problems (12) and (13) consist of finding the stationary points $x = (u,z)$ and $x_m = (u_m, z_m)$ of L and L_m, respectively:

$$x \in X:\ L'(x;y) = 0 \quad \forall y \in X, \tag{15}$$

$$x_m \in X:\ L_m'(x_m;y) = 0 \quad \forall y \in X. \tag{16}$$

Furthermore, the target quantities are given by evaluation of L and L_m at the following stationary points:

$$j(u) = L(x), \quad j(u_m) = L_m(x_m).$$

In order to balance the model and discretization errors, we have to include the discretization error in the analysis. To do this, let $V_{\mathscr{T}} \subset V$ be a finite-dimensional subspace. Typically $V_{\mathscr{T}}$ is a finite element space with respect to a partition $\mathscr{T}$ of the computational domain $\Omega \subset \mathbb{R}^d$, $d \in \{2,3\}$, where possible homogeneous Dirichlet boundary conditions are already included in the choice of the spaces V and $V_{\mathscr{T}}$. Let $a_{\mathscr{T}} :\ V_{\mathscr{T}} \times V_{\mathscr{T}} \to \mathbb{R}$ be a nonlinear form which is different, in general, from the simple restriction of a to $V_{\mathscr{T}} \times V_{\mathscr{T}}$, and denote by $f_{\mathscr{T}} :\ V_{\mathscr{T}} \to \mathbb{R}$ a linear functional which not necessarily coincides with $f|_{V_{\mathscr{T}}}$. For instance, $a_{\mathscr{T}}$ and $f_{\mathscr{T}}$ may result from the finite volume discretization of a, f in (10) according to Sect. 2.

Then $u_{\mathscr{T}m} \in V_{\mathscr{T}}$ is the discrete solution of the problem

$$u_{\mathscr{T}m} \in V_{\mathscr{T}}:\ a_{\mathscr{T}}(u_{\mathscr{T}m};v) = \langle f_{\mathscr{T}}, v\rangle \quad \forall v \in V_{\mathscr{T}} \tag{17}$$

involving both types of error. The difference lies in the definition of the discrete solution $x_{\mathscr{T}m} = (u_{\mathscr{T}m}, z_{\mathscr{T}m}) \in X_{\mathscr{T}} := V_{\mathscr{T}} \times V_{\mathscr{T}}$, where now $u_{\mathscr{T}m}$ satisfies (17) and $z_{\mathscr{T}m}$ is the solution of the following dual problem:

$$z_{\mathscr{T}m} \in V_{\mathscr{T}}:\ a'(u_{\mathscr{T}m};w,z_{\mathscr{T}m}) = j'(u_{\mathscr{T}m};w) \quad \forall w \in V_{\mathscr{T}}. \tag{18}$$

In such a setting, the relations $a(u_{\mathscr{T}m};v) = \langle f, v\rangle$ and $L_m'(x_{\mathscr{T}m};y) = 0$ are no longer valid for all $v \in V_{\mathscr{T}}$ resp. $y \in X_{\mathscr{T}}$.

The target quantities are given by the evaluation of L and $L_{\mathscr{T}m}$, where

$$L_{\mathscr{T}m}(x) := j(u) + \langle f_{\mathscr{T}}, z\rangle - a_{\mathscr{T}}(u;z), \tag{19}$$

at the following stationary points:

$$j(u) = L(x), \quad j(u_m) = L_{\mathcal{T}m}(x_m). \tag{20}$$

For the formulation of the error representation, we use the following notation for the primal and dual residual with respect to the reduced model and for test functions $(w,v) \in X$:

$$\begin{aligned}\rho(u_{\mathcal{T}m}; v) &:= \langle f, v\rangle - a(u_{\mathcal{T}m}; v),\\ \rho^*(u_{\mathcal{T}m}; z_{\mathcal{T}m}, w) &:= j'(u_{\mathcal{T}m}; w) - a'(u_{\mathcal{T}m}; w, z_{\mathcal{T}m}).\end{aligned}$$

Theorem 1. *If $a(u;\cdot)$, $a_\delta(u;\cdot)$ and the functional $j(u)$ are sufficiently differentiable with respect to u, then we have*

$$\begin{aligned} j(u) - j(u_{\mathcal{T}m}) = &-a_\delta(u_{\mathcal{T}m}; z_{\mathcal{T}m}) - \frac{1}{2}\rho(u_{\mathcal{T}m}; z_{\mathcal{T}m} - i_{\mathcal{T}} z)\\ &+ \langle f, z_{\mathcal{T}m}\rangle - \langle f_{\mathcal{T}}, z_{\mathcal{T}m}\rangle - a(u_{\mathcal{T}m}; z_{\mathcal{T}m}) + a_{\mathcal{T}}(u_{\mathcal{T}m}; z_{\mathcal{T}m})\\ &+ \frac{1}{2}\left[\rho(u_{\mathcal{T}m}; z - i_{\mathcal{T}} z) + \rho^*(u_{\mathcal{T}m}; z_{\mathcal{T}m}, u - i_{\mathcal{T}} u)\right]\\ &- \frac{1}{2}\left[a_\delta(u_{\mathcal{T}m}; e_z) + a'_\delta(u_{\mathcal{T}m}; e_u, z_{\mathcal{T}m})\right] - \frac{1}{2}R, \end{aligned}$$

where $e := (e_u, e_z) := (u - u_{\mathcal{T}m}, z - z_{\mathcal{T}m})$, $i_{\mathcal{T}} : V \to V_{\mathcal{T}}$ is an interpolation operator, and the remainder R is defined by $R := \int_0^1 \sigma(1-\sigma) L'''(x_{\mathcal{T}m} + \sigma e; e, e, e)\, d\sigma$.

Proof. By (20),

$$\begin{aligned} j(u) - j(u_{\mathcal{T}m}) &= L(x) - L_{\mathcal{T}m}(x_{\mathcal{T}m})\\ &= L(x) - L_m(x_{\mathcal{T}m}) + L_m(x_{\mathcal{T}m}) - L_{\mathcal{T}m}(x_{\mathcal{T}m})\\ &= L(x) - L_m(x_{\mathcal{T}m})\\ &\quad + \langle f, z_{\mathcal{T}m}\rangle - \langle f_{\mathcal{T}}, z_{\mathcal{T}m}\rangle - a(u_{\mathcal{T}m}; z_{\mathcal{T}m}) + a_{\mathcal{T}}(u_{\mathcal{T}m}; z_{\mathcal{T}m}), \end{aligned}$$

where the last step is a consequence of the definitions (14), (19).

The first difference can be estimated as in the proof of [13, Theorem 2.1]:

$$\begin{aligned} L(x) - L_m(x_{\mathcal{T}m}) &= L(x) - L(x_{\mathcal{T}m}) + L_\delta(x_{\mathcal{T}m})\\ &= \int_0^1 L'(x_{\mathcal{T}m} + \sigma(x - x_{\mathcal{T}m}); x - x_{\mathcal{T}m})\, d\sigma + L_\delta(x_{\mathcal{T}m})\\ &= \frac{1}{2}\left[L'(x_{\mathcal{T}m}; e) + L'(x; e) - R\right] - a_\delta(u_{\mathcal{T}m}; z_{\mathcal{T}m}) \end{aligned}$$

with the above given remainder R of the trapezoidal rule. Since $L'(x;e) = 0$ by (15), we get

$$L(x) - L_m(x_{\mathcal{T}m}) = -a_\delta(u_{\mathcal{T}m}; z_{\mathcal{T}m}) + \frac{1}{2}\left[L'(x_{\mathcal{T}m}; e) - R\right].$$

Furthermore,

$$\begin{aligned} L'(x_{\mathcal{T}m}; e) &= j'(u_{\mathcal{T}m}; e_u) - a'(u_{\mathcal{T}m}; e_u, z_{\mathcal{T}m}) - a'_\delta(u_{\mathcal{T}m}; e_u, z_{\mathcal{T}m}) \\ &\quad + \langle f, e_z\rangle - a(u_{\mathcal{T}m}; e_z) - a_\delta(u_{\mathcal{T}m}; e_z) \\ &= \rho^*(u_{\mathcal{T}m}; z_{\mathcal{T}m}, e_u) - a'_\delta(u_{\mathcal{T}m}; e_u, z_{\mathcal{T}m}) + \rho(u_{\mathcal{T}m}; e_z) - a_\delta(u_{\mathcal{T}m}; e_z). \end{aligned}$$

Since the Galerkin orthogonality is violated, in general, we cannot use the standard argument

$$0 = \rho(u_{\mathcal{T}m}; z_{\mathcal{T}m}) = \rho(u_{\mathcal{T}m}; i_{\mathcal{T}} z)$$

to replace $z_{\mathcal{T}m}$ by $i_{\mathcal{T}} z$ in the third term. Here we can only make use of an analogous property of the dual problem (18), i.e.

$$0 = \rho^*(u_{\mathcal{T}m}; z_{\mathcal{T}m}, u_{\mathcal{T}m}) = \rho^*(u_{\mathcal{T}m}; z_{\mathcal{T}m}, i_{\mathcal{T}} u).$$

(Of course, if the dual problem is approximated by a finite volume method, too, then we have to argue as for the primal problem.) Thus we arrive at

$$\begin{aligned} L'(x_{\mathcal{T}m}; e) &= \rho^*(u_{\mathcal{T}m}; z_{\mathcal{T}m}, u - i_{\mathcal{T}} u) + \rho(u_{\mathcal{T}m}; z - i_{\mathcal{T}} z) \\ &\quad - \rho(u_{\mathcal{T}m}; z_{\mathcal{T}m} - i_{\mathcal{T}} z) - a_\delta(u_{\mathcal{T}m}; e_z) - a'_\delta(u_{\mathcal{T}m}; e_u, z_{\mathcal{T}m}). \end{aligned} \tag{21}$$

This gives the assertion. □

In order to use numerically the error representation derived in Theorem 1, we will neglect the higher-order terms in e, namely the remainder R and the terms $a_\delta(u_{\mathcal{T}m}; e_z)$, $a'_\delta(u_{\mathcal{T}m}; e_u, z_{\mathcal{T}m})$, cf. the related discussion in [13]. Furthermore, we have to approximate the interpolation errors $u - i_{\mathcal{T}} u$ and $z - i_{\mathcal{T}} z$. An efficient possibility for doing this is the recovery process of the computed quantities by patch-wise higher-order interpolation expressed via the operator $i^+_{\mathcal{T}} : V_{\mathcal{T}} \to V^+_{\mathcal{T}}$ formally, where $V^+_{\mathcal{T}}$ is a richer discrete space than $V_{\mathcal{T}}$ (see [10, Sect. 5]). The interpolation errors will be numerically approximated by

$$z - i_{\mathcal{T}} z \approx i^+_{\mathcal{T}} z_{\mathcal{T}m} - z_{\mathcal{T}m}, \quad u - i_{\mathcal{T}} u \approx i^+_{\mathcal{T}} u_{\mathcal{T}m} - u_{\mathcal{T}m}.$$

Without the modeling error and in the case of conforming methods, this approximation is usually observed to be accurate enough.

Taking into account that the residual $\rho^*(u_{\mathcal{T}m}; z_{\mathcal{T}m}, v)$ vanishes with respect to a discrete test function $v \in V_{\mathcal{T}}$, we obtain from Theorem 1 the following approximate estimator consisting of three *indicators*:

$$
\begin{aligned}
j(u) - j(u_{\mathcal{T}m}) \;&\approx\; \eta_{\mathcal{T}} + \eta_m + \eta_{nc},\\
\eta_{\mathcal{T}} &:= \frac{1}{2}\left[\rho(u_{\mathcal{T}m}; i_{\mathcal{T}}^{+} z_{\mathcal{T}m} - z_{\mathcal{T}m}) + \rho^{*}(u_{\mathcal{T}m}; z_{\mathcal{T}m}, i_{\mathcal{T}}^{+} u_{\mathcal{T}m})\right], \qquad (22)\\
\eta_m &:= -a_\delta(u_{\mathcal{T}m}; z_{\mathcal{T}m}),\\
\eta_{nc} &:= \langle f, z_{\mathcal{T}m}\rangle - \langle f_{\mathcal{T}}, z_{\mathcal{T}m}\rangle - a(u_{\mathcal{T}m}; z_{\mathcal{T}m}) + a_{\mathcal{T}}(u_{\mathcal{T}m}; z_{\mathcal{T}m}).
\end{aligned}
$$

The indicator $\eta_{\mathcal{T}}$ of the approximate estimator can be considered as the conforming contribution of the discretization, and the indicator η_m measures the influence of the model. For complex models, the evaluation of η_m may be expensive. Often in practice the decomposition $a + a_\delta$ is changed successively in such a way that portions of a_δ are (locally) shifted to a. The indicator η_{nc} results from the nonconformity of the discretization method caused by the violation of the Galerkin orthogonality. The practical treatment of η_{nc} will be discussed in Sect. 4. In order to use the information (22) for changing locally the model or the discretization parameters (e.g. the mesh size), we have to localize the indicators. After that, an adaptive process has to be designed in order to balance the error sources. Regarding the localization of $\eta_{\mathcal{T}}$ and η_m, so here there are no new aspects. We refer, for instance, to [13].

4 Application to the Finite Volume Method

In the paper [1] an extension of Babuška and Rheinboldt's a posteriori error estimates for finite element methods to finite volume methods for linear diffusion-convection equations has been proposed. In a subsequent paper [2], for a singularly perturbed model problem a modification was introduced with the aim to get two-sided bounds of the error such that the constants occurring in these bounds are independent of the perturbation parameter. In [7, 19], residual-type error estimates for finite volume discretizations of more complicated problems in two and three space dimensions have been presented. A rather general framework for the a posteriori estimation in various finite volume methods can be found in [21]; however, this paper is restricted to linear problems and estimates w.r.t. the energy norm. In [5], *dual-weighted residual error estimators* for finite volume discretizations of linear diffusion-convection equations have been described. Here we apply the results of the previous section to the nonlinear diffusion-convection problem. As a result, we get a posteriori estimates for errors of functionals depending nonlinearly on the solution and for possible modeling errors.

Interpreting $a_{\mathcal{T}}$ and $f_{\mathcal{T}}$ as the finite volume discretizations (4) of the forms a and f in (10), we first observe that the estimators $\eta_{\mathcal{T}}$ and η_m depend only on the computed discrete solution but not directly on the structure of $a_{\mathcal{T}}$ and $f_{\mathcal{T}}$. Therefore, these estimators can be treated as in the (conforming) finite element case and we concentrate on the estimator η_{nc}. To simplify the presentation, we will write $x_{\mathcal{T}} = (u_{\mathcal{T}}, z_{\mathcal{T}})$ instead of $x_{\mathcal{T}m} = (u_{\mathcal{T}m}, z_{\mathcal{T}m})$. Then, by definition, we have that

$$\langle f, z_{\mathscr{T}} \rangle - \langle f_{\mathscr{T}}, z_{\mathscr{T}} \rangle = \sum_{T \in \mathscr{T}} \left\{ (f, z_{\mathscr{T}})_T - \sum_{i \in \Lambda_T} f_i z_{\mathscr{T} i} m_i^T \right\}, \tag{23}$$

where $(f, z_{\mathscr{T}})_T := \int_T f z_{\mathscr{T}} \, dx$ and $m_i^T := \mathrm{meas}_d (\Omega_i \cap T)$. Analogously, with

$$a_{\mathscr{T},T}(u_{\mathscr{T}}; z_{\mathscr{T}}) := \sum_{i \in \Lambda} z_{\mathscr{T} i} \left\{ \sum_{j \in \Lambda_T \setminus \{i\}} \left\{ \mu_{ij} \frac{u_{\mathscr{T} i} - u_{\mathscr{T} j}}{d_{ij}} - \gamma_{ij} (1 - r_{ij}) (u_{\mathscr{T} i} - u_{\mathscr{T} j}) \right\} m_{ij}^T + c_i u_{\mathscr{T} i} m_i^T \right\}$$

and $a_T(u_{\mathscr{T}}; z_{\mathscr{T}})$ resulting from the restriction of all integrals occurring in the expression for $a(u_{\mathscr{T}}; z_{\mathscr{T}})$ to the domain of integration T, we have that

$$a_{\mathscr{T}}(u_{\mathscr{T}}; z_{\mathscr{T}}) - a(u_{\mathscr{T}}; z_{\mathscr{T}}) = \sum_{T \in \mathscr{T}} \left\{ a_{\mathscr{T},T}(u_{\mathscr{T}}; z_{\mathscr{T}}) - a_T(u_{\mathscr{T}}; z_{\mathscr{T}}) \right\}. \tag{24}$$

Putting (23) and (24) together, we obtain computable, localized indicators.

It can be shown that the indicator η_{nc} is *order consistent* with the a priori error estimate [6, Theorem 2] in the following sense :

If $f \in W_q^1(\Omega)$ with some $q > d$ and $u \in W_2^2(\Omega)$, then there is a constant $C_c > 0$ such that

$$\sum_{l=0}^{3} \eta_l \leq C_c h \left[\|u\|_{2,2} + \|f\|_{1,r} \right],$$

see [1, Theorem 4] for a special case.

Acknowledgements The author gratefully acknowledges support from the Visby Program of the Swedish Institute.

References

1. Angermann, L.: An a-posteriori estimation for the solution of elliptic boundary value problems by means of upwind FEM. IMA J. Numer. Anal. **12**, 201–215 (1992)
2. Angermann, L.: Balanced a-posteriori error estimates for finite volume type discretizations of convection-dominated elliptic problems. Computing **55**(4), 305–323 (1995)
3. Angermann, L.: Error estimates for the finite-element solution of an elliptic singularly perturbed problem. IMA J. Numer. Anal. **15**, 161–196 (1995)
4. Angermann, L.: Transport-stabilized semidiscretizations of the incompressible Navier-Stokes equations. Comput. Methods Appl. Math. **6**(3), 239–263 (2006)
5. Angermann, L.: Residual type a posteriori error estimates for upwinding finite volume approximations of elliptic boundary value problems. Mathematik-Bericht 2010/1, Institut für Mathematik, Technische Universität Clausthal (2010)
6. Angermann, L.: A posteriori estimates for errors of functionals on finite volume approximations to solutions of elliptic boundary value problems (2012). e-print arxiv.org/abs/1205.1980

7. Angermann, L., Knabner, P., Thiele, K.: An error estimator for a finite volume discretization of density driven flow in porous media. Appl. Numer. Math. **26**(1–2), 179–191 (1998)
8. Antontsev, S.N., Shmarev, S.I.: Existence and uniqueness of solutions of degenerate parabolic equations with variable exponents of nonlinearity. Fundam. Prikl. Mat. **12**(4), 3–19 (2006). Translation in J. Math. Sci. **150**(5), 2289–2301 (2008)
9. Bangerth, W., Rannacher, R.: Adaptive Finite Element Methods for Differential Equations. Lectures in Mathematics ETH Zürich. Birkhäuser, Basel (2003)
10. Becker, R., Rannacher, R.: An optimal control approach to a posteriori error estimation in finite element methods. Acta Numerica pp. 1–102. Cambridge University Press, Cambridge (2001)
11. Beilina, L., Klibanov, M.V.: A posteriori error estimates for the adaptivity technique for the Tikhonov functional and global convergence for a coefficient inverse problem. Inverse Probl. **26**(4), 045012, 27 (2010)
12. Böhmer, K.: On finite element methods for fully nonlinear elliptic equations of second order. SIAM J. Numer. Anal. **46**(3), 1212–1249 (2008)
13. Braack, M., Ern, A.: A posteriori control of modeling errors and discretization errors. Multiscale Model. Simul. **1**(2), 221–238 (2003) (electronic)
14. Chatzipantelidis, P., Lazarov, R.D.: Error estimates for a finite volume element method for elliptic PDEs in nonconvex polygonal domains. SIAM J. Numer. Anal. **42**(5), 1932–1958, (2005) (electronic)
15. Eymard, R., Fuhrmann, J., Gärtner, K.: A finite volume scheme for nonlinear parabolic equations derived from one-dimensional local Dirichlet problems. Numer. Math. **102**(3), 463–495 (2006)
16. Fuhrmann, J., Langmach, H.: Stability and existence of solutions of time-implicit finite volume schemes for viscous nonlinear conservation laws. Appl. Numer. Math. **37**(1–2), 201–230 (2001).
17. Knabner, P., Angermann, L.: Numerical Methods for Elliptic and Parabolic Partial Differential Equations. Texts in Applied Mathematics, vol. 44. Springer, New York (2003)
18. Oden, J.T., Vemaganti, K.S.: Estimation of local modeling error and goal-oriented adaptive modeling of heterogeneous materials. I. Error estimates and adaptive algorithms. J. Comput. Phys. **164**(1), 22–47 (2000)
19. Thiele, K.: Adaptive finite volume discretization of density driven flows in porous media. Dissertation, Naturwissenschaftliche Fakultät I, Universität Erlangen-Nürnberg (1999)
20. Vemaganti, K.S., Oden, J.T.: Estimation of local modeling error and goal-oriented adaptive modeling of heterogeneous materials. II. A computational environment for adaptive modeling of heterogeneous elastic solids. Comput. Methods Appl. Mech. Eng. **190**(46–47), 6089–6124 (2001)
21. Vohralík, M.: Residual flux-based a posteriori error estimates for finite volume and related locally conservative methods. Numer. Math. **111**(1), 121–158 (2008)

Electromagnetic Wave Propagation in Nonlinear Layered Waveguide Structures: Computational Approach to Determine Propagation Constants

Dmitry V. Valovik

Abstract A plane multilayered waveguide structure is considered. The layers are located between two half-spaces with constant permittivities. The permittivity in each layer can be a constant or nonlinear (depends arbitrarily on modulus of the electric field intensity). We consider propagation of polarized electromagnetic waves in such a structure. The physical problem is reduced to (nonlinear) boundary eigenvalue problem in a multiply-connected domain. We suggest a numerical approach to calculate propagation constants (eigenvalues) for (nonlinear) layered waveguide structures based on numerical solution of a Cauchy problem in each layer. By means of transmission conditions on the layer boundaries we can define initial data for each Cauchy problem. When all Cauchy problems are solved we construct a function that depends on the spectral parameter. The zeros of this function which can be effectively calculated are the sought-for propagation constants (eigenvalues).

1 Introduction

Two problems are considered in the article: propagation of TE and TM electromagnetic waves in plane multilayered nonlinear waveguides. Surface waves are sought for. The layers of the waveguide are located between two half-spaces with constant permittivities. The permittivity inside each layer depends arbitrarily on modulus of the electric field intensity. The problem is to determine propagation constants of electromagnetic waves propagating in the waveguide. Usually, in such problems, the main goal is to obtain a *dispersion equation* (DE) for propagation constants (eigenvalues). In spite of the fact that for such structures formed by layers with

D.V. Valovik (✉)
Penza State University, Krasnaya Str. 40, Penza, 440026, Russia
e-mail: dvalovik@email.ru

L. Beilina and Y.V. Shestopalov (eds.), *Inverse Problems and Large-Scale Computations*, Springer Proceedings in Mathematics & Statistics 52, DOI 10.1007/978-3-319-00660-4_6,

constant permittivities, it is possible to find exact DEs; it is getting tiresome to find them for several layers [17, 19]. It is even harder (if at all possible) to obtain exact DEs for layers with nonlinear permittivities [19]. A lot of different aspects of nonlinear surface polarized wave propagation are given in [2], where important references are also given. For many physically interesting nonlinear permittivities it is far beyond our abilities to obtain and analyze exact DEs. In the article the numerical method to determine propagation constants is suggested.

Such multilayered waveguides can be considered as *nonlinear 1D photonic crystals*. 1D photonic crystals with constant permittivity in each layer (*linear* photonic crystals) are being actively studied recently [7, 9]. For similar problems in one-layer nonlinear waveguides the *dispersion integral equations method* (DIEM) have been suggested [18] and then developed [11, 12, 15, 16]. DE can be easily found if we can solve differential equations of the problem. DIEM allows to find DEs even we cannot solve differential equations of the problem. In the case of TE waves and if the permittivity depends arbitrarily on modulus of the electric field intensity, it is possible to find exact DEs using DIEM. However, in the case of TM waves we have to impose some restrictions for the permittivity (see below). These exact DEs can be studied both analytically and numerically [11, 14]. For Kerr and generalized Kerr nonlinearities and TE waves differential equations can be integrated and DEs are found by means of obtained solutions [10, 13]. It should be noted that the development of computational methods for numerical solution of these exact DEs is not an easy problem [21]. Such calculations can be carried out without great difficulties only for the simplest nonlinearities (like Kerr nonlinearity [11] or nonlinearity with saturation [14]).

It is known that exact DEs for plane-layered (including multilayered) waveguides with constant permittivity in each layer can be derived (see for example [1, 6, 19]). For one-layer waveguides with constant permittivity in the layer DE can be completely studied by analytical methods. However for a nonlinear layer such a study can hide a lot of difficulties. When the number of layers is increasing it is getting difficult to find DEs. Of course in this case it is practically impossible to solve them. In this connection, development of simple, fast, and efficient numerical methods for the discussed problems in one-layer and multilayered waveguides is an urgent task. In the case of one-layer nonlinear waveguides such numerical methods have been developed (see for example [20, 22, 23]). For multilayered structures numerical methods to determine propagation constants are given in this work.

2 TE Waves

2.1 Statement of the Problem

Consider electromagnetic waves propagating through N homogeneous isotropic nonmagnetic dielectric layers. The permittivity in each layer depends arbitrarily on modulus of the electric field intensity. The layers are located between two half-spaces $x < h_0$ and $x > h_N$ in Cartesian coordinate system $Oxyz$. The half-spaces

are filled with homogeneous isotropic nonmagnetic media without any sources and have constant permittivities $\underline{\varepsilon}$ and $\overline{\varepsilon}$, respectively ($\underline{\varepsilon}$ and $\overline{\varepsilon}$ are arbitrary real values). Suppose that everywhere $\mu = \mu_0$ is the permeability of free space.

The fields depend on time harmonically

$$\tilde{\mathbf{E}}(x,y,z,t) = \mathbf{E}_+(x,y,z)\cos\omega t + \mathbf{E}_-(x,y,z)\sin\omega t;$$
$$\tilde{\mathbf{H}}(x,y,z,t) = \mathbf{H}_+(x,y,z)\cos\omega t + \mathbf{H}_-(x,y,z)\sin\omega t,$$

where ω is the circular frequency; $\mathbf{E}_+$, $\mathbf{E}_-$, $\mathbf{H}_+$, $\mathbf{H}_-$ are real sought for functions.

Let $\mathbf{E} = \mathbf{E}_+ + i\mathbf{E}_-$, $\mathbf{H} = \mathbf{H}_+ + i\mathbf{H}_-$ be the complex amplitudes of the fields $\mathbf{E}$, $\mathbf{H}$ [3]. Below the multipliers $\cos\omega t$ and $\sin\omega t$ are omitted.

The electromagnetic field $\mathbf{E}$, $\mathbf{H}$ satisfies the Maxwell equations

$$\operatorname{rot}\mathbf{H} = -i\omega\varepsilon\mathbf{E};\ \ \operatorname{rot}\mathbf{E} = i\omega\mu\mathbf{H}, \tag{1}$$

the continuity condition for the tangential field components on the boundaries $x = h_0$, $x = h_1$, …, $x = h_N$ and the radiation condition at infinity: the electromagnetic field exponentially decays as $|x| \to \infty$ in the domains $x < h_0$ and $x > h_N$.

The permittivity inside each layer has the form

$$\varepsilon = \varepsilon_i + \varepsilon_0 f_i\left(|\mathbf{E}|^2\right), \quad i = \overline{1,N},$$

where ε_i is a constant part of the permittivity ε in the ith layer; ε_0 is the permittivity of free space; $f_i(u)$ is a continuous function.

Geometry of the problem is shown in Fig. 1.

Consider TE waves $\mathbf{E} = (0,E_y,0)^T$, $\mathbf{H} = (H_x,0,H_z)^T$, where $(\ldots)^T$ is the transposition operation. It can be shown that the fields components do not depend on y. Waves propagating along the boundaries z depend harmonically on z. So the fields components have the form

$$E_y = \mathrm{E}_y(x)e^{i\gamma z}, \quad H_x = \mathrm{H}_x(x)e^{i\gamma z}, \quad H_z = \mathrm{H}_z(x)e^{i\gamma z}, \tag{2}$$

where γ is the unknown spectral parameter (propagation constant).

Substituting components (2) into Eq. (1) we obtain

$$\mathrm{E}_y'' = \left(\gamma^2 - \omega^2\mu\varepsilon\right)\mathrm{E}_y,$$

and $\mathrm{H}_x = -\gamma\omega^{-1}\mu^{-1}\mathrm{E}_y$, $\mathrm{H}_z = -i\omega^{-1}\mu^{-1}\mathrm{E}_y'$, where $(\ldots)' \equiv \frac{\partial}{\partial x}$.

Normalizing the latter equation accordingly with the formulae $\tilde{x} = kx$, $\frac{d}{dx} = k\frac{d}{d\tilde{x}}$, $\tilde{\gamma} = \frac{\gamma}{k}$, $\tilde{\varepsilon}_j = \frac{\varepsilon_j}{\varepsilon_0}$, $j = \overline{1,N}$, $\underline{\tilde{\varepsilon}} = \frac{\underline{\varepsilon}}{\varepsilon_0}$, $\overline{\tilde{\varepsilon}} = \frac{\overline{\varepsilon}}{\varepsilon_0}$, where $k^2 = \omega^2\mu_0\varepsilon_0$, denoting by $Y(\tilde{x}) := \mathrm{E}_y(\tilde{x})$ and omitting the tilde, we obtain the equation [11]

$$Y''(x) = \gamma^2 Y(x) - \varepsilon Y(x) \tag{3}$$

It is necessary to find real solutions $Y(x)$ of Eq. (3).

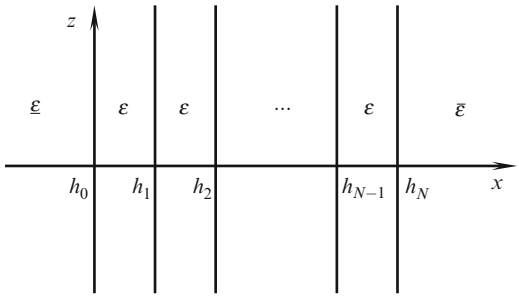

Fig. 1 Geometry of the problem

The value of γ must be real.[1]
It is supposed that

$$\varepsilon = \begin{cases} \underline{\varepsilon}, & x < h_0; \\ \varepsilon_1 + f_1(Y^2), & h_0 < x < h_1; \\ \dots & \\ \varepsilon_N + f_N(Y^2), & h_{N-1} < x < h_N; \\ \overline{\varepsilon}, & x > h_N. \end{cases} \tag{4}$$

The function Y has the properties

$$\begin{aligned} &Y(x) \in C^1(-\infty,\infty) \cap \\ &\quad \cap C^2(-\infty,h_0) \cap C^2(h_0,h_1) \cap \dots \cap C^2(h_{N-1},h_N) \cap C^2(h_N,\infty). \end{aligned} \tag{5}$$

These conditions of continuity and smoothness of the function Y correspond to the physical nature of the problem and will be obtained from the transmission conditions on the boundaries.

2.2 *Differential Equations of the Problem*

Denote by $\underline{k}^2 = \gamma^2 - \underline{\varepsilon}$, $k_i^2 = \varepsilon_i - \gamma^2$, $i = \overline{1,N}$, $\overline{k}^2 = \gamma^2 - \overline{\varepsilon}$.

For the half-space $x < h_0$ we have the permittivity $\varepsilon = \underline{\varepsilon}$. From Eqs. (3) and (4) we obtain the linear equation. Its solution in according to the condition at infinity is

$$Y(x) = Y(h_0 - 0)e^{\underline{k}(x-h_0)}. \tag{6}$$

[1]Let us describe why in this nonlinear problem it is impossible to consider complex values of γ. As $\mathbf{E} = \big(0, \mathrm{E}_y(x)e^{i\gamma z}, 0\big) = e^{i\gamma z}(0, \mathrm{E}_y(x), 0)$, then $|\mathbf{E}|^2 = |e^{i\gamma z}|^2 \cdot |\mathrm{E}_y|^2$. As it is known $|e^{i\gamma z}| = 1$ if $\operatorname{Im}\gamma = 0$. Let $\gamma = \gamma' + i\gamma''$ and $\operatorname{Im}\gamma \neq 0$. Then $|e^{i\gamma z}| = |e^{i\gamma' z}| \cdot |e^{-\gamma'' z}| = e^{-\gamma'' z}$, that is Eq. (3) contains z. This means that the function $Y(x)$ depends on z. This contradicts to the choice of $\mathrm{E}_y(x)$. In the linear problem it is possible to consider complex γ.

For the half-space $x > h_N$ we have the permittivity $\varepsilon = \overline{\varepsilon}$. From Eqs. (3) and (4) we obtain the linear equation. Its solution in according to the condition at infinity is

$$Y(x) = Y(h_N + 0)e^{-\overline{k}(x-h_N)}. \tag{7}$$

In solution (6) the constant $Y(h_0 - 0)$ is defined by initial conditions; in solution (7) the constant $Y(h_N + 0)$ is defined by transmission conditions.

From Eqs. (6) and (7) it is clear that the inequality $\gamma^2 > \max(\underline{\varepsilon}, \overline{\varepsilon})$ holds.

Inside the ith layer $h_{i-1} < x < h_i$, $i = \overline{1,N}$ Eq. (3) takes the form

$$Y'' = -\left(k_i^2 + f_i(Y^2)\right) Y. \tag{8}$$

2.3 Transmission Conditions

Tangential components of an electromagnetic field are known to be continuous at the interfaces. In this case tangential components are E_y and H_z. So we obtain

$$E_y(h_i + 0) = E_y(h_i - 0), \quad H_z(h_i + 0) = H_z(h_i - 0), \quad i = \overline{0,N}.$$

This implies the following conditions for the functions Y and Y'

$$[Y]\big|_{x=h_i} = 0, \quad [Y']\big|_{x=h_i} = 0, \quad i = \overline{0,N} \tag{9}$$

where $[f]\big|_{x=x_0} = \lim\limits_{x \to x_0 - 0} f(x) - \lim\limits_{x \to x_0 + 0} f(x)$.

Using Eqs. (6) and (7) in the half-spaces functions Y, Y' take the form

$$\begin{aligned} Y(x) &= \begin{cases} Y(h_0 - 0)e^{\underline{k}(x-h_0)}, & x < h_0 \\ Y(h_N + 0)e^{-\overline{k}(x-h_N)}, & x > h_N, \end{cases} \\ Y'(x) &= \begin{cases} Y(h_0 - 0)\underline{k}e^{\underline{k}(x-h_0)}, & x < h_0 \\ -Y(h_N + 0)\overline{k}e^{-\overline{k}(x-h_N)}, & x > h_N. \end{cases} \end{aligned} \tag{10}$$

Definition 1. The value $\gamma = \overline{\gamma}$ such that nonzero solution $Y(x)$ of Eq. (8) exists, in the half-spaces $x < h_0$, $x > h_N$ function $Y(x)$ is described by Eq. (10), and in the entire space functions $Y(x)$, $Y'(x)$ satisfy conditions (9) is called an eigenvalue of the problem. The function $Y(x)$ corresponding to the eigenvalue $\gamma = \overline{\gamma}$ is called an eigenfunction of the problem.[2]

[2] Definition 1 is a nonclassical analog of the known definition of the characteristic number of a linear operator function depending nonlinearly on the spectral parameter [5]. This definition, on the one hand, is an extension of the classic definition of an eigenvalue to the case of a nonlinear operator function. On the other hand, it corresponds to the physical nature of the problem.

It is well known that electromagnetic waves in a layer propagate on dedicated frequencies, and there are finite numbers of such frequencies. Fixed values of the spectral parameter γ correspond

Definition 2. The conjugation problem in multiply-connected domain (*problem* P_E) is to determine eigenvalues γ such that there are nonzero functions $Y(x)$ that satisfy the following conditions: if $x < h_0$ and $x > h_N$ then the function Y is defined by Eq. (10), where $Y(h_0 - 0)$ is supposed to be known, and $Y(h_N + 0)$ is defined by Eq. (9); if $h_{i-1} < x < h_i$, $i = \overline{1,N}$ the function Y is a solution of Eq. (8); the functions Y and Y' satisfy transmission conditions (9).[3]

2.4 Existence of Eigenvalues

In this section some theoretical results will be given that are necessary for the correct formulation and proof of convergence of the numerical method. Particularly we will determine the necessary conditions that provide unique solvability of the Cauchy problem for Eq. (3) with the initial conditions

$$Y(h_i + 0), \quad Y'(h_i + 0), \quad i = \overline{0,N-1}. \tag{11}$$

Further we will show the necessary conditions that provide continuous dependence of considered solution on the spectral parameter γ. The question of the *problem* P_E eigenvalue existence will be solved as well.

Since the solutions of Eq. (3) in the half-spaces $x < h_0$ and $x > h_N$ are known, let us go over to the Cauchy problem for nonlinear equation (8).

Rewrite Eq. (8) as a system in the normal form. Let $Y := Y_1$, $Y' := Y_2$, then

$$\begin{cases} Y_1' = Y_2 \\ Y_2' = -\left(k_i^2 + f\left(Y_1^2\right)\right) Y_1. \end{cases} \tag{12}$$

Consider system (12) with initial conditions

$$Y_1(h_i + 0) \quad \text{and} \quad Y_2(h_i + 0), \quad i = \overline{0,N-1}. \tag{13}$$

Let $\sqrt{\max(\underline{\varepsilon}, \overline{\varepsilon})} < \gamma_* < \gamma^* < \infty$, $\gamma \in [\gamma_*, \gamma^*]$ and $b_i < \infty$ be a constant. Define the sets

$$\Pi_{i+1} := \{(Y_1, Y_2) : |Y_1 - Y_1(h_i + 0)| \le b_{i+1}, |Y_2 - Y_2(h_i + 0)| \le b_{i+1}\}, \quad i = \overline{0,N-1}.$$

---

to these dedicated frequencies. This situation takes place both for linear and nonlinear cases. This is the reason why it is necessary (in theory and applications) to determine eigenvalues γ and eigenfunctions.

[3] The nonlinear problem under consideration depends essentially on the initial condition $Y(h_0 - 0)$. Similar problem when the permittivity inside each layer is constant does not depend on an initial condition. This means that in the linear problem the "bundle" of waves with different amplitudes corresponds to each eigenvalue γ. In the nonlinear problem eigenvalues depend on amplitudes.

Let values M_i be such that

$$M_i \geq \max_{(Y_1,Y_2)\in\Pi_i} |Y_2|, \quad M_i \geq \max_{(Y_1,Y_2)\in\Pi_i} \left|\left(k_i^2 + f(Y_1^2)\right) Y_1\right|, \quad i = \overline{1,N}.$$

Theorem 1. *The solution of the Cauchy problem for system (12) with initial conditions (13) exists; in addition this solution is continuously differentiable and unique if* $x \in [0,h]$, *where* $h \leq b_i/M_i$, $i = \overline{1,N}$.

Proof. The proof of this theorem results from Picard theorem (see, for example, [4], p. 165). It is necessary to take into account that system (12) is autonomous and the segment $[h_{i-1},h_i]$ for $i = \overline{1,N}$ can be transformed into the segment $[0,h_i - h_{i-1}]$ by means of change the variable $x = \tilde{x} + h_{i-1}$. In addition it is possible to suppose that $b_i < \infty$ for $i = \overline{1,N}$, as we look for bounded solutions of system (12) only. □

Let $\sqrt{\max(\underline{\varepsilon},\overline{\varepsilon})} < \gamma_* < \gamma^* < \infty$ and $b_i^\gamma < \infty$ be a constant. Define the sets

$$\Pi_{i+1}^\gamma := \left\{(Y_1,Y_2,\gamma) : |Y_1 - Y_1(h_i+0)| \leq b_{i+1}^\gamma, |Y_2 - Y_2(h_i+0)| \leq b_{i+1}^\gamma, \gamma \in [\gamma_*,\gamma^*]\right\},$$

where $i = \overline{0,N-1}$.

Let values M_i^γ be such that

$$M_i^\gamma \geq \max_{(Y_1,Y_2,\gamma)\in\Pi_i^\gamma} |Y_2|, \quad M_i^\gamma \geq \max_{(Y_1,Y_2,\gamma)\in\Pi_i^\gamma} \left|(k_i^2 + f(Y_1^2))Y_1\right|, \quad i = \overline{1,N}.$$

Theorem 2. *The solution* $Y_1(x,\gamma)$, $Y_2(x,\gamma)$ *of the Cauchy problem for system (12) with initial conditions (13) exists, this solution is continuously differentiable w.r.t. x, and in addition this solution is unique for all* $x \in [0,h]$, *where* $h \leq b_i^\gamma/M_i^\gamma, i = \overline{1,N}$ *and continuously depends on* γ, *for all* $\gamma \in [\gamma_*,\gamma^*]$.

Proof. The proof of this theorem results from the theorem of continuity dependence on the parameter of the solution of a Cauchy problem (see, for example, [4], pp. 183–185). It is necessary to take into account that system (12) is autonomous and the segment $[h_{i-1},h_i]$ for $i = \overline{1,N}$ can be transformed into the segment $[0,h_i - h_{i-1}]$ by means of change the variable $x = \tilde{x} + h_{i-1}$. In addition it is possible to suppose that $b_i < \infty$ for $i = \overline{1,N}$, as we look for bounded solutions of system (12) only. □

Pass to the question about eigenvalues existence of the *problem* P_E.

Let us consider the Cauchy problem for system (12) with initial conditions (13).

Using transmission conditions (9) we obtain

$$Y_1(h_i - 0,\gamma) = Y_1(h_i + 0,\gamma), \quad Y_2(h_i - 0,\gamma) = Y_2(h_i + 0,\gamma), \tag{14}$$

where $i = \overline{0,N}$. Taking into account formulae (10) and (14), we get

$$Y_1(h_N - 0,\gamma) = Y(h_N + 0) \text{ and } Y_2(h_N - 0,\gamma) = -\bar{k}Y(h_N + 0). \tag{15}$$

On the other hand, the value $Y(h_N+0)$ is an unknown and must be determined. Then from the first formula (15), we obtain $Y(h_N+0) := Y_1(h_N-0,\gamma)$. Construct the function

$$F(\gamma) := Y_2(h_N-0,\gamma) - Y_2(h_N+0,\gamma).$$

Using Eq. (15) we can rewrite the latter formula as[4]

$$F(\gamma) = Y_2(h_N-0,\gamma) - \overline{k}Y_1(h_N-0,\gamma).$$

If the value $\tilde{\gamma}$ is such that $F(\tilde{\gamma})=0$ then $\tilde{\gamma}$ is an eigenvalue of the *problem* P_E.

Let us formulate criterion of the existence at least one eigenvalue.

Theorem 3. *Let the conditions of Theorems 1 and 2 be satisfied and let the segment* $[\underline{\gamma},\overline{\gamma}] \subset [\gamma_*,\gamma^*]$ *be such that* $F(\underline{\gamma})F(\overline{\gamma}) < 0$. *Then at least one eigenvalue* $\tilde{\gamma}$ *of the problem* P_E *exists and* $\tilde{\gamma} \in (\underline{\gamma},\overline{\gamma})$.

2.5 Numerical Method

The numerical method suggested below allows to determine eigenvalues of the considered problem with any prescribed accuracy. With the help of this method the normalized dependence of the propagation constant (normalized by ω^{-1}) γ w.r.t. the layer's thickness (normalized by ω) h will be depicted as well [so called dispersion curve (DC)].

Let us consider Eq. (8), $i=\overline{1,N}$. In the first layer $h_0 < x < h_1$, the initial conditions are defined by formulae (11).

Let us suppose that the pth layer has a variable thickness that is the value h_{p-1} is a constant and the value h_p is varied from h_* to h^*. The thicknesses of other layers stay constants that is the differences $h_i - h_{i-1} = const$ for all $i=\overline{1,N}$ except $i=p$.

Denote by $h := h_p$. Let $0 < h_* < h^* < \infty$ and $\sqrt{\max(\underline{\varepsilon},\overline{\varepsilon})} < \gamma_* < \gamma^* < \infty$ be constants. Suppose that $h \in [h_*,h^*]$ and $\gamma \in [\gamma_*,\gamma^*]$.

Divide the segments $[h_*,h^*]$ and $[\gamma_*,\gamma^*]$ into n and m pieces, respectively. We get the grid $\left\{h^{(i)},\gamma^{(j)}\right\}$, $i=\overline{0,n}$, $j=\overline{0,m}$ and $h^{(0)}=h_*>0$, $h^{(n)}=h^*$, $\gamma^{(0)}=\gamma_*$, $\gamma_m=\gamma^*$. Then for each pair of indexes (i,j), we obtain a pair of initial conditions

$$\left(Y_{ij}(h_0+0), Y'_{ij}(h_0+0)\right), \tag{16}$$

where $Y_{ij}(h_0+0) = Y(h_0-0)$ and $Y'_{ij}(h_0+0) = \sqrt{(\gamma^{(j)})^2-\underline{\varepsilon}}Y(h_0-0)$.

[4]It is clear that the value of the function F depends on the solutions of considered Cauchy problem only.

Now we can state the Cauchy problem for Eq. (8) in the first layer with initial conditions (16). Solve this problem we obtain the values $Y_{ij}(h_1-0)$ and $Y'_{ij}(h_1-0)$. By virtue of Eq. (9) these values are initial conditions for Eq. (8) in the second layer that is in such a way we state next Cauchy problem. Serially solve the Cauchy problems for each layer, we reach the pth layer. For this layer we have the initial conditions $Y_{ij}(h_{p-1}+0)$ and $Y'_{ij}(h_{p-1}+0)$ defined from the solution on the previous layer. Using these initial conditions we state the Cauchy problem for Eq. (8) in the pth layer. In such a way we reach the final layer. From the solution of the Cauchy problem in the final layer we obtain $Y_{ij}(h_N-0)$ and $Y'_{ij}(h_N-0)$. As the function Y is continuous when $x=h_N$ then we can calculate the value $Y_{ij}(h_N+0):=Y_{ij}(h_N-0)$. Using the second formula (10) and $Y_{ij}(h_N+0)$, we can calculate $Y'_{ij}(h_N+0):=-\sqrt{(\gamma^{(i)})^2-\overline{\varepsilon}}Y_{ij}(h_N+0)$. However we know the value $Y'_{ij}(h_N-0)$ from the solution of the Cauchy problem. Taking into account the continuity of $Y'(x)$ on the boundary $x=h_N$, we construct the function

$$F(\gamma_j)=Y'_{ij}(h_N-0)-Y'_{ij}(h_N+0).$$

Let for given $h_p^{(i)}$ such γ_j and γ_{j+1} exist that $F(\gamma_j)F(\gamma_{j+1})<0$. This means that at least one value $\tilde{\gamma}_j\in(\gamma_j,\gamma_{j+1})$ exists and $\tilde{\gamma}_j$ is an eigenvalue of the *problem* P_E. In addition the thicknesses $h_1,\ldots,h_{p-1},h_p^{(i)},h_{p+1},\ldots,h_N$ correspond to this eigenvalue.

It is clear that the possibility to find eigenvalues is based on the continuity of $F(\gamma)$ w.r.t. γ. The function $F(\gamma)$ is a linear function w.r.t., a solution of the Cauchy problem for Eq. (8) with the initial conditions defined above. Under Theorem 2 the solution of this Cauchy problem is a continuous function w.r.t. γ. This implies that $F(\gamma_j)$ depends continuously on γ.

Let us construct a numerical method to determine an approximate eigenvalue and prove its convergence.

Prescribe the accuracy $\varepsilon>0$ of the eigenvalue $\hat{\gamma}$ calculation. Let the interval $(\underline{\gamma}_1,\overline{\gamma}_1)$ be such that $F(\underline{\gamma}_1)F(\overline{\gamma}_1)<0$.

Determine the center of the segment $\gamma_1=\frac{\underline{\gamma}_1+\overline{\gamma}_1}{2}$ and calculate the value $F(\gamma_1)$. Check the following conditions

1. If $|F(\gamma_1)|<\varepsilon$ then γ_1 is the approximate eigenvalue.
2. If $F(\underline{\gamma}_1)F(\gamma_1)<0$ then $\hat{\gamma}\in(\underline{\gamma}_1,\gamma_1)$. Suppose that $\underline{\gamma}_2:=\underline{\gamma}_1$ and $\overline{\gamma}_2:=\gamma_1$ then $\hat{\gamma}\in(\underline{\gamma}_2,\overline{\gamma}_2)$.
3. If $F(\gamma_1)F(\overline{\gamma}_1)<0$ then $\hat{\gamma}\in(\gamma_1,\overline{\gamma}_1)$. Suppose that $\underline{\gamma}_2:=\gamma_1$ and $\overline{\gamma}_2:=\overline{\gamma}_1$ then $\hat{\gamma}\in(\underline{\gamma}_2,\overline{\gamma}_2)$.

Continuing dichotomy process n times, we obtain that the sought for approximate eigenvalue $\hat{\gamma}_n\in(\underline{\gamma}_n,\overline{\gamma}_n)$. It is clear that $|\overline{\gamma}_n-\underline{\gamma}_n|=2^{-n}|\overline{\gamma}-\underline{\gamma}|$.

Choose n in such a way that $2^{-n}|\overline{\gamma}-\underline{\gamma}|<\varepsilon$. Then we can choose the center of the segment $(\underline{\gamma}_n,\overline{\gamma}_n)$ as the approximate eigenvalue $\hat{\gamma}_n=\frac{\underline{\gamma}_n+\overline{\gamma}_n}{2}$.

Theorem 4. *The iteration process converges to the eigenvalue* $\hat{\gamma}$.

Proof. The sequence $\{\hat{\gamma}_i\}_{i=1}$, where $\hat{\gamma}_i = \frac{\underline{\gamma}_i + \overline{\gamma}_i}{2}$, is a fundamental sequence. Indeed let $p > k > 0$ be integers. Then $|\hat{\gamma}_k - \hat{\gamma}_p| < 2^{-k}|\underline{\gamma} - \overline{\gamma}|$. However $2-^{k}|\underline{\gamma} - \overline{\gamma}| < \varepsilon$ if $k \geq n$. This implies that the sequence above is fundamental. Any fundamental sequence has a limit. Let $\gamma^* = \lim\limits_{n\to\infty} \hat{\gamma}_n$. For any number n the following relations $\hat{\gamma} \in (\underline{\gamma}_n, \overline{\gamma}_n)$ and $\gamma^* \in (\underline{\gamma}_n, \overline{\gamma}_n)$ hold. It follows from the above that $\hat{\gamma} = \gamma^*$. □

Considered problem for Kerr and generalized Kerr nonlinearities can be solved exactly (for a one-layer waveguide [10, 13] and for a double-layer waveguide [17]). Results in [10, 13, 17] agree with the results obtained with the help of numerical method under consideration.

3 TM Waves

3.1 Statement of the Problem

Consider electromagnetic waves propagating through N homogeneous anisotropic nonmagnetic dielectric layers. The permittivity in each layer depends arbitrarily on modulus of the electric field intensity. The layers are located between two half-spaces $x < h_0$ and $x > h_N$ in Cartesian coordinate system $Oxyz$. The half-spaces are filled with homogeneous isotropic nonmagnetic media without any sources and have constant permittivities $\underline{\varepsilon}$ and $\overline{\varepsilon}$, respectively ($\underline{\varepsilon}$ and $\overline{\varepsilon}$ are arbitrary real values). Suppose that everywhere $\mu = \mu_0$ is the permeability of free space.

The fields depends on time harmonically

$$\tilde{\mathbf{E}}(x,y,z,t) = \mathbf{E}_+(x,y,z)\cos\omega t + \mathbf{E}_-(x,y,z)\sin\omega t;$$

$$\tilde{\mathbf{H}}(x,y,z,t) = \mathbf{H}_+(x,y,z)\cos\omega t + \mathbf{H}_-(x,y,z)\sin\omega t,$$

where ω is the circular frequency; $\mathbf{E}_+$, $\mathbf{E}_-$, $\mathbf{H}_+$, $\mathbf{H}_-$ are real sought for functions.

Let $\mathbf{E} = \mathbf{E}_+ + i\mathbf{E}_-$, $\mathbf{H} = \mathbf{H}_+ + i\mathbf{H}_-$ be the complex amplitudes of the fields $\mathbf{E}$, $\mathbf{H}$ [3]. Below the multipliers $\cos\omega t$ and $\sin\omega t$ are omitted.

The electromagnetic field $\mathbf{E}$, $\mathbf{H}$ satisfies the Maxwell equations

$$\operatorname{rot}\mathbf{H} = -i\omega\varepsilon\mathbf{E}; \quad \operatorname{rot}\mathbf{E} = i\omega\mu\mathbf{H}, \tag{17}$$

the continuity condition for the tangential field components on the boundaries $x = h_0$, $x = h_1$, …, $x = h_N$ and the radiation condition at infinity: the electromagnetic field exponentially decays as $|x| \to \infty$ in the domains $x < h_0$ and $x > h_N$.

The permittivity inside each layer is described by the diagonal tensor

$$\tilde{\varepsilon}_i = \begin{pmatrix} \varepsilon_i^{xx} & 0 & 0 \\ 0 & \varepsilon_i^{yy} & 0 \\ 0 & 0 & \varepsilon_i^{zz} \end{pmatrix},$$

where $\varepsilon_i^{xx} = \varepsilon_i^x + \varepsilon_0 f_i\left(|E_x|^2, |E_z|^2\right)$ and $\varepsilon_i^{zz} = \varepsilon_i^z + \varepsilon_0 g_i\left(|E_x|^2, |E_z|^2\right)$, $i = \overline{1,N}$. It is not necessary to define ε_i^{yy} as it is not contained in the equations under consideration (it will be clear below). Here ε_i^x, ε_i^z are constant parts of the permittivities ε_i^{xx}, ε_i^{zz}; ε_0 is the permittivity of free space; $f_i(u,v)$ is a continuously differentiable w.r.t. both variables; $g_i(u,v)$ is a continuous w.r.t. both variables.

Geometry of the problem is shown in Fig. 1.

Consider TM waves $\mathbf{E} = (E_x, 0, E_z)^T$, $\mathbf{H} = (0, H_y, 0)^T$, where $(\ldots)^T$ is the transposition operation. It can be shown that the fields components do not depend on y. Waves propagating along the boundaries z depend harmonically on z. So the fields components have the form

$$E_x = \mathrm{E}_x(x)e^{i\gamma z}, \quad E_z = \mathrm{E}_z(x)e^{i\gamma z}, \quad H_y = \mathrm{H}_y(x)e^{i\gamma z}, \tag{18}$$

where γ is the unknown spectral parameter (propagation constant).

Substituting components (18) into system (17), we obtain

$$\begin{cases} \gamma(i\mathrm{E}_x(x))' - \mathrm{E}_z''(x) = \omega^2\mu\varepsilon_i^{zz}\mathrm{E}_z(x), \\ \gamma^2\left(i\mathrm{E}_x(x)\right) - \gamma\mathrm{E}_z'(x) = \omega^2\mu\varepsilon_i^{xx}\left(i\mathrm{E}_x(x)\right) \end{cases}$$

and $\mathrm{H}_y(x) = \frac{1}{i\omega\mu}\left(i\gamma\mathrm{E}_x(x) - \mathrm{E}_z'(x)\right)$, where $(\ldots)' \equiv \frac{\partial}{\partial x}$.

Normalizing the latter system accordingly with the formulae $\tilde{x} = kx$, $\frac{d}{dx} = k\frac{d}{d\tilde{x}}$, $\tilde{\gamma} = \frac{\gamma}{k}$, $\tilde{\varepsilon}_i^x = \frac{\varepsilon_i^x}{\varepsilon_0}$, $\tilde{\varepsilon}_i^z = \frac{\varepsilon_i^z}{\varepsilon_0}$ $(i = \overline{1,N})$, $\underline{\tilde{\varepsilon}} = \frac{\underline{\varepsilon}}{\varepsilon_0}$, $\tilde{\overline{\varepsilon}} = \frac{\overline{\varepsilon}}{\varepsilon_0}$ where $k^2 = \omega^2\mu_0\varepsilon_0$, denoting by $Z(\tilde{x}) := \mathrm{E}_z$, $X(\tilde{x}) := i\mathrm{E}_x$ and omitting the tilde, we obtain the system [11]

$$\begin{cases} -Z'' + \gamma X' = \varepsilon_i^{zz} Z, \\ -Z' + \gamma X = \gamma^{-1}\varepsilon_i^{xx} X. \end{cases} \tag{19}$$

It is necessary to find real solutions $X(x)$, $Z(x)$ of system (19).

The value γ must be real.[5]

[5]Let us describe why in this nonlinear problem it is impossible to consider complex values of γ. As $\mathbf{E} = \left(\mathrm{E}_x(x)e^{i\gamma z}, 0, \mathrm{E}_z(x)e^{i\gamma z}\right) = e^{i\gamma z}\left(\mathrm{E}_x(x), 0, \mathrm{E}_z(x)\right)$, then $|\mathbf{E}|^2 = \left|e^{i\gamma z}\right|^2 \cdot \left(|\mathrm{E}_x|^2 + |\mathrm{E}_z|^2\right)$. As it is known $|e^{i\gamma z}| = 1$ if $\mathrm{Im}\,\gamma = 0$. Let $\gamma = \gamma' + i\gamma''$ and $\mathrm{Im}\,\gamma \neq 0$. Then $|e^{i\gamma z}| = |e^{i\gamma' z}| \cdot |e^{-\gamma'' z}| = e^{-\gamma'' z}$, that is system (19) contains z. This means that the functions $X(x)$, $Z(x)$ depend on z. This contradicts to the choice of $\mathrm{E}_x(x)$ and $\mathrm{E}_z(x)$. In the linear problem it is possible to consider complex γ.

It is supposed that

$$\varepsilon = \begin{cases} \underline{\varepsilon}, & x < h_0; \\ \tilde{\varepsilon}_1, & h_0 < x < h_1; \\ \cdots & \\ \tilde{\varepsilon}_N, & h_{N-1} < x < h_N; \\ \overline{\varepsilon}, & x > h_N. \end{cases} \tag{20}$$

It should be noticed that in system (19) for the half-spaces $x < 0$ and $x > h$, we suppose that $\varepsilon_{xx} = \varepsilon_{zz} = \text{const}$ and equal to $\underline{\varepsilon}$ or $\overline{\varepsilon}$, respectively.

The functions $X(x)$, $Z(x)$ have the properties

$$\begin{aligned} X(x) \in & C(-\infty,0] \cap C[0,h_1] \cap \ldots \cap C[h_{N-1},h_N] \cap C[h_N,\infty) \cap \\ & \cap C^1(-\infty,0] \cap C^1[0,h_1] \cap \ldots \cap C^1[h_{N-1},h_N] \cap C^1[h_N,\infty), \\ Z(x) \in & C(-\infty,\infty) \cap C^1(-\infty,0] \cap C^1[0,h_1] \cap \ldots \cap C^1[h_{N-1},h_N] \cap C^1[h_N,\infty) \cap \\ & \cap C^2(-\infty,0) \cap C^2(0,h_1) \cap \ldots \cap C^2(h_{N-1},h_N) \cap C^2(h_N,\infty). \end{aligned} \tag{21}$$

These conditions of continuity and smoothness of the functions X, Z correspond to the physical nature of the problem and will be obtained from the transmission conditions on the boundaries.

3.2 *Differential Equations of the Problem*

Denote by $\underline{k}^2 = \gamma^2 - \underline{\varepsilon}$, $(k_i^x)^2 = \varepsilon_i^x - \gamma^2$, $(k_i^z)^2 = \varepsilon_i^z - \gamma^2$, $i = \overline{1,N}$, $\overline{k}^2 = \gamma^2 - \overline{\varepsilon}$.

For the half-space $x < h_0$ we have the permittivity $\varepsilon = \underline{\varepsilon}$. From Eqs. (19) and (20) we obtain the linear system. Its solution in according to the condition at infinity is

$$\begin{cases} X(x) = X(h_0 - 0)e^{\underline{k}(x-h_0)} \\ Z(x) = \gamma^{-1}\underline{k}X(h_0 - 0)e^{\underline{k}(x-h_0)}. \end{cases} \tag{22}$$

For the half-space $x > h_N$ we have the permittivity $\varepsilon = \overline{\varepsilon}$. From Eqs. (19) and (20) we obtain the linear system. Its solution in according to the condition at infinity is

$$\begin{cases} X(x) = X(h_N + 0)e^{-\overline{k}(x-h_N)} \\ Z(x) = -\gamma^{-1}\overline{k}X(h_N + 0)e^{-\overline{k}(x-h_N)}. \end{cases} \tag{23}$$

In solution (22) the constant $X(h_0 - 0)$ is defined by initial conditions; in solution (23) the constant $X(h_N + 0)$ is defined by transmission conditions.

From Eqs. (22) and (23) it is clear that the inequality $\gamma^2 > \max(\underline{\varepsilon}, \overline{\varepsilon})$ holds.

Inside the ith layer $h_{i-1} < x < h_i$, $i = \overline{1,N}$ system (19) takes the form

$$\begin{cases} -Z'' + \gamma X' = (\varepsilon_i^z + g_i) Z, \\ -Z' + \gamma X = \gamma^{-1} (\varepsilon_i^x + f_i) X, \end{cases} \tag{24}$$

further the arguments of the functions f and g will be omitted if there is no confuse.

We can rewrite system (24) in the normal form[6]:

$$\begin{cases} \frac{dX}{dx} = \frac{\gamma^2(\varepsilon_i^z + g_i) + 2(\varepsilon_i^x - \gamma^2 + f_i) X^2 f'_{iv}}{\gamma(2X^2 f'_{iu} + \varepsilon_i^x + f_i)} Z, \\ \frac{dZ}{dx} = \frac{1}{\gamma} (\gamma^2 - \varepsilon_i^x - f_i) X, \end{cases} \tag{25}$$

where $f'_{iu} = \frac{\partial f_i}{\partial X^2}$, $f'_{iv} = \frac{\partial f_i}{\partial Z^2}$, $i = \overline{1,N}$.

3.3 Transmission Conditions

Tangential components of an electromagnetic field are known to be continuous at the interfaces. In this case tangential components are H_y and E_z. So we obtain

$$H_y(h_i + 0) = H_y(h_i - 0), \quad E_z(h_i + 0) = E_z(h_i - 0), \quad i = \overline{0,N}.$$

Normal components of an electromagnetic field have a finite jump at the interface. In this case the normal component is E_x. However the value $\varepsilon_i^{xx} E_x$ is continuous at the interface.

This implies the following conditions for the functions X and Z

$$[\varepsilon_i^{xx} X]|_{x=h_i} = 0, \quad [Z]|_{x=h_i} = 0, \quad i = \overline{0,N} \tag{26}$$

where $[f]|_{x=x_0} = \lim\limits_{x \to x_0 - 0} f(x) - \lim\limits_{x \to x_0 + 0} f(x)$ and if $x < h_0$ then $\varepsilon_0^{xx} \equiv \underline{\varepsilon}$, if $x > h_N$ then $\varepsilon_N^{xx} \equiv \overline{\varepsilon}$.

---

[6] As the functions f_i and g_i are arbitrary it is impossible to integrate system (25). However, there are conditions when the first integral of system (25) can be found. For example, the following condition $\frac{\partial f_i}{\partial(|E_z|^2)} = \frac{\partial g_i}{\partial(|E_x|^2)}$, pointed out in [8], leads to the fact that the equation

$$\frac{dX}{dZ} = \frac{(\gamma^2(\varepsilon_i^z + g_i) + 2(\varepsilon_i^x - \gamma^2 + f_i) X^2 f'_{iv}) Z}{(2X^2 f'_{iu} + \varepsilon_i^x + f_i)(\gamma^2 - \varepsilon_i^x - f_i) X}$$

can be transformed into a total differential equation. Using this condition in [11] allows to find DE.

Taking into account (26) and (22), (23) we get

$$Z(h_0-0)=\gamma^{-1}\underline{k}X(h_0-0),\quad Z(h_N+0)=-\gamma^{-1}\overline{k}X(h_N+0). \tag{27}$$

From transmission conditions (26) we obtain

$$\begin{cases} Z(h_i-0)=Z(h_i+0) \\ (\varepsilon_i^x+f_i(h_i-0))X(h_i-0)=\left(\varepsilon_{i+1}^x+f_{i+1}(h_i+0)\right)X(h_i+0),\ i=\overline{0,N}, \end{cases} \tag{28}$$

where $f_i(h_i-0)=f_i\left(X^2(h_i-0),Z^2(h_i-0)\right)$, $f_i(h_i+0)=f_i\left(X^2(h_i+0),Z^2(h_i+0)\right)$.

As the value $Z(h_0-0)$ is known then solving the second equation in Eq. (28): $\varepsilon X(h_0-0)=(\varepsilon_1^x+f_1(h_0+0))X(h_0+0)$ $(i=0)$, we find $X(h_0+0)$.

Definition 3. The value $\gamma=\overline{\gamma}$ such that nonzero solutions $X(x)$ and $Z(x)$ of system (25) exist, in the half-spaces $x<h_0$, $x>h_N$ functions $X(x)$ and $Z(x)$ are described by Eqs. (22), (23), and in the entire space functions $X(x)$, $Z(x)$ satisfy conditions (26) is called an eigenvalue of the *problem* P_M. The functions $X(x)$ and $Z(x)$ corresponding to the eigenvalue $\gamma=\overline{\gamma}$ are called eigenfunctions of the problem (see the footnote on p. 73).

Definition 4. The conjugation problem in multiply-connected domain (*problem* P_M) is to determine eigenvalues γ such that there are nonzero functions $X(x)$ and $Z(x)$ that satisfy the following conditions: if $x<h_0$ and $x>h_N$ then the function X, Z are defined by Eqs. (22) and (23), where $X(h_0-0)$ is supposed to be known and $X(h_N+0)$ is defined from transmission conditions (26); if $h_{i-1}<x<h_i$, $i=\overline{1,N}$ the functions X, Z are solutions of system (25); the functions X and Z satisfy transmission conditions (26) (see the footnote on p. 74).

3.4 Existence of Eigenvalues

In this section some theoretical results will be given that are necessary for the correct formulation and proof of convergence of the numerical method. Particularly we will determine the necessary conditions that provide unique solvability of the Cauchy problem for Eq. (25) with the initial conditions

$$X(h_i+0),\quad Z(h_i+0),\quad i=\overline{0,N-1}. \tag{29}$$

Further we will show the necessary conditions that provide continuous dependence of considered solution on the spectral parameter γ. The question of the *problem* P_M eigenvalue existence will be solved as well.

Since the solutions of system (19) in the half-spaces $x<h_0$ and $x>h_N$ are known, let us go over to the Cauchy problem for nonlinear system (25).

Consider system (25) with initial conditions (29). As in Sect. 2.4 two analogous theorems can be formulated.

Let $\sqrt{\max(\underline{\varepsilon},\overline{\varepsilon})} < \gamma_* < \gamma^* < \infty$, $\gamma \in [\gamma_*, \gamma^*]$ and $b_i < \infty$ be a constant. Define the sets

$$\Pi_{i+1} := \{(X,Z) : |X - X(h_i+0)| \le b_{i+1}, |Z_2 - Z(h_i+0)| \le b_{i+1}\}, \quad i = \overline{0,N-1}.$$

Let the values M_i be such that

$$M_i \ge \max_{(X,Z)\in\Pi_i} |P|,\ M_i \ge \max_{(X,Z)\in\Pi_i} |Q|,\ \ i = \overline{1,N},$$

where P, Q are right-hand sides of system (25).

Theorem 5. *The solution of the Cauchy problem for system (25) with initial conditions (29) exists, and in addition this solution is continuously differentiable and unique if $x \in [0,h]$, where $h \le b_i/M_i$, $i = \overline{1,N}$.*

Let $\sqrt{\max(\underline{\varepsilon},\overline{\varepsilon})} < \gamma_* < \gamma^* < \infty$ and $b_i^\gamma < \infty$ be a constant. Define the sets

$$\Pi_{i+1}^\gamma := \left\{(X,Z,\gamma) : |X - X(h_i+0)| \le b_{i+1}^\gamma, |Z - Z(h_i+0)| \le b_{i+1}^\gamma, \gamma \in [\gamma_*,\gamma^*]\right\},$$

where $i = \overline{0,N-1}$.

Let the values M_i^γ be such that

$$M_i^\gamma \ge \max_{(X,Z,\gamma)\in\Pi_i^\gamma} |P|,\ M_i^\gamma \ge \max_{(X,Z,\gamma)\in\Pi_i^\gamma} |Q|,\ \ i = \overline{1,N},$$

where P, Q are right-hand sides of system (25).

Theorem 6. *The solution $X(x,\gamma)$, $Z(x,\gamma)$ of the Cauchy problem for system (25) with initial conditions (29) exists, this solution is continuously differentiable w.r.t. x, in addition this solution is unique for all $x \in [0,h]$, where $h \le b_i^\gamma/M_i^\gamma$, $i = \overline{1,N}$ and continuously depends on γ, for all $\gamma \in [\gamma_*,\gamma^*]$.*

Pass to the question about *problem P_M* eigenvalues existence.

Let us consider the Cauchy problem for system (25) with initial conditions (29).

From formulae (27) and (28) we obtain

$$(\varepsilon_N^x + f_N(h_N-0))X(h_N-0) = \overline{\varepsilon}X(h_N+0),\ Z(h_N+0) = -\gamma^{-1}\overline{k}X(h_N+0). \quad (30)$$

On the other hand, the value $X(h_N+0)$ is an unknown and must be determined. Then from Eq. (30) we obtain $X(h_N+0) := \overline{\varepsilon}^{-1}(\varepsilon_N^x + f_N(h_N-0))X(h_N-0)$. Construct the function

$$F(\gamma) := Z(h_N-0,\gamma) - Z(h_N+0,\gamma).$$

Using Eq. (30) we can rewrite the latter formula as[7]

[7]It is clear that the value of the function F depends on the solutions of considered Cauchy problem only.

$$F(\gamma) = Z(h_N - 0, \gamma) + \gamma^{-1}\overline{\varepsilon}^{-1}\overline{k}\,(\varepsilon_N^x + f_N(h_N - 0))\,X(h_N - 0).$$

If the value $\tilde{\gamma}$ is such that $F(\tilde{\gamma}) = 0$ then $\tilde{\gamma}$ is an eigenvalue of the *problem* P_M.

Let us formulate criterion of the existence at least one eigenvalue.

Theorem 7. *Let the conditions of Theorems 5 and 6 be satisfied and let the segment* $[\underline{\gamma}, \overline{\gamma}] \subset [\gamma_*, \gamma^*]$ *be such that* $F(\underline{\gamma})F(\overline{\gamma}) < 0$. *Then at least one eigenvalue* $\tilde{\gamma}$ *of the problem* P_M *exists and* $\tilde{\gamma} \in (\underline{\gamma}, \overline{\gamma})$.

3.5 Numerical Method

The numerical method suggested below allows to determine eigenvalues of the considered problem with any prescribed accuracy. With the help of this method the normalized dependence of the propagation constant (normalized by ω^{-1}) γ w.r.t., the layer's thickness (normalized by ω) h will be depicted as well [so called dispersion curve (DC)].

Let us consider system (25). In the first layer $h_0 < x < h_1$, the initial conditions are defined by formulae (27) and (28).

Let us suppose that the pth layer has a variable thickness that is the value h_{p-1} is a constant, and the value h_p is varied from h_* to h^*. The thicknesses of other layers stay constants that is the differences $h_i - h_{i-1} = const$ for all $i = \overline{1,N}$ except $i = p$.

Denote by $h := h_p$. Let $0 < h_* < h^* < \infty$, and $\sqrt{\max(\underline{\varepsilon}, \overline{\varepsilon})} < \gamma_* < \gamma^* < \infty$ be constants. Suppose that $h \in [h_*, h^*]$ and $\gamma \in [\gamma_*, \gamma^*]$.

Divide the segments $[h_*, h^*]$ and $[\gamma_*, \gamma^*]$ into n and m pieces, respectively. We get the grid $\left\{h^{(i)}, \gamma^{(j)}\right\}$, $i = \overline{0,n}$, $j = \overline{0,m}$ and $h^{(0)} = h_* > 0$, $h^{(n)} = h^*$, $\gamma^{(0)} = \gamma_*$, $\gamma_m = \gamma^*$. Then for each pair of indexes (i, j) we obtain a pair of initial conditions

$$(X_{ij}(h_0 + 0), Z_{ij}(h_0 + 0)), \tag{31}$$

where $X_{ij}(h_0 + 0)$ is defined from the equation

$$\varepsilon X(h_0 - 0) = (\varepsilon_1^x + f_1(h_0 + 0))\,X(h_0 + 0)$$

and $Z_{ij}(h_0 + 0) = (\gamma^{(j)})^{-1}\sqrt{(\gamma^{(j)})^2 - \varepsilon}\,X(h_0 - 0)$.

Now we can state the Cauchy problem for system (25) in the first layer with initial conditions (31). Solve this problem we obtain the values $X_{ij}(h_1 - 0)$, $Z_{ij}(h_1 - 0)$. By virtue of Eq. (26), these values allow to determine initial conditions for system (25) in the second layer that is in such a way we state next Cauchy problem. Serially solve the Cauchy problems for each layer, we reach the pth layer. For this layer we have the initial conditions $X_{ij}(h_{p-1} + 0)$ and $Z_{ij}(h_{p-1} + 0)$ defined from the solution on the previous layer. Using these initial conditions we state the Cauchy problem for system (25) in the pth layer. In such a way we reach

the final layer. From the solution of the Cauchy problem in the final layer we obtain $X_{ij}(h_N-0)$ and $Z_{ij}(h_N-0)$. Using formula (30) we find $X_{ij}(h_N+0) := \overline{\varepsilon}^{-1}(\varepsilon_N^x + f_N(h_N-0))X_{ij}(h_N-0)$. From transmission conditions (26) we know that $Z_{ij}(h_N-0) = Z_{ij}(h_N+0)$. Now using $X_{ij}(h_N+0)$ and formula (30), we can find the value $Z_{ij}(h_N+0) := -(\gamma^{(j)})^{-1}\sqrt{(\gamma^{(j)})^2-\overline{\varepsilon}}X_{ij}(h_N+0)$. Taking into account the continuity of $Z(x)$ on the boundary $x = h_N$ we construct the function

$$F(\gamma_j) = Z_{ij}(h_N-0) - Z_{ij}(h_N+0);$$

we can rewrite this formula in this way

$$F(\gamma_j) = Z_{ij}(h_N-0) + (\gamma^{(j)})^{-1}\overline{\varepsilon}^{-1}\sqrt{(\gamma^{(j)})^2-\overline{\varepsilon}}\,(\varepsilon_N^x + f_N(h_N-0))X_{ij}(h_N-0).$$

Let for given $h_p^{(i)}$ such γ_j and γ_{j+1} exist that $F(\gamma_j)F(\gamma_{j+1}) < 0$. This means that at least one value $\tilde{\gamma}_j \in (\gamma_j, \gamma_{j+1})$ exists and $\tilde{\gamma}_j$ is an eigenvalue of the *problem* P_M. In addition thicknesses $h_1, \ldots, h_{p-1}, h_p^{(i)}, h_{p+1}, \ldots, h_N$ correspond to this eigenvalue.

It is clear that the possibility to find eigenvalues is based on the continuity of $F(\gamma)$ w.r.t. γ. The function $F(\gamma)$ is a linear function w.r.t., a solution of the Cauchy problem for system (25) with the initial conditions defined above. Under Theorem 6 the solution of this Cauchy problem are continuous functions w.r.t. γ. This implies that $F(\gamma_j)$ depends continuously on γ.

The numerical method is constructed in the same way as it is done in Sect. 2.5. The proof of convergence of the numerical method is completely the same.

4 Two-Layer Waveguide: TE Waves

As an example of the technique developed above, let us consider two-layer waveguide. The geometry of the problem is shown in Fig. 1.

In the case when all 4 media have constant permittivities we can derive exact DE.

Inside the first layer $0 < x < h_1$ the solution of Eq. (3) has the form

$$Y(x) = C_{11}\sin k_1 x + C_{12}\cos k_1 x, \quad Y'(x) = k_1(C_{11}\cos k_1 x - C_{12}\sin k_1 x). \tag{32}$$

Inside the second layer $h_1 < x < h_2$ the solution of Eq. (3) has the form

$$Y(x) = C_{21}\sin k_2 x + C_{22}\cos k_2 x, \quad Y'(x) = k_2(C_{21}\cos k_2 x - C_{22}\sin k_2 x). \tag{33}$$

Using transmission conditions (9) and solutions (10) and (32), (33) we obtain

$$\begin{cases} A = C_{12}, \\ A\underline{k} = k_1 C_{11}, \\ C_{11}\sin k_1 h_1 + C_{12}\cos k_1 h_1 = C_{21}\sin k_2 h_1 + C_{22}\cos k_2 h_1, \\ k_1(C_{11}\cos k_1 h_1 - C_{12}\sin k_1 h_1) = k_2(C_{21}\cos k_2 h_1 - C_{22}\sin k_2 h_1), \\ C_{21}\sin k_2 h_2 + C_{22}\cos k_2 h_2 = B, \\ k_2(C_{21}\cos k_2 h_2 - C_{22}\sin k_2 h_2) = -B\overline{k}. \end{cases}$$

Suppose that $\cos k_1 h_1 \neq 0$ and $\cos k_2 h_1 \neq 0$. Then the DE has the form

$$\left(k_1^2\overline{k} + \underline{k}k_2^2\right)\tan k_1 h_1 \tan k_2(h_2 - h_1) - k_2\left(\underline{k}\overline{k} - k_1^2\right)\tan k_1 h_1 + \\ + k_1\left(k_2^2 - \underline{k}\overline{k}\right)\tan k_2(h_2 - h_1) - k_1 k_2\left(\underline{k} + \overline{k}\right) = 0. \quad (34)$$

It is convenient to rewrite Eq. (34) in the following way

$$\tan k_2(h_2 - h_1) = k_2 \frac{\left(\underline{k}\overline{k} - k_1^2\right)\tan k_1 h_1 + k_1\left(\underline{k} + \overline{k}\right)}{\left(k_1^2\overline{k} + \underline{k}k_2^2\right)\tan k_1 h_1 + k_1\left(k_2^2 - \underline{k}\overline{k}\right)},$$

$$\tan k_1 h_1 = k_1 \frac{\left(k_2^2 - \underline{k}\overline{k}\right)\tan k_2(h_2 - h_1) - k_2\left(\underline{k} + \overline{k}\right)}{k_2\left(\underline{k}\overline{k} - k_1^2\right) - \left(\overline{k}k_1^2 + \underline{k}k_2^2\right)\tan k_2(h_2 - h_1)}.$$

Some numerical results for a two-layer waveguide are presented below. For the layers inside the waveguide the permittivities are described by Kerr law:

- $\varepsilon = \varepsilon_1 + \alpha_1\varepsilon_0|\mathbf{E}|^2$ (in the first layer).
- $\varepsilon = \varepsilon_2 + \alpha_2\varepsilon_0|\mathbf{E}|^2$ (in the second layer).

Dispersion curves are depicted in Figs. 2–4. The following parameters are used: $Y(0-0) = 1$ [see Eq. (6)]; $\underline{\varepsilon} = 1$; $\overline{\varepsilon} = 1$; $h_1 = 1$, and h_2 is varied. In each figure below the vertical axis corresponds to γ, the horizontal axis to h_2. Rhombuses are solutions of the nonlinear problem, solid curves (Figs. 2 and 3) are solutions of the linear problem ($\alpha_1 = \alpha_2 = 0$), and horizontal line (Figs. 2 and 3) corresponds to the value $\gamma^2 = \varepsilon_2$ (asymptote for the linear case).

Two interesting points should be indicated. The first one is that the dependence of the propagation constant on h is multi-valued (see Figs. 2–4). The second one is that there is a gap in Fig. 4. Both of these effects take place in the nonlinear case only. It would be interesting to understand and explain what this means from physical point of view.

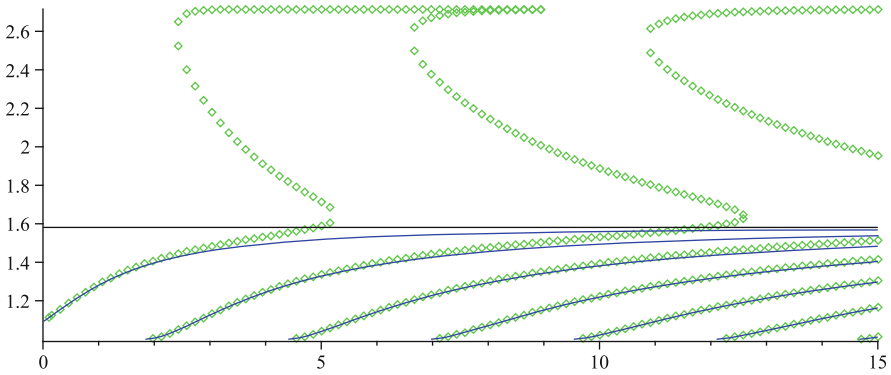

Fig. 2 $\varepsilon_1 = 2, \varepsilon_2 = 2.5, \alpha_1 = 0.02, \alpha_2 = 0.01$

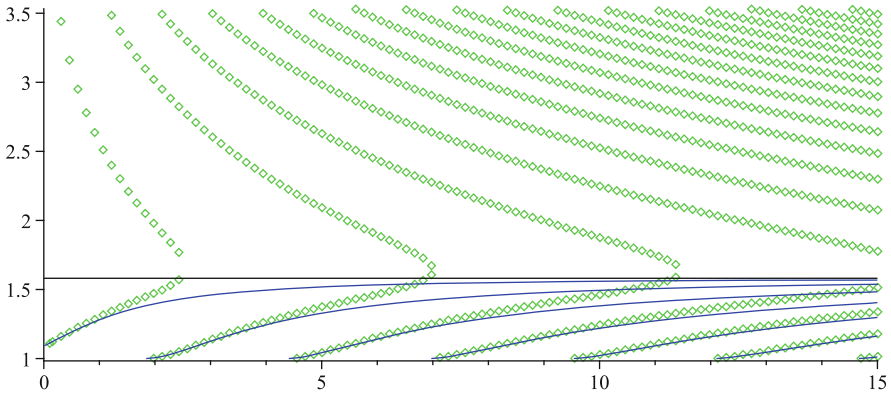

Fig. 3 $\varepsilon_1 = 2, \varepsilon_2 = 2.5, \alpha_1 = 0.02, \alpha_2 = 0.05$

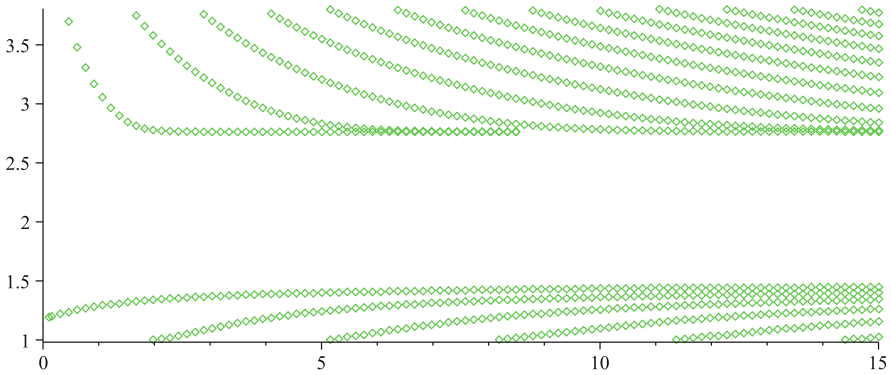

Fig. 4 $\varepsilon_1 = 2.5, \varepsilon_2 = 2, \alpha_1 = 0.02, \alpha_2 = 0.03$

5 Conclusion

Some features of the numerical method should be noticed:

– The method is rather simple to implement (any known mathsoft can solve Cauchy problems).
– The method allows to determine eigenvalues with any prescribed accuracy.
– The method can be applied to study practically arbitrary nonlinearities.

We should also emphasize that suggested numerical method is effective to find isolated eigenvalues. Moreover, let $\hat{\gamma}$ be an eigenvalue of the problem, it is clear that the total derivative of the function $F(\gamma)$ w.r.t. γ must not vanish if $\gamma = \hat{\gamma}$.

With respect to the physical significance it should be noted that the ansatz for the fields and the permittivity is based on the essential assumption that the time dependence of the optical response of the nonlinear media is described by one frequency ω (higher harmonic dependence is not considered). The permittivity depends on ω, it cannot be chosen arbitrarily but must correspond to the frequency of the fields. The difficulty may be hidden by normalization, but, for numeric, permittivity and ω must be chosen consistently. Furthermore the permittivity function represents an approximation. The dipole moment per unit volume and hence the permittivity is not simply controlled by the instant value of the electric (macroscopic) field at the point (x,y,z), due to the time lag of the medium's response. Finally, it is nonlocal in space. The model permittivity in the paper does not incorporate all these features. However considered model permittivity can be chosen to investigate some physical phenomena in nonlinear waveguides.

Acknowledgements I should like to thank Prof. Yu. G. Smirnov for his bright ideas and valuable advice. The work is supported by Russian Federation President Grant (MK-2074.211.1). I also thank to the referee for his remarks.

References

1. Adams, M.J.: An Introduction to Optical Waveguides. Wiley, Chichester (1981)
2. Boardman, A.D., Egan, P., Lederer, F., Langbein, U., Mihalache, D.: Third-order nonlinear electromagnetic TE and TM guided waves. In: Ponath, H.-E., Stegeman, G.I. (eds.) Nonlinear Surface Electromagnetic Phenomena. Elsevier science Publication, Amsterdam (1991)
3. Eleonskii, P.N., Oganes'yants, L.G., Silin, V.P.: Cylindrical nonlinear waveguides. Sov. phys. JETP **35**(1), 44–47 (1972)
4. Erugin, N.P.: Book on the Ordinary Differential Equations Theory. Nauka i Technika, Minsk (1979) (in Russian)
5. Gokhberg, I.Tz., Krein, M.G.: Introduction in the Theory of Linear Nonselfadjoint Operators in Hilbert Space. American Mathematical Society, Providence (1969)
6. Goncharenko, A.M., Karpenko, V.A.: Optical Waveguide Theory. Nauka i Technika, Minsk (1983) (in Russian)
7. Joannopoulos, J.D., Johnson, S.G., Winn, J.N., Meade, R.: Photonic Crystals: Molding the Flow of Light. Princeton University Press, Princeton (2008)

8. Joseph, R.I., Christodoulides, D.N.: Exact field decomposition for TM waves in nonlinear media. Opt. Lett. **12**(10), 826–828 (1987)
9. Lourtioz, J.-M., et al.: Photonic Crystals. Springer, Berlin (2005)
10. Schurmann, H.W., Serov, V.S., Shestopalov, Yu.V.: TE-polarized waves guided by a lossless nonlinear three-layer structure. Phys. Rev. E. **58**(1), 1040–1050 (1998)
11. Smirnov, Y.G., Valovik, D.V.: Electromagnetic Wave Propagation in Nonlinear Layered Waveguide Structures. PSU Press, Penza (2011)
12. Valovik, D.V.: Propagation of TM waves in a layer with arbitrary nonlinearity. Comp. Math. Math. Phys. **51**(9), 1622–1632 (2011). doi: 10.1134/s096554251109017x
13. Valovik, D.V.: Propagation of electromagnetic waves in a nonlinear metamaterial layer. J. Commun. Tech. Electr. **56**(5), 544–556 (2011). doi: 10.1134/s1064226911050111
14. Valovik, D.V.: Propagation of electromagnetic TE waves in a nonlinear medium with saturation. J. Commun. Tech. Electr. **56**(11), 1311–1316 (2011). doi: 10.1134/s1064226911110179
15. Valovik D.V.: Propagation of TE-waves through a nonlinear metamaterial layer with arbitrary nonlinearity. In: PIERS Proceedings, pp. 193–198, Suzhou, China, 12–16 September 2011
16. Valovik, D.V.: Electromagnetic TM wave propagation through a nonlinear metamaterial layer with arbitrary nonlinearity. In: PIERS Proceedings, pp. 1676–1680, Kuala Lumpur, Malaysia, 27–30 March 2012
17. Valovik, D.V.: Conjugation problem for TE waves propagating in a plane nonlinear two-layer dielectric waveguide. Izv. Vyssh. Uchebn. Zaved. Povolzh. Reg. Fiz.-Mat. Nauki. (2), 43–49 (2012) (in Russian)
18. Valovik, D.V., Smirnov, Y.G.: Propagation of TM waves in a Kerr nonlinear layer. Comp. Math. and Math. Phys. **48**(12), 2217–2225 (2008). doi: 10.1134/s0965542508120117
19. Valovik, D.V., Smirnov, Y.G., Shirokova, E.A.: Numerical method in the problem of electromagnetic TE wave propagation in a nonlinear two-layer waveguide. Izv. Vyssh. Uchebn. Zaved. Povolzh. Reg. Fiz.-Mat. Nauki. (1), 66–74 (2012) (in Russian)
20. Valovik, D.V., Zarembo, E.V.: On the Cauchy problem method to solve nonlinear boundary eigenvalue problem for electromagnetic TM wave propagation in a layer with Kerr nonlinearity. J. Commun. Tech. Electr. **58**(1), 62–65 (2013). doi: 10.1134/S1064226913010087
21. Zarembo, E.V.: On the numerical method to solve nonlinear boundary eigenvalue problem for electromagnetic TM wave propgation in a layer with Kerr nonlinearity. Izv. Vyssh. Uchebn. Zaved. Povolzh. Reg. Fiz.-Mat. Nauki. (1), 75–82 (2012) (in Russian)
22. Zarembo, E.V.: Numerical method to solve nonlinear boundary eigenvalue problem for electromagnetic TE wave propagation in a layer with arbitrary nonlinearity. Izv. Vyssh. Uchebn. Zaved. Povolzh. Reg. Fiz.-Mat. Nauki. (2), 59–74 (2012) (in Russian)
23. Zarembo, E.V.: Numerical method to solve nonlinear boundary eigenvalue problem for electromagnetic TM wave propagation in a layer with arbitrary nonlinearity. Izv. Vyssh. Uchebn. Zaved. Povolzh. Reg. Fiz.-Mat. Nauki. (3), 58–71 (2012) (in Russian)

Performance of Multi-cores and Multiprocessor Computers for Some 3D Problems of Nonlinear Optics and Gaseous Dynamics

Vyacheslav A. Trofimov, Olga V. Matusevich, Ivan A. Shirokov, and Mikhail V. Fedotov

Abstract We show that it is significant to take into account the architecture of computer processor and computer platform features to achieve a maximal performance of computer code at parallel computing. With this aim we examine several processor designs, which are used in high-performance computing systems of our faculty. Two problems (SHG—second harmonic generation and laser plume expansion) are chosen as a benchmark. For these problems the optimization technique for a single processor is examined, and the advantages of using the libraries are compared. In some cases the computation reorganization is necessary to take a full advantage of memory hierarchies. Full speedup of computation due to optimizations, suggested at executing in sequential mode of computer code, grows up to 8 times for Intel architectures of computer and up to 5.5 times for IBM architecture of computer.

We discuss also using shared memory at parallel computing the SHG problem. We find out the way for overcoming the performance degradation with increasing a number of processors.

1 Introduction

Nowadays, the multi-core or multiprocessor computer is widely used. Quad-core processors are now normally everywhere, and this requires the new approach for using of advantages of such computers. As a result, the problem has shifted from the hardware challenge to the software challenge: how to use all cores (or processors) to improve the performance of computation. Obviously, this problem was considered in many papers and various solutions of this were proposed (see, e.g., [1–12]).

V.A. Trofimov (✉) • O.V. Matusevich • I.A. Shirokov • M.V. Fedotov
Lomonosov Moscow State University, Leninskie Gory, 119992, Moscow, Russia
e-mail: vatro@cs.msu.ru

L. Beilina and Y.V. Shestopalov (eds.), *Inverse Problems and Large-Scale Computations*, Springer Proceedings in Mathematics & Statistics 52, DOI 10.1007/978-3-319-00660-4_7,

Nevertheless, as it is shown our experience, without taking into account the multi-core nature of the processor and the structure of the computer memory, the performance of an application may not increase essentially. Unexpectedly, in some cases the program can be executed even more slowly while a number of the cores (or a number of processors) increases because of both the lower clock speed and competition for the memory bandwidth and for the cache [13, 14]. Hence, understanding the possibilities of multiprocessor computers or multi-core ones based on various platforms is an actual and hot problem.

In this paper we compare the performance of IBM and Intel servers by analyzing performance of two nonlinear 3D problems. One of them deals with solution of two nonlinear Shrödinger equations (see, for example [15, 16]). The second one is a carbon laser plume expansion problem (see, for example [17–20]). Such simulation was carried out in [20] for 1D case. Good qualitative agreement of the computer simulation results with the experimental results was obtained by the spectroscopic method [19]. In the present paper we also deal with the expansion of a gas bunch in a cell, formed under the action of a nanosecond laser pulse but in a spatially 3D case. Our main goals were (1) to compare the performance acceleration achieved by using message passing interface (MPI) and OpenMP technologies for chosen applications, (2) to compare advantages and disadvantages of using IBM and Intel architectures for multicomputing, (3) to propose a way of the code optimization improving significantly execution speed, and (4) to estimate the speedup for the application performance on different architectures of computers that operate in either 32-bit (Intel Xeon) or 64-bit (Intel Itanium) arithmetic. The study involves also the estimation of the efficiency of using the libraries adapted to the corresponding processors. It should be emphasized that in [16] we have investigated the efficiency of using the double-processor computers with shared memory for solution of the second harmonic generation (SHG) problem in 3D case.

2 Problem Statement

2.1 Second Harmonic Generation Problem

In the axially symmetric case, a propagation of laser pulses in a medium with quadratic and cubic nonlinear response under taking into account the group velocities mismatch, dispersion of group velocity, wave-vector mismatching, and diffraction of optical radiation is described by the following set of dimensionless Shrödinger equations:

$$\frac{\partial A_1}{\partial z} + iD\Delta_r A_1 + iD_1 \frac{\partial^2 A_1}{\partial t^2} + i\gamma A_1^* A_2 e^{-i\Delta kz} + i\alpha_1 A_1 \left(|A_1|^2 + 2|A_2|^2 \right) = 0, \quad (1)$$

$$\frac{\partial A_2}{\partial z} + i\frac{D}{2}\Delta_r A_2 + \nu\frac{\partial A_2}{\partial t} + iD_2\frac{\partial^2 A_2}{\partial t^2} + i\gamma A_1^2 e^{i\Delta kz} + i\alpha_2 A_2\left(2|A_1|^2 + |A_2|^2\right) = 0,$$

$$0 < t < L_t, 0 < r < L_r, 0 < z \le L_z, \alpha_2 = 2\alpha_1 = 2\alpha.$$

Above, well-known notations are used. Functions A_j are complex amplitudes of harmonics ($j = 1,2$) at fundamental (basic) frequency ($j = 1$) and at doubled frequency ($j = 2$). The complex amplitudes are normalized on square root of the maximum intensity of the basic wave in the input section of medium ($z = 0$). t is a dimensionless time in the system of coordinates that accompanies the pulse on the basic frequency. z is a longitudinal coordinate normalized on beam diffraction length $l_d = 2k_1a^2$, a is an initial beam radius of wave with the fundamental frequency, k_1 is its wave number, and $\Delta_r = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial}{\partial r}\right)$ is the Laplace operator on transverse coordinate r that is measured in units of a. Parameters D_j characterize the group velocity dispersion, and D is a diffraction coefficient. Parameter k_j is dimensionless wave number of j-wave correspondingly ($j = 1,2$), and $\Delta k = k_2 - 2k_1$ is the dimensionless mismatching. Parameter γ is a coefficient of nonlinear coupling of the interacting waves, α_j characterizes the self-action of waves due to a cubic nonlinear response. Parameter ν is proportional to the difference of the inverse values of group velocities of the second harmonic wave and the basic one; L_z is a nonlinear medium length. L_r is its transverse size. L_t is the dimensionless time interval, during which the propagation of waves is analyzed.

For Eq. (1) the initial distributions of the complex amplitudes are necessary to define:

$$A_j(t,r,z=0) = A_{0j}(t,r), \quad j = 1,2, \quad 0 \le t \le L_t, \quad 0 \le r \le L_r. \tag{2}$$

One needs to write also the boundary conditions:

$$A_j|_{t=0,L_t=0}, \quad r\frac{\partial A_j}{\partial r}|_{r\to 0} = 0, \quad A_j|_{r=R} = 0. \tag{3}$$

As a rule, in physical experiments the second harmonic wave is absent at the input section of the medium and an amplitude of wave at the fundamental frequency has a Gaussian distribution in space and time. Therefore, the following initial conditions are chosen for computer simulation:

$$A_1|_{z=0} = A_{10}e^{-[(t-L_t/2)/\tau]^2/2}e^{-r^2/2}, A_2|_{z=0} = 0, \tag{4}$$

where τ is a dimensionless pulse duration and A_{10} is dimensionless amplitude of the basic wave. Usually, its value is equal to 1.0.

The set of Eq. (1) has a number of invariants (conservation laws), the values of which should be controlled at computer simulation. The invariants can be written as:

$$E = \int_0^{L_r}\int_0^{L_t}\left(|A_1|^2+|A_2|^2\right)rdrdt,\ I_2 = \int_0^{L_r}\int_0^{L_t}\left(2A_1\frac{\partial A_1^*}{\partial t}+A_2\frac{\partial A_2^*}{\partial t}\right)rdrdt, \quad (5)$$

$$H = \int_0^{L_r}\int_0^{L_t}\left[-2D\left|\frac{\partial A_1}{\partial r}\right|^2-\frac{D}{2}\left|\frac{\partial A_2}{\partial r}\right|^2-2D_1\left|\frac{\partial A_1}{\partial t}\right|^2-D_2\left|\frac{\partial A_2}{\partial t}\right|^2\right]rdrdt$$
$$+\int_0^{L_r}\int_0^{L_t}\left[-\nu A_2\frac{\partial A_2^*}{\partial t}+\gamma\left(A_2A_1^{*2}e^{-i\Delta kz}+A_2^*A_1^2e^{i\Delta kz}\right)\right]rdrdt$$
$$+\int_0^{L_r}\int_0^{L_t}\left[-\Delta k\left(2|A_1|^2+|A_2|^2\right)+\alpha\left(|A_1|^4+|A_2|^4+4|A_1|^2|A_2|^2\right)\right]rdrdt.$$

2.2 Laser Plume Expansion Problem

To describe the laser plume expansion we use a macroscopic approach for which the plume is governed by a set of gas-dynamic equations. In the case of large fluctuations of parameters it is possible to use quasi-gas dynamic approach [21], in which the dissipative terms are added to the macroscopic equations. Let us notice that the problem statement as well as methodology of computer simulation was early discussed in [20] in detail.

Gas is described by three independent functions: $\rho(x,y,z,t)$ is a gas density, $u(x,y,z,t)=\{u_1(x,y,z,t),u_2(x,y,z,t),u_3(x,y,z,t)\}$ is a macroscopic velocity, and $p(x,y,z,t)$ is the gas pressure. Here x,y,z are the spatial Cartesian coordinates. The gas temperature is found out from the ideal gas state equation $p=\rho RT$, where R is the gas constant. Total energy per volume unit and total specific enthalpy is calculated by the formulas: $E=\rho u^2/2+p/(\gamma-1)$ and $H=(E+p)/\rho$, γ is adiabatic power exponent. Sound velocity in the gas is given by formula $c=\sqrt{\gamma RT}$. The dynamics of a viscous heat-conducting gas is described by the quasi-gas dynamic (QGD) set of equations [21]:

$$\frac{\partial}{\partial t}\rho+\nabla_i j_m^i=0,$$

$$\frac{\partial}{\partial t}\rho u^j+\nabla_i(j_m^i u^j)+\nabla^j p=\nabla_i\Pi^{ij}, j=1,2,3,$$

$$\frac{\partial}{\partial t}E+\nabla_i(j_m^i H)+\nabla_i q^i=\nabla_i(\Pi^{ij}u_j), \quad (6)$$

$$j_m^i=\rho(u^i-w^i),\quad w^i=\frac{\tau}{\rho}(\nabla_j u^i u^j+\nabla^i p), i=1,2,3.$$

Using the QGD-set of equations instead of the traditional set of Navier–Stokes (NS) equations was advantageous due to the high computational stability of numerical algorithm constructed on its basis.

The computation is provided in the domain which is a rectangular parallelepiped: $0 \le x \le L_x$, $-L_y \le y \le L_y$, $-L_z \le z \le L_z$. On the boundary $x = 0$, a graphite target with thickness l is located. A laser beam is focused on the target. Its focal spot is an ellipse. Major axis of the ellipse is equal to $l_z = 2.7797 \times 10^{-4}$ m and minor one is equal to $l_y = 2.51927 \times 10^{-4}$ m.

Laser pulse causes the high-carbon plasma formation in the focal spot. Over time, high-temperature region of gas (diatomic carbon) is expanded in the region $x > l$: collapse of strong discontinuity of the gas density takes place. This gas is accelerated to velocities of order 10^4 m/s, and a formation of shock front happens.

3 Finite-Difference Schemes

3.1 Finite-Difference Scheme for Laser Plume Expansion Problem

For computer simulation of the laser plume expansion we introduce the uniform in space and time grid in a computational domain:

$$\Omega_{xyzt} = \omega_{xyz} \times \omega_t, \omega_{xyz} = \{x_i = x_0 + h_x i, i = \overline{0, N_x - 1}, x_0 = -h_x/2; y_j = y_0 + h_y j,$$
$$j = \overline{0, N_y - 1}, y_0 = -h_y/2; z_k = z_0 + h_z k, k = \overline{0, N_k - 1}, z_0 = -h_z/2\};$$
$$\omega_t = \{t_n = t_0 + h_t n, n = \overline{0, N_t - 1}, \quad t_0 = 0\}.$$

The numerical method used in this work is described in detail in [20]. We use the explicit finite-difference scheme. Grid functions for the gas-dynamic variables are defined in the grid points. The spatial derivatives of Eq. (6) are approximated by central finite-differences with the second order in the internal nodes of the grid. Grid steps are chosen equal to $h_x = 5 \cdot 10^{-5} m$, $h_y = 10^{-4} m$, $h_z = 10^{-4} m$, $h_t = 10^{-10} s$.

3.2 Finite-Difference Scheme for SHG Problem

For computer simulation of the SHG problem the conservative finite-difference scheme is used. Let us notice that its application for nonlinear optics problems was early discussed in detail in [22,23]. To write below the conservative finite-difference scheme for the set of Eq. (1), the nonuniform grid $\omega = \omega_t \times \omega_r \times \omega_z$ is introduced for the domain $G = G_t \times G_r \times G_z = \{0 \le t \le L_t\} \times \{0 \le r \le L_r\} \times \{0 \le z \le L_z\}$ in the following manner:

$$\omega_t = \{t_j = j\tau, j = \overline{0,N_t}, \tau = L_t/N_t\},$$
$$\omega_z = \{z_n = nh_z, n = \overline{0,N_z}, h_z = L_z/N_z\},$$
$$\omega_r = \{r_k = (k+0.5)h_r + ((k+0.5)h_r)^3, k = \overline{0,N_r}, h_r = L_r/(N_r+0.5)\}.$$

It should be stressed that the nonuniform grid in the transverse coordinate r is constructed taking into account the properties of the laser pulses interactions.

Let us define grid functions A_{1h},A_{2h} on ω and introduce the following index-free notations:

$$u = A_{1h}(t_j, r_k, z_n),\ \hat{u} = A_{1h}(t_j, r_k, z_{n+1}),\ \overset{0.5}{u} = 0.5(u+\hat{u}),\ |\overset{0.5}{u}|^2 = 0.5(|\hat{u}|^2 + |u|^2),$$

$$v = A_{2h}(t_j, r_k, z_n),\ \hat{v} = A_{2h}(t_j, r_k, z_{n+1}),\ \overset{0.5}{v} = 0.5(v+\hat{v}),\ |\overset{0.5}{v}|^2 = 0.5(|\hat{v}|^2 + |v|^2).$$

Differential operators on time are approximated in the standard way:

$$w_{\overset{o}{t}} = \frac{w(t_{j+1}, r_k, z_n) - w(t_{j-1}, r_k, z_n)}{2\tau},$$
$$w_{\bar{t}t} = \frac{w(t_{j+1}, r_k, z_n) - 2w(t_j, r_k, z_n) + w(t_{j-1}, r_k, z_n)}{\tau^2},$$

where w is one of the grid functions $u, v, \hat{u}, \hat{v}$.

The Laplace operator on the transverse coordinate is approximated as follows:

$$\Lambda_r w = \frac{1}{r_k(r_{k+1} - r_{k-1})}\left[(r_{k+1} + r_k)\frac{w(t_j, r_{k+1}, z_n) - w(t_j, r_k, z_n)}{r_{k+1} - r_k}\right]$$
$$- \frac{1}{r_k(r_{k+1} - r_{k-1})}\left[(r_k + r_{k-1})\frac{w(t_j, r_k, z_n) - w(t_j, r_{k-1}, z_n)}{r_k - r_{k-1}}\right].$$

Using the notations introduced above, for the set of Eq. (1) we can write the following difference scheme:

$$\frac{\hat{u} - u}{h_z} + i\tilde{D}\Lambda_r \overset{0.5}{u} + iD_1 \overset{0.5}{u}_{\bar{t}t} = f(\overset{0.5}{u}, \overset{0.5}{v}), \tag{7}$$

$$\frac{\hat{v} - v}{h_z} + i\frac{\tilde{D}}{2}\Lambda_r \overset{0.5}{v} + \nu \overset{0.5}{v}_{\overset{o}{t}} + iD_2 \overset{0.5}{v}_{\bar{t}t} = g(\overset{0.5}{u}, \overset{0.5}{v}),$$

$$f(\overset{0.5}{u}, \overset{0.5}{v}) \equiv -0.5i\gamma\left(\overset{0.5}{u^* v}\right)\left(e^{-i\Delta k z_n} + e^{-i\Delta k z_{n+1}}\right) - i\alpha \overset{0.5}{u}(|\overset{0.5}{u}|^2 + 2|\overset{0.5}{v}|^2),$$

$$g(\overset{0.5}{u}, \overset{0.5}{v}) \equiv -0.5i\gamma\left(\overset{0.5}{u}\right)^2\left(e^{i\Delta k z_n} + e^{i\Delta k z_{n+1}}\right) - 2i\alpha \overset{0.5}{v}(2|\overset{0.5}{u}|^2 + |\overset{0.5}{v}|^2)$$

in the internal nodes of the grid ω.

In the corresponding boundary points of the grid we write the following finite-difference equations:

$$u_{j,N_r,n} = v_{j,N_r,n} = 0,\ j = 0 \ldots N_t, n = 0 \ldots N_z, \tag{8}$$

$$\frac{\hat{u}_{j,0} - u_{j,0}}{h_z} + i\tilde{D}\frac{\overset{0.5}{u}_{j,1} - \overset{0.5}{u}_{j,0}}{0.5h_r^2} + iD_1 \overset{0.5}{u}_{\bar{t}t} = f(\overset{0.5}{u}, \overset{0.5}{v}),$$

$$\frac{\hat{v}_{j,0} - v_{j,0}}{h_z} + i\frac{\tilde{D}}{2}\frac{\overset{0.5}{v}_{j,1} - \overset{0.5}{v}_{j,0}}{0.5h_r^2} + \nu \overset{0.5}{v}_{\overset{o}{t}} + iD_2 \overset{0.5}{v}_{\bar{t}t} = g(\overset{0.5}{u}, \overset{0.5}{v}).$$

The initial condition for the mesh functions u, v is defined as:

$$u_{j,k,0} = A_{10}(t_j, r_k), \quad v_{j,k,0} = A_{20}(t_j, r_k), \quad j = 0, \ldots, N_t, \quad k = 0, \ldots, N_r. \tag{9}$$

Because the finite-difference scheme, written above, is a nonlinear one then it is necessary to use an iterative process:

$$\frac{\overset{s+1}{\hat{u}} - u}{h_z} + i\tilde{D}\Lambda_r \overset{s+1}{\overset{0.5}{u}} + iD_1 \overset{s+1}{\overset{0.5}{u}}_{\bar{t}t} = f(\overset{s}{\overset{0.5}{u}}, \overset{s}{\overset{0.5}{v}}), \tag{10}$$

$$\frac{\overset{s+1}{\hat{v}} - v}{h_z} + i\frac{\tilde{D}}{2}\Lambda_r \overset{s+1}{\overset{0.5}{v}} + \nu \overset{s+1}{\overset{0.5}{v}}_{\overset{o}{t}} + iD_2 \overset{s+1}{\overset{0.5}{v}}_{\bar{t}t} = g(\overset{s}{\overset{0.5}{u}}, \overset{s}{\overset{0.5}{v}}).$$

At the boundary points the iterative process can be written as follows:

$$\overset{s+1}{u}_{j,N_r,n} = \overset{s+1}{v}_{j,N_r,n} = 0,\ j = 0 \ldots N_t, n = 0 \ldots N_z, \tag{11}$$

$$\frac{\overset{s+1}{\hat{u}}_{j,0} - u_{j,0}}{h_z} + i\tilde{D}\frac{\overset{s+1}{\overset{0.5}{u}}_{j,1} - \overset{s+1}{\overset{0.5}{u}}_{j,0}}{0.5h_r^2} + iD_1 \overset{s+1}{\overset{0.5}{u}}_{\bar{t}t} = f(\overset{s}{\overset{0.5}{u}}, \overset{s}{\overset{0.5}{v}}),$$

$$\frac{\overset{s+1}{\hat{v}}_{j,0} - v_{j,0}}{h_z} + i\frac{\tilde{D}}{2}\frac{\overset{s+1}{\overset{0.5}{v}}_{j,1} - \overset{s+1}{\overset{0.5}{v}}_{j,0}}{0.5h_r^2} + \nu \overset{s+1}{\overset{0.5}{v}}_{\overset{o}{t}} + iD_2 \overset{s+1}{\overset{0.5}{v}}_{\bar{t}t} = g(\overset{s}{\overset{0.5}{u}}, \overset{s}{\overset{0.5}{v}}).$$

The values of functions at zero iteration on the upper layer in z-coordinate are defined as:

$$\overset{s=0}{\hat{u}} = u,\ \overset{s=0}{\hat{v}} = v. \tag{12}$$

The iterative process is terminated, if the following conditions are satisfied:

$$\max_{r_k,t_j}|\overset{s+1}{\hat{u}}-\overset{s}{\hat{u}}|\le\varepsilon_1\max_{r_k,t_j}|\overset{s}{\hat{u}}|+\delta_1,\quad \max_{r_k,t_j}|\overset{s+1}{\hat{v}}-\overset{s}{\hat{v}}|\le\varepsilon_2\max_{r_k,t_j}|\overset{s}{\hat{v}}|+\delta_2. \tag{13}$$

Parameters ε_1, ε_2, δ_1, δ_2 are positive numbers which define the divergence of the iterative process.

It is easy to see that the finite-difference scheme written above has the following approximation order $O(\tau^2+h_r^2/r+h_z^2)$. Indeed, the finite-difference equations (8) approximate the boundary conditions (3) at the point $(t_j,r_0,z_{n+0.5})$ with an order $O(\tau^2+h_r^2/r+h_z^2)$. The initial conditions (2) are approximated explicitly. In the other mesh nodes $(t_j,r_l,z_{n+0.5})$ the corresponding finite-difference equations approximate also the formulated differential equations with an order $O(\tau^2+h_r^2/r+h_z^2)$.

To inverse the finite-difference operator in time we use the pseudo-spectral method [24, 25] based on fast Fourier transform. With this aim we introduce the functions u_l,v_l,f_l,g_l as follows:

$$w_\omega^{(l)}=\sum_{j=0}^{N_t}we^{-i\lambda_l t_j},\quad \lambda_l=\frac{2\pi l}{L_t},l=\overline{0,N_t},$$

where w is one of grid functions $u,v,\hat{u},\hat{v},f,g$.

Then, the finite-difference scheme (10) can be written in the spectral domain as:

$$\frac{\overset{s+1}{\hat{u}}{}_\omega^{(l)}-u_\omega^{(l)}}{h_z}+i\tilde{D}\Lambda_r\overset{s+1}{\underset{0.5}{u}}{}_\omega^{(l)}-iD_1\lambda_l^2\overset{s+1}{\underset{0.5}{u}}{}_\omega^{(l)}=f_\omega^{(l)}(\overset{s}{\underset{0.5}{u}},\overset{s}{\underset{0.5}{v}}), \tag{14}$$

$$\frac{\overset{s+1}{\hat{v}}{}_\omega^{(l)}-v_\omega^{(l)}}{h_z}+i\frac{\tilde{D}}{2}\Lambda_r\overset{s+1}{\underset{0.5}{v}}{}_\omega^{(l)}+iv\lambda_l\overset{s+1}{\underset{0.5}{v}}{}_\omega^{(l)}-iD_2\lambda_l^2\overset{s+1}{\underset{0.5}{v}}{}_\omega^{(l)}=g_\omega^{(l)}(\overset{s}{\underset{0.5}{u}},\overset{s}{\underset{0.5}{v}}).$$

For inversion of the Laplace operator on the transverse coordinate the sweep method is used. To calculate the solution of the finite-difference problem on the upper layer and $s+1$ iteration, we apply inverse Fourier transform to it:

$$w=\frac{1}{(N_t+1)}\sum_{l=0}^{N_t}w_\omega^{(l)}e^{i\lambda_l t_j},$$

where w is one of the grid functions $\hat{u},\hat{v}$.

Finally, the flowchart of an algorithm for solution of the SHG problem is shown in Fig. 1.

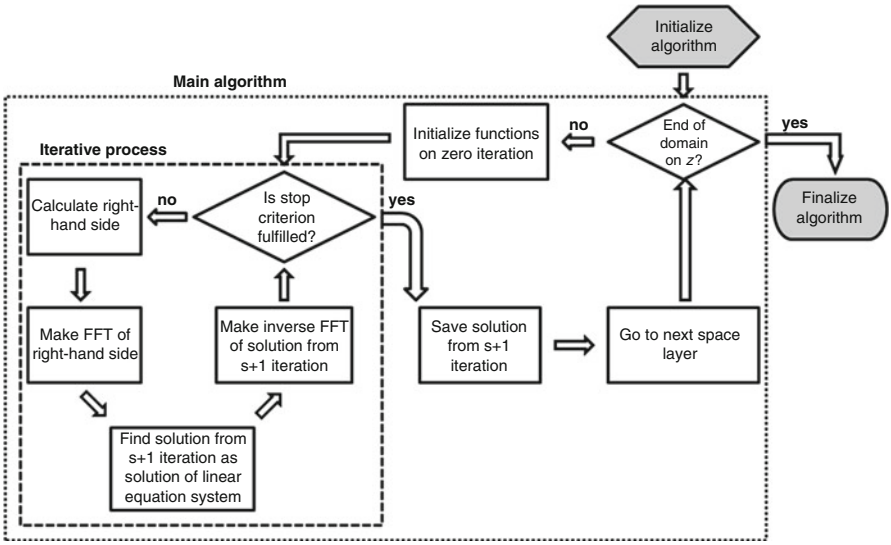

Fig. 1 Flowchart for computer implementation of the algorithm based on using finite-difference scheme for solution of the set of Eq. (1)

4 Hardware and Software Considered

For further analysis one has to introduce some characteristics presented in Table 1. It shows the computer architectures that were considered below and their characteristics such as a number of processors, CPU clock rate, FSB (front-side bus) frequency, and cache memory size.

5 Improving Performance of SHG Problem Implementation on Multicore and Multiprocessor Computers

Because a solution of the 3D problem requires actually a large number of mesh nodes at the computer simulation, it is very important to consider and develop various ways of increasing the computation speed of the program. Therefore, we consider below how the 3D SHG problem can be efficiently implemented on computers with various architectures. Results of the consideration are discussed from two points of view. Firstly, we are interested in how a proposed optimization affects the total execution time. Secondly, after all program improvements were made we use the program as a benchmark to compare the computation speed for various computer architectures.

Before to do the program optimizations, one should make sure that the realized algorithm reuses as much calculated data as possible instead recalculating them.

Table 1 Hardware of computers using for computer simulation

	Processor	Number of processors × number of CPU cores	CPU clock rate (GHz)	FSB frequency	L1/L2 cache memory size
Intel	Core2 Duo E6850	1×2	3.0	1,333 MHz	64 Kb/4 Mb
	Xeon Irwindale	2×1	3.4	800 MHz	16 Kb/2 Mb
	Xeon 5160	2×2	3.0	1,333 MHz	64 Kb/4 Mb
	Itanium 2	4×2	1.6	533 MHz	–/24 Mb
IBM	pSeries 690	4×4	1.3	n/a	n/a
	PowerPC 450	1×4	0.85	n/a	32 Kb/3.5 Mb

Table 2 Software used in the paper

	Processor	Compilers	Performance libraries
Intel	All	Intel® C++ compiler professional edition 11.1 for Windows Intel® visual Fortran compiler professional edition 11.1 for Windows	Intel® math kernel library (Intel® MKL) 10.2 for Windows
IBM	pSeries 690	IBM XL C++ professional/C for AIX V6.0 IBM XL Fortran professional for AIX V8.1	Engineering and scientific subroutine library (ESSL) for AIX V3.3
	PowerPC 450	IBM XL C/C++ advanced edition for Blue Gene/P V9.0 IBM XL Fortran advanced edition for Blue Gene/P V11.1	ESSL for Linux on Power V4.3

Though, it is an obvious statement, its implementation in practice often is not so easy because of limitation of the computer memory, for example. It should be stressed that the starting point of our optimizations is an algorithm that had already been modified to reuse the data obtained during program execution.

Let us refer to Fig. 2. It demonstrates how optimizations considered influences on the execution time of each computational block. The left axis shows the execution time of each block measured in seconds. The right axis shows what part of total execution time in percents (%) takes each algorithmic block. The program execution time without applied optimizations is depicted by white columns with solid line (first column in each group). Comparing these columns allows us to conclude that more than about 90 % of the time or even more is spent in iterative process.

Thus, optimizations, that can reduce the number of iterations, are essential for speeding up the program execution. In [16] it was shown that increasing the

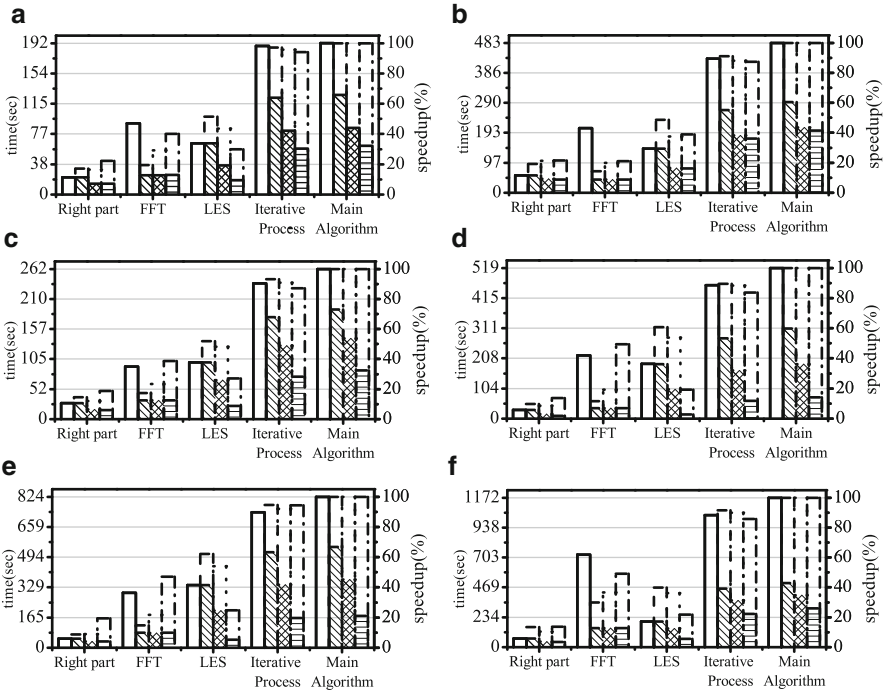

Fig. 2 Execution time of computation of various parts for the SHG problem algorithm in comparison with total execution time for the main algorithm is shown in seconds (*left axis*) or in percents (%) (*right axis*) on Intel Core2 Duo (**a**); Intel Xeon x2 (**b**); Intel Xeon x4 (**c**); Intel Itanium 2 (**d**); IBM p690 Series (**e**); IBM PowerPC 450 (**f**) for various optimizations applied: initial program (*no hatching; solid line*); using corresponding libraries (*inclined hatching; dashed line*); using complex data format (*cross hatching; dotted line*); decreasing cache misses (*horizontal hatching; dash-dotted line*)

accuracy of floating point of data can lead to a reduction in a number of iterations. Therefore, in present paper all calculations are carried out using double-precision arithmetic. Further optimizations are targeted on decreasing the execution time of one iteration.

5.1 Using Performance Libraries

First of all we stress that the FFT calculation block requires much more execution time of the initial program at using the most of processors. As it is hard to suggest an optimization that would decrease the number of FFT calls in the main algorithm, it is reasonably to decrease the execution time of the FFT procedure. Because the FFT is used widely then very effective algorithms were developed [25] and their

realizations can be found in many contemporary mathematical libraries. Therefore, it is preferable to use one of these libraries. But even more important question is about the performance of FFT containing in various libraries adapted to the processor architecture (Table 2).

Columns with inclined hatching (second column in each group) in Fig. 2 demonstrate the advantages of using libraries. Using libraries, one can decrease the FFT computation time from three to five times in comparison with the time computation of its self-written implementation. Consequently, the speedup of the program increases from 30 % to 40 % for Intel processors and from 30 % to 60 % for IBM processors. After this optimization, the most time-consuming block becomes the solution of linear equations by the sweep method. Thus, it becomes the main target of our further optimization.

5.2 Enhancing Data Access

If the program deals with large data volume, the cache memory plays an important role. Understanding the cache work, one can give a hint on how to improve the program performance. Because of this, we provide here a brief overview of how cache works. The detailed information can be found in special literature.

The cache is a smaller and faster memory (in comparison to main memory) which stores copies of the data from the most frequently used in main memory locations. Without the cache memory at every requesting data the CPU would send a request to the main memory which would be sent back then across the memory bus to the CPU. This is a slow process in computing term. Therefore, to increase computation speed, the processor checks firstly whether a copy of the required data is in cache before accessing the main memory. If so, the processor uses immediately the data from the cache. Typically, the time of accessing cache depends on both different microarchitecture and processor implementations and platform components. Nevertheless, a general trend is that the cost of access to the main memory can be expensive in 3–10 times (and more) than accessing data from the cache.

If requested data is not located in the cache, then the processor needs to access the main memory. When it does, it makes a copy of requested data and stores it in the cache with hoping that the data will be required again soon. If the cache memory has space, it will also store data that is close to the requested one. It makes sense at working with arrays: if some element of the array was accessed, then it is very probable that its adjacent elements will be used soon. Consequently, the processor makes also copies of these elements to the cache. From the above one can conclude that to achieve the quickest possible response time to the CPU, it is very important to improve data locality. It can be achieved by several optimization techniques, some of which are described below.

Table 3 Location of data in memory for two real arrays (a) and for complex array (b)

(a)

$Reu_{0,k}$	...	$Reu_{j,k}$	...	$Reu_{N_t,k}$	$Imu_{0,k}$	...	$Imu_{j,k}$	...	$Imu_{N_t,k}$

(b)

$Reu_{0,k}$	$Imu_{0,k}$	...	$Reu_{j,k}$	$Imu_{j,k}$	...	$Reu_{N_t,k}$	$Imu_{N_t,k}$

Listing 1 Right part calculation

```
do i=0,nr-1
    do j=0,nt-1
        r1 = u(j,i,1) * dconjg(u(j,i,1)) + un(j,i,1) *
            dconjg(un(j,i,1))
        r2 = u(j,i,2) * dconjg(u(j,i,2)) + un(j,i,2) *
            dconjg(un(j,i,2))
        f(j,i,1) = -dcmplx(0,1) * (gam(1) * (u(j,i,2) +
             un(j,i,2)) * dconjg(u(j,i,1) + un(j,i,1)) +
             alf(1) * (u(j,i,1) + un(j,i,1)) * (r1 +
            beta1 * r2))
        f(j,i,2) = -dcmplx(0,1) * (gam(2) * (u(j,i,1) *
             u(j,i,1) + un(j,i,1) * un(j,i,1)) + alf(2)
            * (u(j,i,2) + un(j,i,2)) * (beta1 * r1 + r2)
            )
    end do
end do
```

5.3 *Positioning Arrays in Memory*

In dependence of the data location in memory, the execution time of the program can vary significantly. Choice of convenient data structure is very important for current using the data from the cash memory. Our program operates with complex functions, which can be stored in two different ways: as two real arrays for real and imaginary parts, respectively, or as one complex array. The possible way of the data location in memory for each case is shown in Table 3.

Columns with crossed hatching in Fig. 2 (the third column in each group) illustrate the speedup of various blocks of program [such as computation of right-hand side of the algebraic equation and solution of the set of linear equations (SLE)] at using complex data type instead storing real and imaginary part of complex function separately. It can be seen that it is preferable to use the complex data type because it allows to decrease the execution time of the main algorithm by 1.4–1.6 times.

To understand the reasons those lead to increasing performance, we will refer, for example to right-hand side calculation block:

Listing 2 SLE solution

```
do jg=1,ng
    do j=0,nt-1
        do i=0,nr-1
            g(i+1) = f(j,i,jg) + fp(j,i,jg) + p1(i,jg) * h(j,i,jg
                ) * g(i)
        end do
        u(j,nr1,jg) = g(nr) * h(j,nr,jg)
        do i=nr-1,1,-1
            u(j,i-1,jg) = h(j,i,jg) * (g(i) + u(j,i,jg))
        end do
    end do
end do
```

Here, the rightmost dimension of arrays denotes the harmonic number while the leftmost dimension denotes time coordinate. *un* denotes the solution at the previous z layer, u denotes the solution from s iteration at current z layer, and f contains right part of the equations. We can see that a calculation of each element of f array requires both real and imaginary parts of u and *un* arrays. So, reducing the cache can significantly decrease the performance.

We believe that the speedup for Intel architectures is obtained due to enhancing data storage in the cash. If the complex data type is used then the real and imaginary parts of functions are always in the neighboring memory cells. Therefore, when the processor tries to access the real part of an element which is not in the cache, it will put it in the cache with copying several neighboring elements to cache also. Hence, for the memory schema shown in Table 3b it will almost certainly get an imaginary part of this element and can start calculations immediately. However, this is not true for the memory schema shown in Table 3a, as in that case the imaginary part of an element is not stored in the neighboring location to real part and will not be copied to the cache. As a consequence, the processor will have to wait when the imaginary part element is delivered from the main memory.

5.4 Loops Reorganization

Below we give an example of sweep method implementation for a solution of the SLE (see Listing 2). It is well known, that in the Fortran the fastest array access is achieved when arrays are accessed in column-major order. However, in this example the index i varies often, and it is the second one in the array subscript for the expression $x(j,i,jg)$, where x is one of the arrays f, fp, h, u. Therefore, these arrays are accessed in row-major order, which leads to enormous amount of cache misses.

Listing 3 SLE solution

```
do jg=1,ng
    do i=0,nr-1
        do j=0,nt-1
            g(j,i+1) = f(j,i,jg)+fp(j,i,jg)+ p1(i,jg)*h(j,i,jg)*g
                (j,i)
        end do
    end do
    u(:,nr1,jg) = h(:,nr,jg)*g(:,nr)
    do i=nr-1,1,-1
        do j=0,nt-1
            u(j,i-1,jg) = h(j,i,jg)*(g(j,i)+u(j,i,jg))
        end do
    end do
end do
```

Usually, to solve this problem, either the loops on the index i and the index j have to be interchanged or the arrays have to be reorganized so that the index i becomes the leftmost index. In the case under consideration both solutions are unacceptable. If we reorganize arrays then the FFT in further computations should be done on second index of the array. This will result in decreasing the performance significantly. Changing the loops order cannot be done directly because of dependence of $g(i+1)$ on $g(i)$. So, to make the array, that is convenient to natural column-major order, we had to add an additional dimension to array g and change the order of the DO loops so the innermost loop variable corresponds to the leftmost array dimension (Listing 3).

Result of proposed modification is demonstrated in Fig. 2 by the columns with horizontal hatching. We see that the loops reorganization can significantly increase the performance of the SHG program.

5.5 Total SHG Program Speedup on One Core

Here we would like to summarize the speedup achieved by all optimizations for serial execution on one core. From Fig. 2 we can conclude that the greatest speedup is achieved on Itanium2 architecture and make up approximately eight times. Program efficiency on the IBM architectures is increased 4–5 times. It should be stressed that the optimization influences more weakly on the program execution on Intel Xeon Irwindale computer (approximately 2.25 times speedup).

Thus, paying attention to the processor architecture can help to increase the application performance significantly even on one CPU core. Further decreasing of execution time can be achieved by parallelizing application execution.

5.6 Comparing Intel and IBM Architectures

After applying all optimizations we compare the execution time of the SHG program on processors of various vendors and use the program as a benchmark. Comparing total execution time of the main algorithm on the various architectures (Fig. 2) results in several conclusions. Firstly, although Itanium2 CPU clock rate is almost three times smaller than that of Intel Core2, the program executes approximately during the same time on both of these processors.

Next, although the difference in CPU clock rate between Intel Itanium2 and IBM pSeries 690 is not so bigger (1.6 and 1.3 GHz correspondingly), the difference in the program execution time achieves more than 15 times. Our computations also show that the Intel Itanium2 processor is about 4.5 times faster than IBM PowerPC 450 at the SHG program execution. Here we should also note that Intel Xeon Irwindale processor shows significantly lower performance than other Intel architectures considered in this paper. This is due to lower FSB frequency rate and lower cache sizes.

6 Parallelization

6.1 Parallelization of SHG Problem

In this paper we used OpenMP technology [11] to parallelize the SHG program. OpenMP technology supports multi-platform shared-memory parallel programming. Hence, it can be applied to all architectures considered in this paper. However, parallelizing program for architectures with distributed memory (clusters, etc.) one should use other technologies, i.e., MPI. The advantages of OpenMP technology are scalability and ability to use the same code for both sequential and parallel applications. Besides, the OpenMP technology is rather simple to use: most loops can be threaded by inserting only one compiler directive before the loop. Data decomposition is handled automatically by compiler directives.

The maximum performance at using the OpenMP technology is achieved, if it is used for the most time-consuming loops in the application. Therefore, we apply it to the right-hand-side calculation, FFT computation, and solution of set of the linear equations. We expect to get N times increasing the performance (or something close to this dependence) when running program parallelized using OpenMP technology on N processor platform.

However, not all parallelization results in speedup. Generally speaking, N processors in an SMP may have N times the computation power, but the memory bandwidth usually does not scale up N times. Quite often, the original memory path is shared by multiple processors, and performance degradation may be observed when they compete for the shared memory bandwidth [13]. Figure 3a shows the execution time of the SHG code in the parallel mode for different number of

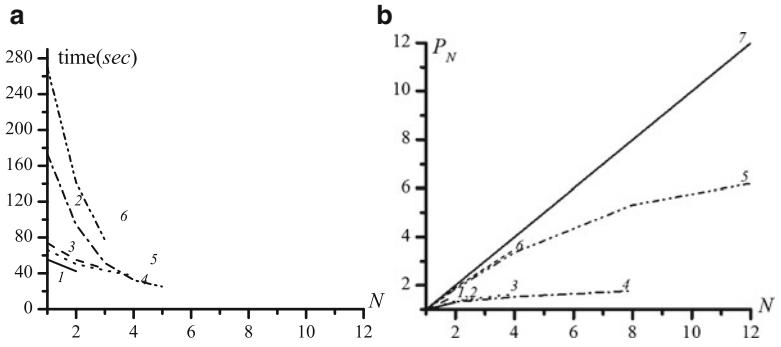

Fig. 3 Comparison of execution time of the SHG problem code (**a**) and performance speedup (**b**) for various number of threads (N) on Intel Core2 Duo (1, *solid line*); Intel Xeon x2 (2, *dashed line*); Intel Xeon x4 (3, *dotted line*); Intel Itanium 2 (4, *dash-dotted line*); IBM pSeries 690 (5, *dash-dot-dotted line*); IBM PowerPC 450 (6, *short dashed line*). *Solid line* (7) shows ideal performance speedup

Listing 4 Two arrays addition

```
do i = 1, n
    d(i)=a(i)+b(i)
end do
```

CPU, and Fig. 3b demonstrates the corresponding performance speedup value $P_N = T_1/T_N$, where T_N is the application execution time corresponding to N processors. From Fig. 3 we can conclude that the IBM servers allow good application scalability with increasing of CPU number which is no greater than 4. For Regatta computer the further increasing CPU number leads to speedup of total execution time, but it differs significantly from linear law.

Scaling Intel architectures is very poor. To clarify this phenomenon we make an additional investigation.

6.2 Parallelization of Arrays Addition Loop

Let's consider the loop for two arrays addition (Listing 4).

Here n is chosen to be a large number (i.e., 2^{26}). The memory for arrays is allocated dynamically. For example, the arrays are initially filled with random numbers. The program is compiled with (Release configuration) or without (Debug configuration) optimizations applied by compiler to produce most efficient binary code.

At first, we measure the execution time of this loop in sequential mode. Then we thread this loop with the help of OpenMP technology and measure the

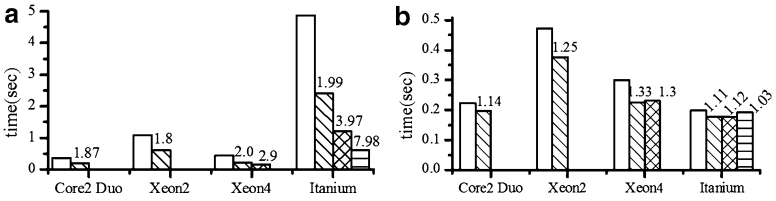

Fig. 4 Loop execution time for different Intel architectures in sequential (*no hatching*) and parallel modes with N = 2 (*inclined hatching*); 4 (*crossed hatching*); 8 (*horizontal hatching*) compiled in Debug (**a**) and Release (**b**) configuration

execution time of this loop in the parallel mode for various numbers of threads. The comparison of program execution time in sequential and parallel modes is presented

Listing 5 Two arrays addition

```
do i = 1, n
    d(i) = a(i)+b(i)
    d(i) = c(i)+a(i)+b(i)
    a(i) = a(i)/b(i)
    ! further four lines can be repeated any number of
       times
    d(i) = d(i) - a(i)
    d(i) = d(i)/b(i)
    d(i) = d(i)+a(i)*b(i)
    d(i) = d(i)/a(i)
end do
```

in Fig. 4. It can be seen that if no compiler optimizations are applied, we can observe almost ideal performance speedup. However, if the program was generated using Release configuration then the faster code is obtained. Nevertheless, the scaling factor does not exceed 1.33, which indicates poor loop scalability.

Next, we increased the number of operations on elements of arrays *a*, *b*, and *d* inside the loop (Listing 5).

+++ Figure 5 illustrates the dependence of the loop execution time on a number of operations in the loop for different number of processors. While a number of operations grow, the efficiency of parallelization increases. All results were obtained for the Release configuration.

From Fig. 5 we can conclude that each architecture has its own number of operations in loop for which the linear scaling is achieved. The biggest number of operations is required for Intel Itanium2, which explains the poor scalability depicted in Fig. 4. Thus, if the program does not contain loops that perform a lot of operations on the same arrays, the parallelization on Intel architectures is not effective.

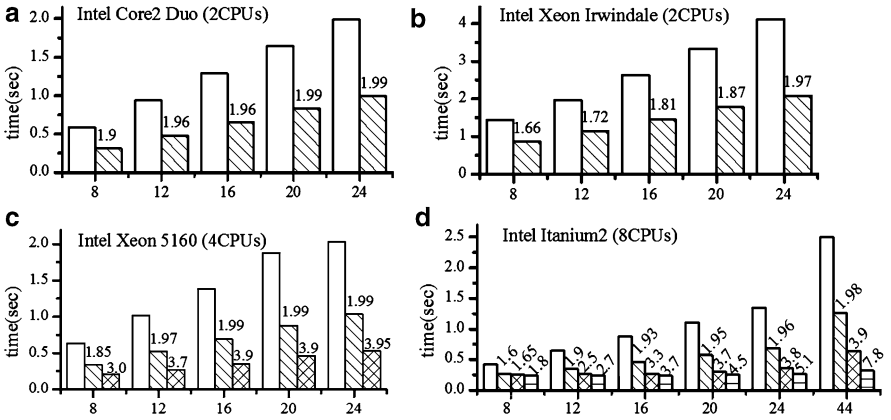

Fig. 5 Loop execution time on a number of operation in the loop for sequential (*no hatching*) and parallel mode with N = 2 (*inclined hatching*); 4 (*crossed hatching*); 8 (*horizontal hatching*) compiled in Release configuration

7 Parallelization of Laser Plume Expansion Problem

In [20], the authors studied the laser plume expansion in detail, and the results obtained were in good agreement with the corresponding experimental data. However, in the present paper we consider the plume expansion problem as a model for parallel computing test only. Briefly, the computational domain is decomposed by n sub-domains with the planes $x = const$. Each processor node calculates time-evolving parameters in one sub-domain. So, the number of n processor nodes can work simultaneously. To ensure the correct work of the finite-difference algorithm, it is necessary to exchange boundary data between nodes of neighbor sub-domains. The exchange procedure is based on MPI protocol.

To measure the performance of parallel computing, the program has been run on different parallel computer systems with varying number of grid nodes and processor number. The initialization time and hard disk data exchange time are eliminated at estimation of execution time of the program. Figure 6a presents the computation time, and Fig. 6b shows the corresponding performance speedup P_n defined above. At computer simulation we use the following number of mesh nodes: $N_x = 80$ (in some cases $N_x = 160$), $N_y = 20$, $N_z = 20$. A number of time steps is equal to $N_t = 500$. The ideal linear performance dependence on processor nodes number n is illustrated with the line 7 (Fig. 6b).

The curve 1 corresponds to the Regatta computer at computation with $N_x = 80$. The curve 3 corresponds to the Blue Gene/P computer at the same number of mesh nodes in x-direction: $N_x = 80$ (Fig. 6a, b). Curves 2 and 4 demonstrate the computation time and performance speedup Regatta and Blue Gene/P computer, respectively, with enlarged spatial grid ($N_x = 160$). Regatta computer is tested with 1–12 processors used, and Blue Gene/P computer is tested on 1–16 processor nodes used. One can see that the performance dependence is close to the linear

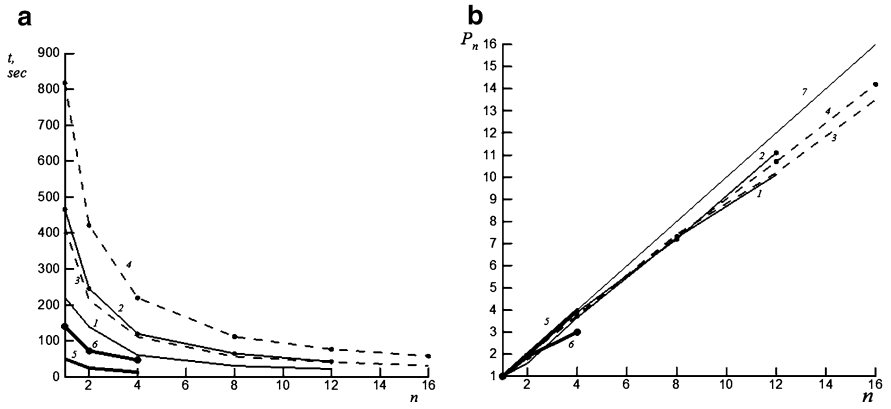

Fig. 6 Comparison of execution time of laser plume expansion problem (**a**) and performance speedup (**b**) for different number of threads (n) on HP Regatta (1, 2, *solid lines*); IBM BlueGene/P (3, 4, *dashed line*); Intel Xeon x4 (5, 6, *solid lines*). *Solid line* (8) shows ideal performance speedup

one. This means that exchange machine time is small compared to the calculations computation time. However, if one increases a number of grid nodes ($N_x = 160$, curves $2,4$), the performance dependence becomes closer to the linear one in comparison with the case of grid nodes $N_x = 80$ (curves $1,3$). This is predictable result, as soon the exchange time stays constant.

Figure 6a, b shows also the performance test results for the four-node platform of Intel Xeon 5160 computer. The numerical test is identical to that used on Regatta and Blue Gene/P (curves 5–6). The curve 5, 6 correspond to $N_x = 80$, $N_x = 160$, respectively. One can see that it is possible to achieve the performance dependence that is close to the linear one. The comparison of time measurement curves (Fig. 6a) shows that the performance of the Intel Xeon 5160 processor node surpasses the performance of the Regatta node by three times and surpasses the Blue Gene/P node by six times.

One can see that the parallel computing algorithm, applied here to the laser plume expansion problem, is well scalable, if the spatial grid nodes is large enough and the performance dependence, that is close to the linear one, can be achieved.

8 Conclusions

Taking into account the processor architecture one can increase the computation performance significantly up to four (or more) times even on one processor.

We show the advantage of using libraries which can significantly reduce the program execution time.

We suggest the way of maximal speedup achievement at reusing of computational results and at using complex and double-precision data types, loops reorganization, and optimization of data storage and data access.

We show how one can be achieved the linear growth of performance with increasing the number of Intel processor, working in optimal mode at using shared memory.

We show that the nonoptimal mode of Intel processor gives the linear growth of performance (as IBM computer) with increasing a number of processors.

One Itanium 2 processor can have performance which is equivalent of using: (a) 15 processor of Regatta (IBM) and (b) 3.5 processor of Blue Gene (IBM).

Acknowledgements This paper was partly financially supported by Russian Foundation for Basic Research (grant number 12-01-00682).

References

1. Paulavicius, R., Gilinskas, J.: Parallel branch and bound algorithm with combination of Lipschitz bounds over multidimensional simplices for multicore computers. Parallel Sci. Comput. Optim. **27**, 93 (2009)
2. Starikovicius, V., Henty, D., Iliev, O., Lakdawala, Z.: A parallel solver for the 3D simulation of flows through oil filters. Parallel Sci. Comput. Optim. **27**, 181 (2009)
3. Hamid, N.A.W.A., Paul, C.: Comparison of MPI benchmark programs on shared memory and distributed memory machines (point-to-point communication). Int. J. High Perform. Comput. Appl. **24**, 469 (2010)
4. Toshiyuki, I., Takuma, K., Susumu, Y., Masahiko, O., Masahiko, M.: High-performance quantum simulation for coupled Josephson junctions on the earth simulator: a challenge to the Schrodinger equation on 2564 grids. Int. J. High Perform. Comput. Appl. **24**, 319 (2010)
5. Pavan, B., Anthony, C., William, G., Rajeev, T., Ewing, L.: The importance of non-data-communication overheads in MPI. Int. J. High Perform. Comput. Appl. **24**, 5 (2010)
6. Keyes, D.: Partial differential equation-based applications and solvers at extreme scale. Int. J. High Perform. Comput. Appl. **23**, 366 (2009)
7. Heise, B., Jung, M.: Parallel solvers for nonlinear elliptic problems based on domain decomposition ideas. Parallel Comput. **22**, 1527 (1997)
8. Chau, M., El Baz, D., Guivarch, R., Spiteri, P.: MPI implementation of parallel subdomain methods for linear and nonlinear convection-diffusion problems. J. Parallel Distrib. Comput. **67**, 581 (2007)
9. Llorente, I.M., Tirado, F., Vazquez, L.: Some aspects about the scalability of scientific applications on parallel architectures. Parallel Comput. **22**, 1169 (1996)
10. Bourgeade, A., Nkonga, B.: Dynamic load balancing computation of pulses propagating in a nonlinear medium. J. Supercomput. **28**, 279 (2004)
11. OpenMP: Application program interface. Available from internet: www.openmp.org
12. Voevodin, V.V., Voevodin, V.V.: Parallel Computing. BHV-Petersburg, St. Petersburg (2002) (In Russian)
13. Drepper, U.: What Every Programmer Should Know About Memory. Red Hat, Inc., Chester (2007)
14. Tanenbaum, A.S.: Structured Computer Organization. Pearson Prentice Hall, Upper Saddle Rivar (2006)

15. Ashihara, S., et al.: Soliton compression of femtosecond pulses in quadratic media. J. Opt. Soc. Am. B **19**, 2505 (2002)
16. Matusevich, O.V., Trofimov, V.A.: The efficiency of application of dual-processor computers for the analysis of the three-dimensional second harmonic generation problem. Moscow Univ. Comput. Math. Cybern. **31**, 97 (2007)
17. Singh, R.K., Narayan, J.: Pulsed-laser evaporation technique for deposition of thin films: Physics and theoretical model. Phys. Rev. B **41**, 8843 (1990)
18. Itina, T.E., Hermann, J., Delaporte, P., Sentis, M.: Combined continuous-microscopic modeling of laser plume expansion. Appl. Surf. Sci. **27** 208–209 (2003)
19. Kuzyakov, Y.Y., Lednev, V.N., Nol'de, S.E.: Evolution of laser plume upon graphite ablation in vacuum and nitrogen. High Energy Chem. **39**, 413 (2005)
20. Kuzyakov, Y.Y., Trofimov, V.A., Shirokov, I.A.: Computer simulation of graphite target ablation under the action of a nanosecond laser pulse. Tech. Phys. **53**, 154 (2008)
21. Elizarova, T.G.: Quasi-Gas Dynamic Equations. Springer, Berlin (2009)
22. Zakharova, I.G., Karamzin, Y.N., Trofimov, V.A., Veremeenko, T.V.: Calculation of the process of thermal self-action of two-dimensional light beams in clear and cloudy media. In: Computer Software Library of Applied Programs BIM-M, vol. 21, p. 123. Institute of Mathematics AN BSSR, Minsk (1985) (In Russian)
23. Karamzin, Y.N., Sukhorukov, A.P., Trofimov, V.A.: Mathematical Modeling in Nonlinear Optics. Moscow University Press, Moscow (1989) (in Russian)
24. Samarskii, A.A., Nikolaev, E.S.: Numerical Methods for Grid Equations. Birkhäuser Press, Basel (1989)
25. Brigham, E.O.: The Fast Fourier Transform. Prentice-Hall, New York (2002)

Modeling of Electromagnetic Wave Propagation in Guides with Inhomogeneous Dielectric Inclusions

Alexander Smirnov, Alexey Semenov, and Yury Shestopalov

Abstract We consider scattering in the time domain of electromagnetic waves from inhomogeneous dielectric inclusions in a 3D waveguide of rectangular cross section. All electromagnetic field components are calculated, and transport of energy in the guide is investigated using finite difference time domain (FDTD) method in different frequency ranges. An efficient 3D FDTD *EMWSolver3D* solver for the nonstationary Maxwell equation system is used. The model computations are performed for the H10-mode scattering from parallelepiped-shaped dielectric inclusions. Attenuation and propagation factors are calculated for the transmitted modes and field distributions are visualized. The present method can be used for a wide class of waveguide problems that meet substantial difficulties as far as numerical solution by conventional FDTD methods is concerned due to complex geometries or computational requirements. The solver employs algorithms of parallel computations and is implemented on supercomputers of last generation for solving large-scale problems with characteristic matrix dimensions achieving 10^{12}.

1 Introduction

Recently a great deal of attention has been paid to the analysis of scattering in the time domain of electromagnetic waves from inhomogeneous dielectric inclusions in 3D waveguides. Theoretical investigation into a waveguide of rectangular cross section with layered dielectric is presented in [1].

A. Smirnov (✉) • A. Semenov
Lomonosov Moscow State University, Moscow, Russian Federation
e-mail: sap@cs.msu.ru; semenov.aleksey.msu@gmail.com

Y. Shestopalov
Karlstad University, Karlstad, Sweden
e-mail: shestop@hotmail.com

L. Beilina and Y.V. Shestopalov (eds.), *Inverse Problems and Large-Scale Computations*, Springer Proceedings in Mathematics & Statistics 52, DOI 10.1007/978-3-319-00660-4_8,

Computational modeling of electromagnetic wave propagation in waveguides requires a numerical solver for a 3D system of Maxwell equations that can handle inhomogeneous dielectric inclusions and obtain numerical solution for complex structures. Problems in an infinite waveguide require special conditions on the boundaries. In this work we use perfectly matched layer conditions with an artificial layer which absorbs scattered waves.

One of the most widespread numerical techniques for solving Maxwell equations is finite-difference time-domain (FDTD) method based upon Yee lattice, Maxwell equations in integral form and special differential relations from which finite-difference approximation is obtained.

The approach presented in this work can be applied to a wide class of waveguide problems that meet substantial difficulties as far as conventional FDTD numerical solution is concerned due to complex geometries or computational requirements. An efficient 3D FDTD *EMW-Solver3D* applied for numerical solution of nonstationary Maxwell equations employs algorithms of parallel computations and is implemented on IBM *BlueGene/P* [2].

2 Mathematical Model

We solve numerically Maxwell's equations that describe scattering from a dielectric body in waveguide:

$$\begin{cases} rot\mathbf{E} = -\dfrac{\partial \mathbf{B}}{\partial t} - \mathbf{M}_s \\ rot\mathbf{H} = \dfrac{\partial \mathbf{D}}{\partial t} + \mathbf{J}_s \\ div\mathbf{B} = 0 \\ div\mathbf{D} = 0 \\ \mathbf{B} = \mu\mu_0\mathbf{H} \\ \mathbf{D} = \varepsilon\varepsilon_0\mathbf{E}, \end{cases} \tag{1}$$

where $\varepsilon = \varepsilon(\mathbf{r})$ is a scalar function and $\mathbf{J}_s$ and $\mathbf{M}_s$ are sources. So the solution is sought for isotropic non-dispersive media.

Consider the waveguide $P = \{x : 0 < x_1 < a, 0 < x_2 < b, -\infty < x_3 < \infty\}$ of rectangular cross section presented in Fig. 1 with the perfectly conducting boundary surface ∂P given in the Cartesian coordinate system. A three-dimensional body Q ($Q \subset P$ is a domain) with a constant magnetic permeability μ_0 and variable permittivity $\varepsilon(x)$ is placed in the waveguide. Function $\varepsilon(x)$ is bounded in $\bar{Q}$, $\varepsilon \in L_\infty(Q)$, and $\varepsilon^{-1} \in L_\infty(Q)$. The boundary ∂Q of domain Q is piecewise smooth. In the provided numerical experiment electric permittivity, $\varepsilon(x)$, was set constant and equals 4.

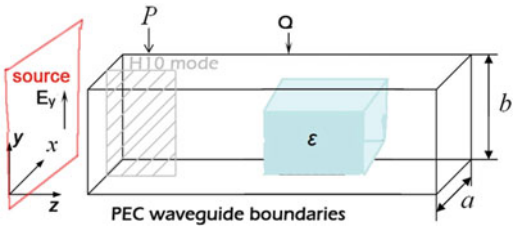

Fig. 1 Waveguide scheme

We assume that the electromagnetic field $\mathbf{E},\mathbf{H} \in L_{2,loc}(P)$ in the waveguide is excited by an external field with the time dependence $e^{-i\omega t}$; the source of the external field is the soft source plane (see Fig. 1).

3 Numerical Model

FDTD method is an explicit difference scheme that allows to evaluate the values of electromagnetic field components on the next time layer with the values on the current time layer. FDTD method is based on the use of Yee lattice [3]; the finite difference grid consists of Yee cells. Let the node of the spatial grid be

$$(i,j,k) = (i\Delta x, j\Delta y, k\Delta z), \tag{2}$$

where $\Delta x, \Delta y, \Delta z$ are steps of the uniform grid in directions x, y, z and $i = 0..M$, $j = 0\ldots,N$, $k = 0\ldots,P$.

Let the value of grid function u in the node (i,j,k) at the moment t_n be

$$u(i\Delta x, j\Delta y, k\Delta z, n\Delta t) = u^n(i,j,k), \tag{3}$$

where Δt is timestep. All the components of electromagnetic field E_x, E_y, E_z, H_x, H_y, H_z are evaluated at the different points of the cell: the components of magnetic and electric fields H and E are evaluated in the centers of faces and edges, respectively.

Assume that dielectric permittivity is continuous inside a Yee cell and magnetic permeability is constant in a cell shifted by $-1/2$ with respect to x, y, and z. The value $\varepsilon(i,j,k) = \tilde{\varepsilon}(i+\frac{1}{2}, j+\frac{1}{2}, k+\frac{1}{2})$ is known in the center of cells and $\mu(i,j,k) = \tilde{\mu}(i,j,k)$ is known in the nodes. Herewith the media interfaces for ε can be on the faces of Yee lattice and for μ in the surfaces passing through the centers of faces parallel to the coordinate axis. Therefore, ε can be discontinuous on the surfaces $x = x_i, y = y_j, z = z_k$, where $x_i = i\Delta x, y_j = j\Delta y, z_k = z\Delta z$, and μ can be discontinuous at $x = x_{i+\frac{1}{2}}, y = y_{j+\frac{1}{2}}, z = z_{k+\frac{1}{2}}$.

We consider Maxwell equations in integral form:

$$\frac{\partial}{\partial t}\int_s \mathbf{D}ds = \oint_l \mathbf{H}dl; \quad -\frac{\partial}{\partial t}\int_s \mathbf{B}ds = \oint_l \mathbf{E}dl. \tag{4}$$

At the $(i+\frac{1}{2},j,k)$ node ε is discontinuous in directions y and z. In view of the continuity of the tangential component E_x on contour l_1 and Eq. (4) and the arrangement of the field components in the Yee cell, we get

$$\frac{\partial}{\partial t}\int_{\Delta z(k-0.5)}^{\Delta z(k+0.5)}\int_{\Delta y(j-0.5)}^{\Delta y(j+0.5)} D_x\,(y,z)dydz \approx \{\ \text{under} D_x = \varepsilon E_x \text{ we get}\ \}$$

$$\approx <\varepsilon_{(i,j,k)}> \frac{\partial}{\partial t}E_x\,\Delta y\Delta z \tag{5}$$

$$\frac{\partial}{\partial t}\int_{s_1}\mathbf{D_x}ds_1 \cong (H_{y\ i+0.5,j,k+0.5} - H_{y\ i+0.5,j,k-0.5})\Delta z$$

$$-(H_{z\ i+0.5,j+0.5,k} - H_{z\ i+0.5,j-0.5,k})\Delta y, \tag{6}$$

where

$$<\tilde{\varepsilon}_(i,j,k)> = (\tilde{\varepsilon}_{(i+1/2,j-1/2,k-1/2)} + \tilde{\varepsilon}_{(i+1/2,j-1/2,k+1/2)}$$

$$+\tilde{\varepsilon}_{(i+1/2,j+1/2,k-1/2)} + \tilde{\varepsilon}_{(i+1/2,j+1/2,k+1/2)})/4. \tag{7}$$

Expressions for other components of field E and H can be evaluated in a similar way according to [4,5].

4 Numerical

We performed calculations for the waveguide presented in Fig. 1 where $a = 2\lambda$, $b = \lambda$, $\varepsilon = 3$, $c = \frac{\lambda}{4}$, $z1 = 2\lambda$, and $z2 = 4\lambda$. Parameters of the problem are $1000 \times 1000 \times 3600$ for the finite-difference grid, and 25τ is the maximum time step at which the stable state is achieved.

As we can see in Fig. 2 only the E_y component propagates in the waveguide while other components attenuate significantly. The E_y component preserves sinusoidal form within the entire length of the waveguide including the dielectric slab.

Using the discrete Fourier transform applied to the E_y in the time domain, we can evaluate the values of its amplitude. In the frequency domain let us take a look at the amplitude at the source frequency: according to Fig. 3, variations in the amplitude

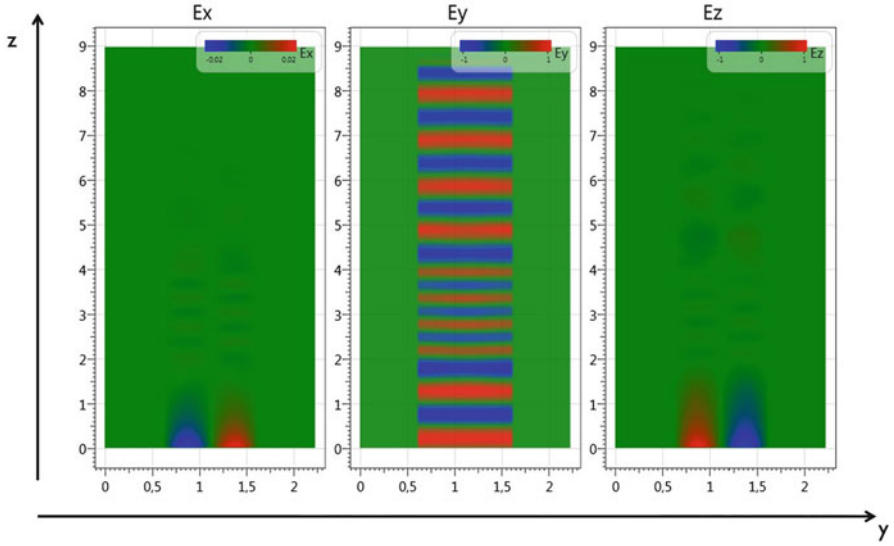

Fig. 2 The slice view of the electric field distribution at $t = 25\tau$

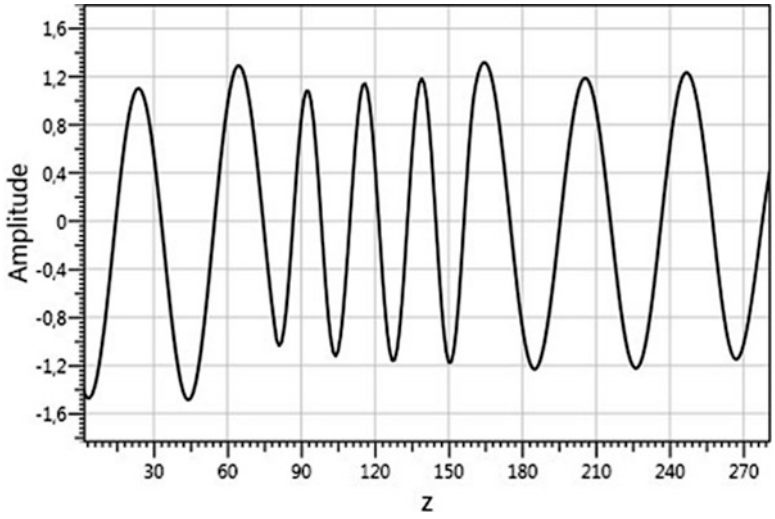

Fig. 3 E_y field amplitude

before and after the dielectric inclusion are minimal. The calculated attenuation factor $p = \mathrm{Re}\left\{\dfrac{\ln \dfrac{A_0}{A_z}}{z}\right\}$ equals 0.049 where A_0 is the source amplitude and A_z is the amplitude at $x = \lambda$, $y = \dfrac{\lambda}{2}$, and $z = 6\lambda$.

5 Conclusion

The scattering in the time domain of electromagnetic waves from inhomogeneous dielectric inclusions in a 3D waveguide of rectangular cross section is considered. The solution is calculated and its properties are investigated using FDTD in different frequency ranges. An efficient 3D FDTD *EMWSolver3D* for the nonstationary Maxwell equation system is used. The model computations are performed for the H10-mode scattering from a parallelepiped-shaped dielectric inclusion. The attenuation factor is calculated for the transmitted modes and field distributions are visualized. The results agree well with theoretical investigation [1]. The computations are performed using supercomputer IBM Bluegene /P installed at Moscow State University [2].

References

1. Smirnov, Y.G., Shestopalov, Y.V., Derevyanchuk, E.D.: Permittivity reconstruction of layered dielectrics in a rectangular waveguide from the transmission coefficients at different frequencies. Workshop on Large-Scale Modelling, Karlshtad University (2012)
2. Semenov, A.N.: Parallelnaya realizacia chislennogo resheniya uravneniy Maksvella FDTD metodom dly bolshih zadach. Nauchniy servis v seti internet, 530, Novorossiysk, Russian Federation (2011)
3. Yee, K.: Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media. IEEE Trans. Antenn. Propag. **14**, 302–307 (1966)
4. Taflove, A., Hagness, S.C.: Computational Electrodynamics: The Finite–Difference Time–Domain Method. Artech House, Norwood (2000)
5. Smirnov, A.P., Semenov, A.N.: Full wave Maxwell's equations solver EMWSolver3D. Progress in Electromagnetics Research Symposium, PIERS 2012, 252–255, Moscow (2012)

Integral Equation Methods in Optical Waveguide Theory

Alexander Frolov and Evgeny Kartchevskiy

Abstract Optical waveguides are regular dielectric rods having various cross-sectional shapes where generally the permittivity may vary in the waveguide's cross section. The permittivity of the surrounding medium may be a step-index function of coordinates. The eigenvalue problems for natural modes (surface and leaky eigenmodes) of inhomogeneous optical waveguides in the weakly guiding approximation formulated as problems for Helmholtz equations with partial radiation conditions at infinity in the cross-sectional plane. The original problems are reduced with the aid of the integral equation method (using appropriate Green functions) to nonlinear spectral problems with Fredholm integral operators. Theorems on the spectrum localization are proved. It is shown that the sets of all eigenvalues of the original problems may consist of isolated points on the Riemann surface and each eigenvalue depends continuously on the frequency and permittivity and can appear or disappear only at the boundary of the Riemann surface. The original problems for surface waves are reduced to linear eigenvalue problems for integral operators with real-valued symmetric polar kernels. The existence, localization, and dependence on parameters of the spectrum are investigated. The collocation method for numerical calculations of the natural modes is proposed, the convergence of the method is proved, and some results of numerical experiments are discussed.

1 Introduction

Optical fibers are dielectric waveguides (DWs), i.e., regular dielectric rods, having various cross-sectional shapes, and where generally the refractive index of the dielectric may vary in the waveguide's cross section [10]. Historically, the first DWs

A. Frolov • E. Kartchevskiy (✉)
Kazan (Volga Region) Federal University, Kremlevskaya 18, Kazan, Russia
e-mail: Alexander_ksu@mail.ru; ekarchev@yandex.ru

L. Beilina and Y.V. Shestopalov (eds.), *Inverse Problems and Large-Scale Computations*, Springer Proceedings in Mathematics & Statistics 52, DOI 10.1007/978-3-319-00660-4_9,

to be studied were step-index waveguides with circular cross section; interior to the waveguide, the refractive index was either homogeneous or coaxial-layered. In these cases, by using separation of variables, modal eigenvalue problems are easily reduced to families of transcendental dispersion equations associated with the azimuthal indices (see, e.g., [6, 10]). The study of the source-free electromagnetic fields, called natural modes, that can propagate on DWs necessitates that longitudinally the rod extend to infinity. Since often DWs are not shielded, the medium surrounding the waveguide transversely forms an unbounded domain, typically taken to be free space. This fact plays an extremely important role in the mathematical analysis of natural waveguide modes and brings into consideration a variety of possible formulations. They differ in the form of the condition imposed at infinity in the cross-sectional plane, and hence in the functional class of the natural-mode field. This also restricts the localization of the eigenvalues in the complex plane of the eigenparameter [3]. During recent years partial condition has been widely used for statements of various wave propagation problems [9]. All of the known natural-mode solutions (i.e., guided modes, leaky modes, and complex modes) satisfy partial condition at infinity. The wavenumbers β may be generally considered on the appropriate logarithmic Riemann surface. For real wavenumbers on the principal ("proper") sheet of this Riemann surface, one can reduce partial condition to either the Sommerfeld radiation condition or to the condition of exponential decay. Partial condition may be considered as a generalization of the Sommerfeld radiation condition and can be applied for complex wavenumbers. This condition may also be considered as the continuation of the Sommerfeld radiation condition from a part of the real axis of the complex parameter β to the appropriate logarithmic Riemann surface.

In this paper we consider the problem of determination of eigenwaves propagating along inhomogeneous optical waveguides with piecewise continuous permittivity in the weakly guiding approximation when the hybrid-mode character of the normal waves is neglected. In this approximation, the considered problem on eigenwaves is virtually equivalent to the determination of eigenoscillations of the cylindrical resonators having the same cross section as the guides under study [4].

We reduce the analysis to the boundary eigenvalue problems for Helmholtz equations with partial radiation conditions at infinity in the cross-sectional plane and latter to finding characteristic numbers of integral equations.

The methods of the spectral theory of integral operator-valued functions were applied to the study of the oscillations in slotted resonators [8] and to the study of the normal waves of slotted waveguides [7].

2 Statement of the Problem

We consider the natural modes of an inhomogeneous optical fiber. Let the three-dimensional space be occupied by an isotropic source-free medium, and let the refractive index be prescribed as a positive real-valued function $n = n(x_1, x_2)$

independent of the longitudinal coordinate x_3 and equal to a constant n_∞ outside a cylinder. The axis of the cylinder is parallel to the x_3–axis, and its cross section is a bounded domain Ω on the plane $R^2 = \{(x_1, x_2) : -\infty < x_1, x_2 < \infty\}$ with a boundary Γ belongs to the class $C^{1,\alpha}$. Denote by Ω_∞ the unbounded domain $\Omega_\infty = R^2 \backslash \overline{\Omega}$, and denote by n_+ the maximum of the function n in the domain Ω, where $n_+ > n_\infty$. Let the function n belong to the space of real-valued twice continuously differentiable in Ω functions. It is supposed that U is the space of twice continuously differentiable in Ω and Ω_∞, continuous and continuously differentiable in $\bar{\Omega}$ and $\bar{\Omega}_\infty$ real-valued functions. The modal problem for the weakly guiding optical fiber can be formulated [2] as an eigenvalue problem for a Helmholtz equation:

$$\left[\Delta + \left(k^2 n^2 - \beta^2\right)\right] u = 0, \quad x \in R^2 \backslash \Gamma. \tag{1}$$

Here $k^2 = \omega^2 \varepsilon_0 \mu_0$; ε_0, μ_0 are the free-space dielectric and magnetic constants, respectively. We consider the propagation constant β as a complex parameter and radian frequency ω as a positive parameter. We seek nonzero solutions u of equation (1) in the space U. Functions u have to satisfy the conjugation conditions:

$$u^+ = u^-, \quad \frac{\partial u^+}{\partial \nu} = \frac{\partial u^-}{\partial \nu}, \quad x \in \Gamma. \tag{2}$$

Here ν is the normal vector. We say that function u satisfies partial condition if u can be represented for all $|x| > R$ as

$$u = \sum_{l=-\infty}^{\infty} a_l H_l^{(1)}(\chi r) \exp(il\varphi), \tag{3}$$

where $H_l^{(1)}$ is the Hankel function of the first kind and index l, (r, φ) are the polar coordinates of the point x and $\chi(\beta) = \sqrt{k^2 n_\infty^2 - \beta^2}$. The series in (3) should converge uniformly and absolutely.

The Hankel functions $H_l^{(1)}(\chi(\beta) r)$ are many-valued functions of the variable β. If we want to consider these functions as holomorphic functions, it is seen that β should be considered on the set Λ, which is the Riemann surface of the function $\ln(\chi(\beta) r)$. This is due to the fact that Hankel functions can be represented as

$$H_l^{(1)}(\chi(\beta) r) = c_l^{(1)}(\chi r) \ln(\chi r) + R_l^{(1)}(\chi r), \tag{4}$$

where $c_l^{(1)}(\chi r)$ and $R_l^{(1)}(\chi r)$ are holomorphic single-valued functions [1]. The Riemann surface Λ is infinitely sheeted, with each sheet having two branch points, $\beta = \pm k n_\infty$. More precisely, due to the branching of $\chi(\beta)$ itself, we consider an infinite number of logarithmic branches Λ_m, $m = 0, \pm 1, \ldots$, each consisting of two square-root sheets of the complex variable β: $\Lambda_m^{(1)}$ and $\Lambda_m^{(2)}$. By $\Lambda_0^{(1)}$ denote the principal ("proper") sheet of Λ, which is specified by the following conditions:

$$-\pi/2 < \arg \chi(\beta) < 3\pi/2, \quad \operatorname{Im}(\chi(\beta)) \geq 0, \quad \beta \in \Lambda_0^{(1)}. \tag{5}$$

The "improper" sheet $\Lambda_0^{(2)}$ is specified by the conditions

$$-\pi/2 < \arg\chi(\beta) < 3\pi/2, \quad \mathrm{Im}\,(\chi(\beta)) < 0, \quad \beta \in \Lambda_0^{(2)}. \tag{6}$$

Denote also the whole real axis of $\Lambda_0^{(1)}$ as $R_0^{(1)}$ and that of $\Lambda_0^{(2)}$ as $R_0^{(2)}$. All of the other pairs of sheets $\Lambda_{m\neq 0}^{(1),(2)}$ differ from $\Lambda_0^{(1),(2)}$ by shift in $\arg\chi(\beta)$ equal to $2\pi m$ and satisfy the conditions

$$-\pi/2 < \arg\chi(\beta) < 3\pi/2, \quad \mathrm{Im}\,(\chi(\beta)) \geq 0, \quad \beta \in \Lambda_m^{(1)}, \tag{7}$$

$$-\pi/2 < \arg\chi(\beta) < 3\pi/2, \quad \mathrm{Im}\,(\chi(\beta)) < 0, \quad \beta \in \Lambda_m^{(2)}. \tag{8}$$

Hence on $\Lambda_0^{(1)}$ there is only a pair of branch-cuts dividing it from $\Lambda_0^{(2)}$; they run along the real axis at $|\beta| < kn_\infty$ and along the imaginary axis. On $\Lambda_0^{(2)}$, additionally, there is a pair of branch-cuts dividing it from $\Lambda_{\pm 1}^{(1)}$; they run along the real axis at $|\beta| > kn_\infty$.

Definition 1. A nonzero function $u \in U$ is referred to as an eigenfunction (generalized mode) of the problem (1)–(3) corresponding to some eigenvalues $\beta \in \Lambda$ and $\omega > 0$ if the relations of problem (1)–(3) are valid. The set of all eigenvalues of the problem (1)–(3) is called the spectrum of this problem (Fig. 1).

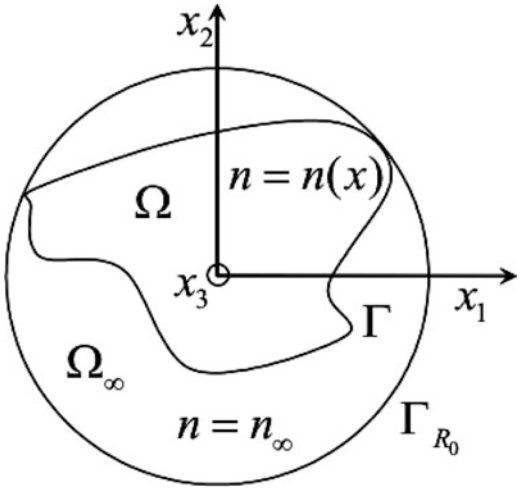

Fig. 1 Cross section of the waveguide which is in free space

Let us describe the geometry and give the problem statement for a waveguide in the half-space. Denote by Ω_∞ the unbounded domain $\Omega_\infty = \{x \in R^2 : x_1 \in R, x_2 > 0\}\backslash\overline{\Omega}$. The refractive index n_∞ of Ω_∞ is very different from the refractive index of the bottom half-space n_b. Suppose that the refractive indices of Ω and

Ω_∞ are approximately equal. So we can suggest that $u = 0$ for $x_2 = 0$ and use the approximation of weakly guiding waveguide (Fig. 2).

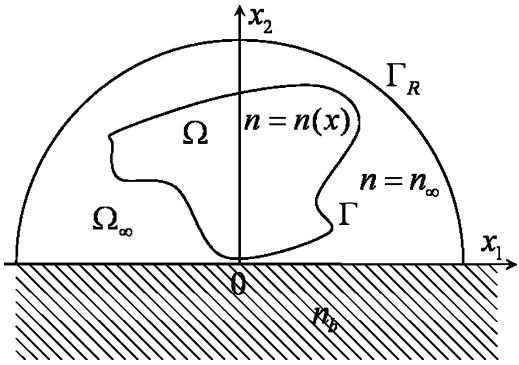

Fig. 2 Cross section of the waveguide which is in the half-space

The problem of the waveguide in the half-space yields as an equivalent Helmholtz equation

$$\left[\Delta + \left(k^2 n^2 - \beta^2\right)\right] u = 0, \quad x \in \Omega \cup \Omega_\infty. \tag{9}$$

Functions u also have to satisfy the following conjugation conditions:

$$u^+ = u^-, \quad \frac{\partial u^+}{\partial \nu} = \frac{\partial u^-}{\partial \nu}, \quad x \in \Gamma. \tag{10}$$

In this case partial condition means that the function u can be represented for all $|x| > R$ as

$$u = \sum_{l=-\infty}^{\infty} a_l H_l^{(1)}(\chi r) \sin(l\varphi). \tag{11}$$

3 Spectrum Properties

In this section the original problems will be reduced to the spectral problems for integral operators. Some results of integral operators spectrum properties will be formulated. If u is an eigenfunction of problem (1)–(3) corresponding to some eigenvalues $\beta \in \Lambda$ and $\omega > 0$, then [2] the following integral presentation is valid:

$$v(x) = \lambda \int_\Omega K_{1,2}(x,y) v(y) dy, \quad x \in \Omega, \tag{12}$$

where

$$K_{1,2}(x,y) = \Phi_{1,2}(\beta;x,y)p(x)p(y), \quad v(x) = u(x)p(x),$$
$$p(x) = \sqrt{((n(x))^2 - n_\infty^2)/(n_+^2 - n_\infty^2)}, \quad \lambda = k^2(n_+^2 - n_\infty^2),$$
$$\Phi_1 = \frac{i}{4}H_0^{(1)}(\chi(\beta)\,|x-y|).$$

An equivalent integral presentation for a waveguide in the half-space is also valid. Note that function

$$\Phi_2 = \frac{i}{4}(H_0^{(1)}(\chi(\beta)\,|x-y|) - H_0^{(1)}(\chi(\beta)\,|x-y^*|))$$

is Green's function of the problem (9)–(11). Here y^* is $(y_1, -y_2)$.

The original problem (1)–(3) is spectrally equivalent [2] to the problem (12). Let the frequency ω have a fixed positive value. Rewrite problem (12) in the form of spectral problem for operator-valued function

$$A(\beta)v = 0, \tag{13}$$

where $A(\beta) = I - \lambda T(\beta) : L_2(\Omega) \to L_2(\Omega)$, T is the operator, defined by the right side of the Eq. (12), and I is the identical operator.

Definition 2. Let $\omega > 0$ be a fixed parameter. A nonzero vector $v \in L_2(\Omega)$ is called an eigenvector of operator-valued function $A(\beta)$ corresponding to an eigenvalue $\beta \in \Lambda$ if the relation (13) is valid. The set of all $\beta \in \Lambda$ for which the operator $A(\beta)$ does not have a bounded inverse operator in $L_2(\Omega)$ is called the spectrum of operator-valued function $A(\beta)$. Denote by $\mathrm{sp}(A) \subset \Lambda$ the spectrum of operator-valued function $A(\beta)$.

Theorem 1. *The following assertions hold:*

1. *For all $\omega > 0$ and $\beta \in \Lambda$ the operator $T(\beta)$ is compact.*
2. *If ω has a fixed positive value, then the spectrum of the operator-valued function $A(\beta)$ can be only a set of isolated points on Λ, moreover on the principal sheet $\Lambda_0^{(1)}$ it can belong only to the set*

$$G = \left\{\beta \in R_0^{(1)} : kn_\infty < |\beta| < kn_+\right\}.$$

3. *Each eigenvalue β of the operator-valued function $A(\beta)$ depends continuously on $\omega > 0$ and can appear and disappear only at the boundary of Λ, i.e., at $\beta = \pm kn_\infty$ and at infinity on Λ.*

This theorem was proved in [3]. The equivalent theorem for the second problem is valid. The proof of this theorem is based on the spectral theory of operator-valued Fredholm holomorphic functions.

The well-known surface modes satisfy to $\beta \in G$. In this case $\chi(\beta) = i\sigma(\beta)$, where $\sigma(\beta) = \sqrt{\beta^2 - k^2 n_\infty^2} > 0$. Let transverse wavenumber σ have a fixed positive value. Rewrite problem (12) in the form of usual liner spectral problem with integral compact operator

$$v = \lambda T(\sigma)v, \quad T: \ L_2(\Omega) \to L_2(\Omega). \tag{14}$$

Definition 3. Let $\sigma > 0$ be a fixed parameter. A nonzero function $v \in L_2(\Omega)$ is called an eigenfunction of the operator T corresponding to a characteristic value λ if the relation (14) is valid. The set of all characteristic values of the operator T is called the spectrum and is denoted by $\mathrm{sp}(T)$.

Theorem 2. *For all positive σ the following statements are valid:*

1. *There exist the denumerable set of positive characteristic values λ_l, $l = 1, 2, \ldots$, with only cumulative point at infinity.*
2. *The set of all eigenfunctions v_l, $l = 1, 2, \ldots$, can be chosen as the orthonormal set.*
3. *The smallest characteristic value λ_1 is positive and simple, corresponding eigenfunction v_1 is positive.*
4. *Each eigenvalue λ_l, $l = 1, 2, \ldots$, depends continuously on $\sigma > 0$,*
5. *$\lambda_1 \to 0$ as $\sigma \to 0$.*

This theorem was proved in [3]. The proving of this theorem is based on the combination of three equivalent statements: original statement, statement in form of spectral problem with integral operator with symmetric weakly polar kernel and on the special variational formulation on the plane and on the half-plane. The corresponding integral operators are self-adjoint and compact; therefore (see, e.g., [5]) there exists a denumerable set of λ_l, $l = 1, 2, \ldots$, with only cumulative point at infinity. We use special variational statement and equivalence of variational and original statements for proving positiveness of these integral operators. Then we obtain that all characteristic values are positive. Moreover, the minimal value λ_1 is simple (it means that multiplicity of λ_1 is equal to one) and $\lambda_1 \to 0$ as $\sigma \to 0$.

However, the last assertion for the problem (9)–(11) has the other form. In particular, $\lambda_1 \to const > 0$ as $\sigma \to 0$. The well-known fundamental mode satisfies to the smallest characteristic value λ_1. We can conclude that the fundamental mode exists for all $\omega > 0$ in case of a waveguide in free space. The fundamental mode will appear from the certain value of ω for a waveguide in the half-space. If some values of the parameters λ and σ are known, then β and ω can be calculated by evidence formulas.

4 Collocation Method

Let us consider the collocation method [11] for numerical approximation of the integral equation (12). We suppose that $\sigma > 0$ is a fixed parameter. We cover Ω with small triangles $\Omega_{j,h}$ such that

$$\max_{1\le j\le N_h} \operatorname{diam}(\Omega_{j,h}) \le h$$

and $\Omega_{i,h} \bigcap \Omega_{j,h} = \emptyset$, if $i \neq j$. Denote by Ω_h the sub-domain $\Omega_h = \bigcup_{j=1}^{N_h} \Omega_{j,h} \subseteq \Omega$. Let $\Xi_h = \{\xi_{j,h}\}_{j=1}^{N_h}$ be a grid for Ω (a finite number of points of Ω) such that $\xi_{j,h}$ is a centroid of $\Omega_{j,h}$, $j = 1,\dots,N_h$ and

$$\operatorname{dist}(x,\Xi_h) \to 0, \quad h \to 0, \quad \forall x \in \Omega.$$

It is well known that each solution of the equation (12) belongs to a space $E = C(\overline{\Omega})$ [11] with norm

$$||v||_E = \sup_{x\in\Omega} |v(x)|.$$

Introduce the space $E_h = C(\Xi_h)$ of functions on the grid Ξ_h with the norm

$$||v_h||_{E_h} = \max_{1\le j\le N_h} |v_h(\xi_{j,h})|, \quad v_h \in E_h.$$

Define $p_h \in L(E,E_h)$ as the operator restricting functions $v \in E$ to the grid Ξ_h: $p_h v \in E_h$ is a grid function with the values

$$(p_h v)(\xi_{j,h}) = v(\xi_{j,h}), \quad j = 1,\dots,N_h.$$

Then the discrete convergence v_h to v means that

$$\max_{1\le j\le N_h} |v_h(\xi_{j,h}) - v(\xi_{j,h})| \to 0, \quad h \to 0.$$

We represent an approximate solution of the integral equation (12) as a piecewise constant function $\tilde{v}_h(x) = \sum_{j=1}^{N_h} v_{j,h} f_{j,h}(x)$, $x \in \Omega_h$, where $f_{j,h}$ are basis functions, $f_{j,h}(x) = 1$, if $x \in \Omega_{j,h}$, $f_{j,h}(x) = 0$, if $x \notin \Omega_{j,h}$. In the integral equation (12) we approximate the domain of integration Ω by Ω_h:

$$v(x) = \lambda \int_{\Omega_h} K(x,y)v(y)dy. \tag{15}$$

Replacing v by $\tilde{v}_h$ and collocating at points $\xi_{i,h}$, we obtain a system of linear algebraic equations to find the values $v_{j,h}$, namely,

$$v_{i,h} = \lambda \sum_{j=1}^{N_h} \int_{\Omega_{j,h}} K(\xi_{i,h}, y) v_{j,h} dy, \quad i = 1, \ldots N_h. \tag{16}$$

Let us introduce a discrete analogue of operator T:

$$(T_h \tilde{v}_h)(\xi_{i,h}) = \sum_{j=1}^{N_h} \int_{\Omega_{j,h}} K(\xi_{i,h}, y) \tilde{v}_h(\xi_{i,h}) dy.$$

Therefore, using collocation method for solving linear spectral problem for the integral equation (12), we obtain finite-dimensional linear spectral problem.

Let us formulate the convergence theorem for the linear case.

Theorem 3. *The following assertions hold:*

1. *If* $0 \neq \lambda_0 \in sp(T)$ *then there exists* $\lambda_h \in sp(T_h)$ *such that* $\lambda_h \to \lambda_0$ *as* $h \to 0$.
2. *Conversely, if* $sp(T_h) \ni \lambda_h \to \lambda_0$ *as* $h \to 0$ *then* $\lambda_0 \in sp(T)$.
3. *The convergence rate for a simple characteristic value is estimated as follows: for* $sp(T_h) \ni \lambda_h \to \lambda_0 \in sp(T)$, $\lambda_0 \neq 0$

$$|\lambda_h - \lambda_0| \leq ch^2.$$

The proof of this theorem is based on the discrete convergence theory [11].

Let us describe calculation of integrals in (16). Taking into account that the diagonal elements have singularities, we obtain these formulas:

$$a_{ii} = \frac{p^2(\xi_i)}{2\pi} \left(\frac{\pi R_i^2}{2} - \ln R_i |\Omega_{i,h}| - \ln \frac{\sigma\gamma}{2} |\Omega_{i,h}| \right),$$

where R_i is the minimal distance from the centroid of triangle to triangle's sides. Nondiagonal elements were calculated by following formulas

$$a_{ij} = \frac{|\Omega_{j,h}|}{2\pi} K_0(\sigma|\xi_i - \xi_j|) p(\xi_i) p(\xi_j),$$

where K_0 is McDonald's function.

The latter formulas for a waveguide in the half-space take the forms

$$a_{ii} = \frac{p^2(\xi_i)}{2\pi} \left(\frac{\pi {R_i}^2}{2} - \ln R_i |\Omega_{i,h}| - \ln \frac{\sigma\gamma}{2} |\Omega_{i,h}| - K_0(2\sigma|\xi_2^i|) |\Omega_{i,h}| \right),$$

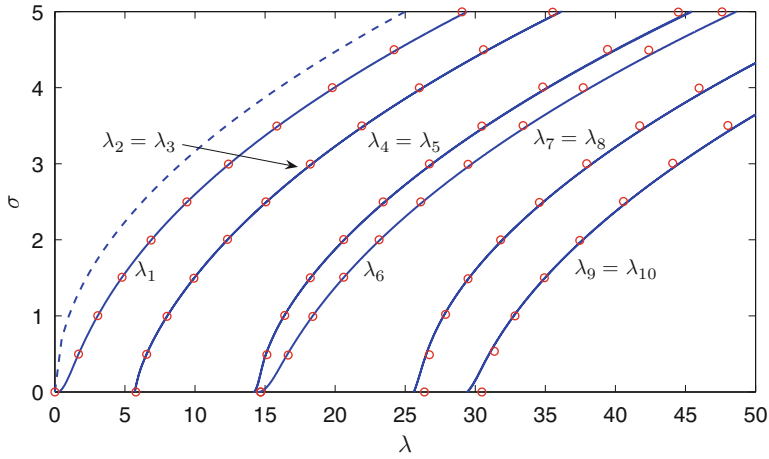

Fig. 3 The first ten dispersion curves for surface modes of circular step-index fiber calculated by the collocation method (*plotted by solid lines*) with comparison to exact solutions (*marked by circles*); $n_+ = \sqrt{2}$, $n_\infty = 1$

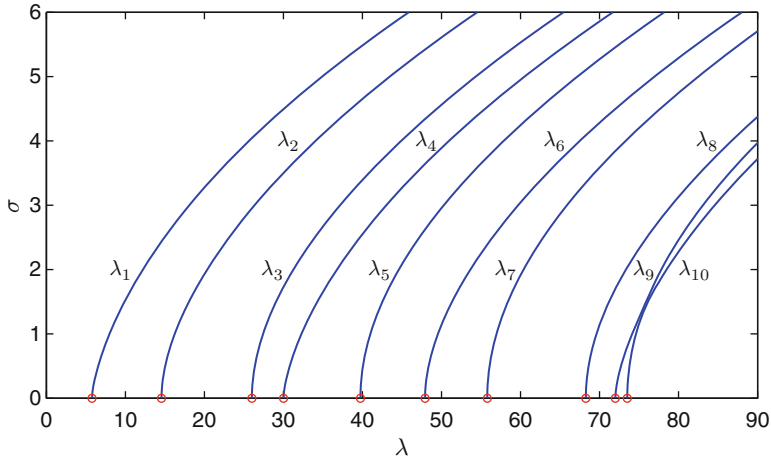

Fig. 4 The first ten dispersion curves for surface modes of semicircle step-index fiber in the half-space calculated by the collocation method; $n_+ = \sqrt{2}$, $n_\infty = 1$

$$a_{ij} = \frac{p(\xi_i)p(\xi_j)|\Omega_j|}{2\pi}\left(K_0(\sigma|\xi_i - \xi_j|) - K_0(\sigma|\xi_i - \xi_j^*|)\right).$$

Now we describe numerical results based on the collocation method. Dispersion curves show dependence for σ of λ. They are presented on the Fig. 3 for the circular waveguide in free space and on the Fig. 4 for the semicircle waveguide in the half-space, respectively. Figures 5 and 6 show the eigenfunction isolines

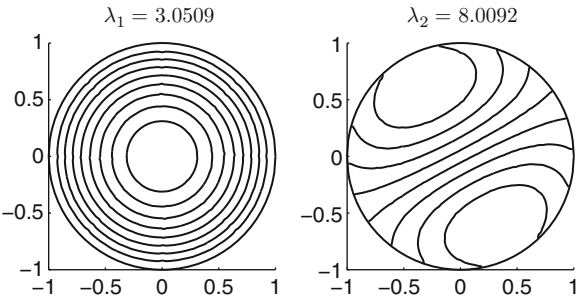

Fig. 5 Eigenfunction isolines for surface waves of circular waveguide in free space; $n_+ = \sqrt{2}$, $n_\infty = 1$

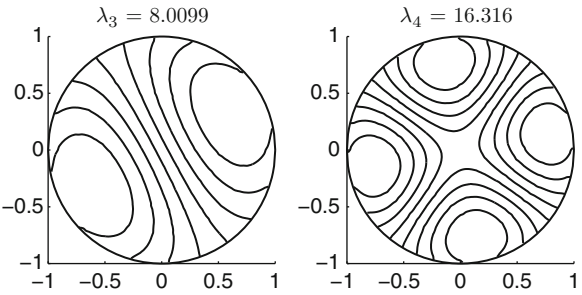

Fig. 6 Eigenfunction isolines for surface waves of circular waveguide in free space; $n_+ = \sqrt{2}$, $n_\infty = 1$

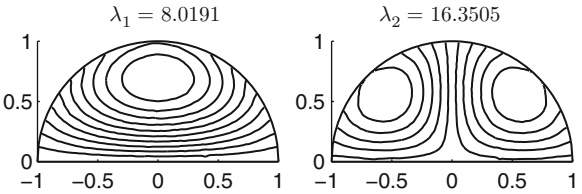

Fig. 7 Eigenfunction isolines for surface waves of semicircle waveguide in the half-space; $n_+ = \sqrt{2}$, $n_\infty = 1$

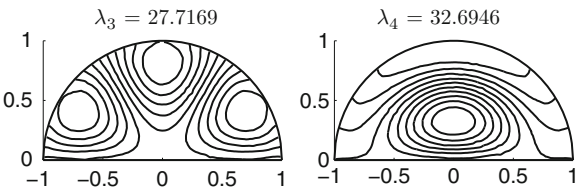

Fig. 8 Eigenfunction isolines for surface waves of semicircle waveguide in the half-space; $n_+ = \sqrt{2}$, $n_\infty = 1$

Table 1 Numerical results for eigenvalue λ_6 ($\sigma = 1$) of circular waveguide in free space; $n_+ = \sqrt{2}$, $n_\infty = 1$

N	64	256	512	1032	2304	4128	6528
h	0.4856	0.2573	0.1551	0.1217	0.0800	0.0618	0.0491
$\tilde{\lambda}_6$	16.0095	17.7529	18.1083	18.2402	18.3337	18.3842	18.4020
e	0.5576	0.5572	0.7315	0.7041	0.8384	0.6868	0.6881
ε	0.1315	0.0369	0.0176	0.0104	0.0054	0.0026	0.0017

Table 2 Numerical results for eigenvalue λ_6 ($\sigma = 1$) of semicircle waveguide in the half-space; $n_+ = \sqrt{2}$, $n_\infty = 1$

N	61	240	506	1059	2024	4236
h	0.3531	0.1693	0.1210	0.0863	0.0605	0.0432
$\tilde{\lambda}_6$	39.3336	48.0972	49.5528	50.2392	50.5952	50.7702
e	1.8172	1.8956	1.7561	1.6377	1.4209	0.9432
ε	0.2266	0.0543	0.0257	0.0122	0.0052	0.0018

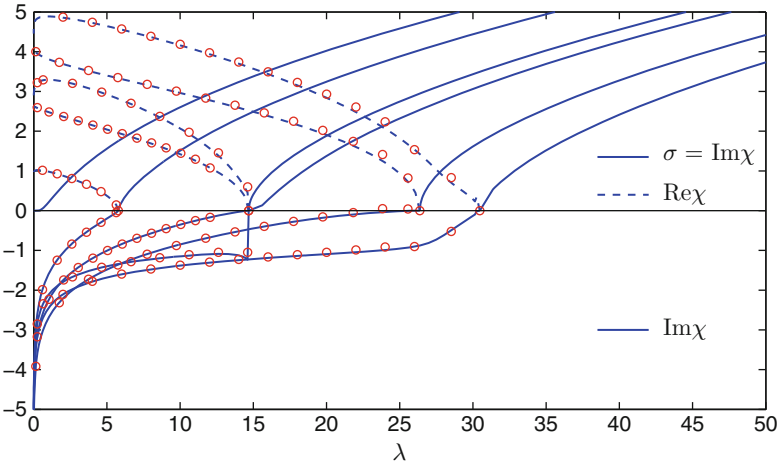

Fig. 9 Dispersion curves for surface and leaky modes of the circular step-index fiber calculated by collocation method (*marked by circles*) with comparison to exact solutions (*plotted by sold lines*); $n_+ = \sqrt{2}$, $n_\infty = 1$

for surface waves of circular waveguide in free space. Figures 7 and 8 show the eigenfunction isolines for surface waves of semicircle waveguide in the half-space. We present a Table 1 for circular waveguide that evaluates dependence for relative error $\varepsilon = |\lambda_6 - \tilde{\lambda}_6|/\lambda_6$ and $e = \varepsilon/(h/R)^2$ of N_h with $\sigma = 1$. Here $\lambda_6 = 18.4324$ is the exact value, $\tilde{\lambda}_6$ is the approximate value, R is the radius of the circular fiber.

The Table 2 describes the behaviour of inner convergence for semicircle waveguide in the half-space. We compare $\tilde{\lambda}_6$ with $\lambda_6 = 50.8596$ which is calculated for $N_h = 8096$. We also applied this method for solving nonlinear problem (13). In this

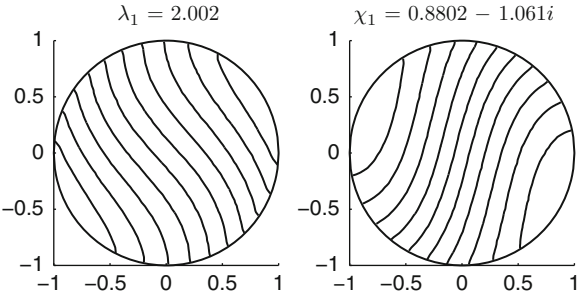

Fig. 10 Isolines for real and imaginary part of the first eigenfunction of circular waveguide; $n_+ = \sqrt{2}, n_\infty = 1$

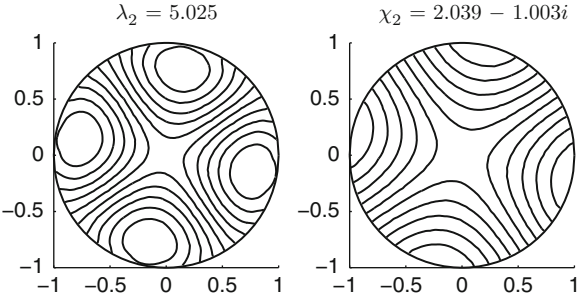

Fig. 11 Isolines for real and imaginary part of the fourth eigenfunction of circular waveguide; $n_+ = \sqrt{2}, n_\infty = 1$

Table 3 Numerical results for eigenvalue χ_3 for $\lambda_3 = 10.02$ of circular waveguide; $n_+ = \sqrt{2}, n_\infty = 1$

N	512	1032	2304	4128
h	0.1551	0.1217	0.0800	0.0618
$\tilde{\chi}_3$	5.8018-1.0489i	5.8030-1.0655i	5.8047-1.0702i	5.8050-1.0726i
e	0.1859	0.1124	0.1300	0.1099
ε	0.0045	0.0017	0.0008	0.0004

case we fixed value for parameter ω and therefore for λ and find values of β. Let us formulate the convergence theorem for the nonlinear case.

Theorem 4. *Let $A_h(\beta) = I - \lambda T_h(\beta)$. The following assertions hold:*

1. *If $\beta_0 \in sp(A)$ then there exists $\beta_h \in sp(A_h)$ such that $\beta_h \to \beta_0$ as $h \to 0$.*
2. *If $\beta_h \in sp(A_h)$ and $\beta_h \to \beta_0 \in \Lambda$ as $h \to 0$ then $\beta_0 \in sp(A)$.*

The proof of this theorem is based on the discrete convergence theory [11].

The dispersion curves for surface and leaky modes of the circular step-index fiber calculated by collocation method in comparation with exact solutions are presented at Fig. 9. Figures 10 and 11 show isolines of the first and second eigenfunctions for

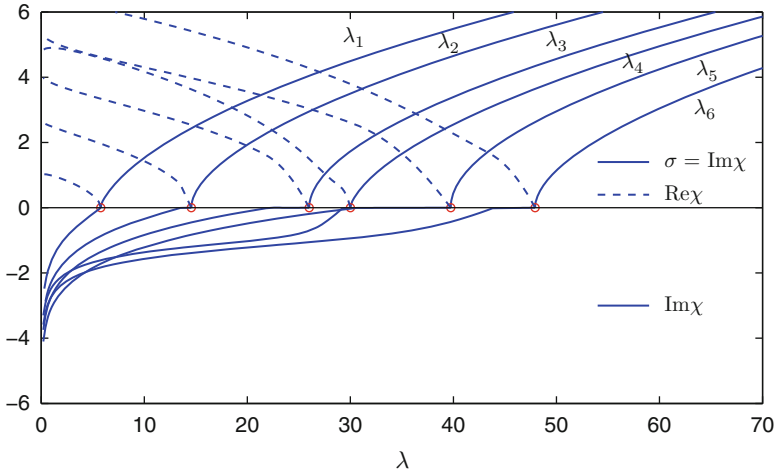

Fig. 12 Dispersion curves for surface and leaky modes of the semicircle step-index waveguide in the half-space; $n_+ = \sqrt{2}$, $n_\infty = 1$

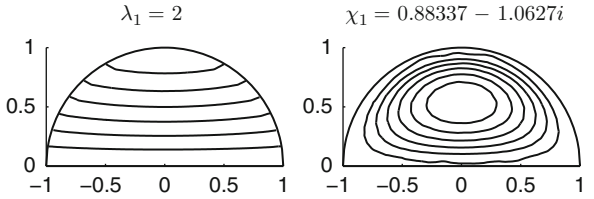

Fig. 13 Isolines for real and imaginary part of the first eigenfunction of semicircle waveguide in the half-space; $n_+ = \sqrt{2}$, $n_\infty = 1$

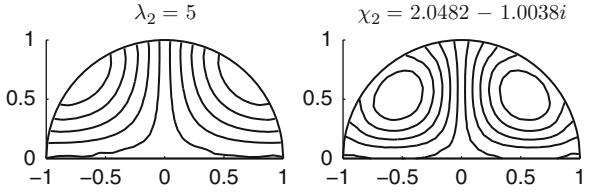

Fig. 14 Isolines for real and imaginary part of the fourth eigenfunction of semicircle waveguide in the half-space; $n_+ = \sqrt{2}$, $n_\infty = 1$

Table 4 Numerical results for eigenvalue χ_4 for $\lambda_4 = 20.02$ of semicircle waveguide; $n_+ = \sqrt{2}$, $n_\infty = 1$

N	240	506	1059	2024
h	0.1693	0.1210	0.0863	0.0605
$\tilde{\chi}_4$	2.7616−0.9311i	2.7897− 1.0195i	2.7978−1.0556i	2.8020−1.0715i
e	1.8019	1.4209	1.1408	0.8241
ε	0.0516	0.0208	0.0085	0.0030

leaky waves for circular waveguide, respectively. The Table 3 shows dependence for relative error $\varepsilon = |\chi_3 - \tilde{\chi}_3|/|\chi_3|$ and $e = \varepsilon/(h/R)^2$ of N_h with $\lambda_3 = 10.02$. Here $\chi_3 = 2.96\text{-}0.8469i$ is the exact value, $\tilde{\chi}_3$ is the approximate value.

We provided the same calculations for the semicircle waveguide in the half-space. The dispersion curves for surface and leaky modes of the semicircle step-index waveguide are presented at Fig. 12. Figures 13 and 14 show isolines of the first and second eigenfunctions for leaky waves for circular waveguide, respectively. The Table 4 shows dependence for relative error ε, value e of N_h. Here $\chi_4 = 2.8042 - 1.0803i$ is the value which is calculated for $N = 4236$ and $\lambda = 20.2$. Our numerical calculations show that the collocation method has the second rate of convergence. This is consistent with the theoretical estimates.

References

1. Abramovitz, M, Stegun, I.:Handbook of Mathematical Functions. Dover, New York (1965)
2. Karchevskii, E.M., Solovi'ev, S.I.: Investigation of a spectral problem for the Helmholtz operator on the plane. Differ equat. **36**, 631–634 (2000)
3. Kartchevski, E.M., Nosich, A.I., Hanson, G.W.: Mathematical analysis of the generalized natural modes of an inhomogeneous optical fiber. SIAM J. Appl. Math. **65**(6), 2003–2048 (2005)
4. Koshparenok, V.N., Melezhik, P.N., Poedinchuk, A.E., Shestopalov, V.P.: Spectral theory of two-dimensional open resonators with dielectric inserts. USSR Comput. Math. Math. Phys. **25**(2), 151–161 (1985)
5. Kress, R.: Linear Integral Equations. Springer, New York (1999)
6. Marcuse, D.: Theory of Dielectric Optical Waveguides. Academic Press, New York (1974)
7. Shestopalov, Yu.V., Kotik, N.Z.: Interaction and propagation of waves in slotted waveguides. New J. Phys. **4**, 40.1–40.16 (2002)
8. Shestopalov, Yu.V., Okuno, Y., Kotik, N.Z.: Oscillations in Slotted Resonators with Several Slots: Application of Approximate Semi-Inversion, vol. 39, pp. 193–247, Progress In Electromagnetics Research (PIER), Moscow (2003)
9. Shestopalov, Yu.V., Smirnov, Yu.G., and Chernokozhin, E.V.: Logarithmic Integral Equations in Electromagnetics. VSP, Leiden, The Netherlands (2000)
10. Snyder, A.W., Love, J.D.: Optical Waveguide Theory. Chapman and Hall, London (1983)
11. Vainikko, G.: Multidimensional Weakly Singular Integral Equations. Springer, Berlin (1993)

Guaranteed Estimates of Functionals from Solutions and Data of Interior Maxwell Problems Under Uncertainties

Yury Podlipenko and Yury Shestopalov

Abstract We are looking for linear with respect to observations optimal estimates of solutions and right-hand sides of Maxwell equations called minimax or guaranteed estimates. We develop constructive methods for finding these estimates and estimation errors which are expressed in terms of solutions to special variational equations and prove that Galerkin approximations of the obtained variational equations converge to their exact solutions.

1 Introduction

Problems of optimal reconstruction of solutions and right-hand sides of Maxwell equations under incomplete data are investigated. Depending on a character of an a priori information, stochastic or deterministic approach is possible. The choice is determined by nature of the parameters in the problem, which can be random or not. Moreover the optimality of estimations depends on a criterion with respect to which a given value is evaluated.

The first reference to the statement of minimax estimation problems for ordinary differential equations can be found in [1]. The approach was developed in [2–4] and then in [5–7] as applied to estimation problems for partial differential equations. Essential contribution to these studies was made in [8, 9].

In the analysis of complex processes of electromagnetic wave scattering described by Maxwell equations, an important problem is the optimal reconstruction

Y. Podlipenko (✉)
Kiev University, Kiev 01601, Ukraine
e-mail: yourip@mail.ru

Y. Shestopalov
Karlstad University, Karlstad 65188, Sweden
e-mail: youri.shestopalov@kau.se

L. Beilina and Y.V. Shestopalov (eds.), *Inverse Problems and Large-Scale Computations*, Springer Proceedings in Mathematics & Statistics 52, DOI 10.1007/978-3-319-00660-4_10,

(estimation) of parameters of the equations, like values of some functionals on their solutions or right-hand sides, from observations which depend on the same solutions. In spite of considerable amount of publications dealing with estimation for partial and ordinary differential equations, there is an important class of Maxwell problems for which estimation problems remain unsolved. The present work is aimed to develop estimation techniques for interior Maxwell problems under uncertain data.

We assume that right-hand sides of Maxwell equations are unknown and belong to the given bounded subsets of the space of all square integrable functions in the considered domain and for solving the estimation problems we must have supplementary data (observations) depending on solutions of these equations. We suppose that observation errors (noises) are realizations of the stochastic processes, with unknown moment functions of the second order also belonging to certain given subsets.

Our approach is as follows. Let D be a domain bounded by a perfect conductor and occupied by a "dissipative" dielectric with unknown electric current density J belonging to a certain bounded set of vector-functions square integrable in D. The considered problem of estimation consists in the following: by observations of electric and magnetic fields $\mathbf{E}$ and $\mathbf{H}$ to estimate linear continuous functionals from $\mathbf{E}$, $\mathbf{H}$, and $\mathbf{J}$ under the assumption that observations are liner transformations of $\mathbf{E}$ and $\mathbf{H}$ perturbed by additive random noises with unknown second moments and belonging to a certain given set in the corresponding Hilbert space. We determine linear with respect to observations optimal estimates of solutions and right-hand sides of Maxwell equations from the condition of minimum of maximal mean square error of estimation taken over the above subsets. We develop the methods for obtaining such estimates, which is expressed in terms of solutions of special variational equations.

Beyond purely theoretical interest, the mentioned problems can find applications in automatized measurement data processing systems and for interpretation of electromagnetic observations.

In our previous studies [10, 11] we considered the estimation problems for Helmholtz problems with incomplete data using a similar approach.

It should be noted that problems of guaranteed estimation for other types of partial differential equations were investigated in [12–15].

2 Preliminaries and Auxiliary Results

Let us introduce the notations and definitions that will be used in this work.

We denote matrices and vectors by bold letters; $x = (x_1, x_2, x_3)$ denotes a spatial variable in an open domain $D \subset \mathbb{R}^3$ with Lipschitzian boundary Γ; $\chi(M)$ is a characteristic function of the set $M \subset \mathbb{R}^3$.

Introduce a Hilbert space $H(\mathrm{rot}, D) := \{\mathbf{v} \in L^2(D)^3 : \mathrm{rot}\,\mathbf{v} \in L^2(D)^3\}$, with inner product $(\cdot,\cdot)_{H(\mathrm{rot},D)}$ and corresponding norm $\|\cdot\|_{H(\mathrm{rot}D)}$ defined by

$$(\mathbf{u},\mathbf{v})_{H(\mathrm{rot},D)} = (\mathrm{rot}\,\mathbf{u},\mathrm{rot}\,\mathbf{v})_{L^2(D)^3} + (\mathbf{u},\mathbf{v})_{L^2(D)^3}, \quad \|\mathbf{u}\|^2_{H(\mathrm{rot},D)} = (\mathbf{u},\mathbf{u})_{H(\mathrm{rot},D)}.$$

For any function $\mathbf{u} \in H(\mathrm{rot},D)$ it is possible to define a tangential trace $\mathbf{n}\times\mathbf{v}$ on Γ as an element of $H^{-1/2}(\Gamma)^3$; all the functions from $H(\mathrm{rot},D)$ with zero tangential traces form its subspace

$$H_0(\mathrm{rot},D) := \{\mathbf{v} \in H(\mathrm{rot},D) : \mathbf{n}\times\mathbf{v}|_\Gamma = 0\} = \{\overline{(\mathscr{D}(D))^3}^{H(\mathrm{rot},D)}\},$$

where $\mathbf{n}$ is the unit normal to Γ and $\mathscr{D}(D)$ is the set of infinitely differentiable functions on D having compact support (see [16]).

Let H be a Hilbert space over the set of complex numbers $\mathbb{C}$ with the inner product $(\cdot,\cdot)_H$ and norm $\|\cdot\|_H$. By $L^2(\Omega,H)$ we denote the Bochner space composed of random[1] variables $\xi = \xi(\omega)$ defined on a certain probability space $(\Omega,\mathscr{B},P)$ with values in H such that

$$\|\xi\|^2_{L^2(\Omega,H)} = \int_\Omega \|\xi(\omega)\|^2_H dP(\omega) < \infty. \tag{1}$$

In this case there exists the Bochner integral $\mathbb{E}\xi := \int_\Omega \xi(\omega)\,dP(\omega) \in H$ called the mathematical expectation or the mean value of random variable $\xi(\omega)$ which satisfies the condition

$$(h,\mathbb{E}\xi)_H = \int_\Omega (h,\xi(\omega))_H\,dP(\omega) \quad \forall h \in H. \tag{2}$$

Being applied to random variable ξ with values in $\mathbb{C}$ or $\mathbb{R}$, this expression leads to a usual definition of its mathematical expectation because the Bochner integral reduces to a Lebesgue integral with probability measure $dP(\omega)$.

In $L^2(\Omega,H)$ one can introduce the inner product

$$(\xi,\eta)_{L^2(\Omega,H)} := \int_\Omega (\xi(\omega),\eta(\omega))_H\,dP(\omega) \quad \forall \xi,\eta \in L^2(\Omega,H). \tag{3}$$

Applying the sign of mathematical expectation, one can write relationships (1)–(3) as

$$\|\xi\|^2_{L^2(\Omega,H)} = \mathbb{E}\|\xi(\omega)\|^2_H, \tag{4}$$

$$(h,\mathbb{E}\xi)_H = \mathbb{E}(h,\xi(\omega))_H \quad \forall h \in H, \tag{5}$$

$$(\xi,\eta)_{L^2(\Omega,H)} := \mathbb{E}(\xi(\omega),\eta(\omega))_H \quad \forall \xi,\eta \in L^2(\Omega,H). \tag{6}$$

$L^2(\Omega,H)$ equipped with norm (4) and inner product (6) is a Hilbert space.

[1]Random variable ξ with values in Hilbert space H is considered as a function $\xi : \Omega \to H$ mapping random events $E \in \mathscr{B}$ to Borel sets in H (Borel σ-algebra in H is generated by open sets in H).

If H_0 is a Hilbert space over $\mathbb{C}$ with inner product $(\cdot,\cdot)_{H_0}$ and norm $\|\cdot\|_{H_0}$, then $\Lambda_{H_0} \in \mathscr{L}(H_0, H_0')$ denotes the Riesz operator acting from H_0 to its adjoint H_0' and determined by the equality[2] $(v,u)_{H_0} = < v, \Lambda_{H_0} u >_{H_0 \times H_0'}$ $\forall u, v \in H_0$, where $< x, f >_{H_0 \times H_0'} := f(x)$ for $x \in H_0$, $f \in H_0'$.

3 Statement of the Problem

Let D be a Lipschitzian domain bounded by a perfect conductor and occupied by a "dissipative" dielectric with permittivity $\varepsilon = \varepsilon(x)$ and permeability $\mu = \mu(x)$ which are bounded and measurable functions in D, $\operatorname{Im}\varepsilon, \operatorname{Im}\mu \geq \alpha = \text{const} > 0$, and electric and magnetic current densities $\mathbf{J}(x)$ and $\mathbf{M}(x)$, $\mathbf{J}, \mathbf{M} \in L^2(D)^3$ (time dependence is $e^{-i\omega t}$).

Electromagnetic field $(\mathbf{E}, \mathbf{H})$ created by these currents, which is of finite energy in D, i.e.,

$$\mathbf{E}, \mathbf{H} \in H(\operatorname{rot}, D),$$

satisfies the interior Maxwell problem:

$$-\operatorname{rot}\mathbf{E} + i\omega\mu\mathbf{H} = \mathbf{M}, \quad \operatorname{rot}\mathbf{H} + i\omega\varepsilon\mathbf{E} = \mathbf{J} \quad \text{in} \quad D, \tag{7}$$

$$\mathbf{n} \times \mathbf{E}|_\Gamma = 0, \tag{8}$$

where $\mathbf{n}$ is a unit outward normal to the boundary $\Gamma = \partial D$ and $\times$ denotes the vector product.

Further, we will use the variational statement equivalent to interior Maxwell problem (7)–(8). Introduce the following sesquilinear form in $H_0(\operatorname{rot}, D)$:

$$a(\mathbf{E}, \mathbf{E}') = \int_D \left(\left(-\frac{1}{i\omega\mu} \operatorname{rot}\mathbf{E}, \operatorname{rot}\mathbf{E}' \right)_{\mathbb{C}^3} - (i\omega\varepsilon\mathbf{E}, \mathbf{E}')_{\mathbb{C}^3} \right) dx \quad \forall \mathbf{E}, \mathbf{E}' \in H_0(\operatorname{rot}, D). \tag{9}$$

Then problem (7)–(8) is equivalent to finding $\mathbf{E} \in H_0(\operatorname{rot}, D)$ from the variational equation

$$a(\mathbf{E}, \mathbf{E}') = \int_D \left(\left(\frac{1}{i\omega\mu} \mathbf{M}, \operatorname{rot}\mathbf{E}' \right)_{\mathbb{C}^3} - (\mathbf{J}, \mathbf{E}')_{\mathbb{C}^3} \right) dx \quad \forall \mathbf{E}' \in H_0(\operatorname{rot}, D) \tag{10}$$

and, in line with the Lax–Milgram lemma, has a unique solution $(\mathbf{E}, \mathbf{H})$, with $\mathbf{E} \in H_0(\operatorname{rot}, D)$, $\mathbf{H} \in H(\operatorname{rot}, D)$ for any $\mathbf{J} \in L^2(D)^3$ and $\mathbf{M} \in L^2(D)^3$ since, according to our assumptions on ε and μ, sesquilinear form (9) is coercive in $H_0(\operatorname{rot}, D)$ and the right-hand side of (10) is an antilinear continuous form in $H_0(\operatorname{rot}, D)$ (see [16]).

We suppose that $\mathbf{M} = 0$ and function $\mathbf{J}(x)$ is not known exactly. The estimation problem consists in the following: from the observations

[2] This operator exists according to the Riesz theorem.

$$y_1 = C_1\mathbf{E} + \eta_1, \quad y_2 = C_2\mathbf{H} + \eta_2 \tag{11}$$

find optimal in a certain sense estimate of the functional[3]

$$l(\mathbf{E},\mathbf{H}) = \int_D (\mathbf{E}(x), \mathbf{l}_1(x))_{\mathbb{C}^3} dx + \int_D (\mathbf{H}(x), \mathbf{l}_2(x))_{\mathbb{C}^3} dx \tag{12}$$

in the class of estimates linear w.r.t. observations (11),

$$\widehat{l(\mathbf{E},\mathbf{H})} = (y_1, u_1)_{H_0} + (y_2, u_2)_{H_0} + c \tag{13}$$

under the assumption that errors $\eta_1 = \eta_1(\omega)$ and $\eta_2 = \eta_2(\omega)$ in observations (11) are realizations of random variables defined on a certain probability space $(\Omega, \mathscr{B}, P)$ with values in a Hilbert space H_0 over $\mathbb{C}$, belong to the set G_1, and $\mathbf{J} \in G_0$. By G_1 we denote the set of random variables $\tilde{\eta}_1$ and $\tilde{\eta}_2 \in L^2(\Omega, H_0)$ with zero means satisfying the inequalities

$$\mathbb{E}(Q_1\tilde{\eta}_1, \tilde{\eta}_1)_{H_0} \le \varepsilon_1, \quad \mathbb{E}(Q_2\tilde{\eta}_2, \tilde{\eta}_2)_{H_0} \le \varepsilon_2, \tag{14}$$

and

$$G_0 = \left\{ \tilde{\mathbf{J}} \in L^2(D)^3 : \left(Q(\tilde{\mathbf{J}} - \mathbf{J}_0), \tilde{\mathbf{J}} - \mathbf{J}_0\right)_{L^2(D)^3} \le \varepsilon_3 \right\}, \tag{15}$$

where $\varepsilon_k > 0$, $k = 1,2,3$, are given constants; $u_1, u_2 \in H_0$; $c \in \mathbb{C}$; $(\cdot,\cdot)_{H_0}$ is inner product in H_0; $\mathbf{l}_1, \mathbf{l}_2, \mathbf{J}_0 \in L^2(D)^3$ are given complex-valued functions; C_1 and $C_2 \in \mathscr{L}(L^2(D)^3, H_0)$ are linear continuous operators; and Q, Q_1, and Q_2 are Hermitian positive definite operators in $L^2(D)^3$ for which there exist bounded inverse operators Q^{-1}, Q_1^{-1}, and Q_2^{-1}. Further, without loss of generality we may set $\varepsilon_k = 1$, $k = 1,2,3$. We also assume that random variables $\tilde{\eta}_1$ and $\tilde{\eta}_2$ are uncorrelated,

$$\mathbb{E}(\tilde{\eta}_1, u_1)_{H_0} \overline{(\tilde{\eta}_2, u_2)_{H_0}} = 0, \quad \forall u_1, u_2 \in H_0. \tag{16}$$

Definition 1. An estimate

$$\widehat{\widehat{l(\mathbf{E},\mathbf{H})}} = (y_1, \hat{u}_1)_{H_0} + (y_2, \hat{u}_2)_{H_0} + \hat{c} \tag{17}$$

is called a minimax estimate of $l(\mathbf{E},\mathbf{H})$ if elements $\hat{u}_1, \hat{u}_2 \in H_0$ and a number $\hat{c}$ are determined from the condition

$$\inf_{u_1,u_2 \in H_0, c \in \mathbb{C}} \sigma(u_1, u_2, c) = \sigma(\hat{u}_1, \hat{u}_2, \hat{c}),$$

[3]For vectors $\mathbf{V}^{(1)} = \left(V_1^{(1)}, V_2^{(1)}, V_3^{(1)}\right)$, $\mathbf{V}^{(2)} = \left(V_1^{(2)}, V_2^{(2)}, V_3^{(2)}\right) \in \mathbb{C}^3$ we set $\left(\mathbf{V}^{(1)}, \mathbf{V}^{(2)}\right)_{\mathbb{C}^3} = \sum_{i=1}^3 V_i^{(2)} \bar{V}_i^{(2)}$.

where

$$\sigma(u_1,u_2,c) := \sup_{\tilde{\mathbf{J}}\in G_0,(\tilde{\eta}_1,\tilde{\eta}_2)\in G_1} \mathbb{E}|l(\tilde{\mathbf{E}},\tilde{\mathbf{H}}) - \widehat{l(\tilde{\mathbf{E}},\tilde{\mathbf{H}})}|^2,$$

$(\tilde{\mathbf{E}},\tilde{\mathbf{H}})$ is a solution to the problem (7)–(8) when $\mathbf{J}(x)=\tilde{\mathbf{J}}(x)$,

$$\widehat{l(\tilde{\mathbf{E}},\tilde{\mathbf{H}})} = (\tilde{y}_1,u_1)_{H_0} + (\tilde{y}_2,u_2)_{H_0} + c, \tag{18}$$

and

$$\tilde{y}_1 = C_1\tilde{\mathbf{E}} + \tilde{\eta}_1, \quad \tilde{y}_2 = C_2\tilde{\mathbf{H}} + \tilde{\eta}_2. \tag{19}$$

The quantity

$$\sigma = [\sigma(\hat{u}_1,\hat{u}_2,\hat{c})]^{1/2} \tag{20}$$

is called the error of the minimax estimation of $l(\mathbf{E},\mathbf{H})$.

Thus, the minimax estimate is an estimate minimizing the maximal mean square estimation error calculated for the "worst" implementation of perturbations.

4 Representation of Guaranteed Estimates of Functionals from Solutions of Interior Maxwell Problems

Introduce a sesquilinear form $a^*(\mathbf{E},\mathbf{E}')$ in $H_0(\mathrm{rot},D)$ by

$$a^*(\mathbf{E},\mathbf{E}') = \int_D \left(\left(\frac{1}{i\omega\bar{\mu}}\mathrm{rot}\,\mathbf{E}, \mathrm{rot}\,\mathbf{E}' \right)_{\mathbb{C}^3} + (i\omega\bar{\varepsilon}\mathbf{E},\mathbf{E}')_{\mathbb{C}^3} \right) dx \quad \forall \mathbf{E},\mathbf{E}' \in H_0(\mathrm{rot},D). \tag{21}$$

Then this form is adjoint of $a(\mathbf{E},\mathbf{E}')$, that is,

$$a^*(\mathbf{E},\mathbf{E}') = \overline{a(\mathbf{E}',\mathbf{E})} \quad \forall \mathbf{E},\mathbf{E}' \in H_0(\mathrm{rot},D). \tag{22}$$

Let the function $\mathbf{Z}(x;u) \in H_0(\mathrm{rot},D)$ be a unique solution of the problem[4]

$$a^*(\mathbf{Z}(\cdot;u),\mathbf{E}')) = \int_D \left(-\left(\frac{1}{i\omega\bar{\mu}}(\mathbf{l}_2 - C_2^*\Lambda_{H_0}u_2), \mathrm{rot}\,\mathbf{E}' \right)_{\mathbb{C}^3} + (\mathbf{l}_1(x) - C_1^*\Lambda_{H_0}u_1(x),\mathbf{E}')_{\mathbb{C}^3} \right) dx \quad \forall \mathbf{E}' \in H_0(\mathrm{rot},D), \tag{23}$$

[4]This problem is uniquely solvable since, owing to (22), the sesquilinear form $a^*(\cdot,\cdot)$ is also coercive in $H_0(\mathrm{rot},D)$.

where $u = (u_1, u_2) \in H_0 \times H_0$, C_i^* is adjoint of C_i, determined by

$$< C_i \mathbf{v}, w >_{H_0 \times H_0'} = \int_D (\mathbf{v}(x), C_i^* w(x))_{\mathbb{C}^3}\, dx$$

for all $\mathbf{v} \in L^2(D)^3$, $w \in H_0'$, $i = 1, 2$.

Then the following result holds.

Lemma 1. *The problem of minimax estimation of the functional (12) (i.e., the determination of $\hat{u} = (\hat{u}_1, \hat{u}_2)$ and $\hat{c}$) is equivalent to the problem of optimal control of the system described by equation (23) with a cost function*

$$I(u) = \int_D (Q^{-1}\mathbf{Z}(x;u), \mathbf{Z}(x;u))_{\mathbb{C}^3}\, dx + (Q_1^{-1}u_1, u_1)_{H_0} + (Q_2^{-1}u_2, u_2)_{H_0} \to \inf_{u \in H_0 \times H_0}.$$

Proof. From relation (12) at $\mathbf{E} = \tilde{\mathbf{E}}$ and $\mathbf{H} = \tilde{\mathbf{H}}$ and (18) and (19), we have

$$\begin{aligned}
l(\tilde{\mathbf{E}}, \tilde{\mathbf{H}}) - \widehat{l(\tilde{\mathbf{E}}, \tilde{\mathbf{H}})} &= \int_D (\tilde{\mathbf{E}}(x), \mathbf{l}_1(x))_{\mathbb{C}^3} dx + \int_D (\tilde{\mathbf{H}}(x), \mathbf{l}_2(x))_{\mathbb{C}^n} dx \\
&\quad - (C_1\tilde{\mathbf{E}}, u_1)_{H_0} - (C_2\tilde{\mathbf{H}}, u_2)_{H_0} - (\tilde{\eta}_1, u_1)_{H_0} - (\tilde{\eta}_2, u_2)_{H_0} - c \\
&= \int_D (\tilde{\mathbf{E}}(x), \mathbf{l}_1(x))_{\mathbb{C}^3} dx + \int_D \left(\frac{1}{i\omega\mu} \operatorname{rot} \tilde{\mathbf{E}}, \mathbf{l}_2(x) \right)_{\mathbb{C}^n} dx \\
&\quad - (C_1\tilde{\mathbf{E}}, u_1)_{H_0} - \left(C_2 \frac{1}{i\omega\mu} \operatorname{rot} \tilde{\mathbf{E}}, u_2 \right)_{H_0} - (\tilde{\eta}_1, u_1)_{H_0} - (\tilde{\eta}_2, u_2)_{H_0} - c \\
&= \int_D (\tilde{\mathbf{E}}(x), \mathbf{l}_1(x))_{\mathbb{C}^3} dx + \int_D \left(\frac{1}{i\omega\mu} \operatorname{rot} \tilde{\mathbf{E}}, \mathbf{l}_2(x) \right)_{\mathbb{C}^n} dx \\
&\quad - < C_1\tilde{\mathbf{E}}, \Lambda_{H_0} u_1 >_{H_0 \times H_0'} - < C_2 \frac{1}{i\omega\mu} \operatorname{rot} \tilde{\mathbf{E}}, \Lambda_{H_0} u_2 >_{H_0 \times H_0'} \\
&\quad - (\tilde{\eta}_1, u_1)_{H_0} - (\tilde{\eta}_2, u_2)_{H_0} - c \\
&= \int_D (\tilde{\mathbf{E}}(x), \mathbf{l}_1(x))_{\mathbb{C}^3} dx + \int_D \left(\frac{1}{i\omega\mu} \operatorname{rot} \tilde{\mathbf{E}}, \mathbf{l}_2(x) \right)_{\mathbb{C}^n} dx \\
&\quad - \int_D (\tilde{\mathbf{E}}(x), C_1^* \Lambda_{H_0} u_1(x))_{\mathbb{C}^3} dx - \int_D \left(\frac{1}{i\omega\mu} \operatorname{rot} \tilde{\mathbf{E}}(x), C_2^* \Lambda_{H_0} u_2 \right)_{\mathbb{C}^3} dx \\
&\quad - (\tilde{\eta}_1, u_1)_{H_0} - (\tilde{\eta}_2, u_2)_{H_0} - c \\
&= \int_D (\tilde{\mathbf{E}}(x), \mathbf{l}_1(x) - C_1^* \Lambda_{H_0} u_1(x))_{\mathbb{C}^3} dx \\
&\quad + \int_D \left(\operatorname{rot} \tilde{\mathbf{E}}, -\frac{1}{i\omega\bar{\mu}} (\mathbf{l}_2(x) - C_2^* \Lambda_{H_0} u_2(x)) \right)_{\mathbb{C}^n} dx \\
&\quad - (\tilde{\eta}_1, u_1)_{H_0} - (\tilde{\eta}_2, u_2)_{H_0} - c. \qquad (24)
\end{aligned}$$

Set $\mathbf{E}' = \tilde{\mathbf{E}}$ in (22) and $\mathbf{E} = \tilde{\mathbf{E}}$, $\mathbf{E}' = \mathbf{Z}(\cdot;u)$ in (10), respectively. Then we have

$$a^*(\mathbf{Z}(\cdot;u),\tilde{\mathbf{E}}) = \int_D \left(-\left(\frac{1}{i\omega\bar{\mu}} (\mathbf{l}_2 - C_2^* \Lambda_{H_0} u_2), \operatorname{rot} \tilde{\mathbf{E}} \right)_{\mathbb{C}^3} + (\mathbf{l}_1(x) - C_1^* \Lambda_{H_0} u_1(x), \tilde{\mathbf{E}})_{\mathbb{C}^3} \right) dx \quad (25)$$

and

$$a(\tilde{\mathbf{E}}, \mathbf{Z}(\cdot;u)) = -\int_D (\mathbf{J}, \mathbf{Z}(\cdot;u))_{\mathbb{C}^3}\, dx. \quad (26)$$

Since

$$\overline{a^*(\mathbf{Z}(\cdot;u),\tilde{\mathbf{E}})} = a(\tilde{\mathbf{E}}, \mathbf{Z}(\cdot;u)), \quad (27)$$

from (24)–(27), we obtain

$$\begin{aligned} l(\tilde{\mathbf{E}},\tilde{\mathbf{H}}) - \widehat{l(\tilde{\mathbf{E}},\tilde{\mathbf{H}})} &= -\int_D (\tilde{\mathbf{J}}(x), \mathbf{Z}(x;u))_{\mathbb{C}^3}\, dx - (\tilde{\eta}_1, u_1)_{H_0} - (\tilde{\eta}_2, u_2)_{H_0} - c \\ &= -\int_D (\tilde{\mathbf{J}}(x) - \mathbf{J}_0(x), \mathbf{Z}(x;u))_{\mathbb{C}^3}\, dx - (\tilde{\eta}_1, u_1)_{H_0} - (\tilde{\eta}_2, u_2)_{H_0} \\ &\quad - \int_D (\mathbf{J}_0(x), \mathbf{Z}(x;u))_{\mathbb{C}^3}\, dx - c. \end{aligned}$$

Taking into consideration (5) and the relationship $\mathbb{D}\xi = \mathbb{E}|\xi - \mathbb{E}\xi|^2 = \mathbb{E}|\xi|^2 - |\mathbb{E}\xi|^2$ that couples dispersion $\mathbb{D}\xi$ of the complex random variable $\xi = \xi_1 + i\xi_2$ and its expectation $\mathbb{E}\xi = \mathbb{E}\xi_1 + i\mathbb{E}\xi_2$, we obtain from the last formulas

$$\mathbb{E}\left| l(\tilde{\mathbf{E}},\tilde{\mathbf{H}}) - \widehat{l(\tilde{\mathbf{E}},\tilde{\mathbf{H}})} \right|^2 = \left| -\int_D (\tilde{\mathbf{J}}(x), \mathbf{Z}(x;u))_{\mathbb{C}^3}\, dx - c \right|^2 + \mathbb{E}\left| (\tilde{\eta}_1, u_1)_{H_0} + (\tilde{\eta}_2, u_2)_{H_0} \right|^2.$$

Therefore,

$$\inf_{c\in\mathbb{C}} \sup_{\tilde{\mathbf{J}}\in G_0, (\tilde{\eta}_1,\tilde{\eta}_2)\in G_1} \mathbb{E}|l(\tilde{\mathbf{E}},\tilde{\mathbf{H}}) - \widehat{l(\tilde{\mathbf{E}},\tilde{\mathbf{H}})}|^2$$

$$= \inf_{c\in\mathbb{C}} \sup_{\tilde{\mathbf{J}}\in G_0} \left| \int_D (\tilde{\mathbf{J}}(x), \mathbf{Z}(x;u))_{\mathbb{C}^3}\, dx + c \right|^2 + \sup_{(\tilde{\eta}_1,\tilde{\eta}_2)\in G_1} \mathbb{E}\left| (\tilde{\eta}_1, u_1)_{H_0} + (\tilde{\eta}_2, u_2)_{H_0} \right|^2. \quad (28)$$

In order to calculate the first term on the right-hand side of (28), make use of the generalized Cauchy–Bunyakovsky inequality and (15). We have

$$\inf_{c\in\mathbb{C}}\sup_{\tilde{\mathbf{J}}\in G_0}\left|\int_D(\tilde{\mathbf{J}}(x),\mathbf{Z}(x;u))_{\mathbb{C}^3}\,dx+c\right|^2$$

$$=\inf_{c\in\mathbb{C}}\sup_{\tilde{\mathbf{J}}\in G_0}\left|\int_D(\tilde{\mathbf{J}}(x)-\mathbf{J}_0(x),\mathbf{Z}(x;u))_{\mathbb{C}^3}dx+\int_D(\mathbf{J}_0(x),\mathbf{Z}(x;u))_{\mathbb{C}^3}dx+c\right|^2$$

$$\le\int_D\left(\mathcal{Q}^{-1}\mathbf{Z}(x;u),\mathbf{Z}(x;u)\right)_{\mathbb{C}^3}dx\int_D\left(\mathcal{Q}(\tilde{\mathbf{J}}-\mathbf{J}_0),\tilde{\mathbf{J}}-\mathbf{J}_0\right)_{\mathbb{C}^3}dx$$

$$\le\int_D\left(\mathcal{Q}^{-1}\mathbf{Z}(x;u),\mathbf{Z}(x;u)\right)_{\mathbb{C}^3}dx. \tag{29}$$

The direct substitution shows that inequality (29) is transformed to an equality on the element

$$\tilde{\mathbf{J}}^{(0)}(x):=\frac{1}{d}\mathcal{Q}^{-1}\mathbf{Z}(x;u)+\mathbf{J}_0(x),$$

where

$$d=\left(\int_D\left(\mathcal{Q}^{-1}\mathbf{Z}(x;u),\mathbf{Z}(x;u)\right)_{\mathbb{C}^3}dx\right)^{1/2}.$$

Therefore,

$$\inf_{c\in\mathbb{C}}\sup_{\tilde{\mathbf{J}}\in G_0}\left|\int_D(\tilde{\mathbf{J}}(x),\mathbf{Z}(x;u))_{\mathbb{C}^3}\,dx+c\right|^2=\int_D\left(\mathcal{Q}^{-1}\mathbf{Z}(x;u),\mathbf{Z}(x;u)\right)_{\mathbb{C}^3}dx \tag{30}$$

with

$$c=-\int_D(\mathbf{J}_0(x),\mathbf{Z}(x;u))_{\mathbb{C}^3}\,dx.$$

In order to calculate the second term on the right-hand side of (28), note that the Cauchy–Bunyakovsky inequality yields

$$\left|(\tilde{\eta}_1,u_1)_{H_0}+(\tilde{\eta}_2,u_2)_{H_0}\right|^2=$$

$$=|(\tilde{\eta}_1,u_1)_{H_0}|^2+(\tilde{\eta}_2,u_2)_{H_0}\overline{(\tilde{\eta}_2,u_2)_{H_0}}$$

$$+(\tilde{\eta}_1,u_1)_{H_0}\overline{(\tilde{\eta}_2,u_2)_{H_0}}+|(\tilde{\eta}_2,u_1)_{H_0}|^2$$

$$\le(Q_1\tilde{\eta}_1,\tilde{\eta}_1)_{H_0}(Q_1^{-1}u_1,u_1)_{H_0}+(\tilde{\eta}_2,u_2)_{H_0}\overline{(\tilde{\eta}_2,u_2)_{H_0}}+$$

$$+(\tilde{\eta}_1,u_1)_{H_0}\overline{(\tilde{\eta}_2,u_2)_{H_0}}+(Q_2\tilde{\eta}_2,\tilde{\eta}_2)_{H_0}(Q_2^{-1}u_2,u_2)_{H_0}.$$

Taking into account (14) and the fact that random variables η_1 and η_1 are uncorrelated (see (16)), we obtain from the latter formula

$$\sup_{(\tilde{\eta}_1,\tilde{\eta}_2)\in G_1} \mathbb{E}\left|(\tilde{\eta}_1,u_1)_{H_0}+(\tilde{\eta}_2,u_2)_{H_0}\right|^2 \le (Q_1^{-1}u_1,u_1)_{H_0}+(Q_2^{-1}u_2,u_2)_{H_0}. \quad (31)$$

It is easy to see that (31) becomes an equality at

$$\tilde{\eta}_1^{(0)}=\frac{\nu_1 Q_1^{-1}u_1}{\{(Q_1^{-1}u_1,u_1)_{H_0}\}^{1/2}}, \quad \tilde{\eta}_2^{(0)}=\frac{\nu_2 Q_2^{-1}u_2}{\{(Q_2^{-1}u_2,u_2)_{H_0}\}^{1/2}},$$

where ν_1, ν_2 are uncorrelated random variables with $\mathbb{E}\nu_1=\mathbb{E}\nu_2=0$ and $\mathbb{E}|\nu_1|^2=\mathbb{E}|\nu_2|^2=1$. Therefore,

$$\sup_{(\tilde{\eta}_1,\tilde{\eta}_2)\in G_1} \mathbb{E}\left|(\tilde{\eta}_1,u_1)_{H_0}+(\tilde{\eta}_2,u_2)_{H_0}\right|^2 = (Q_1^{-1}u_1,u_1)_{H_0}+(Q_2^{-1}u_2,u_2)_{H_0}, \quad (32)$$

which proves the required assertion. The validity of Lemma 2 follows now from relationships (28), (30), and (32). □

Theorem 1. *The minimax estimate of $L(\mathbf{E},\mathbf{H})$ has the form*

$$\widehat{\widehat{l(\mathbf{E},\mathbf{H})}}=(y_1,\hat{u}_1)_{H_0}+(y_2,\hat{u}_2)_{H_0}+\hat{c}, \quad (33)$$

where

$$\hat{c}=-\int_D(\mathbf{J}_0(x),\hat{\mathbf{Z}}(x))_{\mathbb{C}^3}\,dx, \quad \hat{u}_1=Q_1C_1\mathbf{P}, \quad \hat{u}_2=Q_2C_2\left(\frac{1}{i\omega\mu}\mathrm{rot}\,\mathbf{P}\right), \quad (34)$$

and the functions $\hat{\mathbf{Z}}$ and $\mathbf{P}\in H_0(\mathrm{rot},D)$ are determined as a solution of the following uniquely solvable problem:

$$a^*(\hat{\mathbf{Z}},\mathbf{E}')=\int_D\left(\left(-\frac{1}{i\omega\bar{\mu}}\left(\mathbf{l}_2-C_2^*\Lambda_{H_0}Q_2C_2\left(\frac{1}{i\omega\mu}\mathrm{rot}\,\mathbf{P}\right)\right),\mathrm{rot}\,\mathbf{E}'\right)_{\mathbb{C}^3}\right.$$
$$\left.+(\mathbf{l}_1(x)-C_1^*\Lambda_{H_0}Q_1C_1\mathbf{P},\mathbf{E}')_{\mathbb{C}^3}\right)dx \quad \forall\mathbf{E}'\in H_0(\mathrm{rot},D), \quad (35)$$

$$a(\mathbf{P},\mathbf{E}')=\int_D(Q^{-1}\hat{\mathbf{Z}},\mathbf{E}')_{\mathbb{C}^3}\,dx \quad \forall\mathbf{E}'\in H_0(\mathrm{rot},D). \quad (36)$$

The error of estimation σ is given by an expression

$$\sigma=\left(\int_D\left(\left(\frac{1}{i\omega\mu}\mathrm{rot}\,\mathbf{P}(x),\mathbf{l}_2(x)\right)_{\mathbb{C}^3}+(\mathbf{P}(x),\mathbf{l}_1(x))_{\mathbb{C}^3}\right)dx\right)^{1/2}. \quad (37)$$

Proof. Taking into account the coercivity of the sesquilinear form $a^*(\cdot,\cdot)$ in $H_0(\mathrm{rot},D)$, one can easily verify that $I(\mathbf{u})$ is a strictly convex lower semicontinuous functional on H. Also

$$I(u) \geq (Q_1^{-1}u_1,u_1)_{H_0} + (Q_2^{-1}u_2,u_2)_{H_0} \geq c\|u\|^2_{H_0\times H_0} \quad \forall u \in H_0\times H_0,\ \text{c=const.}$$

Then, by Remark 1.2 to Theorem 1.1 (see [17]), there exists one and only one element $\hat{u} = (\hat{u}_1,\hat{u}_2) \in H_0\times H_0$ such that $I(\hat{u}) = \inf_{u\in H_0\times H_0} I(u)$.

Therefore, for any $\tau \in \mathbb{R}$ and $v = (v_1,v_2) \in H_0\times H_0$, the following relations are valid:

$$\frac{d}{d\tau}I(\hat{u}+\tau v)\Big|_{\tau=0} = 0 \quad \text{and} \quad \frac{d}{d\tau}I(\hat{u}+i\tau v)\Big|_{\tau=0} = 0, \tag{38}$$

where $i=\sqrt{-1}$. Since $\mathbf{Z}(x;\hat{u}+\tau v) = \mathbf{Z}(x;\hat{u}) + \tau\tilde{\mathbf{Z}}(x;v)$, where $\tilde{\mathbf{Z}}(x;v)$ is the unique solution to (23) at $u=v$ and $\mathbf{l}_1=\mathbf{l}_2=0$, the first relation in (38) yields

$$\begin{aligned}
0 &= \frac{1}{2}\frac{d}{d\tau}I(\hat{u}+\tau v)|_{\tau=0} = \\
&= \lim_{\tau\to 0}\frac{1}{2\tau}\Big\{\Big[(Q^{-1}\mathbf{Z}(\cdot;\hat{u}+\tau v),\mathbf{Z}(\cdot;\hat{u}+\tau v))_{L^2(D)^3} - (Q^{-1}\mathbf{Z}(\cdot;\hat{u}),\mathbf{Z}(\cdot;\hat{u}))_{L^2(D)^3}\Big] \\
&\quad + \Big[(Q_1^{-1}(\hat{u}_1+\tau v_1),\hat{u}_1+\tau v_1)_{H_0} - (Q_1^{-1}\hat{u}_1,\hat{u}_1)_{H_0}\Big] \\
&\quad + \Big[(Q_2^{-1}(\hat{u}_2+\tau v_2),\hat{u}_2+\tau v_2)_{H_0} - (Q_2^{-1}\hat{u}_2,\hat{u}_2)_{H_0}\Big]\Big\} \\
&= \mathrm{Re}\Big\{(Q^{-1}\mathbf{Z}(\cdot;\hat{u}),\tilde{\mathbf{Z}}(\cdot;v))_{L^2(D)^3} + (Q_1^{-1}\hat{u}_1,v_1)_{H_0} + (Q_2^{-1}\hat{u}_2,v_2)_{H_0}\Big\}.
\end{aligned}$$

Similarly, taking into account that $Z(x;\hat{u}+i\tau v) = Z(x;\hat{u}) + i\tau\tilde{Z}(x;v)$, we find from the second relation in (38)

$$\begin{aligned}
0 &= \frac{1}{2}\frac{d}{d\tau}I(\hat{u}+i\tau v)|_{\tau=0} \\
&= \mathrm{Im}\Big\{(Q^{-1}\mathbf{Z}(\cdot;\hat{u}),\tilde{\mathbf{Z}}(\cdot;v))_{L^2(D)^3} + (Q_1^{-1}\hat{u}_1,v_1)_{H_0} + (Q_2^{-1}\hat{u}_2,v_2)_{H_0}\Big\};
\end{aligned}$$

consequently,

$$\int_D (Q^{-1}\mathbf{Z}(x;\hat{u}),\tilde{\mathbf{Z}}(x;v)_{\mathbb{C}^3})dx + (Q_1^{-1}\hat{u}_1,v_1)_{H_0} + (Q_2^{-1}\hat{u}_2,v_2)_{H_0} = 0. \tag{39}$$

Introduce a function $\mathbf{P} \in H_0(\mathrm{rot},D)$ as the unique solution of the problem

$$a(\mathbf{P},\mathbf{E}') = \int_D (Q^{-1}\mathbf{Z}(x,u),\mathbf{E}')_{\mathbb{C}^3}\,dx \quad \forall \mathbf{E}' \in H_0(\mathrm{rot},D) \tag{40}$$

Setting in (40) $\mathbf{E}' = \tilde{\mathbf{Z}}(\cdot;v)$, we obtain

$$a(\mathbf{P},\tilde{\mathbf{Z}}(\cdot;v)) = \int_D (Q^{-1}\mathbf{Z}(x,\hat{u}),\tilde{\mathbf{Z}}(x;v))_{\mathbb{C}^3}\,dx. \tag{41}$$

Taking into account the fact that $\tilde{\mathbf{Z}}(\cdot;v)$ satisfies the variational equation

$$a^*(\tilde{\mathbf{Z}}(\cdot;v),\mathbf{E}') = \int_D \left(\left(\tfrac{1}{i\omega\bar{\mu}} C_2^* \Lambda_{H_0} v_2(x), \operatorname{rot}\mathbf{E}'(x)\right)_{\mathbb{C}^3} - (C_1^*\Lambda_{H_0}v_1(x),\mathbf{E}')_{\mathbb{C}^3}\right) dx$$
$$\forall \mathbf{E}'(x) \in H_0(\operatorname{rot},D) \tag{42}$$

and putting in (42) $\mathbf{E}' = \mathbf{P}$, we have

$$a^*(\tilde{\mathbf{Z}}(\cdot;v),\mathbf{P}) = \int_D \left(\left(\frac{1}{i\omega\bar{\mu}} C_2^* \Lambda_{H_0} v_2(x), \operatorname{rot}\mathbf{P}(x)\right)_{\mathbb{C}^3} - (C_1^*\Lambda_{H_0}v_1(x),\mathbf{P}(x))_{\mathbb{C}^3}\right) dx$$

Since $\overline{a^*(\tilde{\mathbf{Z}}(\cdot;v),\mathbf{P})} = a(\mathbf{P},\tilde{\mathbf{Z}}(\cdot;v))$, we obtain from (41) and the latter equality

$$\begin{aligned}
&\int_D (Q^{-1}\ \mathbf{Z}\ (x;\hat{u}),\tilde{\mathbf{Z}}(x;v))_{\mathbb{C}^3}dx \\
&\quad = \int_D \left(\left(-\frac{1}{i\omega\mu}\operatorname{rot}\mathbf{P}(x), C_2^*\Lambda_{H_0}v_2(x)\right)_{\mathbb{C}^3} - (\mathbf{P}(x), C_1^*\Lambda_{H_0}v_1(x))_{\mathbb{C}^3}\right) dx \\
&\quad = < C_2\left(-\frac{1}{i\omega\mu}\operatorname{rot}\mathbf{P}\right), \Lambda_{H_0}v_2 >_{H_0\times H_0'} - < C_1\mathbf{P}, \Lambda_{H_0}v_1 >_{H_0\times H_0'} \\
&\quad = -\left(C_2\left(\frac{1}{i\omega\mu}\operatorname{rot}\mathbf{P}\right), v_2\right)_{H_0} - (C_1\mathbf{P},v_1)_{H_0}.
\end{aligned} \tag{43}$$

Relations (43) and (39) imply

$$\left(C_2\left(\frac{1}{i\omega\mu}\operatorname{rot}\mathbf{P}\right), v_2\right)_{H_0} + (C_1\mathbf{P},v_1)_{H_0} = (Q_1^{-1}\hat{u}_1, v_1)_{H_0} + (Q_2^{-1}\hat{u}_2, v_2)_{H_0};$$

hence,

$$\hat{u}_1 = Q_1C_1\mathbf{P}, \quad \hat{u}_2 = Q_2C_2\left(\frac{1}{i\omega\mu}\operatorname{rot}\mathbf{P}\right). \tag{44}$$

Substituting these expressions into (23) and (40) and denoting $\hat{\mathbf{Z}} := \mathbf{Z}(\cdot;\hat{u})$, we establish that functions $\hat{\mathbf{Z}}$ and $\mathbf{P}$ satisfy (35) and (36) and the validity of equalities (33) and (34); the unique solvability of the problem (35)–(36) follows from the existence of the unique minimum point $\hat{u}$ of functional $I(u)$.

Now let us establish the validity of formula (37). From (20) at $u = \hat{u}$ and (44), it follows

$$\sigma^2 = I(\hat{u}) = \int_D (Q^{-1}\hat{\mathbf{Z}}(x), \hat{\mathbf{Z}}(x))_{\mathbb{C}^3} dx + (Q_1^{-1}\hat{u}_1, \hat{u}_1)_{H_0} + (Q_2^{-1}\hat{u}_2, \hat{u}_2)_{H_0}$$
$$= \int_D (Q^{-1}\hat{\mathbf{Z}}(x), \hat{\mathbf{Z}}(x))_{\mathbb{C}^3} dx + (C_1\mathbf{P}, Q_1C_1\mathbf{P})_{H_0}$$
$$+ \left(C_2\left(\frac{1}{i\omega\mu}\operatorname{rot}\mathbf{P}\right), Q_2C_2\left(\frac{1}{i\omega\mu}\operatorname{rot}\mathbf{P}\right)\right)_{H_0}$$

Transform the first term. Make use of equality (36) to obtain

$$a(\mathbf{P}, \hat{\mathbf{Z}}) = \int_D (Q^{-1}\hat{\mathbf{Z}}(x), \hat{\mathbf{Z}}(x))_{\mathbb{C}^3}\, dx; \tag{45}$$

hence,

$$\sigma^2 = a(\mathbf{P}, \hat{\mathbf{Z}}) + (C_1\mathbf{P}, Q_1C_1\mathbf{P})_{H_0} + \left(C_2\left(\frac{1}{i\omega\mu}\operatorname{rot}\mathbf{P}\right), Q_2C_2\left(\frac{1}{i\omega\mu}\operatorname{rot}\mathbf{P}\right)\right)_{H_0}. \tag{46}$$

Setting in (23) $u = \hat{u}$ and $E' = P$, we find

$$a^*(\hat{\mathbf{Z}}, \mathbf{P}) = \int_D \left(\left(-\frac{1}{i\omega\bar{\mu}}(\mathbf{l}_2(x) - C_2^*\Lambda_{H_0}\hat{u}_2(x)\right), \operatorname{rot}\mathbf{P}(x))_{\mathbb{C}^3}\right.$$
$$\left. + (\mathbf{l}_1(x) - C_1^*\Lambda_{H_0}\hat{u}_1(x), \mathbf{P}(x))_{\mathbb{C}^3}\right) dx. \tag{47}$$

From the latter relations, the formula $\overline{a^*(\hat{\mathbf{Z}}, \mathbf{P})} = a(\mathbf{P}, \hat{\mathbf{Z}})$, and (46), it follows that

$$\sigma^2 = \int_D \left(\left(\frac{1}{i\omega\mu}\operatorname{rot}\mathbf{P}(x), \mathbf{l}_2(x) - C_2^*\Lambda_{H_0}\hat{u}_2(x)\right)_{\mathbb{C}^3} + (\mathbf{P}(x), \mathbf{l}_1(x) - C_1^*\Lambda_{H_0}\hat{u}_1(x))_{\mathbb{C}^3}\right) dx$$
$$+ (C_1\mathbf{P}, Q_1C_1\mathbf{P})_{H_0} + \left(C_2\left(\frac{1}{i\omega\mu}\operatorname{rot}\mathbf{P}\right), Q_2C_2\left(\frac{1}{i\omega\mu}\operatorname{rot}\mathbf{P}\right)\right)_{H_0}$$
$$= \int_D \left(\frac{1}{i\omega\mu}\operatorname{rot}\mathbf{P}(x), \mathbf{l}_2(x))_{\mathbb{C}^3} + (\mathbf{P}(x), \mathbf{l}_1(x))_{\mathbb{C}^3}\right) dx.$$

The theorem is proved. □

Obtain now another representation for the minimax mean square estimate of quantity $l(\mathbf{E}, \mathbf{H})$ which is independent of $\mathbf{l}_1$ and $\mathbf{l}_2$. To this end, introduce vector-functions $\hat{\mathbf{P}}, \hat{\mathbf{E}} \in H_0(\operatorname{rot}, D)$ as solution to the problem

$$a^*(\hat{\mathbf{P}},\mathbf{E}') = \int_D \left(-\frac{1}{i\omega\bar{\mu}} \left(C_2^* \Lambda_{H_0} Q_2 \left(y_2 - C_2 \left(\frac{1}{i\omega\mu} \operatorname{rot} \hat{\mathbf{E}} \right) \right), \operatorname{rot} \mathbf{E}' \right)_{\mathbb{C}^3} \right.$$
$$\left. + (C_1^* \Lambda_{H_0} Q_1 (y_1 - C_1 \hat{\mathbf{E}}), \mathbf{E}')_{\mathbb{C}^3} \right) dx \quad \forall \mathbf{E}' \in H_0(\operatorname{rot}, D), \tag{48}$$

$$a(\hat{\mathbf{E}},\mathbf{E}') = \int_D (Q^{-1}\hat{\mathbf{P}} - \mathbf{J}_0, \mathbf{E}')_{\mathbb{C}^3} \, dx \quad \forall \mathbf{E}' \in H_0(\operatorname{rot}, D) \tag{49}$$

at realizations y_1 and y_2 that belong with probability 1 to space H_0.

Note that unique solvability of the problem (48)–(49) at every realization can be proved similarly to the case of problem (35)–(36). Namely, setting $\mathbf{d}_1 = C_1^* \Lambda_{H_0} Q_1 y_1$ and $\mathbf{d}_2 = C_2^* \Lambda_{H_0} Q_2 y_2$, one can show that solutions to the problem of optimal control of the system

$$a^*(\hat{\mathbf{P}}(\cdot;v),\mathbf{E}') = \int_D \left(-\left(\frac{1}{i\omega\bar{\mu}} (\mathbf{d}_2 - C_2^* \Lambda_{H_0} v_2), \operatorname{rot} \mathbf{E}' \right)_{\mathbb{C}^3} \right.$$
$$\left. + (\mathbf{d}_1(x) - C_1^* \Lambda_{H_0} v_1(x), \mathbf{E}')_{\mathbb{C}^3} \right) dx \quad \forall \mathbf{E}' \in H_0(\operatorname{rot}, D),$$

with the cost function

$$\tilde{I}(v) = \int_D (Q^{-1}(\hat{\mathbf{P}}(x;v) - Q\mathbf{J}_0(x)), \hat{\mathbf{P}}(x;v) - Q\mathbf{J}_0(x))_{\mathbb{C}^3} dx$$
$$+ (Q_1^{-1} v_1, v_1)_{H_0} + (Q_2^{-1} v_2, v_2)_{H_0} \to \inf_{v \in H_0 \times H_0},$$

can be reduced to the solution of problem (48)–(49) where the optimal control $\hat{v} = (\hat{v}_1, \hat{v}_2)$ is expressed via solution to this problem as $v_1 = Q_1 C_1 \hat{\mathbf{E}}$, $\hat{v}_2 = Q_2 C_2 \left(\frac{1}{i\omega\mu} \operatorname{rot} \hat{\mathbf{E}} \right)$.

Theorem 2. *The minimax estimate* $\widehat{\widehat{l(\mathbf{E},\mathbf{H})}}$ *of functional* $l(\mathbf{E},\mathbf{H})$ *has the form*

$$\widehat{\widehat{l(\mathbf{E},\mathbf{H})}} = l(\hat{\mathbf{E}},\hat{\mathbf{H}}), \tag{50}$$

where $\hat{\mathbf{H}} = \frac{1}{i\omega\mu} \operatorname{rot} \hat{\mathbf{E}}$, *and function* $\hat{\mathbf{E}} \in H_0(\operatorname{rot}, D)$ *is determined from the solution to problem (48)–(49).*

The random fields $\hat{\mathbf{P}}(x,t)$ *and* $\hat{\mathbf{E}}(x,t)$, *whose realizations satisfy problem (48)–(49), belong to the space* $L^2(\Omega, H_0(\operatorname{rot}, D))$.

Proof. By virtue of (34), (48), and (49),

$$\begin{aligned}
\overbrace{l(\mathbf{E},\mathbf{H})} &= (y_1,\hat{u}_1)_{H_0} + (y_2,\hat{u}_2)_{H_0} + \hat{c} \\
&= (y_1, Q_1C_1\mathbf{P})_{H_0} + \left(y_2, Q_2C_2\left(\frac{1}{i\omega\mu}\mathrm{rot}\mathbf{P}\right)\right)_{H_0} - \int_D (\mathbf{J}_0,\hat{\mathbf{Z}})_{\mathbb{C}^3}\,dx \\
&= \int_D \left((C_1^*\Lambda_{H_0}Q_1y_1,\mathbf{P})_{\mathbb{C}^3} + \left(-\frac{1}{i\omega\bar{\mu}}C_2^*\Lambda_{H_0}Q_2y_2,\mathrm{rot}\mathbf{P}\right)_{\mathbb{C}^3}\right)dx \\
&\quad - \int_D (\mathbf{J}_0,\hat{\mathbf{Z}})_{\mathbb{C}^3}\,dx = a^*(\hat{\mathbf{P}},\mathbf{P}) + \int_D \Bigg((C_1^*\Lambda_{H_0}Q_1C_1\hat{\mathbf{E}},\mathbf{P})_{\mathbb{C}^3} \\
&\quad + \left(C_2^*\Lambda_{H_0}Q_2C_2\left(\frac{1}{i\omega\mu}\mathrm{rot}\hat{\mathbf{E}}\right),\frac{1}{i\omega\mu}\mathrm{rot}\mathbf{P}\right)_{\mathbb{C}^3}\Bigg)dx - \int_D (\mathbf{J}_0,\hat{\mathbf{Z}}))_{\mathbb{C}^3}\,dx \\
&= \overline{a(\mathbf{P},\hat{\mathbf{P}})} + (C_1\hat{\mathbf{E}},Q_1C_1\mathbf{P})_{H_0} + \left(C_2\left(\frac{1}{i\omega\mu}\mathrm{rot}\hat{\mathbf{E}}\right),Q_2C_2\left(\frac{1}{i\omega\mu}\mathrm{rot}\mathbf{P}\right)\right)_{H_0} \\
&\quad - \int_D (\mathbf{J}_0,\hat{\mathbf{Z}})_{\mathbb{C}^3}\,dx. \qquad (51)
\end{aligned}$$

But from (36) and (49), it follows

$$\overline{a(\mathbf{P},\hat{\mathbf{P}})} = \int_D \overline{(Q^{-1}\hat{\mathbf{Z}},\hat{\mathbf{P}})_{\mathbb{C}^3}}\,dx = \int_D \overline{(\hat{\mathbf{Z}},Q^{-1}\hat{\mathbf{P}})_{\mathbb{C}^3}}\,dx = \int_D (Q^{-1}\hat{\mathbf{P}},\hat{\mathbf{Z}})_{\mathbb{C}^3}\,dx,$$

$$a(\hat{\mathbf{E}},\hat{\mathbf{Z}}) = \int_D (Q^{-1}\hat{\mathbf{P}},\hat{\mathbf{Z}})_{\mathbb{C}^3}\,dx - \int_D (\mathbf{J}_0,\hat{\mathbf{Z}})_{\mathbb{C}^3}\,dx,$$

$$\begin{aligned}
a^*(\hat{\mathbf{Z}},\hat{\mathbf{E}}) = \int_D \Bigg(&-\frac{1}{i\omega\bar{\mu}}\left(\mathbf{l}_2 - C_2^*\Lambda_{H_0}Q_2C_2\left(\frac{1}{i\omega\mu}\mathrm{rot}\mathbf{P}\right),\mathrm{rot}\hat{\mathbf{E}}\right)_{\mathbb{C}^3} \\
&+ (\mathbf{l}_1 - C_1^*\Lambda_{H_0}Q_1C_1\mathbf{P},\hat{\mathbf{E}}_{\mathbb{C}^3})\Bigg)\,dx,
\end{aligned}$$

and from $\overline{a^*(\hat{\mathbf{Z}},\hat{\mathbf{E}})} = a(\hat{\mathbf{E}},\hat{\mathbf{Z}})$, we obtain

$$\begin{aligned}
\overline{a(\mathbf{P},\hat{\mathbf{P}})} &= a(\hat{\mathbf{E}},\hat{\mathbf{Z}}) + \int_D (\mathbf{J}_0,\hat{\mathbf{Z}})_{\mathbb{C}^3}\,dx = \overline{a^*(\hat{\mathbf{Z}},\hat{\mathbf{E}})} + \int_D (\mathbf{J}_0,\hat{\mathbf{Z}})_{\mathbb{C}^3}\,dx \\
&= \int_D \left(\left(\frac{1}{i\omega\mu}\mathrm{rot}\hat{\mathbf{E}},\mathbf{l}_2\right)_{\mathbb{C}^3}dx + \int_D (\hat{\mathbf{E}},\mathbf{l}_1)_{\mathbb{C}^3}dx + \int_D (\mathbf{J}_0,\hat{\mathbf{Z}})_{\mathbb{C}^3}\,dx\right. \qquad (52) \\
&\quad \left. - \int_D \left(\frac{1}{i\omega\mu}\mathrm{rot}\hat{\mathbf{E}},C_2^*\Lambda_{H_0}Q_2C_2\left(\frac{1}{i\omega\mu}\mathrm{rot}\mathbf{P}\right)\right)_{\mathbb{C}^3}dx - \int_D (\hat{\mathbf{E}},C_1^*\Lambda_{H_0}Q_1C_1\mathbf{P})_{\mathbb{C}^3}\,dx\right).
\end{aligned}$$

Representation (50) follows from (51) and (52).

Considering system (48) and (49) at realizations y_1 and y_2, it is easy to see that its solution is continuous with respect to the right-hand side. This property enables us to conclude, using the general theory of linear continuous transformations of random processes (see [18]), that functions $\hat{\mathbf{P}}, \hat{\mathbf{E}} \in L^2(\Omega, H_0(\mathrm{rot}, D))$. The theorem is proved. □

Remark 1. Notice that in representation $l(\hat{\mathbf{E}}, \hat{\mathbf{H}})$ for minimax estimate $\widehat{\widehat{l(\mathbf{E}, \mathbf{H})}}$, the functions $\hat{\mathbf{E}}, \hat{\mathbf{H}}$ which are defined from equations (48) and (49) do not depend on specific form of functional l and hence can be taken as a good estimate for unknown solution $\mathbf{E}, \mathbf{H}$ of interior Maxwell problem (7)–(8).

5 Numerical Aspects

Using the Galerkin method for solving the aforementioned equations, we obtain approximate estimates via solutions of linear algebraic equations and show their convergence to the optimal estimates.

Introduce a sequence of finite-dimensional subspaces V^h in $H_0(\mathrm{rot}, D)$, defined by an infinite set of parameters $h_1, h_2, \ldots$ with $\lim_{k\to 0} h_k = 0$.

We say that sequence $\{V^h\}$ is complete in $H_0(\mathrm{rot}, D)$, if for any $\mathbf{E} \in H_0(\mathrm{rot}, D)$ and $\varepsilon > 0$ there exists an $\hat{h} = \hat{h}(\mathbf{E}, \varepsilon) > 0$ such that $\inf_{\mathbf{w}\in V^h} \|\mathbf{E} - \mathbf{w}\|_{H(\mathrm{rot},D)} < \varepsilon$ for any $h < \hat{h}$. In other words, the completeness of sequence $\{V^h\}$ means that any element $\mathbf{E} \in H_0(\mathrm{rot}, D)$ may be approximated with any degree of accuracy by elements of $\{V^h\}$.

Take an approximate minimax estimate of $l(\mathbf{E}, \mathbf{H})$ as

$$\widehat{l^h(\mathbf{E}, \mathbf{H})} = (y_1, \hat{u}_1^h)_{H_0} + (y_2, \hat{u}_2^h)_{H_0} + \hat{c}^h,$$

where

$$\hat{c} = -\int_D (\mathbf{J}_0(x), \mathbf{Z}^h(x;u))_{\mathbb{C}^3}\, dx, \quad \hat{u}_1^h = Q_1 C_1 \mathbf{P}^h, \quad \hat{u}_2 = Q_2 C_2 \left(\frac{1}{i\omega\mu}\mathrm{rot}\,\mathbf{P}^h\right),$$

and functions $\hat{\mathbf{Z}}^h, \mathbf{P}^h \in V^h$ are determined from the following uniquely solvable system of variational equalities

$$a^*(\hat{\mathbf{Z}}^h, \mathbf{E}') = \int_D \left(\left(-\frac{1}{i\omega\bar{\mu}} \left(\mathbf{l}_2 - C_2^* \Lambda_{H_0} Q_2 C_2 \left(\frac{1}{i\omega\mu}\mathrm{rot}\,\mathbf{P}^h \right) \right), \mathrm{rot}\,\mathbf{E}' \right)_{\mathbb{C}^3} + (\mathbf{l}_1 - C_1^* \Lambda_{H_0} Q_1 C_1 \mathbf{P}^h, \mathbf{E}')_{\mathbb{C}^3} \right) dx \quad \forall \mathbf{E}' \in V^h, \tag{53}$$

$$a(\mathbf{P}^h, \mathbf{E}') = \int_D (Q^{-1}\hat{\mathbf{Z}}^h, \mathbf{E}')_{\mathbb{C}^3}\, dx \quad \forall \mathbf{E}' \in V^h. \tag{54}$$

Theorem 3. *Approximate minimax estimate of* $\widehat{l^h(\mathbf{E},\mathbf{H})}$ *of* $l(\mathbf{E},\mathbf{H})$ *tends to a minimax estimate* $\widehat{\widehat{l(\mathbf{E},\mathbf{H})}}$ *of this expression as* $h \to 0$ *in the sense that*

$$\lim_{h\to 0} \mathbb{E}|\widehat{l^h(\mathbf{E},\mathbf{H})} - \widehat{\widehat{l(\mathbf{E},\mathbf{H})}}|^2 = 0, \tag{55}$$

and

$$\lim_{h\to 0} \mathbb{E}|\widehat{l^h(\mathbf{E},\mathbf{H})} - l(\mathbf{E},\mathbf{H})|^2 = \mathbb{E}|\widehat{\widehat{l(\mathbf{E},\mathbf{H})}} - l(\mathbf{E},\mathbf{H}))|^2. \tag{56}$$

Proof. Denote by $\{h_n\}$ any sequence of positive numbers such that $h_n \to 0$ when $n \to \infty$. Let $\mathbf{Z}^{h_n}(\cdot;u) \in V^{h_n}$ be a solution of the problem

$$a^*(\mathbf{Z}^{h_n}(\cdot;u),\mathbf{E}^{h_n}) = \int_D \left(-\left(\frac{1}{i\omega\bar{\mu}}(\mathbf{l}_2 - C_2^*\Lambda_{H_0}u_2), \operatorname{rot}\mathbf{E}^{h_n}\right)_{\mathbb{C}^3} + (\mathbf{l}_1 - C_1^*\Lambda_{H_0}u_1, \mathbf{E}^{h_n})_{\mathbb{C}^3}\right) dx \quad \forall \mathbf{E}^{h_n} \in V^{h_n}.$$

Then

$$\|\mathbf{Z}(\cdot;u) - \mathbf{Z}^{h_n}(\cdot;u)\|_{H(\mathrm{rot},D)} \to 0 \tag{57}$$

when $n \to \infty$. In fact, from the relationship

$$\begin{aligned}\alpha\|\mathbf{Z}^{h_n}(\ \cdot\ ;u)\|^2_{H(\mathrm{rot},D)} &\le \operatorname{Re} a^*(\mathbf{Z}^{h_n}(\cdot;u),\mathbf{Z}^{h_n}(\cdot;u)) \\ &\le \|\mathbf{Z}^{h_n}(\cdot;u)\|_{H(\mathrm{rot},D)} \int_D \left(|\frac{1}{i\omega\bar{\mu}}(\mathbf{l}_2 - C_2^*\Lambda_{H_0}u_2)|^2_{\mathbb{C}^3} + |\mathbf{l}_1 - C_1^*\Lambda_{H_0}u_1|^2_{\mathbb{C}^3}\right) dx\end{aligned}$$

it follows the existence of a subsequence n_k such that $\mathbf{Z}_{n_k}(\cdot;u)$ converges weakly to $\mathbf{Z}(\cdot;u)$ when $k \to \infty$. Therefore,

$$\begin{aligned}a^*(\mathbf{Z}^{h_{n_k}}(\ \cdot\ ;u),\mathbf{E}^{h_{n_k}}) &= \int_D \left(-\left(\frac{1}{i\omega\bar{\mu}}(\mathbf{l}_2 - C_2^*\Lambda_{H_0}u_2), \operatorname{rot}\mathbf{E}^{h_{n_k}}\right)_{\mathbb{C}^3} + (\mathbf{l}_1 - C_1^*\Lambda_{H_0}u_1, \mathbf{E}^{h_{n_k}})_{\mathbb{C}^3}\right) dx \\ &\to \int_D \left(-\left(\frac{1}{i\omega\bar{\mu}}(\mathbf{l}_2 - C_2^*\Lambda_{H_0}u_2), \operatorname{rot}\mathbf{E}'\right)_{\mathbb{C}^3} + (\mathbf{l}_1 - C_1^*\Lambda_{H_0}u_1, \mathbf{E}')_{\mathbb{C}^3}\right) dx \\ &= a^*(\mathbf{Z}(\cdot;u),\mathbf{E}')).\end{aligned}$$

Here we make use of completeness of sequence of subspaces $\{V^h\}$ in $H_0(\mathrm{rot}, D)$. Due to uniqueness of solution to equation (22), we obtain that the whole sequence $\{\mathbf{Z}^{h_n}(\cdot;u)\}$ weakly converges to $\mathbf{Z}(\cdot;u)$.

Further, notice that

$$\begin{aligned}
\alpha\|\mathbf{Z}^{h_n}(\cdot;u) &- \mathbf{Z}(\cdot;u)\|^2_{H(\mathrm{rot},D)} \\
&\le \mathrm{Re}\, a^*(\mathbf{Z}^{h_n}(\cdot;u)-\mathbf{Z}(\cdot;u), \mathbf{Z}^{h_n}(\cdot;u)-\mathbf{Z}(\cdot;u)) \\
&\le |a^*(\mathbf{Z}^{h_n}(\cdot;u), \mathbf{Z}^{h_n}(\cdot;u)) - a^*(\mathbf{Z}(\cdot;u), \mathbf{Z}^{h_n}(\cdot;u))| \\
&+ |a^*(\mathbf{Z}(\cdot;u), \mathbf{Z}^{h_n}(\cdot;u)) - a^*(\mathbf{Z}(\cdot;u), \mathbf{Z}(\cdot;u))| \\
&= |a^*(\mathbf{Z}(\cdot;u), \mathbf{Z}^{h_n}(\cdot;u)) - a^*(\mathbf{Z}(\cdot;u), \mathbf{Z}(\cdot;u))| \\
&= |a^*(\mathbf{Z}(\cdot;u), \mathbf{Z}(\cdot;u) - \mathbf{Z}^{h_n}(\cdot;u))|. \qquad (58)
\end{aligned}$$

From weak convergence of sequence $\{\mathbf{Z}^{h_n}(\cdot;u)\}$ to $\mathbf{Z}(\cdot;u)$, we have

$$|a^*(\mathbf{Z}(\cdot;u), \mathbf{Z}(\cdot;u) - \mathbf{Z}^{h_n}(\cdot;u))| \to 0 \quad \text{as} \quad n \to \infty,$$

and from (58), it follows (57).

Prove now that

$$\lim_{n\to\infty} \|u^{h_n} - \hat{u}\|_{H_0\times H_0} = \lim_{n\to\infty} \left(\|u_1^{h_n} - \hat{u}_1\|_{H_0} + \|u_2^{h_n} - \hat{u}_2\|_{H_0} \right) = 0, \qquad (59)$$

where $u^{h_n} = (u_1^{h_n}, u_2^{h_n})$, $\hat{u} = (\hat{u}_1, \hat{u}_2)$.
Set

$$I_n(u) = (Q^{-1}\mathbf{Z}^{h_n}(\cdot;u), \mathbf{Z}^{h_n}(\cdot;u))_{L^2(D)} + (Q_1^{-1}u_1, u_1)_{H_0} + (Q_2^{-1}u_2, u_2)_{H_0}. \qquad (60)$$

It is clear that

$$\inf_{u\in H_0\times H_0} I_n(u) = I_n(u^{h_n})$$

and

$$I_n(u^{h_n}) \le I_n(\hat{u}).$$

From strong convergence of $\mathbf{Z}^{h_n}(\cdot;\hat{u})$ to $\mathbf{Z}(\cdot;\hat{u})$ in the space $H(\mathrm{rot}, D)$, we have

$$\lim_{n\to\infty} I_n(\hat{u}) = I(\hat{u}),$$

and, hence, $\overline{\lim}_{n\to\infty} I_n(u^{h_n}) \le I(\hat{u})$. Since $I_n(u^{h_n}) \ge (Q_1^{-1}u_1^{h_n}, u_1^{h_n})_{H_0} + (Q_2^{-1}u_2^{h_n}, u_2^{h_n})_{H_0}$, then from sequence $\{u^{h_n}\}$ one can extract a subsequence $\{u^{h_{n_k}}\}$ such that $u^{h_{n_k}} \xrightarrow{\text{weakly}}$

$\tilde{u}$ in $H_0 \times H_0$. From lower semicontinuity of functional $I(u)$ in a weak topology of the space $H_0 \times H_0$ it follows that

$$\underline{\lim}_{k\to\infty} I_{n_k}(u^{h_{n_k}}) \geq I(\tilde{u}).$$

Taking into account the uniqueness of an element on which the minimum of functional $I(u)$ is attained and inequalities

$$I(\tilde{u}) \leq \underline{\lim}_{k\to\infty} I_{n_k}(u^{h_{n_k}}) \leq \overline{\lim}_{k\to\infty} I_n(u^{h_{n_k}}) \leq I(\hat{u}),$$

we find that $\tilde{u} = \hat{u}$. This means

$$\mathbf{Z}^{h_n}(\cdot; u^{h_n}) \xrightarrow{\text{weakly}} \mathbf{Z}(\cdot;\hat{u}) = \hat{\mathbf{Z}} \quad \text{and} \quad \mathbf{P}^{h_n} \xrightarrow{\text{weakly}} \mathbf{P} \quad \text{in} \quad H(\mathrm{rot}, D).$$

Additionally, from above reasoning it follows that $I_n(u^{h_n}) \to I(\hat{u})$. But then

$$(Q_1^{-1}u_1^{h_n}, u_1^{h_n})_{H_0} + (Q_2^{-1}u_2^{h_n}, u_2^{h_n})_{H_0} \to (Q_1^{-1}u_1, u_1)_{H_0} + (Q_2^{-1}u_2, u_2)_{H_0},$$

that together with weak convergence of sequence $\{u^{h_n}\}$ to $\hat{u}$ implies its strong convergence in space $H_0 \times H_0$. Equality (59) is established.

Using this equality, let us prove (55) and (56). We have

$$\begin{aligned}
\mathbb{E}|\widehat{l^{h_n}(\mathbf{E},\mathbf{H})} &- \widehat{\widehat{l(\mathbf{E},\mathbf{H})}}|^2 \\
&= \mathbb{E}[(\hat{u}_1^{h_n}, y_1)_{H_0} + (\hat{u}_2^{h_n}, y_2)_{H_0} + \hat{c}^{h_n} - (\hat{u}_1, y_1)_{H_0} - (\hat{u}_2, y_2)_{H_0} - \hat{c}]^2 \\
&= \mathbb{E}[(\hat{u}_1^{h_n} - \hat{u}_1, y_1)_{H_0} + (\hat{u}_2^{h_n} - \hat{u}_2, y_2)_{H_0} + \hat{c}^{h_n} - \hat{c}]^2 \\
&= [(\hat{u}_1^{h_n} - \hat{u}_1, C_1\mathbf{E})_{H_0} + (\hat{u}_2^{h_n} - \hat{u}_2, C_2\mathbf{H})_{H_0} + \hat{c}^{h_n} - \hat{c}]^2 \\
&+ \mathbb{E}[(\hat{u}_1^{h_n} - \hat{u}_1, \eta_1)_{H_0} + (\hat{u}_2^{h_n} - \hat{u}_2, \eta_2)_{H_0}]. \qquad (61)
\end{aligned}$$

Taking into account that $\mathbf{Z}^{h_n}$ weakly converges to $\hat{\mathbf{Z}} = \mathbf{Z}(\cdot;\hat{u})$ in the space $L^2(D)^3$ and hence $\hat{c}^{h_n} \to \hat{c}$ when $n \to \infty$ and the fact that $\mathbf{J} \in G_0$, we see that the last expression in the chain of inequalities

$$\begin{aligned}
&[(\hat{u}_1^{h_n} - \hat{u}_1, C_1\mathbf{E})_{H_0} + (\hat{u}_2^{h_n} - \hat{u}_2, C_2\mathbf{H})_{H_0} + \hat{c}^{h_n} - \hat{c}]^2 \\
&\quad \leq C\left(\|\hat{u}_1^{h_n} - \hat{u}_1\|_{H_0}^2 + \|\hat{u}_2^{h_n} - \hat{u}_2\|_{H_0}^2 + (\hat{c}^{h_n} - \hat{c})^2\right)\left(\|\mathbf{E}\|_{H(\mathrm{rot},D)}^2 + \|\mathbf{H}\|_{H(\mathrm{rot},D)}^2\right) \\
&\quad \leq \tilde{C}\left(\|\hat{u}^{h_n} - \hat{u}\|_{H_0\times H_0}^2 + (\hat{c}^{h_n} - \hat{c})^2\right)\|\mathbf{J}\|_{L^2(D)}^2 \quad (C, \tilde{C} = \text{const})
\end{aligned}$$

tends to zero as $n \to \infty$. Convergence to zero of the last term

$$\mathbb{E}[(\hat{u}_1^{h_n} - \hat{u}_1, \eta_1)_{H_0} + (\hat{u}_2^{h_n} - \hat{u}_2, \eta_2)_{H_0}]$$

in the right-hand side of (61) as $n \to \infty$ can be proved in a similar way. The validity of the theorem follows now from the inequality

$$\mathbb{E}|l^{h_n}\widehat{(\mathbf{E},\mathbf{H})} - l(\mathbf{E},\mathbf{H})|^{1/2} = \mathbb{E}[l^{h_n}\widehat{(\mathbf{E},\mathbf{H})} - \widehat{\widehat{l(\mathbf{E},\mathbf{H})}} + \widehat{\widehat{l(\mathbf{E},\mathbf{H})}} - l(\mathbf{E},\mathbf{H})|^{1/2}$$
$$\leq \left\{\mathbb{E}|l^{h_n}\widehat{(\mathbf{E},\mathbf{H})} - \widehat{\widehat{l(\mathbf{E},\mathbf{H})}}|^2\right\}^{1/2} + \left\{\mathbb{E}[\widehat{\widehat{l(\mathbf{E},\mathbf{H})}} - l(\mathbf{E},\mathbf{H})]^2|\right\}^{1/2}.$$

□

Let us formulate a similar result in the case when an estimate of the state $\mathbf{E},\mathbf{H}$ is directly determined from the solution to problem (48)–(49).

Theorem 4. *Let* $(\hat{\mathbf{E}}^h,\hat{\mathbf{H}}^h) = (\hat{\mathbf{E}}^h, \frac{1}{i\omega\mu}\mathrm{rot}\,\hat{\mathbf{E}}^h) \in V^h \times V^h$ *be an approximate estimate of the vector-functions* $(\hat{\mathbf{E}},\hat{\mathbf{H}})$ *determined from the solution to the variational problem*

$$a^*(\hat{\mathbf{P}}^h,E') = \int_D \left(-\frac{1}{i\omega\bar{\mu}}\left(C_2^*\Lambda_{H_0}Q_2\left(y_2 - C_2\left(\frac{1}{i\omega\mu}\mathrm{rot}\hat{\mathbf{E}}^h\right)\right),\mathrm{rot}\mathbf{E}'\right)_{\mathbb{C}^3}\right.$$
$$\left. + \left(C_1^*\Lambda_{H_0}Q_1\left(y_1 - C_1\hat{\mathbf{E}}^h\right),\mathbf{E}'\right)_{\mathbb{C}^3}\right)dx \quad \forall\mathbf{E}' \in V^h, \tag{62}$$

$$a(\hat{\mathbf{E}}^h,\mathbf{E}') = \int_D (Q^{-1}\hat{\mathbf{P}}^h - \mathbf{J}_0,\mathbf{E}')_{\mathbb{C}^3}\,dx \quad \forall\mathbf{E}' \in V^h. \tag{63}$$

Then

$$\|\hat{\mathbf{E}} - \hat{\mathbf{E}}^h\|_{H(rot,D)} + \|\hat{\mathbf{H}} - \hat{\mathbf{H}}^h\|_{H(rot,D)} \to 0 \quad \text{as} \quad h \to 0.$$

The problem of finding $\hat{\mathbf{Z}}^h$, $\mathbf{P}^h \in V^h$ and $\hat{\mathbf{P}}^h$, $\hat{\mathbf{E}}^h \in V^h$ from (53), (54), (62), and (63), respectively, is equivalent to determination of coefficients in the expansion $\hat{\mathbf{Z}}^h$, $\mathbf{P}^h$, $\hat{\mathbf{P}}^h$, $\hat{\mathbf{E}}^h$ by basis elements of the space V^h from the corresponding system of linear algebraic equations.

Introducing the basis in the space V^h, the problem (53)–(59) can be rewritten as a system of liner algebraic equations. To do this, let us denote the elements of the basis by ξ_i $(i = 1,\ldots,N)$ where $N = \dim V^h$. The fact that $\hat{\mathbf{Z}}^h$ and $\mathbf{P}^h$ belong to the space V^h means the existence of constants $\hat{z}_i$ and p_i such that

$$\hat{\mathbf{Z}}^h = \sum_{j=1}^N \hat{z}_j\xi_j, \quad \mathbf{P}^h = \sum_{j=1}^N p_j\xi_j. \tag{64}$$

Setting in (53) and (54) $\mathbf{E}' = \xi_i$ $(i = 1,\ldots,N)$, we obtain that finding $\hat{\mathbf{Z}}^h$, $\mathbf{P}^h$ is equivalent to solving the following system of linear algebraic equations with respect to coefficients $\hat{z}_j$, p_j of expansions (64):

$$\sum_{j=1}^{N} a_{ij}^{*}\hat{z}_j + \sum_{j=1}^{N} b_{ij}\hat{p}_j = f_i, \quad i = 1,\dots,N,$$

$$\sum_{j=1}^{N} a_{ij}\hat{p}_j + \sum_{j=1}^{N} d_{ij}\hat{z}_j = 0, \quad i = 1,\dots,N,$$

where

$$a_{ij}^{*} = a^{*}(\xi_j,\xi_i), \qquad a_{ij} = a(\xi_j,\xi_i) = \overline{a_{ji}^{*}},$$

$$b_{ij} = \int_D \left(\left(C_1^{*}\Lambda_{H_0} Q_1 C_1 \xi_j, \xi_i \right)_{\mathbb{C}^3} - \left(\frac{1}{i\omega\bar{\mu}} C_2^{*}\Lambda_{H_0} Q_2 C_2 \left(\frac{1}{i\omega\mu}\operatorname{rot}\xi_j \right), \operatorname{rot}\xi_i \right)_{\mathbb{C}^3} \right) dx,$$

$$d_{ij} = -\int_D (Q^{-1}\xi_j,\xi_i)_{\mathbb{C}^3} dx, \quad i,j = 1,\dots,N,$$

and

$$f_i = \int_D \left((\mathbf{l}_1,\xi_i)_{\mathbb{C}^3} - \left(\frac{1}{i\omega\bar{\mu}}\mathbf{l}_2, \operatorname{rot}\xi_i \right)_{\mathbb{C}^3} \right) dx, \quad i = 1,\dots,N.$$

5.1 Corollary from the Obtained Results for Integral Observation Operators

As an example, we consider the case when $H_0 = L^2(D_1)^3 \times \dots L^2(D_j)^3 \times \dots L^2(D_n)^3$. Then $\Lambda_{H_0} = I_{H_0}$, where I_{H_0} is the unit operator in H_0,

$$y_i(x) = \left(\mathbf{y}_1^i(x),\dots,\mathbf{y}_j^i(x),\dots,\mathbf{y}_n^i(x) \right),$$

$$\eta_i(x) = \left(\eta_1^i(x),\dots,\eta_j^i(x),\dots,\eta_n^i(x) \right),$$

where

$$\mathbf{y}_j^i(x) = (y_{j,1}^i(x), y_{j,2}^i(x), y_{j,3}^i(x))^T \in L^2(D_j)^3,$$
$$\eta_j^i(x) = (\eta_{j,1}^i(x), \eta_{j,2}^i(x), \eta_{j,3}^i(x))^T,$$

is a stochastic vector process with components $\eta_{j,l}^{(i)}(x)$ $(l = 1,2,3,\ j = 1,\dots,n)$ that are stochastic processes with zero expectations and finite second moments.

Let in observations (11) the operators $C_i : (L^2(D)^3)^n \to L^2(D_1)^3 \times \ldots L^2(D_j)^3 \times \ldots L^2(D_n)^3$, $i = 1,2$, be defined by

$$C_i \mathbf{v}(x) = \left(C_1^i \mathbf{v}(x), \ldots, C_j^i \mathbf{v}(x), \ldots, C_n^{(1)} \mathbf{v}(x) \right), \quad j = 1, \ldots, n,$$
$$\mathbf{v}(x) = (v_1,(x) v_2(x), v_3(x))^T,$$

where $C_j^i : L^2(D)^3 \to L^2(D_j)^3$ is an integral operator defined by

$$C_j^i \mathbf{v}(x) := \int_{D_j} \mathbf{K}_j^i(x,\xi)\mathbf{v}(\xi)\, d\xi,$$

$\mathbf{K}_j^i(x,\xi) = \{k_{rs}^{(i,j)}(x,\xi)\}_{r,s=1}^3$ is a matrix with entries $k_{rs}^{(i.j)} \in L^2(D_j) \times L^2(D_j)$. As a result, observations y_1 and y_2 in (11) take the form

$$y_1 = \left(\mathbf{y}_1^{(1)}(x), \ldots, \mathbf{y}_j^{(1)}(x), \ldots, \mathbf{y}_n^{(1)}(x) \right),$$

$$y_2 = \left(\mathbf{y}_1^{(2)}(x), \ldots, \mathbf{y}_j^{(2)}(x), \ldots, \mathbf{y}_n^{(2)}(x) \right),$$

where

$$\mathbf{y}_j^{(1)}(x) = \int_{D_j} \mathbf{K}_{1,j}(x,\xi)\mathbf{E}(\xi)\, d\xi + \eta_j^{(1)}(x), \tag{65}$$

$$\mathbf{y}_j^{(2)}(x) = \int_{D_j} \mathbf{K}_{2,j}(x,\xi)\mathbf{H}(\xi)\, d\xi + \eta_j^{(2)}(x), \quad j = 1, \ldots n, \tag{66}$$

and the operators $\tilde{Q}_i \in \mathscr{L}(L^2(D_1)^3 \times \ldots L^2(D_j)^3 \times \ldots L^2(D_n)^3, L^2(D_1)^3 \times \ldots L^2(D_j)^3 \times \ldots L^2(D_n)^3)$, $i = 1,2$, in (14), which is contained in the definition of set G_1, are given by

$$Q_i \tilde{\eta}_i = (\mathbf{Q}_1^i(x)\tilde{\eta}_1^i(x), \ldots, \mathbf{Q}_j^i(x)\tilde{\eta}_j^i(x), \ldots, \mathbf{Q}_n^i(x)\tilde{\eta}_n^i(x))$$

where $\mathbf{Q}_j^i(x)$ are square Hermitian positive definite 3×3 matrices with entries $q_{rs}^{(i,j)} \in C(\bar{D}_j)$,[5] $r,s = 1,2,3$, $\tilde{\eta}_j^i \in L^2(\Omega, L^2(D_j)^3)$, $j = 1, \ldots, n$, $i = 1,2$.

In this case condition (14) takes the form [6]

[5] Here and below we denote by $C(\bar{D}_j)$ a class of functions continuous in the domain $\bar{D}_j$.

[6] By

$$\mathrm{Sp}\left(\mathbf{Q}_j^i(x)\tilde{\mathbf{R}}_{r_1}^{(i)}(x,x) \right)$$

$$\sum_{j=1}^{n}\int_{D_j}\mathrm{Sp}(\mathbf{Q}_j^1(x)\tilde{\mathbf{R}}_j^{(1)}(x,x))\,dt\le 1,\quad \sum_{j=1}^{n}\int_{D_j}\mathrm{Sp}(\mathbf{Q}_j^2(x)\tilde{\mathbf{R}}_j^{(2)}(x,x))\,dt\le 1,$$

where by $\tilde{\mathbf{R}}_j^{(i)}(x,y)=[\tilde{b}_{r,s}^{(i,j)}(x,y)]_{r,s=1}^3$ we denote the correlation matrix of the vector process

$$\tilde{\eta}_j^i(x)=(\tilde{\eta}_{j,1}^i(x),\tilde{\eta}_{j,2}^i(x),\tilde{\eta}_{j,3}^i(x))^T$$

with components

$$\tilde{b}_{r,s}^{(i.j)}(x,y)=\mathbb{E}\Big(\tilde{\eta}_{j,r}^i(x)\bar{\tilde{\eta}}_{j,s}^i(y)\Big),$$

$j=1,\dots,n$, $i=1,2$.

Uncorrelatedness of random variables $\tilde{\eta}_1$ and $\tilde{\eta}_2$ reduces in this case to the condition

$$\mathbb{E}\left(\tilde{\eta}_k^1,\mathbf{u}_k^{(1)}\right)_{L^2(D_k)^3}\overline{(\tilde{\eta}_r^2,\mathbf{u}_r^{(2)})_{L^2(D_r)^3}}=0\ \forall\mathbf{u}_k^{(1)}\in L^2(D_k)^3,\mathbf{u}_r^{(2)}\in L^2(D_r)^3,\quad k,r=\overline{1,n},\tag{67}$$

and hence, the set G_1 is described by the formula

$$G_1=\Bigg\{\tilde{\eta}=(\tilde{\eta}_1,\tilde{\eta}_2):\tilde{\eta}_1=(\tilde{\eta}_1^i,\dots,\tilde{\eta}_j^i,\dots\tilde{\eta}_n^i),\tilde{\eta}_j^i=(\tilde{\eta}_{j,1}^i,\tilde{\eta}_{j,2}^i,\tilde{\eta}_{j,3}^i)\in L^2(\Omega,L^2(D_j)^3),$$

$$\mathbb{E}\tilde{\eta}_k^i(x)=0,\ \tilde{\eta}_k^1(x)\text{ and }\tilde{\eta}_r^{(2)}(x)\text{ satisfy (67)},\ k,r=1,\dots,n,\ i=1,2;$$

$$\sum_{j=1}^{n}\int_{D_j}\mathrm{Sp}\left(\mathbf{Q}_j^1(x)\tilde{\mathbf{R}}_j^{(1)}(x,x)\right)dx\le 1,\quad \sum_{j=1}^{n}\int_{D_j}\mathrm{Sp}\left(\mathbf{Q}_j^2(x)\tilde{\mathbf{R}}_j^{(2)}(x,x)\right)dx\le 1,\Bigg\}\tag{68}$$

A class of linear with respect of observations (65) and (66) estimates $l(\mathbf{E},\mathbf{H})$ will take the form

$$\widehat{l(\mathbf{E},\mathbf{H})}=\sum_{i=1}^{n}\int_{D_i}\left(\mathbf{y}_i^{(1)}(x),\mathbf{u}_i^{(1)}(x)\right)_{\mathbb{C}^3}dx+\sum_{i=1}^{n}\int_{D_i}\left(\mathbf{y}_i^{(2)}(x),\mathbf{u}_i^{(2)}(x)\right)_{\mathbb{C}^3}dx+c.$$

Thus, performing elementary calculations and using the above analysis, we obtain that under assumptions (15) and (68), the following result is valid for integral observation operators as a corollary from Theorems 2 and 3.

---

we denote the traces of matrices $\mathbf{Q}_j^i(x)\tilde{\mathbf{R}}_j^{(i)}(x,x)$, i.e. the sum of diagonal elements of these matrices, $i=1,2$.

Theorem 5. *The minimax estimate* $\widehat{\widehat{l(\mathbf{E},\mathbf{H})}}$ *of* $l(\mathbf{E},\mathbf{H})$ *is determined by the formula*

$$\widehat{\widehat{l(\mathbf{E},\mathbf{H})}}=\sum_{i=1}^{n}\int_{D_i}\left(\mathbf{y}_i^{(1)}(x),\hat{\mathbf{u}}_i^{(1)}(x)\right)_{\mathbb{C}^3}dx+\sum_{i=1}^{n}\int_{D_i}\left(\mathbf{y}_i^{(2)}(x),\hat{\mathbf{u}}_i^{(2)}(x)\right)_{\mathbb{C}^3}dx+\hat{c}=l(\hat{\mathbf{E}},\hat{\mathbf{H}}),$$

where

$$\begin{aligned}\hat{c} &= -\int_D(\mathbf{J}_0(x),\hat{\mathbf{Z}}(x))_{\mathbb{C}^3}\,dx,\\ \hat{\mathbf{u}}_j^{(1)}(x) &= \mathbf{Q}_j^1(x)\int_{D_j}\mathbf{K}_{1,j}(x,\eta)\mathbf{P}(\eta)\,d\eta,\\ \hat{\mathbf{u}}_j^{(2)}(x) &= \mathbf{Q}_j^2(x)\int_{D_j}\mathbf{K}_{2,j}(x,\eta)\frac{1}{i\omega\mu}\mathrm{rot}\mathbf{P}(\eta)\,d\eta,\quad j=1,\dots,n,\end{aligned}$$

$\hat{\mathbf{H}}=\frac{1}{i\omega\mu}\mathrm{rot}\hat{\mathbf{E}}$, *and functions* $P,\hat{\mathbf{Z}},\hat{\mathbf{E}}\in H_0(\mathrm{rot},D)$ *are found from solution to systems of variational equations*

$$\begin{aligned}a^*(\hat{\mathbf{Z}},\mathbf{E}') = \int_D\Bigg(\Bigg(&-\frac{1}{i\omega\bar{\mu}}\Bigg(\chi_{\omega_2}(x)\mathbf{l}_2(x)\\ &-\sum_{j=1}^{n}\chi_{D_j}(x)\int_{D_j}\tilde{\mathbf{K}}_{2,j}(x,\xi_1)\frac{1}{i\omega\mu}\mathrm{rot}\mathbf{P}(\xi_1)\,d\xi_1\Bigg),\mathrm{rot}\mathbf{E}'(x)\Bigg)_{\mathbb{C}^3}\\ &+\left(\chi_{\omega_1}(x)\mathbf{l}_1(x)-\sum_{j=1}^{n}\chi_{D_j}(x)\int_{D_j}\tilde{\mathbf{K}}_{1,j}(x,\xi_1)\mathbf{P}(\xi_1)\,d\xi_1,\mathbf{E}'(x)\right)_{\mathbb{C}^3}\Bigg)dx,\end{aligned}\tag{69}$$

$$a(\mathbf{P},\mathbf{E}') = \int_D(Q^{-1}\hat{\mathbf{Z}},\mathbf{E}')_{\mathbb{C}^3}\,dx,\qquad\forall\mathbf{E}'\in H_0(\mathrm{rot},D)\tag{70}$$

and

$$\begin{aligned}a^*(\hat{\mathbf{P}},\mathbf{E}') = \int_D\Bigg(\Bigg(&-\frac{1}{i\omega\bar{\mu}}\Bigg(\mathbf{d}_2(x)\\ &-\sum_{j=1}^{n}\chi_{D_j}(x)\int_{D_j}\tilde{\mathbf{K}}_{2,j}(x,\xi_1)\frac{1}{i\omega\mu}\mathrm{rot}\hat{\mathbf{E}}(\xi_1)\,d\xi_1\Bigg),\mathrm{rot}\mathbf{E}'(x)\Bigg)_{\mathbb{C}^3}\\ &+\left(\mathbf{d}_1(x)-\sum_{j=1}^{n}\chi_{D_j}(x)\int_{D_j}\tilde{\mathbf{K}}_{1,j}(x,\xi_1)\hat{\mathbf{E}}(\xi_1)\,d\xi_1,\mathbf{E}'(x)\right)_{\mathbb{C}^3}\Bigg)dx,\end{aligned}\tag{71}$$

$$a(\hat{\mathbf{E}},\mathbf{E}') = \int_D(Q^{-1}\hat{\mathbf{P}}-\mathbf{J}_0,\mathbf{E}')_{\mathbb{C}^3}\,dx\quad\forall\mathbf{E}'\in H_0(\mathrm{rot},D),\tag{72}$$

respectively. Here $\hat{\mathbf{P}} \in H_0(\mathrm{rot}, D)$,

$$\tilde{\mathbf{K}}_{i,j}(x,\xi_1) = \int_D \left(\mathbf{K}^{(i)}_{i,j}(\xi,x)\right)^* \mathbf{Q}^i_j(\xi)\mathbf{K}^{(i)}_{i,j}(\xi,\xi_1)d\xi, \quad i = 1,2^7$$

and

$$\mathbf{d}_i(x) = \sum_{j=1}^{n} \chi_{D_j}(x) \int_D (\mathbf{K}_{i,j}(\xi,x))^* \mathbf{Q}^i_j(\xi)\mathbf{y}^{(i)}_j(\xi)\,d\xi, \quad i = 1,2.$$

Problems (69)–(70) *and* (71)–(72) *are uniquely solvable.*

The estimation error σ *is given by the expression*

$$\sigma = l\left(\mathbf{P}, \frac{1}{i\omega\mu}\mathrm{rot}\,\mathbf{P}\right)^{1/2}.$$

5.2 *Minimax Estimation of the Right-Hand Sides of Equations* (7)*: Representations for Minimax Estimates and Estimation Errors*

The problem is to determine a minimax estimate of the value of the functional

$$l(\mathbf{J}) = \int_D (\mathbf{J}(x), \mathbf{l}_0(x))_{C^3}\,dx \tag{73}$$

from the observations (11) in the class of estimates linear with respect to observations

$$\widehat{l(\mathbf{J})} = (y_1, u_1)_{H_0} + (y_2, u_2)_{H_0} + c,$$

where $u_1, u_2 \in H_0$, $c \in \mathbb{C}$, and $\mathbf{l}_0 \in L^2(D)^3$ is a given function, under the assumption that $\mathbf{J} \in G_0$ and the errors (η_1, η_2) in observations (11) belong to G_1, where sets G_0 and G_1 are defined by (15) and (14), respectively.

Definition 2. The estimate of the form

$$\widehat{\widehat{l(\mathbf{J})}} = (y_1, \hat{u}_1)_{H_0} + (y_2, \hat{u}_2)_{H_0} + \hat{c} \tag{74}$$

[7]We use the following notation: if $\mathbf{A}(\xi) = [a_{ij}(\xi)]_{i,j=1}^N$ is a matrix depending on variable ξ that varies on measurable set Ω, then we define $\int_\Omega \mathbf{A}(\xi)\,d\xi$ by the equality

$$\int_\Omega \mathbf{A}(\xi)\,d\xi = \left[\int_\Omega a_{ij}(\xi)\,d\xi\right]_{i,j=1}^N.$$

will be called the minimax estimate of $l(\mathbf{J})$ if the element $\hat{u} = (\hat{u}_1, \hat{u}_2) \in H_0 \times H_0$ and number $\hat{c} \in \mathbb{C}$ are determined from the condition

$$\inf_{u_1,u_2 \in H_0, c \in \mathbb{C}} \sigma(u_1, u_2, c) = \sigma(\hat{u}_1, \hat{u}_2, \hat{c}),$$

where

$$\sigma(u_1, u_2, c) := \sup_{\tilde{\mathbf{J}} \in G_0, (\tilde{\eta}_1, \tilde{\eta}_2) \in G_1} \mathbf{E} |l(\tilde{\mathbf{J}}) - \widehat{l(\tilde{\mathbf{J}})}|^2,$$

$$\widehat{l(\tilde{\mathbf{J}})} = (\tilde{y}_1, u_1)_{H_0} + (\tilde{y}_2, u_2)_{H_0} + c, \tag{75}$$

$$\tilde{y}_1 = C_1 \tilde{\mathbf{E}} + \tilde{\eta}_1, \quad \tilde{y}_2 = C_2 \tilde{\mathbf{H}} + \tilde{\eta}_2,$$

and $(\tilde{\mathbf{E}}, \tilde{\mathbf{H}})$ is a solution to the problem (7)–(8) when $\mathbf{J}(x) = \tilde{\mathbf{J}}(x)$.

The quantity

$$\sigma = [\sigma(\hat{u}_1, \hat{u}_2, \hat{c})]^{1/2}$$

is called the error of the minimax estimation of $l(\mathbf{J})$.

Lemma 2. *Finding the minimax estimate of $l(\mathbf{J})$ is equivalent to the problem of optimal control of a system described by the problem*

$$\mathbf{Z}(x; u) \in H_0(\mathrm{rot}, D), \tag{76}$$

$$a^*(\mathbf{Z}(\cdot; u), \mathbf{E}') = \int_D \left(\left(\frac{1}{i\omega\bar{\mu}} C_2^* \Lambda_{H_0} u_2, \mathrm{rot}\, \mathbf{E}' \right)_{\mathbb{C}^3} - (C_1^* \Lambda_{H_0} u_1(x), \mathbf{E}')_{\mathbb{C}^3} \right) dx$$
$$\forall\ \mathbf{E}' \in H_0(\mathrm{rot}, D), \tag{77}$$

with the quality criterion

$$I(u) = \int_D (Q^{-1}(\mathbf{l}_0(x) - \mathbf{Z}(x; u)), \mathbf{l}_0(x) - \mathbf{Z}(x; u))_{\mathbb{C}^3} dx$$
$$+ (Q_1^{-1} u_1, u_1)_{H_0} + (Q_2^{-1} u_2, u_2)_{H_0} \to \inf_{u \in H_0 \times H_0}. \tag{78}$$

Proof. Taking into account (73) at $J = \tilde{J}$, (75), and (11), we obtain

$$
\begin{aligned}
l(\tilde{\mathbf{J}}) - \widehat{l(\tilde{\mathbf{J}})} &= \int_D (\tilde{\mathbf{J}}(x), \mathbf{l}_0(x))_{\mathbb{C}^3} dx \\
&- (C_1\tilde{\mathbf{E}}, u_1)_{H_0} - (C_2\tilde{\mathbf{H}}, u_2)_{H_0} - (\tilde{\eta}_1, u_1)_{H_0} - (\tilde{\eta}_2, u_2)_{H_0} - c \\
&= \int_D (\tilde{\mathbf{J}}(x), \mathbf{l}_0(x))_{\mathbb{C}^3} dx - (C_1\tilde{\mathbf{E}}, u_1)_{H_0} \\
&- \left(C_2 \frac{1}{i\omega\mu} \mathrm{rot}\,\tilde{\mathbf{E}}, u_2 \right)_{H_0} - (\tilde{\eta}_1, u_1)_{H_0} - (\tilde{\eta}_2, u_2)_{H_0} - c \\
&= \int_D (\tilde{\mathbf{J}}(x), \mathbf{l}_0(x))_{\mathbb{C}^3} dx - < C_1\tilde{\mathbf{E}}, \Lambda_{H_0} u_1 >_{H_0 \times H_0'} \\
&- < C_2 \frac{1}{i\omega\mu} \mathrm{rot}\,\tilde{\mathbf{E}}, \Lambda_{H_0} u_2 >_{H_0 \times H_0'} - (\tilde{\eta}_1, u_1)_{H_0} - (\tilde{\eta}_2, u_2)_{H_0} - c \\
&= \int_D (\tilde{\mathbf{J}}(x), \mathbf{l}_0(x))_{\mathbb{C}^3} dx - \int_D (\tilde{\mathbf{E}}, C_1^* \Lambda_{H_0} u_1(x))_{\mathbb{C}^3} dx \\
&+ \int_D \left(\mathrm{rot}\,\tilde{\mathbf{E}}, \frac{1}{i\omega\bar{\mu}} C_2^* \Lambda_{H_0} u_2(x) \right)_{\mathbb{C}^3} dx - (\tilde{\eta}_1, u_1)_{H_0} - (\tilde{\eta}_2, u_2)_{H_0} - c. \quad (79)
\end{aligned}
$$

Set $\mathbf{E}' = \tilde{\mathbf{E}}$ in (76) and $\mathbf{E} = \tilde{\mathbf{E}}$, $\mathbf{E}' = \mathbf{Z}(\cdot; u)$ in (10), respectively. Then we have

$$
a^*(\mathbf{Z}(\cdot; u), \tilde{\mathbf{E}}) = \int_D \left(\left(\frac{1}{i\omega\bar{\mu}} C_2^* \Lambda_{H_0} u_2, \mathrm{rot}\,\tilde{\mathbf{E}} \right)_{\mathbb{C}^3} - (C_1^* \Lambda_{H_0} u_1(x), \tilde{\mathbf{E}})_{\mathbb{C}^3} \right) dx, \quad (80)
$$

and

$$
a(\tilde{\mathbf{E}}, \mathbf{Z}(\cdot; u)) = - \int_D (\mathbf{J}, \mathbf{Z}(\cdot; u))_{\mathbb{C}^3}\, dx. \quad (81)
$$

Since

$$
\overline{a^*(\mathbf{Z}(\cdot; u), \tilde{\mathbf{E}})} = a(\tilde{\mathbf{E}}, \mathbf{Z}(\cdot; u)), \quad (82)
$$

from (79) to (82), we obtain

$$
\begin{aligned}
l(\tilde{\mathbf{J}}) - \widehat{l(\tilde{\mathbf{J}})} &= \int_D (\tilde{\mathbf{J}}(x), \mathbf{l}_0(x))_{\mathbb{C}^3} dx \\
&- \int_D (\tilde{\mathbf{J}}(x), \mathbf{Z}(x; u))_{\mathbb{C}^3}\, dx - (\tilde{\eta}_1, u_1)_{H_0} - (\tilde{\eta}_2, u_2)_{H_0} - c \\
&= \int_D (\tilde{\mathbf{J}}(x) - \mathbf{J}_0(x), \mathbf{l}_0(x) - \mathbf{Z}(x; u))_{\mathbb{C}^3}\, dx \\
&- (\tilde{\eta}_1, u_1)_{H_0} - (\tilde{\eta}_2, u_2)_{H_0} + \int_D (\mathbf{J}_0(x), \mathbf{l}_0(x) - \mathbf{Z}(x; u))_{\mathbb{C}^3}\, dx - c
\end{aligned}
$$

Beginning from this place, we apply the same reasoning as in the proof of Lemma 1 (replacing $\mathbf{Z}(x, u)$ by $\mathbf{l}_0(x) - \mathbf{Z}(x, u)$) to obtain

$$\inf_{c\in\mathbb{C}}\sup_{\tilde{\mathbf{J}}\in G_0,(\tilde{\eta}_1,\tilde{\eta}_2)\in G_1}\mathbf{E}|l(\tilde{\mathbf{J}})-\widehat{l(\tilde{\mathbf{J}})}|^2=\inf_{c\in\mathbb{C}}\sigma(u_1,u_2,c)=I(u),$$

where $I(u)$ is determined by formula (78) for

$$c=\int_D(\mathbf{J}_0(x),\mathbf{l}_0(x)-\mathbf{Z}(x;u))_{\mathbb{C}^3}\,dx.$$

The validity of Lemma 2 is established. □

Theorem 6. *The minimax estimate of the functional $l(\mathbf{J})$ has the form*

$$\widehat{\widehat{l(\mathbf{J})}}=(y_1,\hat{u}_1)_{H_0}+(y_2,\hat{u}_2)_{H_0}+\hat{c},$$

where

$$\hat{c}=\int_D(\mathbf{J}_0(x),\mathbf{l}_0(x)-\hat{\mathbf{Z}}(x))_{\mathbb{C}^3}\,dx,\quad \hat{u}_1=-Q_1C_1\mathbf{P},\quad \hat{u}_2=-Q_2C_2\left(\frac{1}{i\omega\mu}\mathrm{rot}\mathbf{P}\right),\tag{83}$$

and the functions $\hat{\mathbf{Z}}$ and $\mathbf{P}\in H_0(\mathrm{rot},D)$ are determined as a solution of the following problem:

$$a^*(\hat{\mathbf{Z}},\mathbf{E}')=\int_D\left(-\left(\frac{1}{i\omega\bar{\mu}}C_2^*\Lambda_{H_0}Q_2C_2\left(\frac{1}{i\omega\mu}\mathrm{rot}\mathbf{P}\right),\mathrm{rot}\mathbf{E}'\right)_{\mathbb{C}^3}\right.$$
$$\left.+(C_1^*\Lambda_{H_0}Q_1C_1\mathbf{P},\mathbf{E}')_{\mathbb{C}^3}\right)dx,\quad\forall\mathbf{E}'\in H_0(\mathrm{rot},D)\tag{84}$$

$$a(\mathbf{P},\mathbf{E}')=\int_D(Q^{-1}(\mathbf{l}_0-\hat{\mathbf{Z}}),\mathbf{E}')_{\mathbb{C}^3}\,dx\quad\forall\mathbf{E}'\in H_0(\mathrm{rot},D)\tag{85}$$

The error of estimation σ is given by the expression

$$\sigma=\left(\int_D(\tilde{\mathbf{P}}(x),\mathbf{l}_0(x))_{\mathbb{C}^3}dx\right)^{1/2},\tag{86}$$

where $\tilde{\mathbf{P}}=Q^{-1}(\mathbf{l}_0-\hat{\mathbf{Z}})$.

Proof. The existence of the unique element $\hat{u}\in H_0\times H_0$ such that

$$I(\hat{u})=\inf_{u\in H_0\times H_0}I(u)$$

follows from the reasoning similar to that in the proof of Theorem 1. Therefore, for any $\tau\in\mathbb{R}$ and $v\in H_0\times H_0$, the relations

$$\frac{d}{d\tau}I(\hat{u}+\tau v)\Big|_{\tau=0}=0 \quad \text{and} \quad \frac{d}{d\tau}I(\hat{u}+i\tau v)\Big|_{\tau=0}=0 \tag{87}$$

hold. Taking into account that functions $\mathbf{Z}(x;\hat{u}+\tau v)$ and $\mathbf{Z}(x;\hat{u}+i\tau v)$ can be written, respectively, as $\mathbf{Z}(x;\hat{u}+\tau v)=\mathbf{Z}(x;\hat{u})+\tau\mathbf{Z}(x;v)$ and $\mathbf{Z}(x;\hat{u}+i\tau v)=z(x;\hat{u})+i\tau\mathbf{Z}(x;v)$, where $\mathbf{Z}(x;v)$ is the unique solution to problem (76),(77) at $u=v$, we deduce from (87) that

$$0=\frac{1}{2}\frac{d}{d\tau}I(\hat{u}+\tau v)|_{\tau=0}=-\mathrm{Re}\left\{\int_D(Q^{-1}(\mathbf{l}_0(x)-\mathbf{Z}(x;\hat{u})),\mathbf{Z}(x;v))_{\mathbb{C}^3}dx\right\}$$
$$+\mathrm{Re}\,\{(Q_1^{-1}\hat{u}_1,v_1)_{H_0}+(Q_2^{-1}\hat{u}_2,v_2)_{H_0}\}.$$

$$0=\frac{1}{2}\frac{d}{d\tau}I(\hat{u}+i\tau v)|_{\tau=0}=-\mathrm{Im}\left\{\int_D(Q^{-1}(\mathbf{l}_0(x)-\mathbf{Z}(x;\hat{u})),\mathbf{Z}(x;v))_{\mathbb{C}^3}dx\right\}$$
$$+\mathrm{Im}\,\{(Q_1^{-1}\hat{u}_1,v_1)_{H_0}+(Q_2^{-1}\hat{u}_2,v_2)_{H_0}\}=0.$$

Hence,

$$-(Q^{-1}(\mathbf{l}_0-\mathbf{Z}(\cdot;\hat{u})),\mathbf{Z}(\cdot;v))_{L^2(D)^3}+(Q_1^{-1}\hat{u}_1,v_1)_{H_0}+(Q_2^{-1}\hat{u}_2,v_2)_{H_0}=0. \tag{88}$$

Introduce the function $\mathbf{P}\in H_0(\mathrm{rot},D)$ as the unique solution to the variational problem

$$a(\mathbf{P},\mathbf{E}')=\int_D(Q^{-1}(\mathbf{l}_0-\mathbf{Z}(\cdot;\hat{u})),\mathbf{E}')_{\mathbb{C}^3}\,dx \quad \forall\mathbf{E}'\in H_0(\mathrm{rot},D). \tag{89}$$

Setting in (89) $\mathbf{E}'=\mathbf{Z}(\cdot;v)$, we obtain

$$a(\mathbf{P},\mathbf{Z}(\cdot;v))=\int_D(Q^{-1}(\mathbf{l}_0-\mathbf{Z}(\cdot;\hat{u})),\mathbf{Z}(\cdot;v))_{\mathbb{C}^3}\,dx. \tag{90}$$

Taking into account the fact that $\mathbf{Z}(\cdot;v)$ satisfies Eq. (77) with $u=v$ and setting there $\mathbf{E}'=\mathbf{P}$, we have

$$a^*(\mathbf{Z}(\cdot;v),\mathbf{P})=\int_D\left(\left(\frac{1}{i\omega\bar{\mu}}C_2^*\Lambda_{H_0}v_2,\mathrm{rot}\,\mathbf{P}\right)_{\mathbb{C}^3}-(C_1^*\Lambda_{H_0}v_1(x),\mathbf{P})_{\mathbb{C}^3}\right)dx. \tag{91}$$

Relations (88)–(91) imply

$$(Q_1^{-1}\hat{u}_1, v_1)_{H_0} + (Q_2^{-1}\hat{u}_2, v_2)_{H_0} = a(\mathbf{P}, \mathbf{Z}(\cdot; v)) = \overline{a^*(\mathbf{Z}(\cdot; v), \mathbf{P})}$$
$$= -\int_D (\mathbf{P}, C_1^* \Lambda_{H_0} v_1)_{\mathbb{C}^3} dx - \int_D \left(\frac{1}{i\omega\mu}\mathrm{rot}\mathbf{P}, C_2^* \Lambda_{H_0} v_2\right)_{\mathbb{C}^3} dx$$

Hence,

$$\hat{u}_1 = -Q_1 C_1 \mathbf{P}, \quad \hat{u}_2 = -Q_2 C_2 \left(\frac{1}{i\omega\mu}\mathrm{rot}\mathbf{P}\right).$$

Substituting these expressions into (77) and (89) and setting $\hat{\mathbf{Z}} := \mathbf{Z}(\cdot; \hat{u}))$, we establish the validity of equalities (83) and that functions $\hat{\mathbf{Z}}$ and $\mathbf{P}$ satisfy (84) and (85).

Let us prove representation (86). From (78) at $u = \hat{u}$ and (83), it follows

$$\sigma^2 = I(\hat{u}) = \int_D (Q^{-1}(\mathbf{l}_0(x) - \mathbf{Z}(x; \hat{u})), \mathbf{l}_0(x) - \mathbf{Z}(x; \hat{u}))_{\mathbb{C}^3} dx$$
$$+ (Q_1^{-1}\hat{u}_1, \hat{u}_1)_{H_0} + (Q_2^{-1}\hat{u}_2, \hat{u}_2)_{H_0}$$
$$= \int_D (Q^{-1}(\mathbf{l}_0(x) - \hat{\mathbf{Z}}(x)), \mathbf{l}_0(x) - \hat{\mathbf{Z}}(x))_{\mathbb{C}^3} dx$$
$$+ (C_1\mathbf{P}, Q_1 C_1 \mathbf{P})_{H_0} + \left(C_2\left(\frac{1}{i\omega\mu}\mathrm{rot}\mathbf{P}\right), Q_2 C_2 \left(\frac{1}{i\omega\mu}\mathrm{rot}\mathbf{P}\right)\right)_{H_0}$$
$$= -\int_D (Q^{-1}(\mathbf{l}_0(x) - \hat{\mathbf{Z}}(x)), \hat{\mathbf{Z}}(x))_{\mathbb{C}^3} dx + \int_D (Q^{-1}(\mathbf{l}_0(x) - \hat{\mathbf{Z}}(x)), \mathbf{l}_0(x))_{\mathbb{C}^3} dx$$
$$+ (C_1\mathbf{P}, Q_1 C_1 \mathbf{P})_{H_0} + \left(C_2\left(\frac{1}{i\omega\mu}\mathrm{rot}\mathbf{P}\right), Q_2 C_2 \left(\frac{1}{i\omega\mu}\mathrm{rot}\mathbf{P}\right)\right)_{H_0}. \tag{92}$$

Transform the first term. Make use of equalities (84) and (85) to obtain

$$\int_D (Q^{-1}(\mathbf{l}_0 - \mathbf{Z}(\cdot; \hat{u})), \mathbf{Z}(\cdot; v))_{\mathbb{C}^3}\, dx = a(\mathbf{P}, \hat{\mathbf{Z}}) = \overline{a^*(\hat{\mathbf{Z}}, \mathbf{P})}$$
$$= \int_D \overline{\left(-\left(\frac{1}{i\omega\bar{\mu}} C_2^* \Lambda_{H_0} Q_2 C_2 \left(\frac{1}{i\omega\mu}\mathrm{rot}\mathbf{P}\right), \mathrm{rot}\mathbf{P}\right)_{\mathbb{C}^3} + (C_1^* \Lambda_{H_0} Q_1 C_1 \mathbf{P}, \mathbf{P})_{\mathbb{C}^3}\right)} dx$$
$$= \overline{(Q_1 C_1 \mathbf{P}, C_1 \mathbf{P})_{H_0} + \left(Q_2 C_2 \left(\frac{1}{i\omega\mu}\mathrm{rot}\mathbf{P}\right), C_2 \left(\frac{1}{i\omega\mu}\mathrm{rot}\mathbf{P}\right)\right)_{H_0}}$$
$$= (C_1\mathbf{P}, Q_1 C_1 \mathbf{P})_{H_0} + \left(C_2\left(\frac{1}{i\omega\mu}\mathrm{rot}\mathbf{P}\right), Q_1 C_2 \left(\frac{1}{i\omega\mu}\mathrm{rot}\mathbf{P}\right)\right)_{H_0}. \tag{93}$$

From (92) and (93), it follows

$$\sigma^2 = \int_D (Q^{-1}(\mathbf{l}_0(x) - \hat{\mathbf{Z}}(x)), \mathbf{l}_0(x))_{\mathbb{C}^3} dx = l(\tilde{\mathbf{P}}).$$

The theorem is proved. □

Theorem 7. *The minimax estimate* $\widehat{\widehat{l(\mathbf{J})}}$ *of* $l(\mathbf{J})$ *has the form*

$$\widehat{\widehat{l(\mathbf{J})}} = l(\hat{\mathbf{J}}), \tag{94}$$

where $\hat{\mathbf{J}} = \mathbf{J}_0 - Q^{-1}\hat{\mathbf{P}}$, *and function* $\hat{\mathbf{P}} \in H_0(\mathrm{rot}, D)$ *is determined from the solution to the problem (48)–(49).*

Proof. Let us prove representation (94). Taking into notice (83), (48), and (49), we obtain

$$\begin{aligned}
\widehat{\widehat{l(\mathbf{J})}} &= (y_1, \hat{u}_1)_{H_0} + (y_2, \hat{u}_2)_{H_0} + \hat{c} \\
&= -(y_1, Q_1 C_1 \mathbf{P})_{H_0} - \left(y_2, Q_2 C_2 \left(\frac{1}{i\omega\mu}\mathrm{rot}\,\mathbf{P}\right)\right)_{H_0} + \int_D (\mathbf{J}_0(x), \mathbf{l}_0(x) - \hat{\mathbf{Z}}(x))_{\mathbb{C}^3}\, dx \\
&= -\int_D \left((C_1^* \Lambda_{H_0} Q_1 y_1, \mathbf{P})_{\mathbb{C}^3} - \left(\frac{1}{i\omega\bar{\mu}} C_2^* \Lambda_{H_0} Q_2 y_2, \mathrm{rot}\,\mathbf{P}\right)_{\mathbb{C}^3}\right) dx \\
&\quad + \int_D (\mathbf{J}_0(x), \mathbf{l}_0(x) - \hat{\mathbf{Z}}(x))_{\mathbb{C}^3}\, dx \\
&= -a^*(\hat{\mathbf{P}}, \mathbf{P}) - \int_D \left((C_1^* \Lambda_{H_0} Q_1 C_1 \hat{\mathbf{E}}, \mathbf{P})_{\mathbb{C}^3} - \left(C_2^* \Lambda_{H_0} Q_2 C_2 \left(\frac{1}{i\omega\mu}\mathrm{rot}\,\hat{\mathbf{E}}\right), \frac{1}{i\omega\mu}\mathrm{rot}\,\mathbf{P}\right)_{\mathbb{C}^3}\right) dx \\
&\quad + \int_D (\mathbf{J}_0(x), \mathbf{l}_0(x) - \hat{\mathbf{Z}}(x))_{\mathbb{C}^3}\, dx \\
&= -\overline{a(\mathbf{P}, \hat{\mathbf{P}})} - (C_1 \hat{\mathbf{E}}, Q_1 C_1 \mathbf{P})_{H_0} - (C_2(\tfrac{1}{i\omega\mu}\mathrm{rot}\,\hat{\mathbf{E}}), Q_2 C_2(\tfrac{1}{i\omega\mu}\mathrm{rot}\,\mathbf{P}))_{H_0} \\
&\quad + \int_D (\mathbf{J}_0(x), \mathbf{l}_0(x) - \hat{\mathbf{Z}}(x))_{\mathbb{C}^3}\, dx.
\end{aligned} \tag{95}$$

But from (84) and (49) it follows $a(\mathbf{P}, \hat{\mathbf{P}}) = \int_D (Q^{-1}(\mathbf{l}_0 - \hat{\mathbf{Z}}), \hat{\mathbf{P}})_{\mathbb{C}^3}\, dx$,

$$\begin{aligned}
\overline{a(\mathbf{P}, \hat{\mathbf{P}})} &= \int_D \overline{(Q^{-1}(\mathbf{l}_0 - \hat{\mathbf{Z}}), \hat{\mathbf{P}})_{\mathbb{C}^3}}\, dx = \int_D \overline{(\mathbf{l}_0 - \hat{\mathbf{Z}}, Q^{-1}\hat{\mathbf{P}})_{\mathbb{C}^3}}\, dx = \int_D (Q^{-1}\hat{\mathbf{P}}, \mathbf{l}_0 - \mathbf{Z})_{\mathbb{C}^3} dx \\
a(\hat{\mathbf{E}}, \hat{\mathbf{Z}}) &= \int_D (Q^{-1}\hat{\mathbf{P}}, \hat{\mathbf{Z}})_{\mathbb{C}^3}\, dx - \int_D (\mathbf{J}_0, \hat{\mathbf{Z}})_{\mathbb{C}^3}\, dx, \\
a^*(\hat{\mathbf{Z}}, \hat{\mathbf{E}}) &= \int_D \left(-\left(\frac{1}{i\omega\bar{\mu}} C_2^* \Lambda_{H_0} Q_2 C_2 \left(\frac{1}{i\omega\mu}\mathrm{rot}\,\mathbf{P}\right), \mathrm{rot}\,\hat{\mathbf{E}}\right)_{\mathbb{C}^3} + (C_1^* \Lambda_{H_0} Q_1 C_1 \mathbf{P}, \hat{\mathbf{E}})_{\mathbb{C}^3}\right) dx.
\end{aligned}$$

From the equality $\overline{a^*(\hat{\mathbf{Z}},\hat{\mathbf{E}})} = a(\hat{\mathbf{E}},\hat{\mathbf{Z}})$, we obtain

$$\overline{a(\mathbf{P},\hat{\mathbf{P}})} = -a(\hat{\mathbf{E}},\hat{\mathbf{Z}}) - \int_D (\mathbf{J}_0,\hat{\mathbf{Z}})_{\mathbb{C}^3}\,dx + \int_D (Q^{-1}\hat{\mathbf{P}},\mathbf{l}_0)_{\mathbb{C}^3}\,dx$$

$$= -\overline{a^*(\hat{\mathbf{Z}},\hat{\mathbf{E}})} - \int_D (\mathbf{J}_0,\hat{\mathbf{Z}})_{\mathbb{C}^3}\,dx + \int_D (Q^{-1}\hat{\mathbf{P}},\mathbf{l}_0)_{\mathbb{C}^3}\,dx$$

$$- \int_D \left(\frac{1}{i\omega\mu}\operatorname{rot}\hat{\mathbf{E}}, C_2^*\Lambda_{H_0}Q_2C_2\left(\frac{1}{i\omega\mu}\operatorname{rot}\mathbf{P}\right)\right)_{\mathbb{C}^3} dx - \int_D (\hat{\mathbf{E}}, C_1^*\Lambda_{H_0}Q_1C_1\mathbf{P})_{\mathbb{C}^3}\,dx$$

$$- \int_D (\mathbf{J}_0,\hat{\mathbf{Z}})_{\mathbb{C}^3}\,dx + \int_D (Q^{-1}\hat{\mathbf{P}},\mathbf{l}_0)_{\mathbb{C}^3}\,dx. \tag{96}$$

The representation (94) follows from (95) and (96). □

Remark 2. Notice that in representation $l(\hat{\mathbf{J}})$ for minimax estimate $\widehat{\widehat{l(\mathbf{J})}}$, the function $\hat{\mathbf{J}} = \mathbf{J}_0 - Q^{-1}\hat{\mathbf{P}}$, where $\hat{\mathbf{P}}$ is defined from equations (48) and (49), can be taken as a good estimate for unknown function $\mathbf{J}$ entering the right-hand side of Eq. (7) (for explanations, see Remark 1).

Remark 3. If $\varepsilon = \varepsilon(x) = \text{const} > 0$ and $\mu = \mu(x) = \text{const} > 0$ in D and $k^2 = \varepsilon\mu\omega^2$ is not an eigenvalue of the interior Maxwell problem, then all the results obtained above are also valid.

6 Conclusion

We have developed analytical and numerical techniques for finding guaranteed estimates of solutions and right-hand sides of Maxwell equations from observations that depend on the same solutions and boundary data. The results can be applied to modeling and analysis of data processing systems, processing and interpretation of electromagnetic observations of various nature, and solution to inverse problems with noisy data. The technique can be extended to the estimation of exterior Maxwell problems.

Acknowledgements This work is supported by the Visby program of the Swedish Institute.

References

1. Krasovskii, N.N.: Theory of Motion Control. Nauka, Moscow (1968)
2. Kurzhanskii, A.B.: Control and Observation under Uncertainties. Nauka, Moscow (1977)
3. Kurzhanskii, A.B.: Dynamic control system estimation under uncertainty conditions. Probl. Control Inform. Theory **9**(6), 395–401 (1980)
4. Kurzhanskii, A.B.: Dynamic control system estimation under uncertainty conditions. Probl. Control Inform. Theory **10**(1), 33–48 (1981)

5. Nakonechnyi, O.G.: Minimax Estimation of Functionals of Solutions to Variational Equations in Hilbert Spaces. Kiev State University, Kiev (1985)
6. Nakonechnyi, O.G.: Optimal Control and Estimation for Partial Differential Equations. Kyiv University, Kyiv (2004)
7. Nakonechnyi, O.G.: Minimax Estimates in Systems with Distributed Parameters. Acad. Sci. USSR, Inst. Cybernetics, Kyiv (1979), pp. 1–55 Preprint 79
8. Bensoussan, A.: Filtrage Optimale les Systems Lineares. Dunod, Paris (1971)
9. Bencala K.E., Seinfield J.H.: Distributed parameter filtering: Boundary noise and discrete observations. Int. J. Syst. Sch. **10**(5), 493–512 (1979)
10. Shestopalov, Y., Podlipenko, Y., Prishlyak V.: Estimation of solutions of Helmholtz problems with uncertain data. In Proceedings 2010 International Symposium on Electromagnetic Theory EMTS 2010, Berlin, August 15–19, 2010 (pp. 621–623)
11. Shestopalov, Y., Podlipenko, Y., Prishlyak V.: Estimation under uncertainties of acoustic and electromagnetic fields from noisy observations. arXiv:0910.2331 (2009)
12. Nakonechnyi, O.G., Pavluchenko, O.G., Podlipenko, Yu.K.: On prediction of solutions to hyperbolic equations. Probl. Upravl. Inform. **1**, 98–113 (1995)
13. Kirichenko, N.F., Nakonechnyi, O.G.: A minimax approach to recurrent estimation of the states of linear dynamical systems. Kibernetika **4**, 52–55 (1977)
14. Podlipenko, Yu.K., Ryabikova, A.V.: Optimal estimation of parameters of noeter boundary value problems for linear ordinary differential equations of order n under uncertainties. Dopovidi Acad. Nauk Ukrainy **11**, 59–67 (2005)
15. Podlipenko, Yu.K., Grishchuk, N.V.: Minimax estimation of solutions to degenerated Neumann boundary value problems for elliptic equations. Syst. Dosl. Inf. Techn. **2**, 104–128 (2004)
16. Cessenat, M.: Mathematical Methods in Electromagnetism. Linear Theory and Applications. World Scientific, Singapore (1996)
17. Lions, J.L.: Contrôle Optimal de Systèmes Gouvernés par des Équations Aux Dérivées Partielles. Dunod, Paris (1968)
18. Schwab, C., Gittelson, C.J.: Sparse tensor discretization of high-dimensional parametric and stochastic PDEs. Acta Numerica. doi: 10.1017/S0962492911000055

Permittivity Reconstruction of Layered Dielectrics in a Rectangular Waveguide from the Transmission Coefficients at Different Frequencies

Yu. G. Smirnov, Yu. V. Shestopalov, and E. D. Derevyanchuk

Abstract Determination of electromagnetic parameters of dielectric bodies of complicated structure is an urgent problem. However, as a rule, these parameters cannot be directly measured (because of composite character of the material and small size of samples), which leads to the necessity of applying methods of mathematical modeling and numerical solution of the corresponding forward and inverse electromagnetic problems. It is especially important to develop the solution techniques when the inverse problem for bodies of complicated shape is considered in the resonance frequency range. In this paper we develop a method of solution to the inverse problem of reconstructing (complex) permittivity of layered dielectrics in the form of diaphragms in a waveguide of rectangular cross section from the transmission coefficients measured at different frequencies. The method enables in particular obtaining solutions in a closed form in the case of one-sectional diaphragm. In the case of an n-sectional diaphragm we solve the inverse problem using numerical solution of a nonlinear equation system of n complex variables. Solvability and uniqueness of the system are studied and convergence of the method is discussed. Numerical results of calculating (complex) permittivity of the layers are presented. The case of metamaterials is also considered. The results of solution to the inverse problem can be applied in nanotechnology, optics, and design of microwave devices.

Yu.G. Smirnov (✉) • E.D. Derevyanchuk
Penza State University, Penza, Russia
e-mail: smirnovyug@mail.ru; catherinderevyanchuk@mail.ru

Yu.V. Shestopalov
Karlstad University, Karlstad, Sweden
e-mail: youri.shestopalov@kau.se

L. Beilina and Y.V. Shestopalov (eds.), *Inverse Problems and Large-Scale Computations*, Springer Proceedings in Mathematics & Statistics 52, DOI 10.1007/978-3-319-00660-4_11,

1 Introduction

Determination of electromagnetic parameters of dielectric bodies that have complicated geometry or structure is an urgent problem arising, e.g., when nanocomposite or artificial materials and media are used as elements of various devices. However, as a rule, these parameters cannot be directly measured (because of composite character of the material and small size of samples), which leads to the necessity of applying methods of mathematical modeling and numerical solution of the corresponding forward and inverse electromagnetic problems. It is especially important to develop the solution techniques when the inverse problem for bodies of complicated shape is considered in the resonance frequency range, which is the case when permittivity of nanocomposite materials must be reconstructed [19, 20].

One of possible applications of composites is the creation of radio absorbing materials that can be used in systems that provide electromagnetic compatibility of modern electronic devices and in 'Stealth'-type systems aimed at damping and decreasing reflectivity of microwave electromagnetic radiation from objects to be detected [16, 17]. When calculating reflection and absorption characteristics of electromagnetic microwave radiation of radio absorbing materials, researchers use models employing the data on the material constants (permittivity, permeability, conductivity) of these materials in the microwave range. Such composites often contain carbon particles, short carbon fibers, carbon nanofibers, and multilayer carbon nanotubes as fillers for polymer dielectric matrices [12, 16, 17]. The use of carbon nanotubes enables one to achieve a significant (up to 10 dB) absorption of microwave electromagnetic radiation at relatively thin composite layers and low volume fractions of nanotubes and hence a small weight, in a broad frequency range (up to 5 GHz). Such characteristics are caused by both the geometrical sizes of individual nanotubes and their electrophysical properties; among the most important parameters here are permittivity and electric conductivity (which can vary over very wide ranges).

It is important to determine permittivity and conductivity not only of a composite as a solid body (as in [16, 17]) but also of its components, e.g., nanotubes, whose physical characteristics can vary substantially in the process of composite formation.

The forward scattering problem for a diaphragm in a parallel-plane waveguide was considered in [13]. In papers [1, 3, 7, 8, 15, 22–24] the inverse problem of reconstructing complex permittivity was analyzed from the measurements of the transmission coefficient; in [7, 8, 14] the artificial neural networks method was applied.

Several techniques for the permittivity determination of homogeneous materials loaded in a waveguide are reported [1, 3, 6]. The permittivity reconstruction of inhomogeneous structures is not as widely investigated, and only a few studies exist for multilayered materials [2, 9]. Note a recently developed advanced approach [5] that can be also applied to numerical solution of this inverse problem.

However, the solution in closed form to the inverse problem of permittivity determination of materials loaded in a waveguide is not available in the literature,

to the best of our knowledge, even for the simplest configuration of a parallel-plane dielectric insert in a guide of rectangular cross section. This fact dictates the aim of this work: to develop a method of solution to the inverse problem of reconstructing effective permittivity of layered dielectrics in the form of diaphragms in a waveguide of rectangular cross section that would enable both obtaining solution in a closed form for benchmark problems and efficient numerical implementation. We note that the corresponding forward problem for a one-sectional diaphragm is considered in [10] and [21].

2 Statement of the Problem

Assume that a waveguide $P = \{x : 0 < x_1 < a, 0 < x_2 < b, -\infty < x_3 < \infty\}$ with the perfectly conducting boundary surface ∂P is given in Cartesian coordinate system. A three-dimensional body Q $(Q \subset P)$

$$Q = \{x : 0 < x_1 < a, 0 < x_2 < b, 0 < x_3 < l\}$$

is placed in the waveguide; the body has the form of a diaphragm (an insert), namely, a parallelepiped separated into n sections adjacent to the waveguide walls (Fig. 1). Domain $P \backslash \bar{Q}$ is filled with an isotropic and homogeneous layered medium having constant permeability $(\mu_0 > 0)$ in whole waveguide P; the sections of the diaphragm

$$\begin{aligned} Q_0 &= \{x : 0 < x_1 < a, 0 < x_2 < b, -\infty < x_3 < 0\} \\ Q_j &= \{x : 0 < x_1 < a, 0 < x_2 < b, l_{j-1} < x_3 < l_j\}, j = 1, \ldots, n \\ Q_{n+1} &= \{x : 0 < x_1 < a, 0 < x_2 < b, l < x_3 < +\infty\} \end{aligned}$$

are filled each with a medium having constant permittivity $\varepsilon_j > 0$; $l_0 := 0$, $l_n := l$.

The electromagnetic field inside and outside of the object in the waveguide is governed by Maxwells' equations with harmonic dependence on the time:

$$\begin{aligned} \mathrm{rot}\,\mathbf{H} &= -i\omega\varepsilon\mathbf{E} + \mathbf{j}_E^0 \\ \mathrm{rot}\,\mathbf{E} &= i\omega\mu_0\mathbf{H}, \end{aligned} \tag{1}$$

where $\mathbf{E}$ and $\mathbf{H}$ are the vectors of the electric and magnetic field intensity, $\mathbf{j}$ is the electric polarization current, and ω is the circular frequency.

Assume that $\pi/a < k_0 < \pi/b$, where k_0 is the wavenumber, $k_0^2 = \omega^2\varepsilon_0\mu_0$ [11]. In this case, only one wave H_{10} propagates in the waveguide without attenuation (we have a single-mode waveguide [11]).

The incident electrical field is

$$\mathbf{E}^0 = \mathbf{e}_2 A \sin\left(\frac{\pi x_1}{a}\right) e^{-i\gamma_0 x_3} \tag{2}$$

with a known A and $\gamma_0 = \sqrt{k_0^2 - \pi^2/a^2}$.

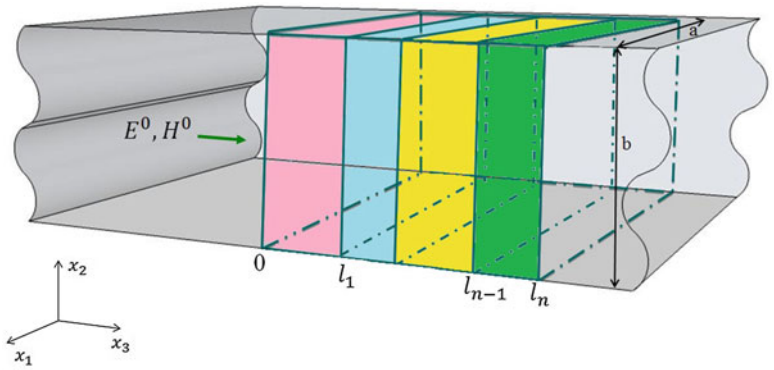

Fig. 1 Multilayered diaphragms in a waveguide

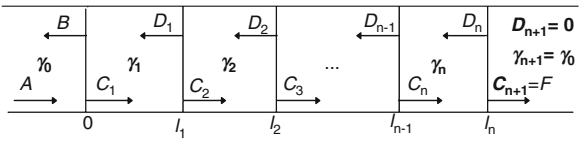

Fig. 2 Scheme of the propagation of wave through the diaphragms

Solving the forward problem for Maxwell's equations with the aid of (1) and the propagation scheme in Fig. 1, we obtain explicit expressions for the field inside every section of diaphragm Q and outside the diaphragm (Fig. 2):

$$E_{(0)} = \sin\left(\frac{\pi x_1}{a}\right)\left(Ae^{-i\gamma_0 x_3} + Be^{i\gamma_0 x_3}\right), \quad x \in Q_0, \tag{3}$$

$$E_{(j)} = \sin\left(\frac{\pi x_1}{a}\right)\left(C_j e^{-i\gamma_j x_3} + D_j e^{i\gamma_j x_3}\right), \tag{4}$$

$$j = 1, \ldots, n+1; \qquad D_{n+1} = 0, \quad x \in Q_j,$$

where $\gamma_j = \sqrt{k_j^2 - \pi^2/a^2}$ and $k_j^2 = \omega^2 \varepsilon_j \mu_0$, $\gamma_{n+1} = \gamma_0$.

From the conditions on the boundary surfaces of the diaphragm sections

$$[E_{(j)}] = [E_{(j+1)}] = 0; [\frac{\partial E_{(j)}}{\partial x_3}] = [\frac{\partial [E_{(j+1)}]}{\partial x_3}] = 0, \; j = 0, \ldots, n+1, \tag{5}$$

where square brackets [] denote function jump via boundary surfaces, of applied to (3) and (4), we obtain using conditions (5) a system of equations for the unknown coefficients (Fig. 2).

$$\begin{cases} A+B=C_1+D_1 \\ \gamma_0(B-A)=\gamma_1(D_1-C_1) \\ C_j e^{-i\gamma_j l_j}+D_j e^{i\gamma_j l_j}=C_{j+1}e^{-i\gamma_{j+1}l_j}+D_{j+1}e^{i\gamma_{j+1}l_j} \\ \gamma_j(D_j e^{i\gamma_j l_j}-C_j e^{-i\gamma_j l_j})=\gamma_{j+1}(D_{j+1}e^{i\gamma_{j+1}l_j}-C_{j+1}e^{-i\gamma_{j+1}l_j}),\ \ j=1,\ldots,n, \end{cases} \tag{6}$$

where $C_{n+1}=F$, $D_{n+1}=0$. In system (6), coefficients $A, B, C_j, D_j, \varepsilon_j, (j=1,\ldots,n)$ are supposed to be complex.

We can express C_j, D_j from C_{j+1}, D_{j+1} in order to obtain a recurrent formula that couples amplitudes A and F.

We obtained the formula

$$A=\frac{1}{2\prod\limits_{j=0}^{n}\gamma_j}(\gamma_n p_{n+1}+\gamma_0 q_{n+1})Fe^{-i\gamma_0 l_n}, \tag{7}$$

where p and q are denoted by such recurrent formulas

$$p_{j+1}=\gamma_{j-1}p_j\cos\alpha_j+\gamma_j q_j i\sin\alpha_j;\quad p_1:=1, \tag{8}$$

$$q_{j+1}=\gamma_{j-1}p_j i\sin\alpha_j+\gamma_j q_j\cos\alpha_j;\quad q_1:=1. \tag{9}$$

Here $\alpha_j=\gamma_j(l_j-l_{j-1})$, $j=1,\ldots,n$. Note that similar formulas are obtained in classical monographs dealing with wave propagation in layered media, e,g, in [4].

3 Inverse Problem for Multisectional Diaphragm

Formulate the inverse problem for a multisectional diaphragm that will be addressed in this work.

Inverse problem P: *find (complex) permittivity ε_j of each section from the known amplitude of the incident wave and amplitude F of the transmitted wave at different frequencies.*

It is reasonable to consider the right-hand side of (7) as a complex-valued function with respect to n variables ε_j. For n sections we must know amplitudes A and F for each of n frequency values to have a consistent system of n equations with respect to n unknown permittivity values ε_j. This system is then solved to obtain the sought-for permittivity values.

Let us rewrite Eq. (7) in the form

$$G(h)=H,\quad H:=\frac{2A\gamma_0 e^{i\gamma_0 l_n}}{F}, \tag{10}$$

where

$$G(h) := \frac{1}{\prod_{j=1}^{n} \gamma_j} (\gamma_n p_{n+1} + \gamma_0 q_{n+1}) \tag{11}$$

and $h := (\varepsilon_1, \ldots, \varepsilon_n)$.

We will consider (11) as a complex function of n complex variables. It follows from (8) and (9) that

$$\begin{pmatrix} p_{j+1} \\ q_{j+1} \end{pmatrix} = \begin{pmatrix} \cos\alpha_j & i\sin\alpha_j \\ i\sin\alpha_j & \cos\alpha_j \end{pmatrix} \begin{pmatrix} \gamma_{j-1} & 0 \\ 0 & \gamma_j \end{pmatrix} \begin{pmatrix} p_j \\ q_j \end{pmatrix} \tag{12}$$

$(j = 1, \ldots, n)$. Thus we can represent p_{n+1}, q_{n+1} via finite multiplication of matrices by formula (12). From representation (12) we select, for every fixed j, only the matrices depending on γ_j . Finally we obtain

$$\begin{pmatrix} \gamma_j & 0 \\ 0 & \gamma_{j+1} \end{pmatrix} \begin{pmatrix} \cos\alpha_j & i\sin\alpha_j \\ i\sin\alpha_j & \cos\alpha_j \end{pmatrix} \begin{pmatrix} \gamma_{j-1} & 0 \\ 0 & \gamma_j \end{pmatrix} = \begin{pmatrix} \gamma_j\gamma_{j-1}\cos\alpha_j & i\gamma_j^2\sin\alpha_j \\ i\gamma_{j+1}\gamma_{j-1}\sin\alpha_j & \gamma_j\gamma_{j+1}\cos\alpha_j \end{pmatrix} \tag{13}$$

Dividing matrix (13) by γ_j we have

$$\begin{pmatrix} \gamma_{j-1}\cos\alpha_j & i\gamma_j\sin\alpha_j \\ i\gamma_{j+1}\gamma_{j-1}\sin\alpha_j/\gamma_j & \gamma_{j+1}\cos\alpha_j \end{pmatrix} \tag{14}$$

Taking into account Taylor series for functions $\sin\alpha_j$ and $\cos\alpha_j$ and that $\alpha_j = \gamma_j(l_j - l_{j-1})$ (14) we see that each coefficient of this matrix depends on γ_j^2. Since $\gamma_j^2 = \varepsilon_j\mu_0\omega^2 - \pi^2/a^2$ we have that each coefficient of matrix (14) is an analytical function w.r.t. ε_j. Hence function $G(h)$ depends on ε_j analytically for every j $(j = 1, \ldots, n)$.

Using Hartogs theorem [18] we obtain the following statement.

Theorem 1. *$G(h)$ is holomorphic on $\mathbf{C}^n$ as a function of n complex variables.*

Let us formulate inverse problem P for n-sectional diaphragm in the following form. Consider n different frequencies $\Omega = (\omega_1, \ldots, \omega_n)$ and functions $G_j(h) := G(h, \omega_j)$, $j = 1, \ldots, n$. It is necessary to find a solution to the (nonlinear) system of n equations w.r.t. n variables $\varepsilon_1, \ldots, \varepsilon_n$:

$$G_j(h) = H_j, \quad H_j = H(\omega_j), \quad j = 1, \ldots, n. \tag{15}$$

Theorem 1 implies [18]

Theorem 2. *If Jacobian* $\frac{\partial(G_1,\ldots,G_n)}{\partial(h_1,\ldots,h_n)} \neq 0$ *at the point* h^**, then function* $G(h)$ *is locally invertible in a vicinity of* h^**, and inverse problem P has unique solution for every h from that vicinity.*

4 One-Sectional Diaphragm: Explicit Solution to the Inverse Problem

From (7) for a one-sectional diaphragm we have

$$\begin{aligned} \frac{Ae^{i\gamma_0 l_1}}{F} &= g(z), \\ g(z) &= \cos z + i\left(\frac{z}{2\gamma_0 l_1} + \frac{\gamma_0 l_1}{2z}\right)\sin z, \\ z &= \gamma_1 l_1 = l_1\sqrt{k_1^2 - \frac{\pi^2}{a^2}}, \end{aligned} \tag{16}$$

where z is generally a complex variable. From (16), we obtain a relation for the transmission coefficient

$$F = \frac{Ae^{i\gamma_0 l_1}}{g(z)}, \tag{17}$$

which, together with formulas (3) and (4), gives an explicit solution to the forward problem under study.

When the inverse problem is solved, ε_1 is considered as an unknown quantity that should be determined from Eq. (16) in terms of F.

List the most important properties of $g(z)$ which easily follows from its explicit representation:

(i) $g(z)$ is an entire function.
(ii) $g(z)$ has neither real zeros nor poles. This fact is in line with physical requirements that the transmission coefficient does not vanish and is a bounded quantity at real frequencies.
(iii) $g(z)$, also considered as a function of real τ, is not invertible locally at the origin because it is easy to check that $g'(0) = 0$. Next, the inverse of $g(z)$ is a multi-valued function. In fact, the inverse function does not exist globally according to the statement in Remark concerning violation of uniqueness.
(iv) $g(z)$ is not a fractional-linear function; therefore, $g(z)$ performs one-to-one conformal mappings only of certain regions of the complex plane onto regions of the complex plane.
(v) It is easy to check up that $g'(\tau) \neq 0$ for (real) $\tau \neq 0$. Hence, $g(z)$ is invertible locally at the real point $\tau \neq 0$.

Assuming that ε_1 is real it is reasonable to introduce a real variable

$$\tau = \gamma_1 l_1 = l_1 \sqrt{k_1^2 - \frac{\pi^2}{a^2}} > 0 \tag{18}$$

which may be used for parametrization. Extract the real and imaginary part of $g(\tau)$, denoting them by x and y,

$$\begin{cases} x = \cos\tau, \\ y = h(\tau)\sin\tau, \end{cases} \qquad \text{where } h(\tau) = \frac{\tau}{2C} + \frac{C}{2\tau},\ C = \gamma_0 l_1. \tag{19}$$

Equation (16) is equivalent to the system

$$\begin{cases} \cos\tau = p, & p = \mathrm{Re}\left(\frac{Ae^{-i\gamma_0 l_1}}{F}\right), \\ h(\tau)\sin(\tau) = q, & q = \mathrm{Im}\left(\frac{Ae^{-i\gamma_0 l_1}}{F}\right), \end{cases} \tag{20}$$

where p and q are known values. Using the results of Appendix I we finally obtain from (20) an explicit formula for the sought (real) permittivity

$$\widetilde{\varepsilon_1} = \frac{1}{\omega^2 \mu_0 \varepsilon_0} \left(\left(\frac{\pi}{a}\right)^2 + \left(\frac{\tau}{l_1}\right)^2 \right), \tag{21}$$

where

$$\tau = \tau_1 = C \left(\frac{|q| + \sqrt{p^2 + q^2 - 1}}{\sqrt{1 - p^2}} \right) \tag{22}$$

when $\tilde{\varepsilon_1} > 1$ and

$$\tau = \tau_2 = C \left(\frac{\sqrt{1 - p^2}}{|q| + \sqrt{p^2 + q^2 - 1}} \right) \tag{23}$$

when $\frac{\pi^2}{a^2 k_0^2} < \tilde{\varepsilon_1} < 1$.

Here $\tilde{\varepsilon_1}$ is a relative permeability. Formulas (21)–(23) constitute explicit solution of inverse problem P under study.

Using the reasoning and results of Appendix I we prove the following result stating the existence and uniqueness of solution to the inverse problem of finding permittivity of a one-sectional diaphragm in a waveguide of rectangular cross section.

Theorem 3. *Assume that* $|p| < 1$ *and* $p^2 + q^2 \geq 1$. *Then inverse problem P has only one solution expressed by (22) if* $\frac{\tau_1}{C} > 1$, $\cos\tau_1 = p$, *and* $\mathrm{sign}(q) = \mathrm{sign}(\sin(\tau_1))$. *If* $\frac{\tau_2}{C} < 1$, $\cos\tau_2 = p$, *and* $\mathrm{sign}(q) = \mathrm{sign}(\sin(\tau_2))$, *inverse problem P has only one solution expressed by (23). Otherwise, inverse problem P has no solution.*

Remark 1. If $p = 1$, then q must be equal zero and $\tau = 2\pi n, n \in Z$. If $p = -1$, then q must be equal to zero and $\tau = \pi + 2\pi n$, $n \in Z$. In these cases inverse problem P has infinitely many solutions; therefore, they are excluded from the theorem.

5 One-Sectional Thin Diaphragm

In the case of thin diaphragms matrix (14) transforms to the following matrix:

$$\begin{pmatrix} \gamma_{j-1} & i\gamma_j^2(l_j - l_{j-1}) \\ i\gamma_{j+1}\gamma_{j-1}(l_j - l_{j-1}) & \gamma_{j+1} \end{pmatrix} \tag{24}$$

It allows us to obtain simple formulas for approximate solution of the inverse problem.

Consider the case of a thin one-sectional diaphragm, i.e, $l_1 \ll 1$ (Fig. 3). Then Eq. (16) can be approximated by the equation

$$g_0(w) := 1 + \frac{il_1}{2}\left(\frac{w}{\gamma_0} + \gamma_0\right) - \frac{Ae^{i\gamma_0 l_1}}{F} = 0, \tag{25}$$

where $w := \gamma_1^2 = k_1^2 - \frac{\pi^2}{a^2}$. The root of Eq. (25) w.r.t. w is

$$w_0 = \gamma_0\left(\frac{2i}{l_1}\left(1 - \frac{A}{F}e^{i\gamma_0 l_1}\right) - \gamma_0\right) \tag{26}$$

which yields

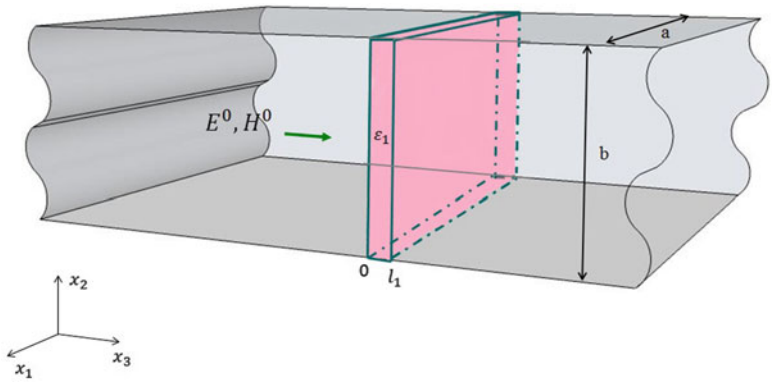

Fig. 3 One-sectional thin diaphragm

$$\tilde{\varepsilon_1} = \left(\gamma_0 \left(\frac{2i}{l_1}\left(1 - \frac{A}{F}e^{i\gamma_0 l_1}\right) - \gamma_0\right) + \frac{\pi^2}{a^2}\right) k_0^{-2} \tag{27}$$

Below we present an example of numerical solutions to inverse problem P for a one-sectional thin diaphragm. The table shows the test results of numerical solution to the inverse problem of reconstructing permittivities of a one-sectional diaphragm. The test values of the transmission coefficient are taken from the solution to the forward problem.

Parameters of the one-sectional diaphragm are $a = 2\,\text{cm}$, $b = 1\,\text{cm}$, $c = 2\,\text{cm}$, and $\tilde{\varepsilon_1} = 1.9$ (exact value used in the solution to forward problem). The first, second, and third calculation triples are performed, respectively, at $f = 11.94\,\text{GHz}$, $f = 8.12\,\text{GHz}$, and $f = 9.55\,\text{GHz}$.

We see that in all examples the error of computations does not exceed 3 % which proves high efficiency of the method.

Value of $\frac{F}{A}$	$\tilde{\varepsilon_1}$	Calculated $\tilde{\varepsilon_1}$	l_1, cm
$0.929 - 0.26i$	$1.772 + 0.567i$	$1.901876 - 0.00094i$	0.2
$0.98 - 0.14i$	$1.865 + 0.289i$	$1.8974 - 0.000824i$	0.1
$0.999 - 0.029i$	$1.903 + 0.054i$	$1.90382 - 0.00933i$	0.02
$0.866 - 0.341i$	$1.877 + 0.134i$	$1.9009 - 0.00041i$	0.2
$0.962 - 0.192i$	$1.895 + 0.069i$	$1.895 + 0.069i$	0.1
$0.998 - 0.04i$	$1.902 + 0.0047i$	$1.9021 - 0.00893i$	0.02
$0.926 - 0.263i$	$1.843 + 0.312i$	$1.89957 - 0.000493i$	0.2
$0.98 - 0.142i$	$1.887 + 0.162i$	$1.9017868 + 0.0033i$	0.1
$0.999 - 0.029i$	$1.898 + 0.0027i$	$1.8988 - 0.0049i$	0.02

6 Conclusion

We have developed a numerical-analytical method of solution to the inverse problem of reconstructing permittivities of n-sectional diaphragms in a waveguide of rectangular cross section. For a one-sectional diaphragm, a solution in the closed form is obtained and the uniqueness is proved. These results make it possible to use the case of a one-sectional diaphragm in a waveguide of rectangular cross section as a benchmark test problem and perform a complete analysis of the inverse scattering problem for arbitrary n-sectional diaphragms.

Acknowledgements This work is partially supported by Russian Foundation of Basic Research 11-07-00330-a and Visby Program of the Swedish Institute.

Appendix 1

Reduce Eq. (16) to a quadratic equation. From (18) it follows that (on the domain of all the functions involved)

$$p^2 + \frac{q^2}{h^2(\tau)} = 1, \quad h(\tau) > 0.$$

From (20), we obtain

$$h^2(\tau) = \frac{q^2}{1-p^2}, \quad |p| < 1. \tag{28}$$

Then

$$h(\tau) = Q, \quad h(\tau) := \frac{\tau}{2C} + \frac{C}{2\tau}, \quad Q := \frac{|q|}{\sqrt{1-p^2}} > 0, \tag{29}$$

and we obtain a quadratic equation

$$\tau^2 - 2CQ\tau + C^2 = 0 \tag{30}$$

which has the roots

$$\tau_1 = C(Q + \sqrt{Q^2-1}), \quad \tau_2 = \frac{C}{Q + \sqrt{Q^2-1}}. \tag{31}$$

$\tau_{1,2}$ are real if $Q \geq 1$; therefore,

$$p^2 + q^2 \geq 1. \tag{32}$$

Inequality (32) constitutes the existence condition for the solution of equation (16). Since $\tau = \gamma_1 l_1$ and $C = \gamma_0 l_1$, we have

$$\frac{\tau}{C} = \frac{\gamma_1}{\gamma_0} = \frac{\sqrt{\omega^2\mu_0\varepsilon_1 - \frac{\pi^2}{a^2}}}{\sqrt{\omega^2\mu_0\varepsilon_0 - \frac{\pi^2}{a^2}}},$$

so that, in view of the assumption $\varepsilon_1 > \varepsilon_0$,

$$\frac{\tau}{C} > 1.$$

Similarly, for $\frac{\pi^2}{a^2\omega^2\mu_0} < \varepsilon_1 < \varepsilon_0$,

$$\frac{\tau}{C} < 1.$$

Thus, for $\varepsilon_1 > \varepsilon_0$ we obtain

$$\frac{\tau_1}{C} = Q + \sqrt{Q^2 - 1} \quad (> 1).$$

For $\frac{\pi^2}{a^2\omega^2\mu_0} < \varepsilon_1 < \varepsilon_0$,

$$\frac{\tau_2}{C} = \frac{1}{Q + \sqrt{Q^2 - 1}} \quad (< 1).$$

Thus, when $\varepsilon_1 > \varepsilon_0$ Eq. (29) has only one root (28) τ_1. Similarly, Eq. (31) has the only one root (29) τ_2 for $\frac{\pi^2}{a^2\omega^2\mu_0} < \varepsilon_1 < \varepsilon_0$.

It should be noted that reduction of (16) to quadratic equation (30) is not an equivalent transformation. It is necessary to complement (30) with one of the equations of system (19), for example, with the first, and take into accounts the signs of p and q. As a result, (16) will be equivalent to the system

$$\begin{cases} \cos\tau = p, \ \text{sign}(q) = \text{sign}(\sin(\tau)), \\ p^2 + \frac{q^2}{h^2(\tau)} = 1. \end{cases}$$

References

1. Akleman, F.: Reconstruction of complex permittivity of a longitudinally inhomogeneous material loaded in a rectangular waveguide. IEEE Microw. Wireless Compon. Lett. **18**(3), 158–160 (2008)
2. Baginski, M.E., Faircloth, D.L., Deshpande M.D.: Comparison of two optimization techniques for the estimation of complex permittivities of multilayered structures using waveguide measurements. IEEE Trans. Microw. Theory Tech. **53**(10), 3251–3259 (2005)
3. Baker-Jarvis, J., Vanzura, E.J.: Improved technique for determining complex permittivity with the transmission-reflection method. IEEE Trans. Microw. Theory Tech. **38**(8), 1096–1103 (1990)
4. Brekhovskih, L.: Waves in Layered Media. Academic Press, New York (1980)
5. Beilina, L., Klibanov, M.: Approximate Global Convergence and Adaptivity for Coefficient Inverse Problems. Springer, New York (2012)
6. Dediu, S., McLaughlin, J.R.: Recovering inhomogeneities in a waveguide using eigensystem decomposition. Inv. Prob. **22**(4), 1227–1246 (2006)
7. Eves, E.E., Kopyt, P., Yakovlev, V.V.: Determination of complex permittivity with neural networks and FDTD modeling. Microwave Opt. Tech. Let. **40**(3), 183–188 (2004)
8. Eves, E.E., Kopyt, P., Yakovlev, V.V.: Practical aspects of complex permittivity reconstruction with neural-network-controlled FDTD modeling of a two-port fixture. J. Microw. Power Electromagn. Energ. **41**(4), 81–94 (2007)
9. Faircloth, D.L., Baginski, M.E., Wentworth, S.M.: Complex permittivity and permeability extraction for multilayered samples using S-parameter waveguide measurements. IEEE Trans. Microw. Theory Tech. **54**(3), 1201–1209 (2006)

10. Grishina, E.E, Derevyanchuk, E.D., Medvedik, M.Y, Smirnov, Y.G.: Numerical and analytical solutions of electromagnetic field diffraction on the two-sectional body with different permittivity in the rectangular waveguide. Izv. Vyssh. Uchebn. Zaved. Povolzh. Region, Fiz.-Mat. Nauki **4**, 73–81 (2010)
11. Jackson, J.D.: Classical Electromagnetics. Wiley, New York (1967)
12. Kelly, J.M., Stenoien, J.O., Isbell, D.E.: Wave-guide measurements in the microwave region on metal powders suspended in paraffin wax. J. Appl. Phys. **24**(3), 258–262 (1953)
13. Lurie, K.A., Yakovlev, V.V.: Optimization of electric field in rectangular waveguide with lossy layer. IEEE Trans. Magn. **36**, 1094–1097 (2000)
14. Lurie, K.A., Yakovlev, V.V.: Control over electric field in traveling wave applicators. J. Eng. Math. **44**(2), 107–123 (2002)
15. Outifa, L., Delmotte, M., Jullien, H.: Dielectric and geometric dependence of electric field and power distribution in a waveguide heterogeneously filled with lossy dielectrics. IEEE Trans. Microw. Theor. Tech. **45**(1), 1154–1161 (1997)
16. De Rosa, I.M., Dinescu, A., Sarasini, F., Sarto, M.S., Tamburrano, A.: Effect of short carbon fibers and MWCNTs on microwave absorbing properties of polyester composites containing nickel-coated carbon fibers. Composites Sci. Techn. **70**, 102–109 (2010)
17. Saib, A., Bednarz, L., Daussin, R., Bailly, C., Lou, X., Thomassin, J.-M., Pagnoulle, C., Detrembleur, C., Jerome, R., Huynen, I.: Carbon nanotube composites for broadband microwave absorbing materials. IEEE Trans. Microw. Theor. Tech. **54**(6), 2745–2754 (2006)
18. Shabat, B.: Introduction to Complex Analysis. Part II: Functions of Several Variables. American Mathematical Society, Providence (2003)
19. Shestopalov, Yu.V., Smirnov, Yu.G., Yakovlev, V.V.: Volume Singular Integral Equations Method for Determination of Effective Permittivity of Meta- and Nanomaterials. Proceedings of Progress in Electromagnetics Research Symposium, Cambridge, USA, 291–292 (2008)
20. Shestopalov, Yu.V., Smirnov, Yu.G., Yakovlev, V.V.: Development of Mathematical Methods for Reconstructing Complex Permittivity of a Scatterer in a Waveguide. Proceedings of 5th International Workshop on Electromagnetic Wave Scattering, Antalya, Turkey (2008)
21. Solymar, L., Shamonina, E.: Waves in Metamaterials. Oxford University Press Inc., New York (2009)
22. Usanov, D.A, Skripal, A.V., Abramov, A.V, Bogolyubov, A.S.: Determination of the metal nanometer layer thickness and semiconductor conductivity in metal-semiconductor structures from electromagnetic reflection and transmission spectra. Tech. Phys. **51**(5), 644–649 (2006)
23. Usanov, D.A, Skripal, A.V., Abramov, A.V, Bogolyubov, A.S.: Complex permittivity of composites based on dielectric matrices with carbon nanotubes. Tech. Phys. **56**(1), 102–106 (2011)
24. Yakovlev, V.V., Murphy, E.K., Eves, E.E.: Neural networks for FDTD-backed permittivity reconstruction. COMPEL **33**(1), 291–304 (2005)

Computer Algorithms for Processing Large Information Volumes to Make Decision on Countermeasures for Multiple Emergencies Occurring Simultaneously

A.S. Samokhina and E.A. Trahtengerts

Abstract In this work we consider the algorithms of computer support facilitating the decision-making process when simultaneous or almost simultaneous emergencies take place. In the presence of huge volumes of incoming information effective algorithms of emergency identification are proposed and developed for the analysis and solution of corresponding large-scale problems. We consider determination algorithms of necessary forces and measures aimed at elimination of emergencies. The issues related to dynamic computer support are also examined.

1 Introduction

An old Russian proverb "Trouble never comes alone" manifests itself in various forms of technogenic and natural disasters that occur simultaneously or within a short period of time. Recent best-known ones are the tsunami that caused the destruction of settlements and roads, conflagration at a nuclear reactor in Japan in 2011, and consequent radioactive contamination; an accident on a drilling rig in the Gulf of Mexico in 2010 which resulted in emissions of very large amounts of oil into the sea, beach pollution, and destruction of flora and fauna; and abnormally hot weather in Russia in 2010 that caused forest fires and resulted in burned-out settlements, smoke formation in metropolises, and the loss of about 30% yield. Each of the above-mentioned disasters led to several catastrophes of various kinds. Emergencies caused by these catastrophes may be further complicated by the superposition of different types of catastrophes, natural and anthropogenic, occurring simultaneously or sequentially. By "emergency situation" we would also define a variety of biological disasters. Catastrophes may occur simultaneously or

A.S. Samokhina (✉) • E.A. Trahtengerts
Institute of Control Sciences of Russian Academy of Sciences, Moscow, Russia
e-mail: assamokhina@yandex.ru; absamokhina@yandex.ru; tract@ipu.ru

L. Beilina and Y.V. Shestopalov (eds.), *Inverse Problems and Large-Scale Computations*, Springer Proceedings in Mathematics & Statistics 52, DOI 10.1007/978-3-319-00660-4_12,

sequentially: tsunami in Japan has caused an accident at a nuclear power plant within a short time period, whereas a biological emergency can take place hours or even days after the disaster. The accumulated experience shows that managing the disasters' consequences is most effective when the rescue, first aid, saving of tangible assets measures, etc. are taken during the initial period, as shown in Fig. 1 [12]; thus the emergencies' control system must support decision-making on a time scale close to real time. Based on these assumptions, there is a strong need for computer control systems that manage elimination of emergency situations, i.e., consequences of disasters occurring simultaneously.

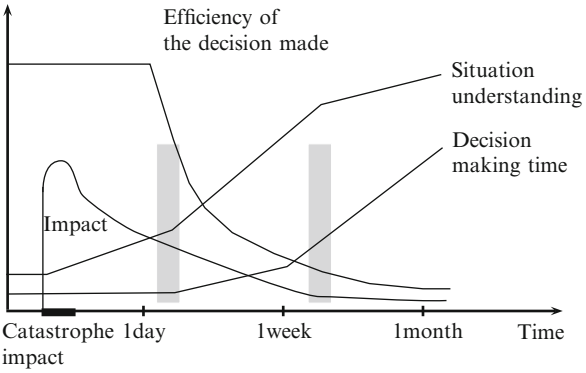

Fig. 1 Efficiency of decision-making depending of time passed after catastrophe

The elimination of emergency situations could be a long time-consuming process and may last for months or even years. In the pre-planning period of this process it is advisable to separate the operational measures, which should be implemented immediately after the accident occurs, and the long-term measures. Long-term measures may be related to reorganization of the existing management system, establishing additional structures, and implementation of new functions on the adoption of sophisticated and expensive measures to eliminate the most serious consequences of emergencies. Operational measures usually consist of generation of plans aimed at eradicating the consequences of possible accidents and managing the dynamics of their implementation.

Preliminary analysis of advantages and disadvantages of different countermeasures (as well as relevant economic costs) is a necessary tool within the planning process and raises the efficiency of decision-making in emergency situations. Manager's efforts could be greatly facilitated by the use of computer modeling, which simulates execution of various countermeasures depending on time and space, and evaluates their advantages and disadvantages [9].

Decision-making model for disaster response can be represented as [1, 15]

$$S = (F : T \times X \times Q \to Y),$$

where S is a set of managerial decisions; other sets in this equation are: T is the time points examined, X is the input data elements that characterize the types of disasters, Q is the possible control actions, Y is the rules of data conversion which takes into account the preferences of a manager, and finally F is a set of rules for ranking.

The forces and the means used to eliminate the multiple disasters occurring at the same time are as follows: M_i is a set of capabilities required for elimination of the i^{th} accident, G_{ij} is a set of parameters characterizing i^{th} accident in interaction with j^{th} accident, and N is the number of accidents occurring simultaneously, $mes\, G_{ij} = 1$. Veil diagram shows interaction of a set of characteristics related to j^{th} accident in the following way (Fig. 2).

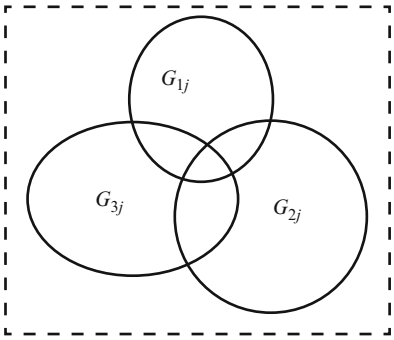

Fig. 2 Interaction of sets of parameters characterizing 1,2,3-accident in relation with j^{th} accident

The total number of parameters characterizing the required capabilities, designed to eliminate the N catastrophes occurring at the same time, can be described using the following formula

$$M = \sum_{j=1}^{N} M_j \left[mes \bigcup_{i=1}^{N} G_{ij} \right].$$

The function of model of the decision-making during an emergency is based on the input received through the monitoring of potentially dangerous objects, which results in us receiving data about the current state of the object and accumulating databases. Analysis of data monitoring helps to promptly detect an emergency situation consequent to catastrophe or multiple catastrophes that occur simultaneously. This is the primary focus of this work.

2 Monitoring of Potentially Hazardous Areas and the Current State of Emergency

Monitoring is the systematic accumulation and data processing of the status and dynamics of the analyzed parameters of the object or process and presenting the results in a suitable form to a manager or an expert. The task of monitoring for integrated management of the elimination of the consequences of different types of emergencies is a timely assessment of risks of each type of emergency, the analysis of the dynamics of their development and their comprehensive assessment [2, 13, 16].

Since we are considering integrated elimination of the consequences of different types of disasters, monitoring should be focused on the parameters characterizing the types of disasters that can take place in the areas—be it chemical or radiological accidents, tsunami and their consequences such as fires, all sorts of explosions, epidemics due to dangerous epidemic diseases, etc,—and thus depend on the prevailing conditions in the same area or group of related areas. Disasters can posses a combined characteristic; they can occur simultaneously or sequentially: explosions, fires, and chemical contamination.

Information sources for monitoring vary from sensor readings and manual data entry to satellite data. There are thousands of different types of local sensors at every potentially dangerous plant (NPP, chemical production, bacteriological laboratory, etc., as well as buoys in the ocean for a tsunami warning). In addition to that, there are also territorial sensors. However, even local sensors are able to monitor global disasters, for example, sensors that measure the radiation level installed at the Novovoronezh Nuclear Power Plant caught radioactive iodine IO9131 released in the accident at the nuclear power plant in Fukusima Power Plant in Japan.

During computer processing of information within the monitoring process, the methods of data mining are widely used. These methods have been widely developed in connection with the widespread use of technology in control systems. The volume of processed data obtained from monitoring can be overwhelming. Therefore, processing of data collected during monitoring may be based on the following approaches:

1. The first one is that the system fixates the expert's experience, which is used per the situation's assessment. Construction of expert systems is based on this approach.
2. The second approach is based on a retrospective analysis of data describing the behavior of the object.
3. Finally, there is a third approach, a combination of the previous two: results received from a retrospective analysis of data are estimated taking into account the experience of the expert. Recently, interest has increased sharply for the third approach. It is due to the new requirements for in-depth analysis of incoming information and historical information stored in the databases that are attributed

to the abrupt increase in the complexity of management tasks. This analysis is performed in real time.

The aim of this work is to create tools that formally describe emergencies caused by disasters, which will allow to automatically identify the type of current situation on the basis of data monitoring.

An essential feature of emergencies associated with various types of disasters is to determine in which type of situation it is required to not only have information on current parameters' values but also history of changes of these parameters.

Moreover, in addition to the data available for monitoring that is stored in a database (known parameters), there are certain parameters that are not available for monitoring (hidden parameters) that with some delay/advance correlate with the data stored in the database. It is these not explicitly monitored parameters that define the type of current emergency situation.

Therefore, in order to correctly identify a type of situation, the requirement lies not only with a database, which stores the results of monitoring, but a knowledge base as well, which describes how and with what time delay (or advance), hidden variables associate with the monitored ones.

In case of any emergency, the information needed for decision-making is never fully represented. During the process of disaster management the volume of initial information increases and it becomes more reliable and comprehensive.

The information can be interpreted ambiguously and contain uncertainty. Sources of uncertainty include:

- Measurement errors, which are determined by the method and means of measurement
- Incorrect use of the measured values
- The limited sample measurements with a statistical interpretation of any size
- Uncertainty of used expert assessments and opinions

The task of assessing the uncertainty of forecasting is to construct a procedure for calculating the uncertainty of predictions based on estimates of the uncertainty of data measurements, expert preferences, and uncertainty modeling. Obviously, such procedure depends on the type used scheme, type of model, and methods for uncertainty representation.

The tasks of data processing and evaluation of prediction uncertainty are interrelated and must be addressed together.

In terms of computer decision support systems (DSS), the task of data processing methods is formulated as follows: based on the available monitoring data, information about the situation, and the preferences of decision makers (DM), it is necessary to choose from the set of models presented in DSS one that is most suitable for the situation forecasting. Expert and knowledge-base systems could be used to solve this problem.

Specification language is used to describe relations between parameters, hidden and measured ones. Since the type of situation is completely determined by the values of hidden variables, then by knowing their values at current moment, it is

possible to determine the type of current situation. The specification itself (text in the specification language) should not be seen at all as a knowledge base. Specification is the source material for building the knowledge base.

Thus, it is necessary to create a specification language that meets the following requirements:

1. The language must be significant enough to describe relationships between hidden and monitored parameters' values.
2. There should be a procedure to build a computer program that restores the values of hidden parameters using specification text and sequence diagram (history of changes) of measured parameters.

3 Specification Language

As a basis for the language of specifications the classic language of monadic second-order weak theories with one successor relation [3, 6, 8, 10] (MTL henceforth) has been chosen. The only difference between the MTL formulas and the formulas of the first-order language is that MTL allows quantification over not only variables but also over monadic predicate symbols. In the MTL language only one unary functional symbol is used, notably **next**. **Next**$(\mathbf{t})$ is interpreted as the time moment following $\mathbf{t}$.

Bellow the expressions (term) of the type $\mathbf{next}(\ldots\mathbf{next}(\mathbf{x}))$ (n times), compiled from variable $\mathbf{x}$ and functional symbol **next**, will be denoted as $\mathbf{x}+\mathbf{n}$.

Definition 1. Suppose P is a set of monadic predicate symbols, while V is a set of variables not intersected with P. Let's call the rows of a $\mathbf{p}(\mathbf{t})$-sort, where $\mathbf{t}$ is a term of the $(\mathbf{x}+\mathbf{n})$-sort compiled from variables $x \in V$ and functional symbol **next** as *atoms over P and V* (the set of all atoms over P and V is labeled **Atoms** $(\mathbf{P},\mathbf{V})$). A set of MTL's formulas over P and V labeled **Formulas** $(\mathbf{P},\mathbf{V})$ relates to the minimal set of rows containing atoms (**Atoms** $(\mathbf{P},\mathbf{V}) \subseteq$ **Formulas** $(\mathbf{P},\mathbf{V})$) and satisfies the following conditions [5]:

If $\mathbf{F},\mathbf{F}' \in \mathbf{Formulas}(\mathbf{P},\mathbf{V})$ and $\mathbf{p} \in \mathbf{P} \cup \mathbf{V}$,
$(\mathbf{F} \vee \mathbf{F}') \in \mathbf{Formulas}(\mathbf{P},\mathbf{V})$ (conjunction),
$(\sim \mathbf{F}) \in \mathbf{Formulas}(\mathbf{P},\mathbf{V})$ (negation),
$\exists \mathbf{pF} \in \mathbf{Formulas}(\mathbf{P},\mathbf{V})$.

Reductions

$$(\mathbf{F} \wedge \mathbf{F}') \stackrel{\text{def}}{=} \sim ((\sim \mathbf{F}) \vee (\sim \mathbf{F}')),$$

$$\mathbf{F} \to \mathbf{F}' \stackrel{\text{def}}{=} \sim \mathbf{F} \vee \mathbf{F}' \forall \mathbf{pF} \stackrel{\text{def}}{=} \sim \exists \mathbf{p}(\sim \mathbf{F}).$$

Variables are interpreted as time moments for which the information on measured parameters is stored in a database.

The relation $t \leq t'$ is satisfied if the time moment t occurs prior to t'.

Monadic predicates are used to specify the values of parameters: each time moment is matched with the set of properties at this moment. The predicate symbols quantifier corresponds to the hidden parameters and non-predicate symbols to the measured ones.

Further on is the Definition 2 of formulas' interpretation, where each formula is assigned a set of its own models (interpretation defines the semantics of the language). This definition is correct only for the so-called canonical formula. Any formula can be reduced to canonical form through renaming the variables bound by a quantifier, so that none variables are related to a quantifier more than once.

Definition 2. Interpretation area of the set of formulas **Formulas** $(\mathbf{P},\mathbf{V})$ in MTL language is denoted **Models** $(\mathbf{P},\mathbf{V})$ and specifies the set of rows (finite sequences) over the alphabet $\mathbf{2^{P\cup V}}$ (sequents are subsets of set $P\cup V$).

For any variable $x\in V$ as set of its models let's call the set **Models** $(\mathbf{x})$ with such rows $\alpha:\{\mathbf{1},\ldots\mathbf{n}\}\to\mathbf{2^{P\cup V}}$ that the inequality $\{\mathbf{i}:\mathbf{i}\in\mathbf{dom}(\alpha)\&\mathbf{x}\in\alpha(\mathbf{i})\}\neq\emptyset$ is satisfied. For any variable and a row $\alpha\in$ **Models** $(\mathbf{x})$ let's call variable's interpretation in model α as the number $\mathbf{varInterpretation}\,(\alpha\mathbf{x})=\min\{\mathbf{i}:\mathbf{i}\in\mathbf{dom}(\alpha)\&\mathbf{x}\in\alpha(\mathbf{i})\}$.

For each expression $\mathbf{t}=\mathbf{x}+\mathbf{i}$ and row $\alpha\in$ **Models** $(\mathbf{P},\mathbf{V})$ of the length $\mathbf{n}$ we can interpret term $\mathbf{t}$ as number

$$\mathbf{TermInterpretation}(\alpha,\mathbf{t})=\min\{\mathbf{n},\mathbf{varInterpretation}\,(\alpha,\mathbf{x})+\mathbf{i}\};$$

Interpretation of the canonical formulas is defined by induction as a mapping $\mathbf{Interpretation}(\mathbf{Formulas}(\mathbf{P},\mathbf{V}))\to\mathbf{2}^{\mathrm{Models}(\mathbf{P},\mathbf{V})}$ from the set of canonical formulas into subsets of rows' set by using the conventional approach:

- For any atomic formula $\mathbf{p}(\mathbf{x})$ and any row

$$\alpha\colon\{\mathbf{1},\ldots\mathbf{n}\}\to\mathbf{2^{P\cup VI}}\alpha\in\mathbf{Interpretation}(\mathbf{p}(\mathbf{x}))$$

  is true if and only if $\alpha\in$ **models** $(\mathbf{P},\mathbf{V})$ and $\mathbf{p}\in\alpha(\mathbf{varInterpretation}(\alpha\mathbf{x}))$.
- $\mathbf{Interpretation}(\mathbf{F}\vee\mathbf{F}')=\mathbf{Interpretation}(\mathbf{F})\cup\mathbf{Interpretation}(\mathbf{F}')$.
- $\mathbf{Interpretation}(\sim\mathbf{F})$ is supplement of the set Interpretation $(\mathbf{F})$ up to the set of all rows.
- $\mathbf{Interpretation}(\exists\mathbf{xF})=\{\alpha:\alpha'\in\mathbf{Interpretation}(\mathbf{F})\&\mathbf{dom}(\alpha)=\mathbf{dom}(\alpha')\&$ $\forall\mathbf{i}\in\mathbf{dom}(\alpha')\alpha(\mathbf{i})=\alpha'(\mathbf{i})\backslash\{\mathbf{x}\}\}$.

  Elements of the set Interpretation (F) are called models of formula F.

3.1 Examples

1. Further on, the following reductions will be used in the formula notation:

$\mathbf{x} \le \mathbf{y} \stackrel{\text{def}}{=} \forall \mathbf{q}(\mathbf{q}(\mathbf{x}) \& (\forall \mathbf{z} \& \mathbf{q}(\mathbf{z}) \to \mathbf{q}(\mathbf{z}+\mathbf{1}))) \to \mathbf{q}(\mathbf{y})$,
$\mathbf{x} = \mathbf{y} \stackrel{\text{def}}{=} \mathbf{x} \le \mathbf{y} \& \mathbf{y} \le \mathbf{x}$,
$\mathbf{x} < \mathbf{y} \stackrel{\text{def}}{=} \mathbf{x} \le \mathbf{y} \& \sim (\mathbf{x} = \mathbf{y})$,
$\mathbf{first}(\mathbf{x}) \stackrel{\text{def}}{=} \forall \mathbf{y} \mathbf{x} \le \mathbf{y}$,
$\mathbf{last}(\mathbf{x}) \stackrel{\text{def}}{=} \forall \mathbf{y} \mathbf{y} \le \mathbf{x}$.
It is easy to verify that interpretation of $\mathbf{Formula} \mathbf{x} \le \mathbf{y}$ coincides with the ordinary non-strict inequality for the set of integers. Accuracy of incidental formulas is subsequent.

2. Models of formula
$\exists \mathbf{p} \forall \mathbf{x} \mathbf{p}(\mathbf{x}) \to \sim \mathbf{p}(\mathbf{x}+\mathbf{1}) \& ((\sim \mathbf{p}(\mathbf{x})) \to \mathbf{p}(\mathbf{x}+\mathbf{1})) \& \mathbf{first}(\mathbf{x}) \to \mathbf{p}(\mathbf{x}) \& \mathbf{last}(\mathbf{x}) \to \mathbf{p}(\mathbf{x})$ are all rows of uneven length (this formula describes all rows with uneven length).
3. Let the set of predicates P be equal to $\{\mathbf{a}, \mathbf{b}\}$.
Formula $\mathbf{F}_{\mathbf{a},\mathbf{b}} = \forall \mathbf{x}(\mathbf{a}(\mathbf{x}) \to \sim \mathbf{b}(\mathbf{x})) \& (\mathbf{b}(\mathbf{x}) \to \sim \mathbf{a}(\mathbf{x})) \& (\mathbf{a}(\mathbf{x}) \vee \mathbf{b}(\mathbf{x}))$ describes all rows over the alphabet $\{\{\mathbf{a}\}, \{\mathbf{b}\}\}$.
Formula $\mathbf{F}_{\mathbf{a},\mathbf{b}} \& \forall \mathbf{x}((\mathbf{a}(\mathbf{x}) \to \mathbf{b}(\mathbf{x}+\mathbf{1})) \& (\mathbf{b}(\mathbf{x}) \to \mathbf{a}(\mathbf{x}+\mathbf{1})))$ describes all sequences of a kind $\{\mathbf{a}\}\{\mathbf{b}\}\{\mathbf{a}\}\{\mathbf{b}\}$.
4. Suppose (Q, F, μ) is a finite automation (unary algebra) with the set of states Q, input signals F, and the diagram of transitions $\mu : F \to (Q \to Q)$ and suppose $q0, q1 \in Q$ are some of the states of automation.

Then the formula

$$\forall \mathbf{x} \forall \mathbf{y} \forall_{\mathbf{Q}} \& \left\{ (\mathbf{q}(\mathbf{x}) \& \mathbf{a}(\mathbf{x})) \to \mathbf{q}'(\mathbf{x}+\mathbf{1}) : \mathbf{q} \in \mathbf{Q} \& \mathbf{a} \in \mathbf{F} \& \mathbf{q}' = \mu(\mathbf{a})(\mathbf{q}) \right\} \& (\mathbf{first}(\mathbf{x}) \to \mathbf{q0}(\mathbf{x})) \to (\mathbf{last}(\mathbf{y}) \to \mathbf{q1}(\mathbf{y})),$$

here $\forall_Q$ describes quantifying over all elements Q and $\{F1, \ldots, Fn\}$—reduction of notation $F1\& \ldots \& Fn$ that describes all such rows over alphabet F where automation transits from state $q0$ to state $q1$.

The choice of MTL as a basis for the specification language was governed by the fact that for any MTL formula, it is possible to construct a finite automation that computes the values of hidden parameters in real time (upon arrival of each recurrent input data in the database). The states of automation correspond to the values (sets of properties) of hidden variables, and input symbols correspond to the measured parameters' values. This automation browses the cyclogram of measured parameters (sequence diagrams can be viewed as a row, symbols of which are sets of properties of the measured parameters).

4 The Technology of Using Specification Language

Pre-processing of initial data can be itemized in the following way: initial description in the MTL language, the MTL language compilation, and building of the monitor program as shown in Fig. 3.

In the MTL language a description of F is given (F is a logical formula in the MTL language) as a sequence of events that correspond to an emergency. Using the text of description (formula F) MTL compiler builds the program P. The program P simulates the finite automation behavior that identifies the set of all models of formula F.

The sequence of signals arriving from the monitored object (deviations of the characteristic properties of monitored object) is passed through program P. Thus, each time when the passed sequence corresponds to an emergency situation (to a situation described in formula F), program P detects this match.

To determine the type of situation, it is necessary in addition to the cyclogram of measured parameters to have a knowledge base containing information about the mutual influence of measured and hidden variables. Therefore, the knowledge base should store transition table for finite automaton that is built using specification text [11].

For assessing the impact of disasters and, consequently, the formation of goals and strategies (scenarios) to eliminate them, one can proceed from different conceptual hierarchies: preparation for maximum possible effects (these are rare but lead to very serious consequences, especially if not envisaged), preparation for "average" effects (according to statistics, forecasts, subjective expert estimates), and preparation for often occurring "relatively small" disasters. Each approach has its advantages and disadvantages; they are obvious. The choice of approach is determined by many factors, including available resources and the mentality of the manager.

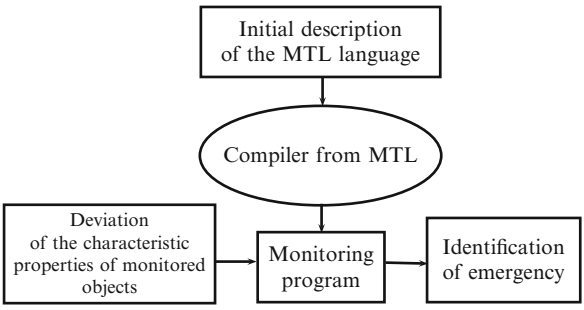

Fig. 3 Scheme of using monitoring program to identify emergency

For threat assessment one should take into account not only the parameters listed above that affect the degree of destruction and casualties listed but also the disaster parameters, such as earthquake magnitude, altitude and speed of a tsunami wave, and the force of explosion. Depending on these parameters, the values of the criteria

for determining the degree of destruction will change. Thus, to range the tsunami risk, taking into account the magnitude of the earthquake, the time lag waves, and some other factors in the areas of potential danger, there are three degrees of alarm. The threat, depending on the values of the disaster's parameters, can be divided into a greater number of levels. Therefore, for each type of occurring disaster it is necessary to plan countermeasures for different sets of parameters for particular disasters. The scheme of monitoring processes is represented accordingly in Fig. 4.

For a comprehensive planning of elimination of different disasters' consequences, one must first define the range of their prospective parameters. Determining the lower limit of these parameters is usually unhindered, but defining the upper limit in many cases is a challenge. For example, the strongest of the registered tsunamis was caused by an earthquake with a magnitude of 9.5 on the Richter scale in Chile in 1960, wave height reached 10–11 m and the speed peaked up to 100 km/h. No other earthquakes leading to tsunamis with such magnitude have ever been observed. Japan earthquake in 1911 was estimated at 8.8–9 on the Richter scale, the wave height was estimated according to various sources from 10 to 40 m, and the speed exceeded 30 km/h. It is important to note that the engineering of nuclear power plants in Japan had not envisaged the possibility of a tsunami of such force that caused the catastrophic consequences: conflagration, explosion at a nuclear reactor, and radioactive contamination of the surrounding area. This example shows that even the most serious studies related to the engineering of high-end nuclear power plants may result in tragic mistakes. Therefore, the determination of the upper limit of parameters of a possible disaster should be performed thoroughly [4,7,14,17,18].

Remark: shading in Fig. 4 shows a number of different sets of parameters' values for one type of disaster.

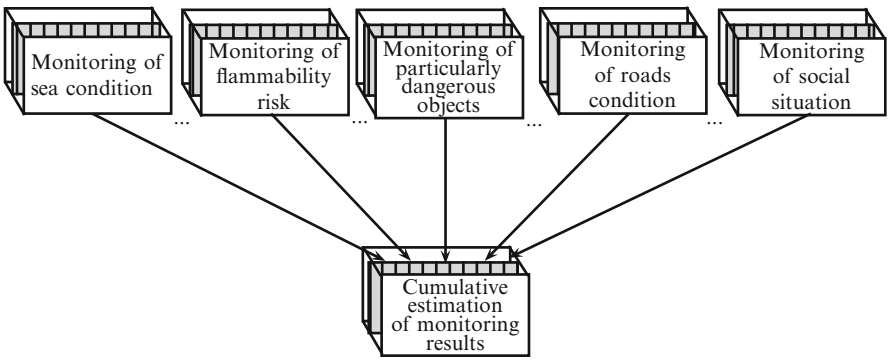

Fig. 4 Scheme of monitoring cumulative estimating

A possible procedure for generating estimates may be as follows.

1. The system offers experts and manager to provide their estimates of a maximum level of parameters, whenever possible in correlation with the previously known emergencies.
2. The system tabulates the obtained data and highlights it on displays. If estimations strongly disperse, the manager arranges their discussion (i.e., transition to p. 3). If they are similar, the system coordinates them, and the procedure terminates.
3. The system offers experts, after discussion, to enter the estimations again, tabulates them, calculates average values, and highlights them on experts' displays.
4. The experts may change the estimations after assessing averages. The system then calculates anew their average values and presents them to the manager for approval).

After defining the boundaries of possible parameters' values for each type of potential disaster, they are divided into ranges. Each range includes the value of a parameter or a set of them that can determine consequences of the type of the accident. Partitioning into ranges can be performed either by the manager or experts; the system will negotiate their proposals. The information received is entered in the knowledge base and used in the identification of disaster using the MTL language.

5 The Use of Specification Language for a Particular Event

Let's assume that the dispatcher has informed a decision maker about the suspicion on aerosol diversion with the use of biological warfare agent (BWA) in the car of a suburban train in connection with the following events.

In a suburban hospital patient *P* has been delivered. The symptoms are similar to the ones caused by one of the known biological warfare agent BWA; let's call this agent "*A*", for allergic persons. For people not subjected to an allergy, the first symptoms of the BWA influence become evident within several days up to a week.

According to eyewitnesses, present in the same carriage as *P*, awhile before *P*'s condition deterioration, an unknown citizen sprayed a certain substance from an aerosol bottle and then disappeared. The latter gives the grounds to assume that sudden condition deterioration of patient *P* was not a random occurrence, and there is a possibility of worsening of the epidemic situation in the region in the upcoming days.

The ability to quickly clarify the situation is complicated by the following factors:

- During reception of patient *P* the data has not been recorded that could allow to identify and find the witnesses of the incident.
- The analysis confirming the presence of agent *A* in blood of the patient will be ready only in a few days.
- Suffering patient *P* is continually unconscious, and even application of drug *R* has not deduced her from a coma.

If it is known that drug R is effective against the influence of agent A at early stages, it is possible to draw a conclusion from the resulted report on little reliability of the influence of a specific biological agent. If the incident has happened in summer, it is more probable to assume that the unknown citizen had used a repellent from mosquitoes (which is an allergen) before leaving the train.

The program of the primary information processing can draw a conclusion on low probability of the biological agent influence only if it is in some way was informed on correlation between occurring events, their chronological sequence, and resulting varying reliability of this or that fact.

The MTL language is *that* formal language in which one can transmit this information to the program.

For example, let $R(t)$ be the fact of application of drug R at the time moment t, $unwell(t)$ indicates the sickness at the time moment t, and $A(t)$ means the fact of influence of agent A. Then the following formula describes the influence of drug R on patient influenced by agent A.

$$\forall \mathbf{tA(t)\&unwell(t+1)\&R(t+2)} \rightarrow\sim \mathbf{unwell(t+3)}.$$

Accordingly, from sequence of events $\mathbf{unwell(t0+1), R(t0+2), unwell(t0+3)}$ fixed in the cyclogram on patient P(client) case, the monitor program can draw a conclusion on the absence of influence of agent A on P at the time moment $\mathbf{t0}$.

This example shows that the MTL is applicable for tracking the history of events of just one client. There is only the time parameter, because in the MTL language, it is not possible to specify to which objects or client the events in question relate. This restriction cannot be ignored as it is because only due to this restriction the text written in MTL can be compiled into the monitor program. Therefore, in the aggregate protocol of the events' sequence, all the monitored clients will be analyzed independently of each other and on each client's case a separate inference will be issued. The substantial number of monitored clients can be useful either to calculate the probability that the inferences made by the program are accurate, or to consider the client group as one object. In the latter case of grouping the following additional independent problems arise:

- Identification of characteristic properties (changing in time) of the groups of clients using a cyclogram of properties of the groups' members
- Distribution and redistribution of clients into groups
- Conflict resolution between the inferences made on different clients

Except for above-specified inconvenience incidental to non-personalized events it is necessary to note the following issues with application of MTL:

1. The MTL language is suitable for formal analysis (analysis of the set of described events), but it is not convenient to describe specifications: it should be considered only as a basis over which the language of specifications can be built, e.g., the programming language "C" uses assembler language. Time-based logics were introduced, where some cumbersome concepts of the MTL language were

replaced by special quantifiers (modalities) and flexibility in use of connectives and quantifiers was reduced. It makes sense to consider using these logics instead of creating a new language based on the MTL. Also, instead of MTL as a specification language one may try to use Presburger arithmetic (a formal system which is a special case of MTL).
2. There are only three values to measure reliability in MTL: "possible", "not possible", and "is true" (negation to "cannot be"). Possibilities of expansion of this scale have not been investigated.

In the above example quantifiers over predicates are not used and there are no mentions of allergies. Consider another similar example and demonstrate that the MTL is applicable only as a basis for the specification language (due the complexity of formulas that define the simplest biographical facts of the client)

Assume that the allergy to agent A is usually incidental to the client's previous ailment from disease C. The fact that before time moment $\mathbf{t0}$ the client had the disease C can be notated using the following formula:

$$\exists \mathbf{xx} < \mathbf{t0} \& \mathbf{C}(\mathbf{t}).$$

The above formula does not explicitly contain predicates. We should take into account that construction $\mathbf{t} < \mathbf{t0}$ is not an element of the MTL language but the reduction for rather long formula which is described through the construction:

$$\mathbf{t} \leq \mathbf{t}' \stackrel{\text{def}}{=} (\forall \mathbf{q}(\mathbf{q}(\mathbf{t}) \& (\forall \mathbf{z} \& \mathbf{q}(\mathbf{z}) \rightarrow \mathbf{q}(\mathbf{z}+\mathbf{1}))) \rightarrow \mathbf{q}(\mathbf{t}`))$$

is described through construction $\mathbf{t} \leq \mathbf{t0} \& \sim (\mathbf{t0} \leq \mathbf{t})$.

Special reductions (generally defined as modalities) have been introduced for structures similar to $t \leq t$ in the temporal logics.

One of the reasons of making the last example is that the necessity to use quantifiers over predicate symbols occurs if, and only if, the time intervals are of arbitrarily large length. If there is no need to analyze long periods of time, then it is not necessary to use quantifiers over predicate symbols, e.g., if there are no medical records and the incident cannot develop longer than for a month. In this case, MTL becomes an ordinary first-order language with monadic predicate symbols and relation of inequality.

Let's give the formal description to the identifying of the hidden parameters by the monitor program.

Let $P(t) = \{P_{s1}, P_{s2}, \ldots P_{sn}, P_{d1}(t), P_{d2}(t), \ldots P_{dm}(t)\}$, where $P_{si},\ i = 1, \ldots n$, and $P_{dj}(t),\ j = 1, \ldots m$, are, respectively, the static and dynamic measured parameters. For the above-mentioned example, existence of the incident (dispersion of aerosol) and identifying of a biological agent are static parameters, while the number of the diseased by days starting from the beginning of epidemic and the area of distribution of patients are dynamic measured parameters. The set Q of the measured and hidden parameters' values that describes this biological emergency situation is stored in the knowledge-base Q.

Let C_s be a static hidden parameter and $C_d(t)$ be the dynamic hidden parameter; then it is possible to conclude

$$C_s = F_1[P_s, P_d(t)], \quad C_d(t) = F_2[P_s, P_d(t)],$$

where F_1 and F_2 are functionals on the set of parameters $P(t)$. An example of the dynamic hidden parameter is the density of distribution of the diseased by days, and the static hidden parameter—disease symptoms. Here $\{C_s\}$, $\{C_d\}$ are the sets of the static and dynamic hidden parameters accordingly. Thus if $\{P_s\} \cup \{P_d(t)\} \cup \{C_s\} \cup \{C_d(t)\} \in Q$, then it is possible to interpret a considered situation as a biological emergency situation.

6 Conclusion

We have developed and analyzed a specification language created on the base of a classic language of the monadic second-order weak theories with one successor relation.

Using the developed language we can describe relationships between hidden and monitored parameters stored in the knowledge base and therefore reliably identify emergency.

Application of specification language during the monitoring of potentially dangerous objects for detection and identification of emergency situations caused by accidents of various kinds, occurring simultaneously or sequentially, will facilitate decision-making process to prevent, combat and disaster relief.

Acknowledgements This work is supported by the RFFI Grant 10-08-00590-a. The authors are thankful to Dr A. V. Babichev for presenting important materials and valuable discussion of the results.

References

1. Andreev, D.K., Kamaev, D.A., Trahtengerts, E.A.: Expert forecasting of consequences of damage of life-support systems. Manag. Big Syst. **25**, 243–293 (2009)
2. Barkalov, S.A., Novikov, D.A., Peskovatov, V.I., Serebryakov, V.I.: Two-channel Model of Active Examination. IPU, Moscow (2000)
3. Buchi, J.R.: Weak second order arithmetic and finite automata. Z. Math. Logik Grundl. Math. **6**, 66–92 (1960)
4. Counteraction to biological terrorism. In: Onischenko, G. (ed.) Practical Guidance on Counter Epidemic Provision. oscow, Bangkok (2003)
5. Dijkstra, E.W.: Cooperating Sequential processes. Program lang. **4**, 43–112 (1968)
6. Ladner, R.E., Semenov, A.L.: The use of model-theoretic games to linear orders and finite automata, in b. The Cybernetic Collection (A New Series), vol. 17, pp. 164–191. The World, Moscow (1980)

7. Novikov, A.M., Novikov, D.A.: Methodology. SINTEG, Moscow (2007)
8. Samokhina, A.S.: General problem of classification and identification in the decision support system for biological emergency. Interuniversity Collection of Scientific Papers "Theoretical Aspects of Computer Science, Software and Information Technologies in the Municipal Sector", pp. 216–221. MIREA, Moscow (2005)
9. Samokhina, A.S.: Analysis of primary data processing schemas in the system to prevent a biological emergency. Probl. Safety Emergen. Situations **2**, 92–106 (2006)
10. Semenov, A.L.: Logical theories of monadic function on integers. DAN **47**(3), 623–658 (1983)
11. Semenov, A.L.: Resolving procedures for logic theories. Cybernetics Comput. Tec. **2**, 134–146. The Science, Moscow (1986)
12. Shershakov, V.M.: Research and development of methods and decision support systems in emergency situations involving contaminated environment. Thesis for the degree of Doctor of Science. IPU, Moscow (2001)
13. Shishkin, E.V., Chhartishvili, A.T.: Mathematical Methods and Models in Management. Business, Moscow (2000)
14. Trahtengerts, E.A.: Software of Parallel Processes. The Science, Moscow (1987)
15. Trahtengerts, E.A., Shershakov, V.M., Kamayev, D.A.: Computer Support of Managing Elimination of Radiating Influence Consequences. SINTEG, Moscow (2004)
16. Trahtengerts, E.A.: Computer Support of the Purposes' and Strategy's Generating. SINTEG, Moscow (2005)
17. Trahtengerts, E.A.: Computer Methods of Implementation of Economic and Information Administrative Decisions. vol. 2, SINTEG, Moscow (2009)
18. The state report of the Ministry of Emergency Measures of Russia on the situation with protection of the population and territories of the Russian Federation from emergency situations of natural and man-made nature in 2002. In: Mahutov, N.A. (ed.) Problems of safety and emergency situations. The Information Collection, pp. 3–186. VINITI, Moscow 2003

System of Nonlinear Boundary-Value Problems and Self-Consistent Analysis of Resonance Scattering and Generation of Oscillations by a Cubically Polarisable Layered Structure

Vasyl V. Yatsyk

Abstract The problem of scattering and generation of waves on an isotropic, non-magnetic, linearly polarised (E-polarisation), nonlinear, layered, cubically polarisable, dielectric structure, which is excited by a packet of plane waves, is investigated in the domain of resonance frequencies. The resulting mathematical model can be represented by a system of one-dimensional nonlinear integral equations. The solution of this problem is approximated numerically by the help of quadrature methods and iterative procedures which require the solution of a linear system in each step. Layers with negative and positive values of the coefficient of cubic susceptibility of the nonlinear medium have fundamentally different scattering and generation properties. Here the investigations are restricted to the third harmonic generated by layers with a negative value of the cubic susceptibility of the medium. In such a case, a decanalisation of the electromagnetic field can be detected. Results of calculations of characteristics of the scattered and generated fields of plane waves are presented, taking into account the influence of weak fields at multiple frequencies on the cubically polarisable layer.

1 Introduction

Nonlinear dielectrics with controllable permittivity are subject of intense studies and begin to find broad applications in device technology and electronics. We develop a model of resonance scattering and generation of waves on an isotropic nonmagnetic nonlinear layered dielectric structure excited by a packet of plane waves in the resonance frequency range in a self-consistent formulation [1, 3, 4]. Here, both the radio [5] and optical [8] frequency ranges are of interest. We consider wave

V.V. Yatsyk (✉)
O.Ya. Usikov Institute for Radiophysics and Electronics of the National Academy of Sciences of Ukraine, 12 Ac. Proskura Str., Kharkov, 61085, Ukraine
e-mail: yatsyk@vk.kharkov.ua; vasyl.yatsyk@rambler.ru

L. Beilina and Y.V. Shestopalov (eds.), *Inverse Problems and Large-Scale Computations*, Springer Proceedings in Mathematics & Statistics 52, DOI 10.1007/978-3-319-00660-4_13,

packets consisting of both strong electromagnetic fields at the excitation frequency of the nonlinear structure, leading to the generation of waves, and of weak fields at the multiple frequencies, which do not lead to the generation of harmonics but influence on the process of scattering and generation of waves by the nonlinear structure. The electromagnetic waves for a nonlinear layer with a cubic polarisability of the medium can be described by an infinite system of nonlinear boundary-value problems. In the study of particular nonlinear effects it proves to be possible to restrict this system to a finite number of problems and also to leave certain terms in the representation of the polarisation coefficients, which characterise the physical problem under investigation [3, 4, 13]. The analysis of the quasi-homogeneous electromagnetic fields of the nonlinear dielectric layered structure made it possible to derive a condition of phase synchronism of waves. If the classical formulation of the problem is supplemented by the condition of phase synchronism, we arrive at a rigorous formulation of a system of boundary-value problems with respect to the components of the scattered and generated fields [3, 4]. Our mathematical model reduces to a system of nonlinear boundary-value problems of Sturm–Liouville type or, equivalently, to a system of nonlinear integral equations. Here the solution to nonlinear boundary-value problems is obtained rigorously in a self-consistent formulation and without using approximations of the preset field, slowly varying amplitudes, etc. The numerical algorithms of the solution of the nonlinear problems are based on iterative procedures which require the solution of a linear system in each step. In this way the approximate solution of the nonlinear problems is described by means of solutions of linear problems with an induced nonlinear permittivity. The analytical continuation to the complex frequency region allows us to turn to the analysis of spectral problems and to reveal various resonance phenomena related to the nonlinearity of the structure; see [3]. We present and discuss results of calculations of the scattered field taking into account the third harmonic generated by nonlinear layer. The presented results of numerical calculations describe properties of the nonlinear permeability of the layers as well as their scattering and generation characteristics. The results indicate a possibility of designing a frequency multiplier and nonlinear dielectrics with controllable permittivity. The transformation of the frequency and angular spectra, the rapid control of amplitude and phase of the waves form the basis of a broad class of technical systems [9].

2 The Scattering Problem and the Generation of the Third Harmonic

In this paper, we consider the problem of scattering and generation of waves on an isotropic, non-magnetic, linearly polarised $\mathbf{E} = (E_1, 0, 0)^\top$, $\mathbf{H} = (0, H_2, H_3)^\top$ (E-polarisation), nonlinear, layered, cubically polarisable, dielectric structure (cf. Fig. 1), which is excited by a plane stationary electromagnetic wave, where the

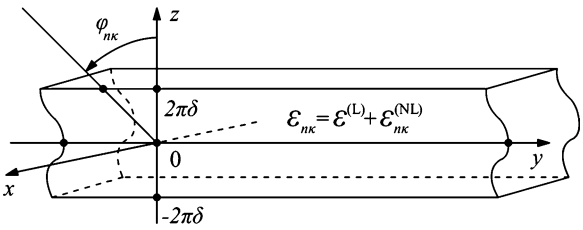

Fig. 1 The nonlinear dielectric layered structure

time dependency of the fields is of the form $\exp(-in\omega t)$ and the vector of cubic polarisation is given as $\mathbf{P}^{(NL)} = \left(P_1^{(NL)}, 0, 0\right)^\top$.

The analysis of the scattering problem for the plane wave packet

$$\left\{E_1^{\text{inc}}(\mathbf{r}, n\kappa) := E_1^{\text{inc}}(n\kappa; y, z) := a_{n\kappa}^{\text{inc}} \exp\left(i\left(\phi_{n\kappa} y - \Gamma_{n\kappa}(z - 2\pi\delta)\right)\right)\right\}_{n=1}^{3}, \quad (1)$$

$z > 2\pi\delta$, $\delta > 0$, with amplitudes $a_{n\kappa}^{\text{inc}}$, angles of incidence $\varphi_{n\kappa}$, $|\varphi| < \pi/2$ (cf. Fig. 1) and $\kappa := \omega/c = 2\pi/\lambda$, $\phi_{n\kappa} := n\kappa \sin\varphi_{n\kappa}$, $\Gamma_{n\kappa} := \sqrt{(n\kappa)^2 - \phi_{n\kappa}^2}$, on the nonlinear structure can be simplified by means of *Kleinman's rule* ([6, 8]) and reduces finally to the following system of boundary-value problems ([2–4, 7, 12]):

$$\begin{aligned} &\Delta E_1(\mathbf{r}, n\kappa) + (n\kappa)^2 \varepsilon_{n\kappa}(z, \alpha(z), E_1(\mathbf{r}, \kappa), E_1(\mathbf{r}, 2\kappa), E_1(\mathbf{r}, 3\kappa)) \\ &= -\delta_{n1}\kappa^2 \alpha(z) E_1^2(\mathbf{r}, 2\kappa)\overline{E}_1(\mathbf{r}, 3\kappa) \\ &\quad - \delta_{n3}(3\kappa)^2 \alpha(z)\left\{\frac{1}{3}E_1^3(\mathbf{r}, \kappa) + E_1^2(\mathbf{r}, 2\kappa)\overline{E}_1(\mathbf{r}, \kappa)\right\}, \; n = 1, 2, 3, \end{aligned} \quad (2)$$

where $\kappa := \frac{\omega}{c} = \frac{2\pi}{\lambda}$, $\varepsilon_{n\kappa} := \begin{cases} 1, & |z| > 2\pi\delta, \\ \varepsilon^{(L)} + \varepsilon_{n\kappa}^{(NL)}, & |z| \le 2\pi\delta, \end{cases}$ and $\varepsilon^{(L)} := 1 + 4\pi\chi_{11}^{(1)}$,

$$\begin{aligned} \varepsilon_{n\kappa}^{(NL)} := \alpha(z)\Big[\sum_{j=1}^{3} |E_1(\mathbf{r}, j\kappa)|^2 + \delta_{n1}\frac{\left[\overline{E}_1(\mathbf{r}, \kappa)\right]^2}{E_1(\mathbf{r}, \kappa)} E_1(\mathbf{r}, 3\kappa) \\ + \delta_{n2}\frac{\overline{E}_1(\mathbf{r}, 2\kappa)}{E_1(\mathbf{r}, 2\kappa)} E_1(\mathbf{r}, \kappa) E_1(\mathbf{r}, 3\kappa)\Big] \end{aligned} \quad (3)$$

with $\alpha(z) := 3\pi\chi_{1111}^{(3)}(z)$, δ_{nj} … Kronecker's symbol. $\chi_{11}^{(1)}$ and $\chi_{1111}^{(3)}$ denote the components of the corresponding media susceptibility tensors in the expansion of the vector of the polarisation moment in terms of the electric field intensity.

Taking into account the following conditions ($n = 1, 2, 3$)

(C1) $E_1(n\kappa; y, z) = U(n\kappa; z)\exp(i\phi_{n\kappa} y)$
(the quasi-homogeneity w.r.t. y),

(C2) $\phi_{n\kappa} = n\phi_\kappa$
(the condition of phase synchronism of waves),
(C3) $\mathbf{E}_{\mathrm{tg}}(n\kappa;y,z)$ and $\mathbf{H}_{\mathrm{tg}}(n\kappa;y,z)$ (i.e. $E_1(n\kappa;y,z)$ and $H_2(n\kappa;y,z)$) are continuous across the interfaces,
(C4) $E_1^{\mathrm{scat}}(n\kappa;y,z) = \left\{ \begin{matrix} a_{n\kappa}^{\mathrm{scat}} \\ b_{n\kappa}^{\mathrm{scat}} \end{matrix} \right\} \exp\left(i\left(\phi_{n\kappa}y \pm \Gamma_{n\kappa}(z \mp 2\pi\delta)\right)\right), z \gtrless \pm 2\pi\delta$
(the radiation condition)

with $\mathfrak{Re}\,\Gamma_{n\kappa} > 0$, $\mathfrak{Im}\,\Gamma_{n\kappa} = 0$, and making use of the following representation for the desired solution ($n = 1,2,3$):

$$
\begin{aligned}
E_1(n\kappa;y,z) &= U(n\kappa;z)\exp(i\phi_{n\kappa}y) \\
&= \begin{cases}
a_{n\kappa}^{\mathrm{inc}}\exp(i(\phi_{n\kappa}y - \Gamma_{n\kappa}(z-2\pi\delta))) \\
\quad + a_{n\kappa}^{\mathrm{scat}}\exp(i(\phi_{n\kappa}y + \Gamma_{n\kappa}(z-2\pi\delta))), & z > 2\pi\delta, \\
U(n\kappa;z)\exp(i\phi_{n\kappa}y), & |z| \le 2\pi\delta, \\
b_{n\kappa}^{\mathrm{scat}}\exp(i(\phi_{n\kappa}y - \Gamma_{n\kappa}(z+2\pi\delta))), & z < -2\pi\delta,
\end{cases}
\end{aligned} \tag{4}
$$

we obtain a nonlinear system of ordinary differential equations and, equivalently, the following system of one-dimensional nonlinear integral equations w.r.t. $U(n\kappa;\cdot) \in L_2(-2\pi\delta, 2\pi\delta)$ (cf. [2–4, 7, 10, 13, 16]):

$$
\begin{aligned}
& U(n\kappa;z) + \frac{i(n\kappa)^2}{2\Gamma_{n\kappa}} \int_{-2\pi\delta}^{2\pi\delta} \exp(i\Gamma_{n\kappa}|z-z_0|) \times \\
& \times [1 - \varepsilon_{n\kappa}(z_0, \alpha(z_0), U(\kappa;z_0), U(2\kappa;z_0), U(3\kappa;z_0))]\, U(n\kappa;z_0)dz_0 \\
&= \delta_{n1}\frac{i(n\kappa)^2}{2\Gamma_{n\kappa}} \int_{-2\pi\delta}^{2\pi\delta} \exp(i\Gamma_{n\kappa}|z-z_0|)\alpha(z_0)U^2(2\kappa;z_0)\overline{U}(3\kappa;z_0)dz_0 \\
& + \delta_{n3}\frac{i(n\kappa)^2}{2\Gamma_{n\kappa}} \int_{-2\pi\delta}^{2\pi\delta} \exp(i\Gamma_{n\kappa}|z-z_0|)\alpha(z_0)\Big\{\frac{1}{3}U^3(\kappa;z_0) \\
& + U^2(2\kappa;z_0)\overline{U}(\kappa;z_0)\Big\}dz_0 \\
& + U^{\mathrm{inc}}(n\kappa;z), \quad |z| \le 2\pi\delta, \quad n = 1,2,3.
\end{aligned} \tag{5}
$$

Here $U^{\mathrm{inc}}(n\kappa;z) = a_{n\kappa}^{\mathrm{inc}}\exp\left[-i\Gamma_{n\kappa}(z-2\pi\delta)\right]$.

3 Numerical Analysis of the Nonlinear Integral Equations and Spectral Problems

The application of suitable quadrature rules to the system (5) as described in [2–4] leads to a system of complex-valued nonlinear algebraic equations:

$$
(\mathbf{I} - \mathbf{B}_{n\kappa}(\mathbf{U}_\kappa, \mathbf{U}_{2\kappa}, \mathbf{U}_{3\kappa}))\mathbf{U}_{n\kappa} = \delta_{n1}\mathbf{C}_\kappa(\mathbf{U}_{2\kappa}, \mathbf{U}_{3\kappa}) + \delta_{n3}\mathbf{C}_{n\kappa}(\mathbf{U}_\kappa, \mathbf{U}_{2\kappa}) + \mathbf{U}_{n\kappa}^{\mathrm{inc}}, \tag{6}
$$

where $\mathbf{U}_{n\kappa} := \{U_l(n\kappa)\}_{l=1}^N \approx \{U(n\kappa; z_l)\}_{l=1}^N$ and $\{z_l\}_{l=1}^N$ is a discrete set of nodes such that $-2\pi\delta =: z_1 < z_2 < \ldots < z_l < \ldots < z_N =: 2\pi\delta$. $\mathbf{I} := \{\delta_{lj}\}_{l,j=1}^N$ is the identity matrix, $\mathbf{B}_{n\kappa}(\mathbf{U}_\kappa, \mathbf{U}_{2\kappa}, \mathbf{U}_{3\kappa}))$; $\mathbf{C}_\kappa(\mathbf{U}_{2\kappa}, \mathbf{U}_{3\kappa})$, $\mathbf{C}_{n\kappa}(\mathbf{U}_\kappa, \mathbf{U}_{2\kappa})$ are the matrices and the right-hand side, resp., generated by the quadrature method. The solution of (6) can be found iteratively, where at each step a system of linearised nonlinear complex-valued algebraic equations is solved.

The system of nonlinear integral equations (5) can be linearised directly by freezing the permittivities $\varepsilon_{n\kappa}$. The analytic continuation of these linearised nonlinear problems into the region of complex values of the frequency parameter allows us to switch to the analysis of spectral problems. That is, the eigenfrequencies and the eigenfields of the homogeneous linear problems with an *induced* nonlinear permittivity are to be determined. Analogously as above but at the discrete level, we obtain a set of independent systems of linear algebraic equations depending nonlinearly on the spectral parameter:

$$(\mathbf{I} - \mathbf{B}_{n\kappa}(\kappa_n))\mathbf{U}_{\kappa_n} = \mathbf{0}, \tag{7}$$

where $\kappa_n \in \Omega_{n\kappa} \subset \mathsf{H}_{n\kappa}$, at $\kappa = \kappa^{\text{inc}}$, $n = 1,2,3$, $\Omega_{n\kappa}$ are the discrete sets of eigenfrequencies and $\mathsf{H}_{n\kappa}$ denote two-sheeted Riemann surfaces (see [4] and Fig. 2). $\mathbf{B}_{n\kappa}(\kappa_n) := \mathbf{B}_{n\kappa}(\kappa_n; \mathbf{U}_\kappa, \mathbf{U}_{2\kappa}, \mathbf{U}_{3\kappa})$ for $\mathbf{U}_{n\kappa}$ given. The spectral problem of finding

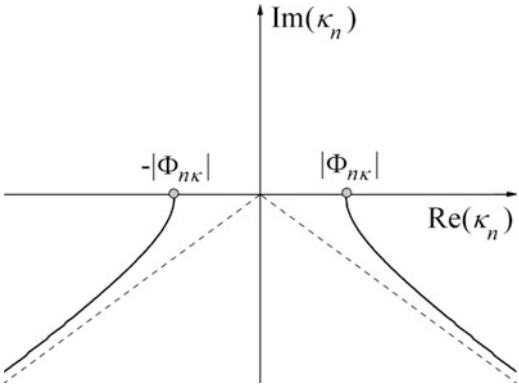

Fig. 2 The two-sheeted Riemann surfaces $\mathsf{H}_{n\kappa}$

the eigenfrequencies κ_n and the corresponding eigenfields $\mathbf{U}_{\kappa_n}$ (i.e. the nontrivial solutions of the linearised homogeneous integral equations) reduces to the following equations:

$$\begin{cases} f_{n\kappa}(\kappa_n) := \det(\mathbf{I} - \mathbf{B}_{n\kappa}(\kappa_n)) = 0, \\ (\mathbf{I} - \mathbf{B}_{n\kappa}(\kappa_n))\mathbf{U}_{\kappa_n} = \mathbf{0}, \\ \kappa := \kappa^{\text{inc}}, \quad \kappa_n \in \Omega_{n\kappa} \subset \mathsf{H}_{n\kappa}, \quad n = 1,2,3. \end{cases} \tag{8}$$

4 Numerical Results: A Single-Layered Structure with a Negative Value of the Cubic Susceptibility

Consider the excitation of the nonlinear structure by a strong incident field at the basic frequency κ and, in addition, by weak incident quasi-homogeneous electromagnetic fields at the double and triple frequencies 2κ, 3κ (see (1)), i.e.,

$$0 < \max\{|a_{2\kappa}^{\mathrm{inc}}|, |a_{3\kappa}^{\mathrm{inc}}|\} \ll |a_{1\kappa}^{\mathrm{inc}}|. \tag{9}$$

The desired solution of the scattering and generation problem (2), (2)–(2) (or of the equivalent problem (5)) is represented as in (4). The solution of (6) is obtained by means of successive approximations using the self-consistent approach based on an iterative algorithm.

In order to describe the scattering and generation properties of the nonlinear structure in the zones of reflection $z > 2\pi\delta$ and transmission $z < -2\pi\delta$, we introduce the following notation:

$$R_{n\kappa} := |a_{n\kappa}^{\mathrm{scat}}|^2 \Big/ \sum_{n=1}^{3} |a_{n\kappa}^{\mathrm{inc}}|^2 \quad \text{and} \quad T_{n\kappa} := |b_{n\kappa}^{\mathrm{scat}}|^2 \Big/ \sum_{n=1}^{3} |a_{n\kappa}^{\mathrm{inc}}|^2, \quad n = 1,2,3.$$

The quantities $R_{n\kappa}$, $T_{n\kappa}$ are called *reflection, transmission* or *generation coefficients* of the waves w.r.t. the total intensity of the incident packet.

We note that, for nonabsorbing media with $\Im\mathfrak{m}\left[\varepsilon^{(L)}(z)\right] = 0$, the energy balance equation

$$\sum_{n=1}^{3} \left[R_{n\kappa} + T_{n\kappa}\right] = 1 \tag{10}$$

is satisfied. This equation generalises the law of conservation of energy which has been treated in [10, 14] for the case of a single incident field and a single equation. If we define by

$$W_{n\kappa} := |a_{n\kappa}^{\mathrm{scat}}|^2 + |b_{n\kappa}^{\mathrm{scat}}|^2 \tag{11}$$

the total energy of the scattered and generated fields at the frequencies $n\kappa$, $n = 1,2,3$, then the energy balance equation (10) can be rewritten as

$$\sum_{n=1}^{3} W_{n\kappa} = \sum_{n=1}^{3} |a_{n\kappa}^{\mathrm{inc}}|^2 .$$

In the numerical experiments, the quantities $W_{3\kappa}/W_{\kappa}$ (which characterises the portion of energy generated in the third harmonic in comparison to the energy scattered in the nonlinear layer) and

$$W^{(\mathrm{Error})} := 1 - \sum_{n=1}^{3} \left[R_{n\kappa} + T_{n\kappa} \right] \tag{12}$$

(which characterises the numerical violation of the energy balance) are of particular interest. We emphasise that in the numerical simulation of scattering and generation processes without any weak fields, i.e., $a^{\mathrm{inc}}_{2\kappa} = a^{\mathrm{inc}}_{3\kappa} = 0$, the residual of the energy balance equation (10) does not exceed the value $|W^{(\mathrm{Error})}| < 10^{-8}$. However, taking into consideration the impact of weak fields in the numerical simulation of the same scattering and generation processes, i.e., $a^{\mathrm{inc}}_{n\kappa} \neq 0$, $n = 2,3$, the error in the balance equation (10) can reach up to several percent. This indicates that the intensities of the exciting weak fields are sufficiently large such that these fields become also sources for the generation of oscillations. For such situations, the presented mathematical model (2), (C1)–(C4) and the linearised nonlinear spectral problems should take into account the complex Fourier amplitudes of the oscillations at the frequencies $n\kappa$ for numbers $n > 3$. Furthermore, we observe, on the one hand, situations in which the influence of a weak field $a^{\mathrm{inc}}_{2\kappa} \neq 0$ on the scattering and generation process of oscillations leads to small errors in the energy balance equation (10) not exceeding 2 % (i.e. $|W^{(\mathrm{Error})}| < 0.02$), and, on the other hand, situations in which the error can reach 10 % (i.e $|W^{(\mathrm{Error})}| < 0.1$ there, where in the region of generation of oscillations, the condition (9) is violated). The scattering, generating, energetic and dielectric properties of the nonlinear layer are illustrated by surfaces in dependence on the parameters of the particular problem. The bottom chart depicts the surface of the value of the residual $W^{(\mathrm{Error})}$ of the energy balance equation (see (12)) and its projection onto the top horizontal plane of the figure. In particular, by the help of these graphs, it is easy to localise that region of parameters of the problem, where the error of the energy balance does not exceed a given value, that is $|W^{(\mathrm{Error})}| < \mathrm{const}$.

The spectral characteristics of the linearised nonlinear problems with the induced dielectric permittivity at the frequencies $n\kappa$, $n = 1,2,3$, of excitation and generation were calculated by means of the algorithm (8). In the graphical illustration of the eigenfields $\mathbf{U}_{\kappa_n}$, we have set $a_{\kappa_n} := 1$ for $\kappa_n \in \Omega_{n\kappa} \subset \mathsf{H}_{n\kappa}$, $n = 1,2,3$. Finally we mention that the later-used classification of scattered, generated or eigenfields of the dielectric layer by the $H_{m,l,p}$-type is identical to that given in [10, 11, 15]. In the case of E-polarisation, $H_{m,l,p}$ (or $TE_{m,l,p}$) denotes the type of polarisation of the wave field under investigation. The subscripts indicate the number of local maxima of $|E_1|$ (or $|U|$, as $|E_1| = |U|$) along the coordinate axes x, y and z (see Fig. 1). Since the considered waves are homogeneous along the x-axis and quasi-homogeneous along the y-axis, we study actually fields of the type $H_{0,0,p}$ (or $TE_{0,0,p}$), where the subscript p is equal to the number of local maxima of the function $|U|$ of the argument $z \in [-2\pi\delta, 2\pi\delta]$.

In what follows we present and discuss results of the numerical analysis of scattering and generation properties as well as the eigenmodes of the dielectric layer with a negative value of the cubic susceptibility of the medium. In more detail, we consider a single-layered structure with a dielectric permittivity

$\varepsilon_{n\kappa}(z,\alpha(z),U(\kappa;z),U(2\kappa;z),U(3\kappa;z)) = \varepsilon^{(L)}(z)+\varepsilon_{n\kappa}^{(NL)}$, $n=1,2,3$, where $\varepsilon^{(L)}(z) := 16$ and $\alpha(z) := -0.01$ for $z \in [-2\pi\delta, 2\pi\delta]$, $\delta := 0.5$, $\kappa^{\text{inc}} := \kappa := 0.375$, and $\varphi_\kappa \in [0°, 90°)$. Figures 3–10 illustrate the following cases of the incident fields:

$a_{2\kappa}^{\text{inc}} = \frac{1}{3}a_\kappa^{\text{inc}}$, $a_{3\kappa}^{\text{inc}} = 0$ ... graphs labeled by "1/3",
$a_{2\kappa}^{\text{inc}} = \frac{2}{3}a_\kappa^{\text{inc}}$, $a_{3\kappa}^{\text{inc}} = 0$ ... graphs labeled by "2/3",
$a_{2\kappa}^{\text{inc}} = a_{3\kappa}^{\text{inc}} = 0$ ... graphs labeled by "0".

The results shown in Fig. 3 allow us to track the dynamic behaviour of the quantity $W_{3\kappa}/W_\kappa$ characterising the ratio of the generated and scattered energies. Figure 3 shows the dependence of $W_{3\kappa}/W_\kappa$ on the angle of incidence φ_κ and on the amplitude a_κ^{inc} of the incident field for different relations between $a_{2\kappa}^{\text{inc}}$ and a_κ^{inc}. It describesthe portion of energy generated in the third harmonic by the nonlinear

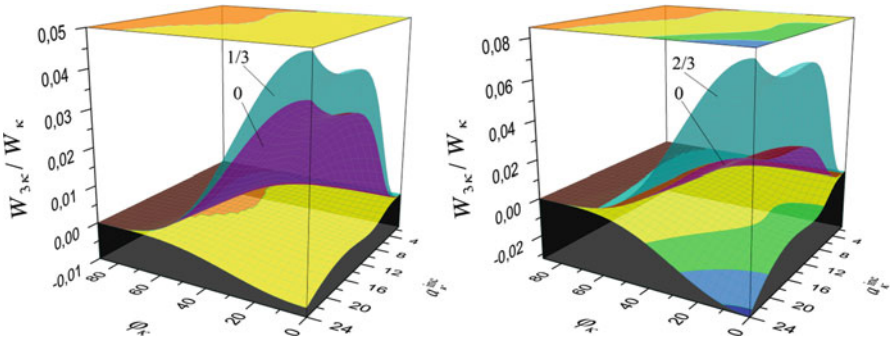

Fig. 3 The portion of energy generated in the third harmonic: $a_{2\kappa}^{\text{inc}} = \frac{1}{3}a_\kappa^{\text{inc}}$ (*left*), $a_{2\kappa}^{\text{inc}} = \frac{2}{3}a_\kappa^{\text{inc}}$ (*right*)

layer when a plane wave at the excitation frequency κ and with the amplitude a_κ^{inc} is passing the layer under the angle of incidence φ_κ. It can be seen that the weaker incident field at the frequency 2κ leads to an increase of $W_{3\kappa}/W_\kappa$ in comparison with the situation where the structure is excited only by a single field at the basic frequency κ. For example, in Fig. 3 the maximum value of $W_{3\kappa}/W_\kappa$ and the value $W^{(\text{Error})}$ are reached at the following parameters $[a_\kappa^{\text{inc}}, a_{2\kappa}^{\text{inc}}, \varphi_\kappa]$: $W_{3\kappa}/W_\kappa = 0.0392$, $W^{(\text{Error})} = 6.00514 \cdot 10^{-9}$, $[a_\kappa^{\text{inc}} = 24, a_{2\kappa}^{\text{inc}} = 0, \varphi_\kappa = 0°]$ ... graph #0 and, taking into consideration the weak field at the double frequency, $W_{3\kappa}/W_\kappa = 0.04937$, $W^{(\text{Error})} = -0.00772$, $[a_\kappa^{\text{inc}} = 24, a_{2\kappa}^{\text{inc}} = \frac{1}{3}a_\kappa^{\text{inc}}, \varphi_\kappa = 0°]$ ... graph #1/3 (left); $W_{3\kappa}/W_\kappa = 0.08075$, $W^{(\text{Error})} = -0.03207$, $[a_\kappa^{\text{inc}} = 24, a_{2\kappa}^{\text{inc}} = \frac{2}{3}a_\kappa^{\text{inc}}, \varphi_\kappa = 0°]$ ... graph #2/3 (right).

The numerical analysis of the processes displayed in Fig. 4 (top) illustrates the portion of energy generated in the third harmonic in dependence on the angle of incidence φ_κ and on the amplitude $a_{2\kappa}^{\text{inc}}$ of the incident field at the double frequency. Here the maximum values of $W_{3\kappa}/W_\kappa$ are reached at $[a_\kappa^{\text{inc}} = 20, a_{2\kappa}^{\text{inc}} = 28, \varphi_\kappa = 0°]$, where we have $W_{3\kappa}/W_\kappa = 0.22277$ and $W^{(\text{Error})} = -0.08986$ (top left), and at $[a_\kappa^{\text{inc}} = 20, a_{2\kappa}^{\text{inc}} = 14, \varphi_\kappa = 0°]$, where we have $W_{3\kappa}/W_\kappa = 0.07336$ and $W^{(\text{Error})} =$

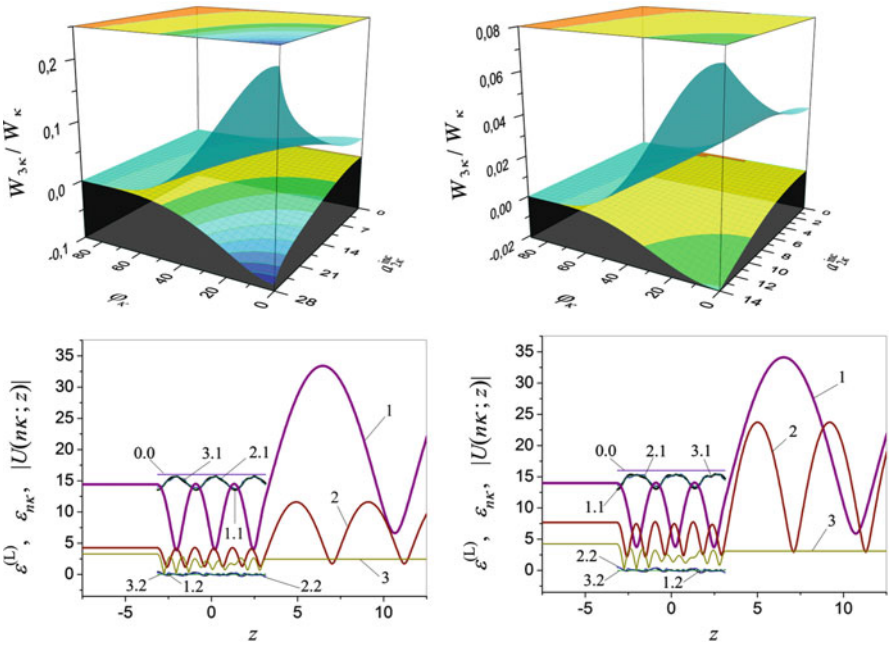

Fig. 4 The dependence of $W_{3\kappa}/W_\kappa$ on φ_κ, $a_{2\kappa}^{\text{inc}}$ for $a_\kappa^{\text{inc}} = 20$ (*top*), some graphs describing the properties of the nonlinear layer for $\varphi_\kappa = 0^\circ$, $a_\kappa^{\text{inc}} = 20$ and $a_{2\kappa}^{\text{inc}} = \frac{1}{3}a_\kappa^{\text{inc}}$ (*bottom left*), $a_{2\kappa}^{\text{inc}} = \frac{2}{3}a_\kappa^{\text{inc}}$ (*bottom right*): #0.0 ... $\varepsilon^{(L)}$, #1 ... $|U(\kappa;z)|$, #2 ... $|U(2\kappa;z)|$, #3 ... $|U(3\kappa;z)|$, #n.1 ... $\Re\mathfrak{e}(\varepsilon_{n\kappa})$, #$n$.2 ... $\Im\mathfrak{m}(\varepsilon_{n\kappa})$

-0.02085 (top right). If the structure is excited by a single field at the basic frequency κ only, then the portion of energy generated in the third harmonic is $\approx 3.4\,\%$, i.e., for $[a_\kappa^{\text{inc}} = 20, a_{2\kappa}^{\text{inc}} = 0, \varphi_\kappa = 0^\circ]$ we have $W_{3\kappa}/W_\kappa = 0.03395$ and $W^{(\text{Error})} = 7.333817 \cdot 10^{-10}$ (top). These data allow us to estimate the increase in the portion of energy generated in the third harmonic. Note also that the violation of condition (9) in the region of generation of oscillations leads to the violation of the energy balance law (10). So $W^{(\text{Error})}[a_\kappa^{\text{inc}} = 20, a_{2\kappa}^{\text{inc}} = 28, \varphi_\kappa = 0^\circ] \approx -0.09$, i.e., the relative error is $\approx 9\,\%$ (Fig. 4 (top left)). If we reduce $a_{2\kappa}^{\text{inc}}$ by half, then (9) is satisfied and we get $W^{(\text{Error})}[a_\kappa^{\text{inc}} = 20, a_{2\kappa}^{\text{inc}} = 14, \varphi_\kappa = 0^\circ] \approx -0.02$, i.e., the relative error is $\approx 2\,\%$ (Fig. 4 (top right)).

The results depicted in Figs. 3 and 4 (top) show that the maximal portion of the total energy generated in the third harmonic is observed in the direction normal to the structure, cf. the behaviour of the surfaces $W_{3\kappa}/W_\kappa$ at $\varphi_\kappa = 0^\circ$. These graphs also show that the weak field with the amplitude $a_{2\kappa}^{\text{inc}}$ only increases the portion of generated energy.

The bottom diagrams in Fig. 4 display some graphs characterising the scattering and generation properties of the nonlinear structure. Graph #0.0 illustrates the value of the linear part $\varepsilon^{(L)} = 16$ of the permittivity of the nonlinear layered structure.

Graphs #n.1 and #n.2 show the real and imaginary parts of the permittivities at the frequencies $n\kappa$, $n = 1,2,3$. The figure also shows the absolute values $|U(\kappa;z)|$, $|U(2\kappa;z)|$ of the amplitudes of the full scattered fields at the frequencies of excitation κ, 2κ (graphs #1, #2) and $|U(3\kappa;z)|$ of the generated field at the frequency 3κ (graph #3). The values $|U(n\kappa;z)|$ are given in the nonlinear layered structure ($|z| \le 2\pi\delta$) and outside it (i.e. in the zones of reflection $z > 2\pi\delta$ and transmission $z < -2\pi\delta$). Here $W^{(\mathrm{Error})} = -4.363084 \cdot 10^{-3}$, i.e., the error in the energy balance is less than 0.44 % (bottom left) and $W^{(\mathrm{Error})} = -1.902471 \cdot 10^{-2}$, i.e., the error in the energy balance is less than 1.9% (bottom right).

Figure 5 shows the numerical results obtained for the scattered and the generated fields in the nonlinear structure and for the residual $W^{(\mathrm{Error})}$ of the energy balance equation (10) for an incident angle $\varphi_\kappa = 0^\circ$ in dependence on the amplitudes a_κ^{inc} at $a_{2\kappa}^{\mathrm{inc}} = \frac{2}{3}a_\kappa^{\mathrm{inc}}$ and $a_{2\kappa}^{\mathrm{inc}} = 0$ of the plane incident waves at the basic frequency κ and at the double frequency 2κ, resp. The figures show the graphs of $\left|U_{n\kappa}\left[a_\kappa^{\mathrm{inc}}, a_{2\kappa}^{\mathrm{inc}}, z\right]\right|$, $n = 1,2,3$, demonstrating the dynamic behaviour of the scattered and the generated fields $|U(n\kappa;z)|$ in the nonlinear layered structure in dependence on increasing amplitudes a_κ^{inc} and $a_{2\kappa}^{\mathrm{inc}}$ for an incident angle $\varphi_\kappa = 0^\circ$ of the plane waves. We mention that, in the range $a_\kappa^{\mathrm{inc}} \in (0,24]$ and $a_{2\kappa}^{\mathrm{inc}} = \frac{2}{3}a_\kappa^{\mathrm{inc}}$ (see Fig. 5) of the amplitudes of the incident fields and for an incident angle $\varphi_\kappa = 0^\circ$ of the plane waves, the scattered field has the type $H_{0,0,4}$ at the frequency κ and $H_{0,0,7}$ at the frequency 2κ. The generated field, observed in the range $a_\kappa^{\mathrm{inc}} \in [4,24]$, is of the type $H_{0,0,10}$ and is converted to the type $H_{0,0,9}$ at the frequency 3κ; see Fig. 5 (bottom left), Fig. 4 (bottom) and [4]. The type conversion $H_{0,0,10} \rightsquigarrow H_{0,0,9}$ of the generated oscillations can occur with an increase of a_κ^{inc} and/or $a_{2\kappa}^{\mathrm{inc}}$; see the surfaces #0 and #2/3 in Fig. 5 (bottom left). This effect is also observed if a weak field with an amplitude $a_{2\kappa}^{\mathrm{inc}}$ in the region of generation of oscillations excites the structure. For example, in Fig. 4, where $a_\kappa^{\mathrm{inc}} = 20$, the graph #3 for $a_{2\kappa}^{\mathrm{inc}} = \frac{1}{3}a_\kappa^{\mathrm{inc}}$ corresponds to the type of oscillation $H_{0,0,10}$ (bottom left), whereas for $a_{2\kappa}^{\mathrm{inc}} = \frac{2}{3}a_\kappa^{\mathrm{inc}}$ it corresponds to the type $H_{0,0,9}$ (bottom right).

The nonlinear parts $\varepsilon_{n\kappa}^{(NL)}$ of the dielectric permittivity at each frequency $n\kappa$ depend on the values $U_{n\kappa} := U(n\kappa;z)$, $n = 1,2,3$, of the fields. The variation of the nonlinear parts $\varepsilon_{n\kappa}^{(NL)}$ of the dielectric permittivity for increasing amplitudes a_κ^{inc} and $a_{2\kappa}^{\mathrm{inc}}$ of the incident fields is illustrated by the behaviour of $\mathfrak{Re}\left(\varepsilon_{n\kappa}\left[a_\kappa^{\mathrm{inc}}, a_{2\kappa}^{\mathrm{inc}}, z\right]\right)$ and $\mathfrak{Im}\left(\varepsilon_{n\kappa}\left[a_\kappa^{\mathrm{inc}}, a_{2\kappa}^{\mathrm{inc}}, z\right]\right)$ at the frequencies $n\kappa$ in Fig. 6 (case $a_{2\kappa}^{\mathrm{inc}} = \frac{2}{3}a_\kappa^{\mathrm{inc}}$). The quantities $\mathfrak{Im}\left(\varepsilon_{n\kappa}\right)$ take both positive and negative values along the height of the nonlinear layer (i.e. in the interval $z \in [-2\pi\delta, 2\pi\delta]$); see Fig. 6 (right). For given amplitudes a_κ^{inc} and $a_{2\kappa}^{\mathrm{inc}}$, the graph of $\mathfrak{Im}\left(\varepsilon_{n\kappa}\left[a_\kappa^{\mathrm{inc}}, a_{2\kappa}^{\mathrm{inc}}, z\right]\right)$ characterises the *loss of energy* in the nonlinear layer at the excitation frequencies $n\kappa$, $n = 1,2$, caused by the *generation* of the electromagnetic field of the third harmonic. Such a situation arises because of the right-hand side of (2) at the triple frequency and the generation which is evoked by the right-hand side of (2) at the basic frequency. In our case $\mathfrak{Im}\left[\varepsilon^{(L)}(z)\right] = 0$ and $\mathfrak{Im}\left[\alpha(z)\right] = 0$; therefore,

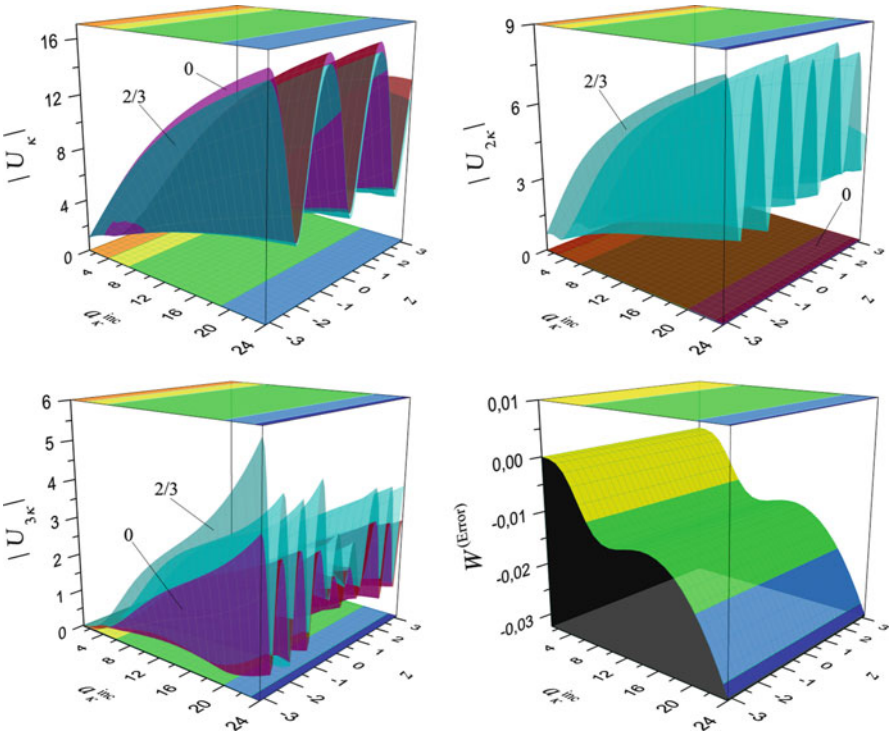

Fig. 5 Graphs of the scattered and generated fields in the nonlinear layered structure in dependence on $\left[a_{\kappa}^{\mathrm{inc}}, a_{2\kappa}^{\mathrm{inc}}, z\right]$ for $\varphi_{\kappa} = 0^{\circ}$ and $a_{2\kappa}^{\mathrm{inc}} = \frac{2}{3} a_{\kappa}^{\mathrm{inc}}$: $|U_{\kappa}|$ (*top left*), $|U_{2\kappa}|$ (*top right*) $|U_{3\kappa}|$ (*bottom left*), and the residual $W^{(\mathrm{Error})}$ (*bottom right*)

$$
\begin{aligned}
&\mathfrak{Im}(\varepsilon_{n\kappa}(z, \alpha(z), U(\kappa;z), U(2\kappa;z), U(3\kappa;z))) \\
&= \alpha(z)\big[\delta_{n1}|U(\kappa;z)||U(3\kappa;z)|\mathfrak{Im}\left(\exp\{i[-3\arg(U(\kappa;z)) + \arg(U(3\kappa;z))]\}\right) \\
&\quad + \delta_{n2}|U(\kappa;z)||U(3\kappa;z)| \\
&\quad \times \mathfrak{Im}\left(\exp\{i[-2\arg(U(2\kappa;z)) + \arg(U(\kappa;z)) + \arg(U(3\kappa;z))]\}\right)\big], \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad n = 1,2,3.
\end{aligned}
\tag{13}
$$

From Fig. 6 (right) we see that small values of $a_{\kappa}^{\mathrm{inc}}$ and $a_{2\kappa}^{\mathrm{inc}}$ induce a small amplitude of the function $\mathfrak{Im}(\varepsilon_{n\kappa})$, i.e., $|\mathfrak{Im}(\varepsilon_{n\kappa})| \approx 0$. The increase of $a_{\kappa}^{\mathrm{inc}}$ corresponds to a strong incident field and leads to the generation of a third harmonic field $U(3\kappa;z)$, and the increase of $a_{2\kappa}^{\mathrm{inc}}$ changes the behaviour of $\varepsilon_{n\kappa}$ (compare the surface #0 with the surface #2/3 in Fig. 6). Figure 6 (right) shows the dynamic behaviour of $\mathfrak{Im}(\varepsilon_{n\kappa})$. It can be seen that $\mathfrak{Im}(\varepsilon_{3\kappa}) = 0$, whereas at the same time the values of $\mathfrak{Im}(\varepsilon_{n\kappa})$, $n = 1,2$, may be *positive or negative* along the height of the nonlinear layer, i.e., in the interval $z \in [-2\pi\delta, 2\pi\delta]$; see (13). The zero values of $\mathfrak{Im}(\varepsilon_{n\kappa})$, $n = 1,2$, are determined by the phase relations between the scattered and the generated fields in the nonlinear layer, namely, at the basic frequency κ by the phase relation

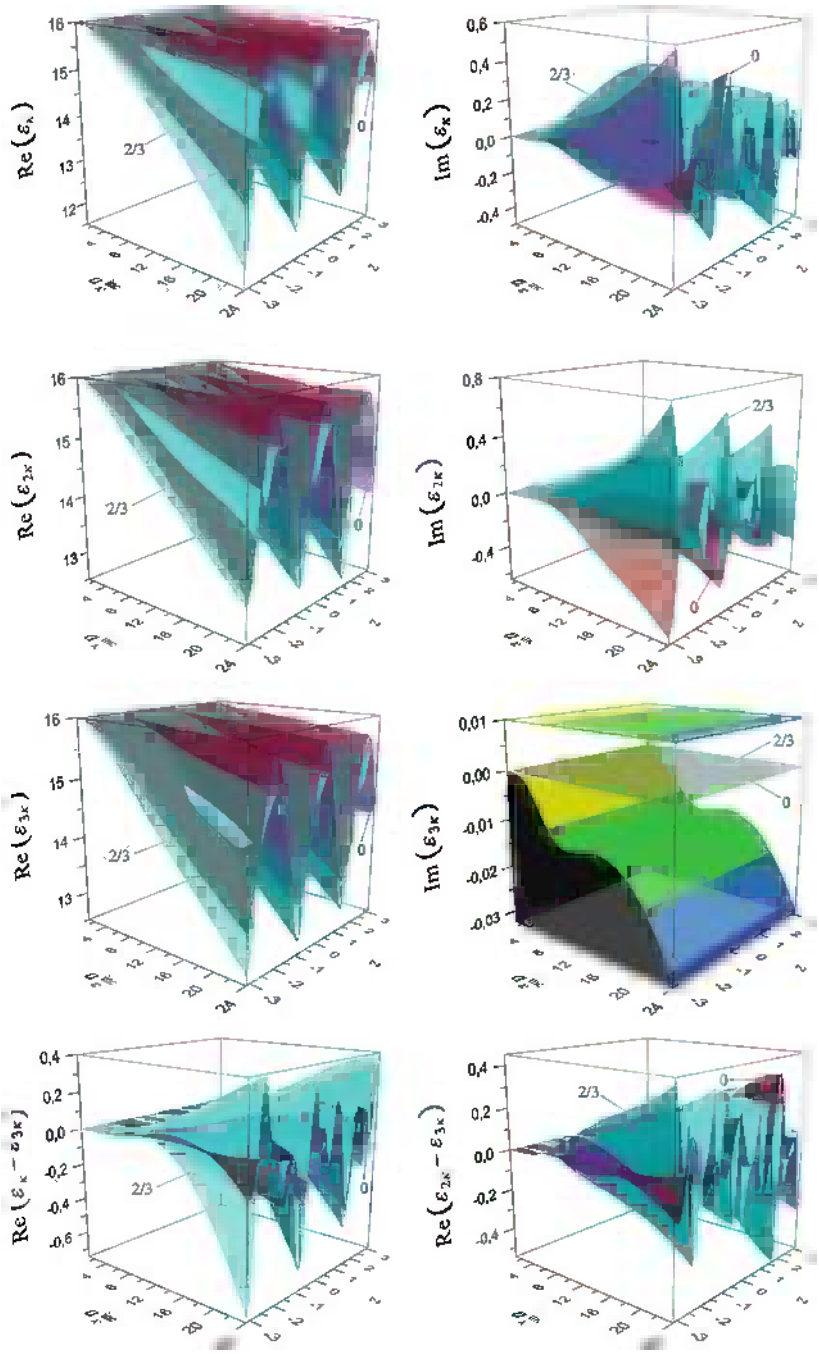

Fig. 6 Graphs characterising the nonlinear dielectric permittivity in dependence on $\left[a_{\kappa}^{\mathrm{inc}}, a_{2\kappa}^{\mathrm{inc}}, z\right]$ for $\varphi_{\kappa} = 0^{\circ}$ and $a_{2\kappa}^{\mathrm{inc}} = \frac{2}{3} a_{\kappa}^{\mathrm{inc}}$: $\mathfrak{Re}\left(\varepsilon_{\kappa}\right)$ (*top left*), $\mathfrak{Im}\left(\varepsilon_{\kappa}\right)$ (*top right*), $\mathfrak{Re}\left(\varepsilon_{2\kappa}\right)$ (second from *top left*), $\mathfrak{Im}\left(\varepsilon_{2\kappa}\right)$ (second from *top right*), $\mathfrak{Re}\left(\varepsilon_{3\kappa}\right)$ (second to the last *left*), $\mathfrak{Im}\left(\varepsilon_{3\kappa}\right)$ (second to the last *right*), $\mathfrak{Re}\left(\varepsilon_{\kappa} - \varepsilon_{3\kappa}\right)$ (*bottom left*), $\mathfrak{Re}\left(\varepsilon_{2\kappa} - \varepsilon_{3\kappa}\right)$ (*bottom right*)

between $U(\kappa;z)$ and $U(3\kappa;z)$ and at the double frequency 2κ by the phases of $\{U(n\kappa;z)\}_{n=1,2,3}$, see (13):

$$\begin{aligned}&\delta_{n1}\left[-3\arg(U(\kappa;z))+\arg(U(3\kappa;z))\right]\\&+\delta_{n2}\left[-2\arg(U(2\kappa;z))+\arg(U(\kappa;z))+\arg(U(3\kappa;z))\right]=p\pi,\\&\qquad\qquad p=0,\pm1,\ldots,\quad n=1,2.\end{aligned}$$

We mention that the behaviour of both the quantities $\mathfrak{Im}(\varepsilon_{n\kappa})$ and

$$\begin{aligned}&\mathfrak{Re}(\varepsilon_{n\kappa}(z,\alpha(z),U(\kappa;z),U(2\kappa;z),U(3\kappa;z))\\&\quad-\varepsilon_{3\kappa}(z,\alpha(z),U(\kappa;z),U(2\kappa;z),U(3\kappa;z)))\\&=\alpha(z)\big[\delta_{n1}|U(\kappa;z)||U(3\kappa;z)|\mathfrak{Re}\left(\exp\{i[-3\arg(U(\kappa;z))+\arg(U(3\kappa;z))]\}\right)\\&\quad+\delta_{n2}|U(\kappa;z)||U(3\kappa;z)|\\&\quad\times\mathfrak{Re}\left(\exp\{i[-2\arg(U(2\kappa;z))+\arg(U(\kappa;z))+\arg(U(3\kappa;z))]\}\right)\big],\\&\qquad\qquad n=1,2,\end{aligned}\tag{14}$$

plays an essential role in the process of third harmonic generation. Figure 6 (bottom) shows the graphs describing the behaviour of $\mathfrak{Re}\left(\varepsilon_{\kappa}\left[a_{\kappa}^{\mathrm{inc}},a_{2\kappa}^{\mathrm{inc}},z\right]-\varepsilon_{3\kappa}\left[a_{\kappa}^{\mathrm{inc}},a_{2\kappa}^{\mathrm{inc}},z\right]\right)$ and $\mathfrak{Re}\left(\varepsilon_{2\kappa}\left[a_{\kappa}^{\mathrm{inc}},a_{2\kappa}^{\mathrm{inc}},z\right]-\varepsilon_{3\kappa}\left[a_{\kappa}^{\mathrm{inc}},a_{2\kappa}^{\mathrm{inc}},z\right]\right)$.

We mention that the impact of a strong electromagnetic field with an amplitude $a_{\kappa}^{\mathrm{inc}}$ even in the absence of a weak field $a_{2\kappa}^{\mathrm{inc}}=0$ (where $U(2\kappa;z)=0$, the surface #0 in Fig. 5 (top right)) induces a nontrivial component of the nonlinear dielectric permittivity at the frequency 2κ. Figure 6 (second from top) shows that the existence of nontrivial values $\mathfrak{Re}(\varepsilon_{2\kappa})\neq\mathfrak{Re}\left(\varepsilon^{(L)}\right)$ and $\mathfrak{Im}(\varepsilon_{2\kappa})\neq0$ is caused by the amplitude and phase characteristics of the fields $U(\kappa;z)$ and $U(3\kappa;z)$. Moreover, the nonlinear component of the dielectric permittivity, which is responsible for the variation of $\mathfrak{Re}(\varepsilon_{n\kappa}-\varepsilon_{3\kappa})$ and $\mathfrak{Im}(\varepsilon_{n\kappa})$, does not depend on the absolute value of the amplitude of the field at the double frequency $|U(2\kappa;z)|$, see (14) and (13). Thus, even a weak field (see #2/3 in Fig. 5 (top right)) includes a mechanism for the redistribution of the energy of the incident wave packet which is consumed for the scattering process and the generation of waves, cf. the dynamics of the surfaces #0 with #2/3 in Figs. 5 and 6.

The scattering and generation properties of the nonlinear structure in the ranges $\varphi_{\kappa}\in[0^{\circ},90^{\circ})$, $a_{\kappa}^{\mathrm{inc}}\in[1,24]$, $a_{2\kappa}^{\mathrm{inc}}=\frac{2}{3}a_{\kappa}^{\mathrm{inc}}$ of the parameters of the incident field are presented in Figs. 7 and 8 (top). The graphs show the dynamics of the scattering ($R_{\kappa}\left[\varphi_{\kappa},a_{\kappa}^{\mathrm{inc}},a_{2\kappa}^{\mathrm{inc}}\right]$, $T_{\kappa}\left[\varphi_{\kappa},a_{\kappa}^{\mathrm{inc}},a_{2\kappa}^{\mathrm{inc}}\right]$, $R_{2\kappa}\left[\varphi_{\kappa},a_{\kappa}^{\mathrm{inc}},a_{2\kappa}^{\mathrm{inc}}\right]$, $T_{2\kappa}\left[\varphi_{\kappa},a_{\kappa}^{\mathrm{inc}},a_{2\kappa}^{\mathrm{inc}}\right]$, see Fig. 7 (top 2)), and generation ($R_{3\kappa}\left[\varphi_{\kappa},a_{\kappa}^{\mathrm{inc}},a_{2\kappa}^{\mathrm{inc}}\right]$, $T_{3\kappa}\left[\varphi_{\kappa},a_{\kappa}^{\mathrm{inc}},a_{2\kappa}^{\mathrm{inc}}\right]$, see Fig. 7 (bottom)) properties of the structure. Figure 8 (top) shows cross sections of the surfaces #0 depicted in Fig. 7 and of the graph #0 of $W_{3\kappa}\left[\varphi_{\kappa},a_{\kappa}^{\mathrm{inc}},a_{2\kappa}^{\mathrm{inc}}\right]$ / $W_{\kappa}\left[\varphi_{\kappa},a_{\kappa}^{\mathrm{inc}},a_{2\kappa}^{\mathrm{inc}}\right]$ (see Fig. 3) by the planes $\varphi_{\kappa}=0^{\circ}$ and $a_{\kappa}^{\mathrm{inc}}=20$.

In Figs. 8–10, a slightly more detailed illustration for the situation of a single incident field (i.e. $a_{2\kappa}^{\mathrm{inc}}=0$) is given, cf. also the graphs #0 in Fig. 7. In the resonant range of wave scattering and generation frequencies, i.e., $\kappa^{\mathrm{scat}}:=\kappa^{\mathrm{inc}}=$

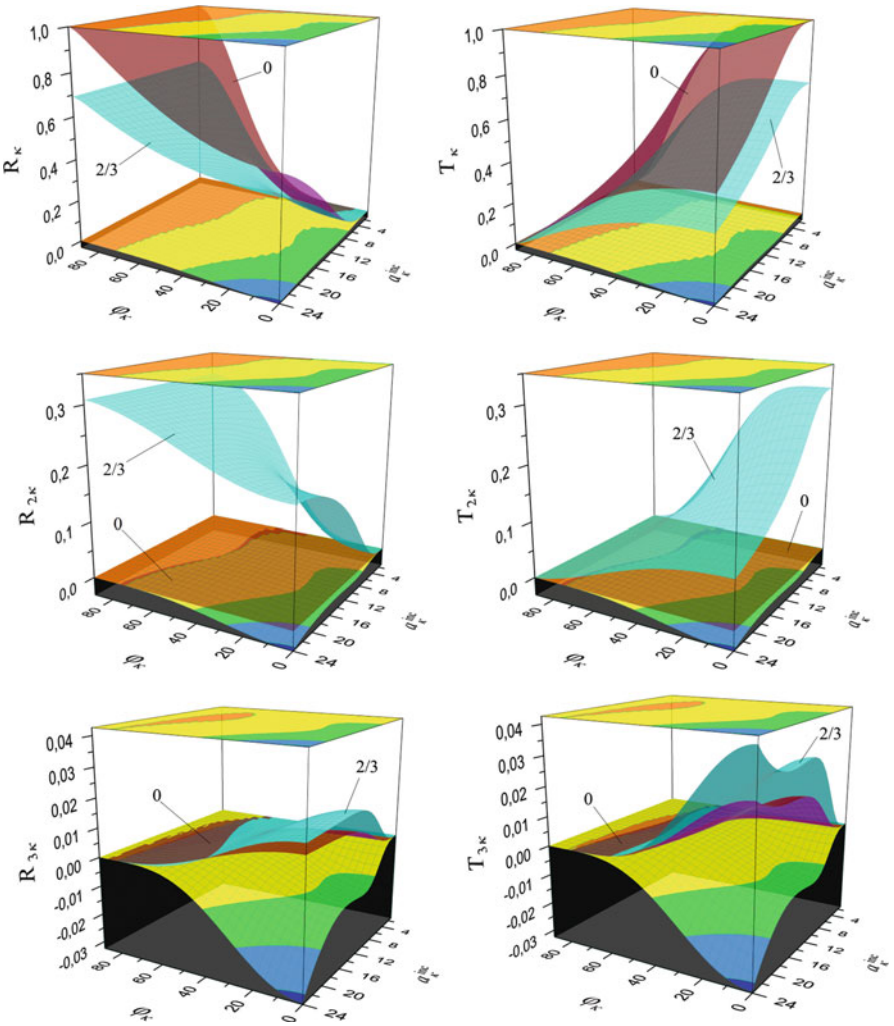

Fig. 7 The scattering and generation properties of the nonlinear structure in dependence on $\left[\varphi_\kappa, a_\kappa^{\text{inc}}, a_{2\kappa}^{\text{inc}}\right]$ for $a_{2\kappa}^{\text{inc}} = \frac{2}{3} a_\kappa^{\text{inc}}$: R_κ, T_κ (*top*), $R_{2\kappa}$, $T_{2\kappa}$ (second from *top*), $R_{3\kappa}$, $T_{3\kappa}$ (*bottom*)

κ and $\kappa^{\text{gen}} = 3\kappa$, resp., the dynamic behaviour of the characteristic quantities depicted in Figs. 8–10 has the following causes. The scattering and generation frequencies are close to the corresponding eigenfrequencies of the linear ($\alpha = 0$) and linearised nonlinear ($\alpha \neq 0$) spectral problems. Furthermore, the distance between the corresponding eigenfrequencies of the spectral problems with $\alpha = 0$ and $\alpha \neq 0$ is small. Thus, the graphs in Fig. 8 (top) can be compared with the dynamic behaviour of the branches of the eigenfrequencies of the spectral problems presented in Fig. 8 (second from top). The graphs of the eigenfields corresponding to the branches of the considered eigenfrequencies are shown in Fig. 8 (bottom).

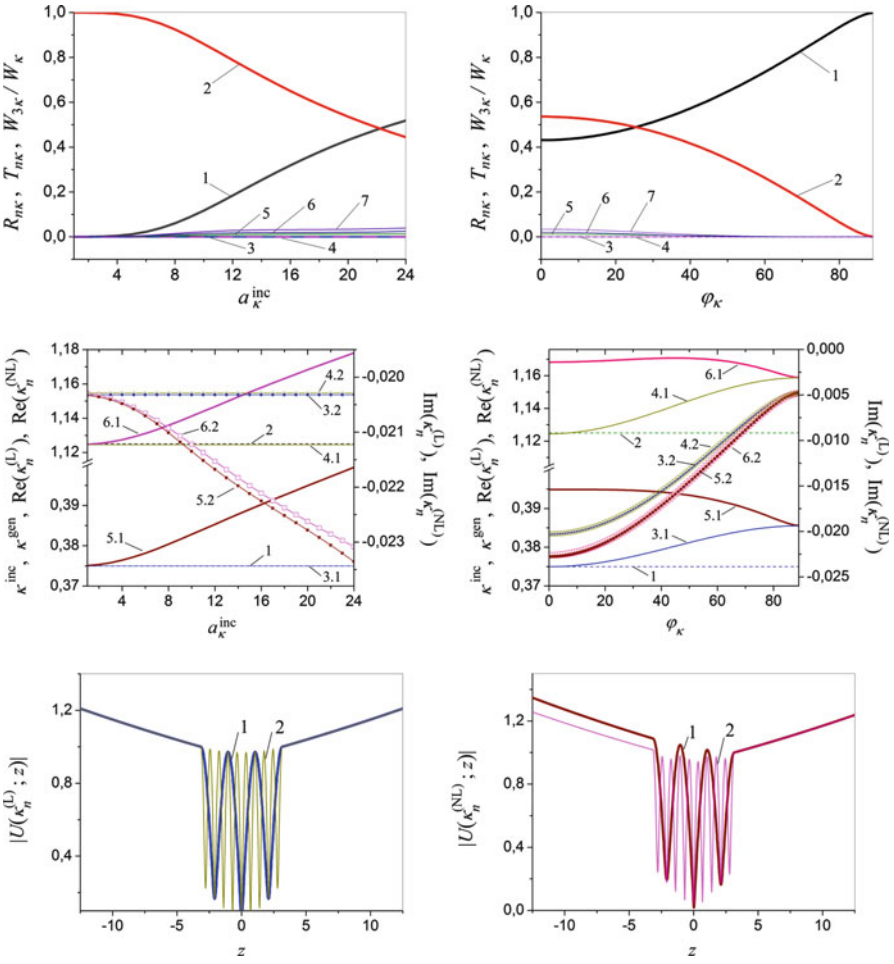

Fig. 8 The curves R_κ (#1), T_κ (#2), $R_{2\kappa}$ (#3), $T_{2\kappa}$ (#4), $R_{3\kappa}$ (#5), $T_{3\kappa}$ (#6), $W_{3\kappa}/W_\kappa$ (#7) for $\varphi_\kappa = 0^\circ$ (*top left*) and $a_\kappa^{\rm inc} = 20$ (*top right*); the curves $\kappa := \kappa^{\rm inc} := 0.375$ (#1), $3\kappa = \kappa^{\rm gen} = 3\kappa^{\rm inc} = 1.125$ (#2), the complex eigenfrequencies $\mathfrak{Re}(\kappa_1^{(L)})$ (#3.1), $\mathfrak{Im}(\kappa_1^{(L)})$ (#3.2), $\mathfrak{Re}(\kappa_3^{(L)})$ (#4.1), $\mathfrak{Im}(\kappa_3^{(L)})$ (#4.2) of the linear problem ($\alpha = 0$) and $\mathfrak{Re}(\kappa_1^{(NL)})$ (#5.1), $\mathfrak{Im}(\kappa_1^{(NL)})$ (#5.2), $\mathfrak{Re}(\kappa_3^{(NL)})$ (#6.1), $\mathfrak{Im}(\kappa_3^{(NL)})$ (#6.2) of the linearised nonlinear problem ($\alpha = -0.01$) for $\varphi_\kappa = 0^\circ$ (second from *top left*) and $a_\kappa^{\rm inc} = 20$ (second from *top right*); the graphs of the eigenfields of the layer for $\varphi_\kappa = 0^\circ$, $a_\kappa^{\rm inc} = 20$. The linear problem ($\alpha = 0$, *bottom left*): $|U(\kappa_1^{(L)}; z)|$ with $\kappa_1^{(L)} = 0.3749822 - i\,0.02032115$ (#1), $|U(\kappa_3^{(L)}; z)|$ with $\kappa_3^{(L)} = 1.124512 - i\,0.02028934$ (#2), the linearised nonlinear problem ($\alpha = -0.01$, *bottom right*): $|U(\kappa_1^{(NL)}; z)|$ with $\kappa_1^{(NL)} = 0.3949147 - i\,0.02278218$ (#1), $|U(\kappa_3^{(NL)}; z)|$ with $\kappa_3^{(NL)} = 1.168264 - i\,0.02262382$ (#2)

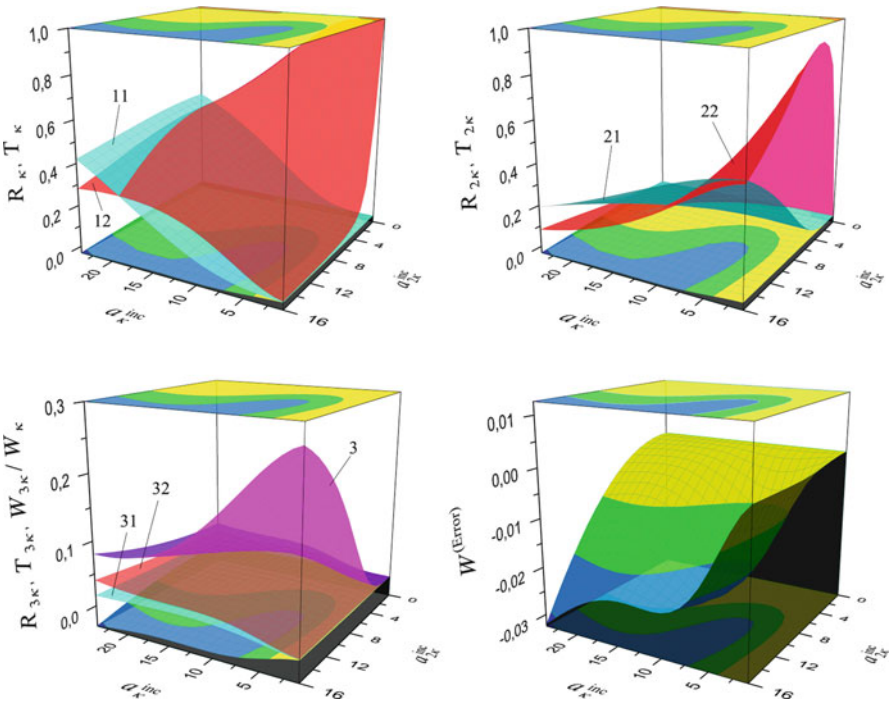

Fig. 9 The scattering and generation properties of the nonlinear structure in dependence on $\left[\varphi_\kappa, a_\kappa^{\mathrm{inc}}, a_{2\kappa}^{\mathrm{inc}}\right]$ for $\varphi_\kappa = 0°$: R_κ, T_κ (#11, #12 *top left*), $R_{2\kappa}$, $T_{2\kappa}$ (#21, #22 *top right*), $W_{3\kappa}/W_\kappa$, $R_{3\kappa}$, $T_{3\kappa}$ (#3, #31, #32 *bottom left*), $W^{(\mathrm{Error})}$ (*bottom right*)

Figure 8 (second from top) illustrates the dispersion characteristics of the linear ($\alpha = 0$) and the linearised nonlinear ($\alpha = -0.01$) layer $\varepsilon_{n\kappa} = \varepsilon^{(L)} + \varepsilon_{n\kappa}^{(NL)}$, $n = 1, 2, 3$. The nonlinear components of the permittivity at the scattering (excitation) frequencies $\kappa^{\mathrm{scat}} := \kappa^{\mathrm{inc}} = \kappa$ and the generation frequencies $\kappa^{\mathrm{gen}} := 3\kappa$ depend on the amplitude a_κ^{inc} and the angle of incidence φ_κ of the incident field. This is reflected in the dynamics of the behaviour of the complex-valued eigenfrequencies of the linear and the linearised nonlinear layer.

We start the analysis of the results of our calculations with the comparison of the dispersion relations given by the branches of the eigenfrequencies (curves #3.1, #3.2 and #5.1, #5.2) near the scattering frequency (curve #1, corresponding to the excitation frequency) and (curves #4.1, #4.2, #6.1, #6.2) near the oscillation frequency (line #2) in the situations presented in Fig. 8 (second from top). The graph #5.1 lies above the graph #3.1 and the graph #6.1 above the graph #4.1. That is, decanalising properties (properties of transparency) of the nonlinear layer occur if $\alpha < 0$.

Comparing the results shown in Fig. 8 (top) and Fig. 8 (second from top), we note the following. The dynamics of the change of the scattering properties R_κ, T_κ of the nonlinear layer (compare the behaviour of curves #1 and #2 in Fig. 8 (top)) depends

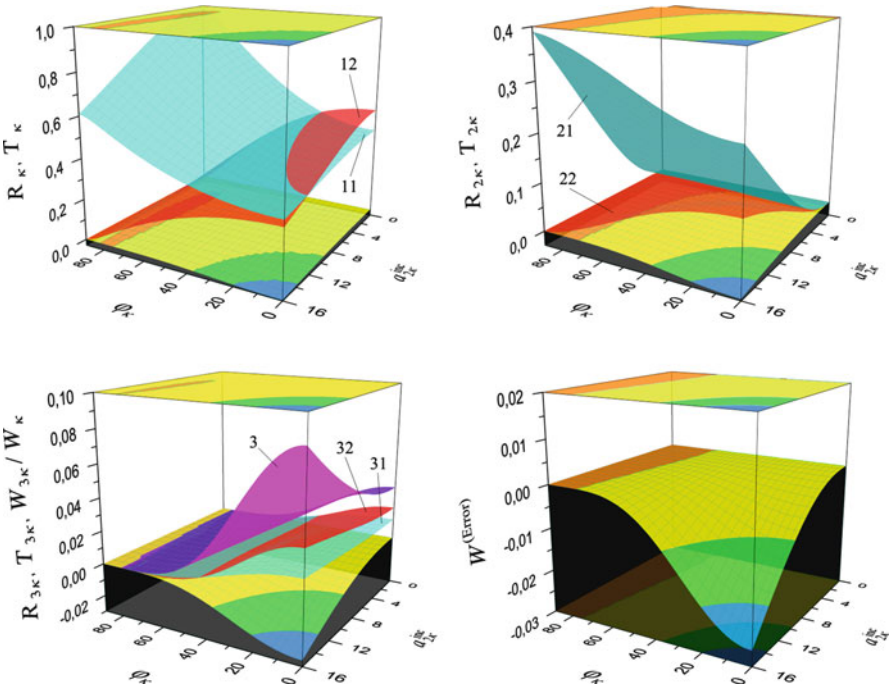

Fig. 10 The scattering and generation properties of the nonlinear structure in dependence on $\left[\varphi_\kappa, a_\kappa^{\text{inc}}, a_{2\kappa}^{\text{inc}}\right]$ for $a_\kappa^{\text{inc}} = 20$: R_κ, T_κ (#11, #12 *top left*), $R_{2\kappa}$, $T_{2\kappa}$ (#21, #22 *top right*), $W_{3\kappa}/W_\kappa$, $R_{3\kappa}$, $T_{3\kappa}$ (#3, #31, #32 *bottom left*), $W^{(\text{Error})}$ (*bottom right*)

on the magnitude of the distance between the curves #3.1 and #5.1 in Fig. 8 (second from top). Decanalising properties of the layer occur when $\alpha < 0$. A previously transparent (Fig. 8 (top left)) or reflective (Fig. 8 (top right)) structure loses its properties. It becomes transparent and the reflection and transmission coefficients become comparable. The greater the distance between the curves #4.1 and #6.1 (see Fig. 8 (second from top)), the greater the values of $R_{3\kappa}$, $T_{3\kappa}$, $W_{3\kappa}/W_\kappa$, characterising the generating properties of the nonlinear layer; see Fig. 8 (top).

The magnitudes of the absolute values of the eigenfields shown in Fig. 8 (bottom) correspond to the branches of the eigenfrequencies of the linear and the linearised nonlinear spectral problems; see Fig. 8 (second from top). The curves in Fig. 8 (bottom) are labeled by #1 for an eigenfield of type $H_{0,0,4}$ and by #2 for an eigenfield of type $H_{0,0,10}$. The loss of symmetry in the eigenfields with respect to the z-axis in Fig. 8 (bottom right) is due to the violation of the symmetry (w.r.t. the axis $z = 0$) in the induced dielectric permittivity at both the scattering (excitation) and the oscillation frequencies; see Fig. 6.

Figures 9 and 10 show the same dependencies as in Fig. 8 (top) but with the additional parameter $a_{2\kappa}^{\text{inc}}$. Here we can track the dynamics of the scattering,

generation and energy characteristics of the nonlinear layer under the influence of the wave package. The incident package consists of a strong and a weak magnetic field with amplitudes a_{κ}^{inc} and $a_{2\kappa}^{\text{inc}}$, resp.

The numerical results presented in this paper were obtained using an approach based on the description of the wave scattering and generation processes in a nonlinear, cubically polarisable layer by a system of nonlinear integral equations (5) and of the corresponding spectral problems by the nontrivial solutions of (8). We have considered an excitation of the nonlinear layer defined by the condition (9). For this case we passed from (5) to (6) and from (7) to (8) by the help of Simpson's quadrature rule. The numerical solution of (6) was obtained using the self-consistent iterative algorithm ([3, 4]). The problem (8) was solved by means of Newton's method. In the investigated range of parameters, the dimension of the resulting systems of algebraic equations was $N = 301$, and the relative error of calculations did not exceed $\xi = 10^{-7}$.

5 Conclusion

We presented results of a computational analysis based on a mathematical model of resonance scattering and generation of waves on an isotropic nonmagnetic nonlinear layered dielectric structure excited by a packet of plane waves in a self-consistent formulation, where the analysis is performed in the domain of resonance frequencies [1, 3, 4]. Here, both the radio [5] and optical [8] frequency ranges are of interest. The wave packets consist of both strong electromagnetic fields at the excitation frequency of the nonlinear structure (leading to the generation of waves) and of weak fields at the multiple frequencies (which do not lead to the generation of harmonics but influence on the process of scattering and generation of waves by the nonlinear structure). The model reduces to a system of nonlinear boundary-value problems which is equivalent to a system of nonlinear integral equations.

The approximate solution of the nonlinear problems was obtained by means of solutions of linear problems with an induced nonlinear dielectric permeability. The analytical continuation of these linear problems into the region of complex values of the frequency parameter allowed us to switch to the analysis of spectral problems. In the frequency domain, the resonant scattering and generation properties of nonlinear structures are determined by the proximity of the excitation frequencies of the nonlinear structures to the complex eigenfrequencies of the corresponding homogeneous linear spectral problems with the induced nonlinear dielectric permeability of the medium.

We presented a collection of numerical results that describe interesting properties of the nonlinear permittivities of the layers as well as their scattering and generation characteristics. In particular, for a nonlinear single-layered structure with decanalising properties, the effect of type conversion of generated oscillations was observed. The results demonstrate the possibility to control the scattering and generating properties of a nonlinear structure via the intensities of its excitation

fields. They also indicate a possibility of designing a frequency multiplier and other electrodynamic devices containing nonlinear dielectrics with controllable permittivity.

Acknowledgements This work was partially supported by the Visby Program of the Swedish Institute and by the joint Russian-Ukrainian RFBR-NASU grant no. 12.02.90425-2012.

References

1. Angermann, L., Shestopalov, Y.V., Yatsyk. V.V.: Modeling and analysis of wave packet scattering and generation for a nonlinear layered structure. In Kiley, E.M., Yakovlev, V.V. (eds.) Multiphysics Modeling in Microwave Power Engineering, pages 21–26, University of Bayreuth, Germany, 2012. 14th Seminar Computer Modeling in Microwave Engineering and Applications, Bayreuth, March pp. 5–6 (2012)
2. Angermann, L., Yatsyk. V.V.: Mathematical models of the analysis of processes of resonance scattering and generation of the third harmonic by the diffraction of a plane wave through a layered, cubically polarisable structure. Int. J. Electromagn. Waves Electron. Syst. **15**(1), 36–49 (2010) In Russian.
3. Angermann, L. Yatsyk, V.V.: Generation and resonance scattering of waves on cubically polarisable layered structures. In: Angermann, L. (ed.) Numerical Simulations–Applications, Examples and Theory, pp. 175–212. InTech, Rijeka/Vienna, Croatia/Austria (2011)
4. Angermann, V. Yatsyk, V.V.: Resonance properties of scattering and generation of waves on cubically polarisable dielectric layers. In: Zhurbenko, V. (ed.) Electromagnetic Waves, pp. 299–340. InTech, Rijeka/Vienna, Croatia/Austria (2011)
5. Chernogor, L.F.: Nonlinear Radiophysics. V.N. Karazin Kharkov National University, Kharkov (2004)
6. Kleinman, D.A.: Nonlinear dielectric polarization in optical media. Phys. Rev. **126**(6), 1977–1979 (1962)
7. Kravchenko, V.F., Yatsyk, V.V.: Effects of resonant scattering of waves by layered dielectric structure with Kerr-type nonlinearity. Int. J. Electromagn. Waves Syst. **12**(12), 17–40 (2007)
8. Miloslavsky, V.K.: Nonlinear Optics. V.N. Karazin Kharkov National University, Kharkov (2008)
9. Shen, Y.R.: The Principles of Nonlinear Optics. Wiley, New York (1984)
10. Shestopalov, V.P., Sirenko, Y.K.: Dynamical Theory of Gratings. Naukova, Dumka, Kiev (1989)
11. Shestopalov, V., Yatsyk, V.V.: Spectral theory of a dielectric layer and the Morse critical points of dispersion equations. Ukrainian J. Phys. **42**(7), 861–869 (1997)
12. Shestopalov, Y.V., Yatsyk, V.V.: Resonance scattering of electromagnetic waves by a Kerr nonlinear dielectric layer. Radiotekhnika i Elektronika (J. Comm. Tech. Electron.) **52**(11), 1285–1300 (2007)
13. Shestopalov, Y.V., Yatsyk V.V.: Diffraction of electromagnetic waves by a layer filled with a Kerr-type nonlinear medium. J. Nonlinear Math. Phys. **17**(3), 311–335 (2010)
14. Vainstein, L.A.: Electromagnetic Waves. Radio i Svyas, Moscow (1988) In Russian.
15. Yatsyk, V.V.: A constructive approach to construction of local equations of irregular dispersion and evolution of fields in a quasi-homogeneous electrodynamic structure. Usp. Sovr. Radioelektroniki **10**, 27–44 (2000) [Translated in: Telecommunications and Radio Engineering, 56(8–9), 89–113 (2001)]
16. Yatsyk, V.V.: About a problem of diffraction on transverse non-homogeneous dielectric layer of Kerr-like nonlinearity. Int. J. Electromagn. Waves Electronic Syst. **12**(1), 59–69 (2007)