EFFICIENT COMPUTATIONAL METHODS FOR ELECTROMAGNETIC IMAGING WITH APPLICATIONS TO 3D MAGNETOTELLURICS

by

Michal Adam Kordy

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Mathematics

The University of Utah

December 2014

Copyright © Michal Adam Kordy 2014

All Rights Reserved

THE UNIVERSITY OF UTAH GRADUATE SCHOOL

STATEMENT OF DISSERTATION APPROVAL

The dissertation of Michał Adam Kordy

has been approved by the following supervisory committee members:

Elena Cherkaev , Chair	09/24/2014
	Date Approved
David Dobson , Member	09/24/2014
	Date Approved
Yekaterina Epshteyn , Member	09/24/2014
	Date Approved
Graeme Milton , Member	09/24/2014
	Date Approved
Philip Wannamaker , Member	09/24/2014
	Date Approved

and by Peter Trapa , Chair of the Department of Mathematics

and by David B. Kieda, Dean of The Graduate School.

ABSTRACT

The motivation for this work is the forward and inverse problem for magnetotellurics, a frequency domain electromagnetic remote-sensing geophysical method used in mineral, geothermal, and groundwater exploration. The dissertation consists of four papers. In the first paper, we prove the existence and uniqueness of a representation of any vector field in H(curl) by a vector lying in H(curl) and H(div). It allows us to represent electric or magnetic fields by another vector field, for which nodal finite element approximation may be used in the case of non-constant electromagnetic properties. With this approach, the system matrix does not become ill-posed for low-frequency. In the second paper, we consider hexahedral finite element approximation of an electric field for the magnetotelluric forward problem. The near-null space of the system matrix for low frequencies makes the numerical solution unstable in the air. We show that the proper solution may obtained by applying a correction on the null space of the curl. It is done by solving a Poisson equation using discrete Helmholtz decomposition. We parallelize the forward code on multicore workstation with large RAM. In the next paper, we use the forward code in the inversion. Regularization of the inversion is done by using the second norm of the logarithm of conductivity. The data space Gauss-Newton approach allows for significant savings in memory and computational time. We show the efficiency of the method by considering a number of synthetic inversions and we apply it to real data collected in Cascade Mountains. The last paper considers a cross-frequency interpolation of the forward response as well as the Jacobian. We consider Pade approximation through model order reduction and rational Krylov subspace. The interpolating frequencies are chosen adaptively in order to minimize the maximum error of interpolation. Two error indicator functions are compared. We prove a theorem of almost always lucky failure in the case of the right hand analytically dependent on frequency. The operator's null space is treated by decomposing the solution into the part in the null space and orthogonal to it.

CONTENTS

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank my advisor, Professor Elena Cherkaev, for many long conversations that directed the course of my research. I would also like to thank my other advisor, Professor Phil Wannamaker, who introduced me to the world of geophysics, supported my work but never put any pressure on me.

I acknowledge my Graduate Committee members, Professor David Dobson, Professor Yekaterina Epshteyn, and Professor Graeme Milton, for many helpful comments and guidance. I am grateful to the faculty members of the Department of Mathematics for teaching many interesting courses in applied mathematics and to fellow graduate students for the company in the graduate student life.

I would like to acknowledge help of Virginia Maris for testing our code and for many helpful suggestions. I would like to thank Graham Hill, for sharing MT data of Mount St. Helens and all the members of electromagnetic geophysical community, for a very warm welcome, many helpful conversations, and great problems on which to work.

Many thanks to my wife, Kasia, for continuous support and to my children, Hania and Pawel, for motivating me to graduate sooner than later.

CHAPTER 1

INTRODUCTION

Magnetotellurics (MT) is a deep electromagnetic sounding used to image the earth's subsurface up to depths of 50km and more. It is used to look deep into the crust to help understand the geological processes. The main source of signal is due to solar wind and lightning, which is approximated by a plane wave traveling downwards. When the wave encounters conducting rock, it induces currents, which in turn act as a source of electromagnetic waves, which are emitted in all directions, also upwards. Thus the electromagnetic field measured on the surface is a superposition of the source wave, traveling downwards and the wave traveling upwards, which carries information about the conductivity of the subsurface. The smaller the frequency of the wave, the deeper the wave penetrates. Thus the MT response is usually considered for a number of frequencies, which give information for a range of depths. Additionally, in a 3D MT survey, a number of receivers are placed in a spatial grid on the earth's surface. Both of those techniques allow one to obtain information of the conductivity of the subsurface as a function of location in 3D space.

As the source of the wave traveling upwards is a conductive rock, MT is intrinsically more sensitive to conductive structures. This is the reason why the main application of this method is in geothermal energy exploration, where the target reservoirs contain hot brine, which is electrically conductive. Thermal changes often cause development of a clay cap above the reservoir, which is even more conductive and may be detected with MT. Recently, there is more interest in using MT and other electromagnetic methods in mineral and hydrocarbon exploration. For the latter, the targets are rocks rich in hydrocarbons that are electrically very resistive, thus as the electromagnetic methods get more sensitive, more hydrocarbon deposits will be found. This makes the field of electromagnetic geophysical methods very dynamic

and exciting.

One of the difficulties the field experiences is a need for fast and stable 3D inversion code. A fast 3D inversion requires a fast 3D forward problem, as it has to be solved many times in the inversion process. This was a motivation for the current work.

The thesis consists of four papers that are put into separate chapters. In the first paper, presented in Chapter 2, we prove a theorem that every member K of $\mathcal{H}_0(\nabla\times)$ for a scalar function κ can be represented by a vector field $F \in \mathcal{H}_0(\nabla\times) \cap \mathcal{H}(\nabla\cdot)$ in a form

$$K = F - \nabla(\kappa\nabla \cdot F)$$

This is an interesting representation that, simplifying a little, says that even if a vector has discontinuous normal component across some surface, it may be represented using a continuous vector field, where the information about the jumps of normal component of K is stored in the divergence of F. The vector field F is called a Shelkunoff potential and we apply it in the numerical approximation of the electromagnetic field at low frequencies, as the electric field E as well as the magnetic field H are members of $\mathcal{H}(\nabla\times)$. In this case, κ is a function of electromagnetic properties that does not need to be constant. The advantage of this approach is that one can use nodal finite element approximation and the system matrix does not suffer from a very large null space as is the case for a curl-curl equation for an electric field.

In the next paper, which is presented in Chapter 3, we consider a finite element approximation of the electric field at low frequency and we apply it to the magnetotelluric forward problem. Particular difficulty arises from the presence of the air in part of the domain. The smallness of conductivity in the air makes the system matrix very ill-conditioned. Because of that, even if a direct solver is used, for low frequencies, the approximation of the electric field in the air consists of almost pure numerical error. We show that this solution still contains the information about the electric field values that may be extracted if a correction on the null space of the curl is applied. Using a compatible approximation through edge elements, using discrete Helmholtz decomposition, the correction may be applied by solving the Poisson equation on the same mesh as is used for approximation of the electric field. The correction makes sure that there are no spurious current sources and is called divergence correction.

We implement a finite element approximation of the electric field on a hexahedral mesh and we parallelize the code on a symmetric multiprocessor workstation with large RAM. We show the accuracy of the approximation by considering a number of magnetotelluric models and compare the approximated field values with the other codes.

The next paper describes the use of this forward code in the inversion of magnetotelluric data, which given MT measurements, seeks a conductivity model in the subsurface that is able to replicate the measurements. We regularize the ill-posed inversion by using the second norm of the gradient of the logarithm of conductivity and consider various spatial weightings for the second norm. To minimize the inversion functional, we employ the Gauss-Newton approach, which is suitable for our quadratic regularization functional. The model update in Gauss-Newton is low rank. The update is in the space of dimension equal to the number of real valued measurements. This gives rise to the data space Gauss-Newton, which allows for significant savings in memory and computational time. We show that it is possible to use the same regularization as for the model space Gauss-Newton, and in our case, it is required to apply an inverse of a matrix to each row of the Jacobian. The considered matrix has a similar non-zero pattern to a matrix arising from finite difference approximation of a Poisson problem, thus the time needed for the calculation is not significant. We also implement inversion for static distortion matrices, which allows us to account for small-scale conductivity anomaly close to an MT receiver. We test the inversion code with a couple of synthetic MT models and apply it to real data collected in the Cascade Mountains.

The last paper considers a speedup of calculation of the forward response and Jacobian, through cross-frequency interpolation. The approximation is done by noticing that in both cases, one needs to evaluate transfer function $(\tilde{A} + \mathrm{i}\omega\tilde{B})^{-1}\tilde{b}$ for a range of frequencies $\omega \in [\omega_{\min}, \omega_{\max}]$. In the case of the forward response, $\tilde{b}$ depends on frequency and in the case of the Jacobian, it does not. We consider a Pade approximation through model order reduction using rational Krylov subspace. We consider an adaptive choice of shifts in a strategy that tries to minimize the maximum error of approximation. For the case of $\tilde{b}$ having an analytic dependence

on frequency, we prove a theorem of almost always lucky failure. We consider two error indicator functions. As in our case, matrix $\tilde{A}$ has a non-empty null space, we propose to decompose $\tilde{b}$ into the part in the null space and orthogonal to it. This is accomplished using a discrete Helmholtz decomposition, in a similar manner to divergence correction, and in the case of $\tilde{b}$ dependent on frequency, it allows us to decrease the relative error of approximation by two orders of magnitude. Overall, for a moderately large MT survey, the described cross-frequency interpolation is able to speed-up the inversion 4 times.

CHAPTER 2

VARIATIONAL FORMULATION FOR MAXWELL'S EQUATIONS WITH LORENTZ GAUGE: EXISTENCE AND UNIQUENESS OF SOLUTION[1]

Kordy M.[2,3], Cherkaev E.[2], and Wannamaker P.[3]

2.1 Abstract

We develop the finite element method based on nodal shape functions for simulation of low-frequency electromagnetic fields in geophysical applications. The existence and uniqueness of the vector-scalar potential for Maxwell's equations with a Lorentz gauge is proven for a conducting medium with piecewise constant properties. A variational formulation based on the Schelkunoff potential is considered for both the electric field E and the magnetic field H. A regularized formulation for the magnetic field is obtained for the case when magnetic permeability μ is constant and thus the magnetic field is divergence free. In the case of non-divergence-free H field, an equation involving scalar and vector potentials is presented. The solution to both problems may be approximated by nodal shape functions in the finite element method with system matrices that remain well-conditioned for low frequencies. A numerical

[1]Submitted to International Journal of Numerical Analysis and Modeling in 2014

[2]Department of Mathematics, University of Utah

[3]Energy & Geoscience, University of Utah

study of a forward problem of computation of electromagnetic fields in the diffusive electromagnetic regime shows the efficiency of the proposed method.

2.2 Introduction

Fast and stable methods are needed for calculating electromagnetic (EM) fields in and over the Earth. Such simulation has applications in imaging of subsurface electrical conductivity structures related to exploration for geothermal, mining, and hydrocarbon resources. Over commonly used frequencies, EM propagation in the Earth is diffusive since conduction dominates over dielectric displacement. The Finite Element Method (FEM) is attractive for this simulation in comparison with other techniques in that it may be easily adapted to complex boundaries between regions of constant EM properties, including topography or bathymetry. The 3D interpretation of geophysical data is numerically expensive, as the simulation (forward problem) needs to be computed many times [1–3].

For large-scale simulation problems, iterative methods have been the ones of choice to solve the linear system matrices resulting from FEM formulations [4–8]. The speed of iterative methods is strongly related to the properties of the variational problem used. Difficulties arise when the computational domain includes high contrast, both non-conducting air and a conducting medium in the Earth subsurface, especially for low frequencies. Furthermore, the Earth's subsurface in general is characterized by spatially changing conductivity, dielectric permittivity and magnetic permeability. These can slow or prevent iteration convergence [9, 10].

There have been multiple approaches to addressing the difficulties encountered with high physical property contrasts and potentially discontinuous EM field variables. One is to apply special finite elements, so-called edge elements, that have a discontinuous normal component of the vector field across elements, while keeping the tangential field component continuous [11–13]. Edge elements are also compatible with the curl operator and are a part of the de Rham diagram [14]. However, if the curl-curl equation for the electric field E is used, and if the conductivity is very small in part of the domain (e.g., in the air) or if frequencies are very low, the problem becomes ill-posed and the system matrix has a very large, near null space. This requires use of sophisticated preconditioners that handle the null space of the curl

properly in order to use iterative solvers. Such preconditioners have been developed (see [15–20]).

An alternative is to not solve directly for the EM fields themselves, but instead to initially solve a well-conditioned equation for a quantity that is continuous. Subsequently, the EM fields are obtained through spatial differentiation with the field discontinuities defined by the property jumps. One such quantity is the Schelkunoff potential [21–24], which we examine in this paper. In general, this potential has both scalar and vector components, and there are both electric and magnetic versions. The scalar potential can be expressed as a function of the vector potential, and as a result, only the vector potential is needed to represent the EM field. The Schelkunoff potential approach for non-constant EM properties, to the best of our knowledge, lacks theoretical justification. It is essential to know under what conditions such a potential exists, and what boundary conditions make it unique, before attempting to approximate it numerically.

In this paper, we show that a continuous Schelkunoff potential exists for both the electric and magnetic fields with assumptions valid for low frequency. Specifically, we show that for spatially varying and complex valued κ, a Schelkunoff potential representation $E = F - \nabla(i\kappa\nabla \cdot F)$ exists for every member E of $\mathcal{H}_0(\nabla\times)$. We discuss the use of this potential for FEM approximation of the EM field. First we consider the possibility of using just the vector potential, which is enough to represent the field. Yet in the case of the electric Schelkunoff potential, when the conductivity σ is not constant and the electric field is not divergence-free, finding a weak equation involving only the vector potential is difficult. In particular, we show that the potential does not satisfy the weak form of Helmholtz equation, sometimes erroneously used as a basis for FEM simulation [23]. For the general case of non divergence-free EM fields, we propose a mixed formulation involving scalar and vector potentials.

We consider also the case of the magnetic Schelkunoff potential. If the magnetic permeability μ is constant, the magnetic field is divergence-free and the vector potential coincides with the magnetic field. We show that the Schelkunoff potential approach leads to a regularized weak equation for the magnetic field involving the divergence term, and as a result, the equation does not suffer from the large near null

space problem.

We show that a bilinear form of the equations for both magnetic vector potential and electric scalar-vector formulations remains coercive at low frequencies. It makes iterative solvers fast even if only standard vector multigrid preconditioners [25] are used. Another advantage is that continuous Schelkunoff potentials (more precisely, a member of $\mathcal{H}(\nabla\times)\cap\mathcal{H}(\nabla\cdot)$) allow us to use nodal shape functions, which have more widely available implementations than edge elements. Edge elements, due to discontinuity of the shape functions across elements boundaries, require postprocessing to get the value of the field at a specific point within an element. In geophysical applications, the domain is a convex polygon, so nodal discretization is dense in $\mathcal{H}_0(\nabla\times)\cap\mathcal{H}(\nabla\cdot)$ or in $\mathcal{H}(\nabla\times)\cap\mathcal{H}_0(\nabla\cdot)$ [14, 26].

Regularization of the curl-curl equation using the divergence term has also been suggested in [26, 27]. The current paper extends these ideas to the case of non-constant, complex valued electromagnetic properties and non-divergence-free fields. In [27], the authors consider existence, uniqueness, and proper boundary conditions for a Schelkunoff-like vector potential only for the case of constant electromagnetic properties. In [26], the authors consider non-constant properties; however, they seek a solution $E\in\mathcal{H}(\nabla\times)$ such that $\sigma E\in\mathcal{H}(\nabla\cdot)$. If σ is not constant, it is difficult to construct a compatible finite element discretization for the space of vector fields of the suggested kind.

In this paper, we consider a different approach. The original vector electric Schelkunoff potential F and the vector electric field E differ by $\nabla\varphi$. The scalar potential φ satisfies a Poisson equation for which the source term is given by the jumps of the normal component of E across boundaries of regions with different EM properties. Separating the discontinuities of the electric field onto φ allows the vector potential to be continuous, or more precisely to lie in the space $\mathcal{H}(\nabla\times)\cap\mathcal{H}(\nabla\cdot)$, which allows us to approximate it using nodal elements.

A representation of an electric field related to ours was considered in ([24] Lorentz gauge #2), where the authors proved the uniqueness of a continuous Schelkunoff potential for a non-lossy medium using a mixed formulation that involved both scalar and vector potentials. The mixed formulation involving scalar and vector potentials

considered in the current paper is a reformulation of this approach for a medium with losses. We prove not only uniqueness, but also existence of the solution (Theorem 6).

The structure of the paper is as follows. In Section 2.3 next, a brief description of the electric Schelkunoff potential is given in the way it typically appears in the literature. We also show that it satisfies the Helmholtz equation if the electromagnetic properties are constant.

In Section 2.4, a general definition of the electric Schelkunoff potential is given for the case when no conditions are imposed on the boundaries between regions of constant electromagnetic properties. Then, an existence and uniqueness theorem for a continuous Schelkunoff potential is formulated and proven. In Section 2.5, the difficulty in obtaining a weak equation involving only the vector electric Schelkunoff potential is presented.

In Section 2.6, the approach suggested in [24] is discussed and reformulated for a lossy medium. As a result, a mixed formulation involving a scalar and vector potential for electric Schelkunoff potential is developed.

In Section 2.7, a different approach is suggested to proceed from the difficulties with the electric potential. A magnetic Schelkunoff potential is defined and, in the situation where magnetic permeability μ is constant, an appealing weak form of the governing equation is derived.

The last Section (2.8) shows results of numerical simulations. We use the developed magnetic Schelkunoff potential approach to calculate the electromagnetic field generated by a conductive brick in a resistive whole space with a plave-wave (magnetotelluric) source. Comparison of the results with calculations done by an independent Integral Equations code [28] is shown. Good agreement between the calculated fields provides a verification of the validity of the method.

2.3 Lorentz gauge formulation of Maxwell's equations

Let us consider the electromagnetic field satisfying Maxwell's equations in the frequency domain, with time dependence $e^{i\omega t}$, with the electric source J^{imp}, in some bounded domain $\Omega \subset \mathbb{R}^3$:

$$\begin{cases} \nabla \times E &= -i\omega\mu H \\ \nabla \times H &= \hat{\sigma}E + J^{imp} \end{cases}, \qquad \hat{\sigma} = \sigma + i\omega\epsilon \tag{2.1}$$

Here, σ and ϵ are conductivity and complex permittivity of the medium, μ is magnetic permeability, and ω is the frequency of the applied field.

The Schelkunoff potential, or electric Schelkunoff potential, is a vector potential F used together with a scalar potential ψ to represent the electric field E [21–24] in a form:

$$E = -i\omega F - \nabla\psi \tag{2.2}$$

A relationship between F and ψ, called the Lorentz gauge, is imposed:

$$\nabla\left(\frac{\nabla \cdot F}{\hat{\sigma}\mu}\right) = -\nabla\psi \tag{2.3}$$

As a result, the electric field is represented as:

$$E = -i\omega F + \nabla\left(\frac{\nabla \cdot F}{\hat{\sigma}\mu}\right) \tag{2.4}$$

Substituting the first equation into the second one in (2.1) and using (2.2) to represent electric field E, in the region of *constant properties* $\hat{\sigma}, \mu$ we obtain:

$$\nabla \times \left(\nabla \times \frac{1}{\mu}F\right) = J^{imp} - \hat{\sigma}i\omega F - \hat{\sigma}\nabla\psi$$

Application of the vector identity (2.49) results in:

$$\nabla\left(\nabla \cdot \frac{1}{\mu}F\right) - \nabla \cdot \left(\nabla\left(\frac{1}{\mu}F\right)\right) = J^{imp} - \hat{\sigma}i\omega F - \hat{\sigma}\nabla\psi$$

If the equation is multiplied by $-\mu$ (again, it is assumed that $\hat{\sigma}, \mu$ are constant), the Lorentz gauge (2.3) is used, and the following vector Helmholtz equation is obtained:

$$\Delta F - i\hat{\sigma}\mu\omega F = -\mu J^{imp} \tag{2.5}$$

Yet the potential satisfies this equation only if the electromagnetic properties are constant. The weak form of the Helmholtz equation, which is a separate equation for each component F_k of the vector field, $k = 1, 2, 3$, for any $A_k \in \mathcal{H}^1(\Omega)$,

$$\int_\Omega \nabla A_k \cdot \nabla F_k + i\omega \int_\Omega \hat{\sigma}\mu F_k \cdot A_k = \int_\Omega \mu J_k^{imp} \cdot A_k \tag{2.6}$$

imposes conditions on the boundaries between regions of different $\hat{\sigma}, \mu$ listed below:

1. F_k is continuous, $k = 1, 2, 3$
2. $\frac{\partial}{\partial n} F_k$ is continuous, $k = 1, 2, 3$

In Section 2.4 existence and uniqueness of an electric Schelkunoff potential satisfying those conditions is investigated. As it turns out, with some reasonable assumptions when $\hat{\sigma}, \mu$ are not constant, a continuous electric Schelkunoff potential (condition 1 satisfied) exists, yet the condition 2 is not satisfied. As a result, there is no electric Schelkunoff potential that satisfies the weak form of Helmholtz equation (2.6), so it should not be used as a basis for the Finite Element Method if the electromagnetic properties are not constant.

2.4 Existence and uniqueness of a *continuous* electric Schelkunoff potential

Let us consider a domain $\Omega \subset \mathbb{R}^3$, which is an open bounded set with Lipschitz boundary. The domain is divided into a finite number of disjoint open regions $V_j, j \in I$ (I is a finite set of indices), such that

$$\bigcup_{j \in I} V_j \subset \Omega \subset \bigcup_{j \in I} \overline{V}_j$$

It is assumed that the properties $\hat{\sigma}, \mu$ are constant in each V_j, but may differ between those regions. In other words, the properties $\hat{\sigma}, \mu$ are piecewise constant in Ω.

It is assumed not only that Maxwell's equations (2.1) are satisfied in strong sense in each V_j but also that they are satisfied in weak sense in Ω.

It is assumed further that all considered functions are $C^\infty(\overline{V_j})$ for $j \in I$, yet the functions may have jumps in value or derivatives across boundaries between sets V_j. These assumptions greatly simplify the analysis and allow us to develop physical intuition about the Schelkunoff potential. A more rigorous reasoning regarding existence and uniqueness of Schelkunoff potential is presented as Theorem 8 in the Appendix (Section 2.9). Schelkunoff potential is a member of $\mathcal{H}(\nabla\times) \cap \mathcal{H}(\nabla\cdot)$, which with our simplifying assumption is equivalent to being continuous.

As a definition of an electric Schelkunoff potential, we will take a vector field that is properly defined in each region V_j, no matter what the boundary conditions between regions V_j are.

Definition 1 *An electric Schelkunoff potential is any vector field $F \in C^\infty(\overline{V_j})$ for all $j \in I$, satisfying*

$$E = -i\omega F + \nabla\left(\frac{\nabla \cdot F}{\hat{\sigma}\mu}\right) \qquad in\ V_j \quad \forall j \in I \tag{2.7}$$

where E is the electric field satisfying Maxwell's Equations (2.1) in weak sense in Ω.

It turns out that if we assume that the source J^{imp} is piecewise divergence free, then the electric field E multiplied by a constant is an electric Schelkunoff potential, which is expressed in the following lemma.

Lemma 2 *If $\nabla \cdot J^{imp} = 0$ in V_j for all $j \in I$, then $\nabla \cdot E = 0$ in V_j for all $j \in I$ and*

$$F_E = -\frac{1}{i\omega}E \tag{2.8}$$

is an electric Schelkunoff potential.

Proof : As Maxwell's Equations (2.1) are satisfied in strong sense in each of V_j, then the following equation is satisfied in strong sense:

$$\nabla \times \left(-\frac{1}{i\omega\mu}\nabla \times E\right) = J^{imp} + \hat{\sigma}E \quad \text{in } V_j$$

Taking divergence of both sides yields

$$\nabla \cdot J^{imp} = -\nabla \cdot (\hat{\sigma}E) = -\hat{\sigma}\nabla \cdot E$$

and the last equality holds as $\hat{\sigma}$ is constant in V_j. As a result, if $\nabla \cdot J^{imp} = 0$, then

$$\nabla \cdot E = 0 \quad \text{in } V_j$$

If we define F_E according to (2.8), then

$$\nabla \cdot F_E = \nabla \cdot \left(-\frac{1}{i\omega}E\right) = -\frac{1}{i\omega}\nabla \cdot E = 0$$

Hence the equation (2.7), defining electric Schelkunoff potential, is satisfied in strong sense in each V_j:

$$-i\omega F_E + \nabla\left(\frac{\nabla \cdot F_E}{\hat{\sigma}\mu}\right) = -i\omega F_E + 0 = -i\omega\left(-\frac{1}{i\omega}E\right) = E$$

This is exactly what is needed for F_E to be the electric Schelkunoff potential according to Definition 1.

The fact that the equation (2.7) defining electric Schelkunoff potential is linear allows us to state the conditions for the vector field F to be an electric Schelkunoff potential as a condition on the difference $F - F_E$. This is expressed in the following theorem.

Lemma 3 *If $\nabla \cdot J^{imp} = 0$ in each V_j, then the following statements are equivalent:*

1. *F is an electric Schelkunoff potential (definition 1)*
2. *$F = F_E + K = \frac{1}{i\omega}E + K$, where K satisfies*

$$-i\omega K + \nabla\left(\nabla \cdot \frac{K}{\hat{\sigma}\mu}\right) = 0 \tag{2.9}$$

   in strong sense in each V_j.

3. *$F = F_E + \nabla\varphi = \frac{1}{i\omega}E + \nabla\varphi$, where φ satisfies*

$$\Delta\varphi - i\omega\hat{\sigma}\mu\,\varphi = 0 \tag{2.10}$$

   in strong sense in each V_j.

Proof : Assume 1. As both F and F_E satisfy the equation (2.7) defining electric Schelkunoff potential, the equation (2.9) for K is obtained by subtracting the equations for F and F_E.

Moreover, from (2.9) it follows that in each V_j, K is a gradient of some scalar function, $K = \nabla\left(\frac{1}{i\omega}\nabla \cdot \frac{K}{\hat{\sigma}\mu}\right) = \nabla\widetilde{\varphi}$. Inserting the latter into (2.9) yields

$$-i\omega\nabla\widetilde{\varphi} + \nabla\left(\nabla \cdot \frac{\nabla\widetilde{\varphi}}{\hat{\sigma}\mu}\right) = 0$$

Hence the following holds:

$$\nabla\left(-i\omega\widetilde{\varphi} + \left(\nabla \cdot \frac{\nabla\widetilde{\varphi}}{\hat{\sigma}\mu}\right)\right) = 0$$

Multiplying this equation by $\hat{\sigma}\mu$ and using the fact that $\hat{\sigma}, \mu$ are constant in V_j, the following equation is obtained:

$$\nabla\left(\nabla \cdot \nabla\widetilde{\varphi} - i\omega\hat{\sigma}\mu\,\widetilde{\varphi}\right) = 0$$

This is equivalent to existence of constants C_j such that

$$\nabla \cdot \nabla \widetilde{\varphi} - i\omega\hat{\sigma}\mu\,\widetilde{\varphi} = C_j$$

in each V_j, so that

$$\nabla \cdot \nabla \widetilde{\varphi} - i\omega\hat{\sigma}\mu\,\left(\widetilde{\varphi} + \frac{C_j}{i\omega\hat{\sigma}\mu}\right) = 0$$

Define $\varphi = \widetilde{\varphi} + \frac{C_j}{i\omega\hat{\sigma}\mu}$. In each V_j, the values of $C_j, \hat{\sigma}, \mu$ are constant, therefore, in each V_j, $\nabla\widetilde{\varphi} = \nabla\varphi$, so $K = \nabla\varphi$, and (2.10) is satisfied in strong sense in each V_j. 1⇒2⇒3 has been proven.

To prove 3⇒1, assume that $F = F_E + \nabla\varphi$ and φ satisfies (2.10). To prove that F is an electric Schelkunoff potential, it is enough to prove that $K = \nabla\varphi$ satisfies (2.9), which is readily obtained if gradient of the equation (2.10) is taken.

Next we consider a continuous electric Schelkunoff potential and investigate conditions for φ imposed by continuity. Consider two regions V_1, V_2 and a boundary between those regions $\partial V_1 \cap \partial V_2$. Let F_1, F_2 be potentials in V_1, V_2, respectively. Let n be a vector normal to the boundary, pointing towards V_2. It is useful to split continuity of F into continuity of the tangential components and continuity of normal components,

$$n \times F_1 = n \times F_2 \tag{2.11}$$

$$n \cdot F_1 = n \cdot F_2 \tag{2.12}$$

If F is an electric Schelkunoff potential, then according to Lemma 3, $F = F_E + \nabla\varphi = -\frac{1}{i\omega}E + \nabla\varphi$. As the tangential component of electric field is continuous, in order for (2.11) to be satisfied, it is needed that $\nabla\varphi$ have continuous tangential components,

$$n \times \nabla\varphi_1 = n \times \nabla\varphi_2$$

The last equation states that the derivative in the tangential direction t is the same on both sides of the boundary,

$$\frac{\partial\varphi_1}{\partial t} = \frac{\partial\varphi_2}{\partial t} \tag{2.13}$$

where $\frac{\partial\varphi_j}{\partial t}$, $j = 1, 2$ is the tangential derivative. The equation (2.13) has to be satisfied for all tangential directions. Therefore, continuity of the tangential component of F is

equivalent to continuity of the tangential derivative of φ as expressed by the equation (2.13).

Let us now focus on a condition more difficult to satisfy – continuity of the normal component, the equation (2.12). It is useful to notice that if $\hat{\sigma}_1 \neq \hat{\sigma}_2$, then the normal component of the electric field is not continuous. Representing the potential in a form $F = -\frac{1}{i\omega}E + \nabla\varphi$ one may see that (2.12) is equivalent to

$$n \cdot \left(-\frac{1}{i\omega}E_1 + \nabla\varphi_1\right) = n \cdot \left(-\frac{1}{i\omega}E_2 + \nabla\varphi_2\right)$$

which is equivalent to

$$n \cdot (\nabla\varphi_1 - \nabla\varphi_2) = n \cdot \left(\frac{1}{i\omega}E_1 - \frac{1}{i\omega}E_2\right) \tag{2.14}$$

or using a different notation:

$$\frac{\partial\varphi_1}{\partial n} - \frac{\partial\varphi_2}{\partial n} = n \cdot \left(\frac{1}{i\omega}E_1 - \frac{1}{i\omega}E_2\right) \tag{2.15}$$

To sum up, the continuity of the normal component of F (2.12) is equivalent to a jump in the normal derivative of φ, dictated by jump in normal component of the electric field E. The jump is given in (2.15).

Armed with those results, we are ready to prove the main theorem of this paper:

Theorem 4 *Suppose that the following assumptions are satisfied:*

- *Maxwell's Equations (2.1) are satisfied for E, H in weak sense in an open, bounded region $\Omega \subset \mathbb{R}^3$ with Lipschitz boundary, subdivided into disjoint open sets V_j, $j \in I$.*

- *$\hat{\sigma}, \mu$ are constant in each V_j, $\hat{\sigma}$ is real and*

$$0 < \sigma_m \leq \hat{\sigma}_j \leq \sigma_M < \infty,\ 0 < \mu_m \leq \mu_j \leq \mu_M < \infty \ \text{ for all } j \in I \tag{2.16}$$

- *$\nabla \cdot J^{imp} = 0$ inside each V_j*

Then there exists a continuous electric Schelkunoff potential F satysfying

$$E = -i\omega F + \nabla\left(\frac{\nabla \cdot F}{\hat{\sigma}\mu}\right) \tag{2.17}$$

$$\frac{\nabla \cdot F}{\hat{\sigma}\mu} \text{ is continuous across boundaries between sets } V_j \tag{2.18}$$

so (2.17) is satisfied in the sense of distributions. Moreover, if

$$n \cdot F|_{\partial\Omega} = -\frac{1}{i\omega}\,(n \cdot E)|_{\partial\Omega} \tag{2.19}$$

or

$$\nabla \cdot F|_{\partial\Omega} = 0 \tag{2.20}$$

then F is unique.

For complex valued $\hat{\sigma}$, assumption (2.16) may be replaced with:

$\mu, \hat{\sigma}$ are piecewise constant,

$$0 < \mu_m \leq \mu_j \leq \mu_M < \infty,\ |\hat{\sigma}| \leq \sigma_M < \infty$$

and one of the following is satisfied:

- *there exists a positive constant γ such that $\sigma - \omega\epsilon \geq \gamma > 0$*
- *in the case of boundary condition (2.20),*

$$0 < \epsilon_m \leq \epsilon \leq \epsilon_M \leq \infty, \quad c > \omega^2\epsilon_M$$

where c is a constant in Poincare inequality (2.52) for functions in $\mathcal{H}_0^1$.

Remark 5 *Assumption $\nabla \cdot J^{imp} = 0$ is equivalent to $\nabla \cdot E = 0$ in each V_j, which in turn is equivalent to absence of current sources inside of V_j. Notice that it is not assumed that $\nabla \cdot E = 0$ in the whole Ω, so there may be an (oscillating) charge deposited on the boundaries between regions V_j.*

Proof : Consider an electric Schelkunoff potential F, which is the vector field $F = F_E + \nabla\varphi$, with φ satisfying (2.10) in strong sense in each V_j. Because $F = F_E + \nabla\varphi = -\frac{1}{i\omega}E + \nabla\varphi$ and $\nabla \cdot E = 0$ in V_j, the following relationship holds

$$\nabla \cdot F = \nabla \cdot \left(-\frac{1}{i\omega}E + \nabla\varphi\right) = \nabla \cdot \nabla\varphi \tag{2.21}$$

so using the equation (2.10),

$$\frac{\nabla \cdot F}{\hat{\sigma}\mu} = \frac{\nabla \cdot \nabla\varphi}{\hat{\sigma}\mu} = i\omega\varphi \tag{2.22}$$

As a result, φ being continuous is equivalent to $\frac{\nabla \cdot F}{\hat{\sigma}\mu}$ being continuous. To ensure (2.18), φ should be continuous, so we request $\varphi \in \mathcal{H}^1(\Omega)$.

Next, a weak equation for φ, that results in a continuous electric Schelkunoff potential F, will be formulated. It was shown that a desired φ should satisfy (2.10), (2.13), (2.15).

It is useful to notice that (2.13) is satisfied if only $\varphi_1 = \varphi_2$ on $\partial V_1 \cap \partial V_2$, so as φ is a continuous function in Ω, then (2.13) is satisfied automatically. Multiplying (2.10) by a conjugated test function and integrating by parts yields

$$\sum_{j\in I}\int_{V_j} \Delta\varphi\overline{\xi} - i\omega\sum_{j\in I}\int_{V_j}\hat{\sigma}\mu\,\varphi\,\overline{\xi} = 0$$

$$-\sum_{j\in I}\int_{V_j}\nabla\varphi\cdot\nabla\overline{\xi} + \sum_{j\in I}\int_{\partial V_j}\left(\frac{\partial}{\partial n_j}\varphi_j\right)\overline{\xi} - i\omega\sum_{j\in I}\int_{V_j}\hat{\sigma}\mu\,\varphi\,\overline{\xi} = 0$$

where n_j denotes a unit outward normal on ∂V_j. Let us now analyze the middle term:

$$\sum_{j\in I}\int_{\partial V_j}\left(\frac{\partial}{\partial n_j}\varphi_j\right)\overline{\xi} = \sum_{j_1\in I,\, j_2\in I, j_1\neq j_2}\int_{\partial V_{j_1}\cap\partial V_{j_2}}\left(\frac{\partial}{\partial n_{j_1}}\varphi_{j_1} + \frac{\partial}{\partial n_{j_2}}\varphi_{j_2}\right)\overline{\xi} + \int_{\partial\Omega}\left(\frac{\partial}{\partial n}\varphi\right)\overline{\xi}$$

and using (2.15), the latter equals to

$$\sum_{j_1\in I,\, j_2\in I, j_1\neq j_2}\int_{\partial V_{j_1}\cap\partial V_{j_2}}\frac{1}{i\omega}\left(n_{j_1}\cdot E_{j_1} + n_{j_2}\cdot E_{j_2}\right)\overline{\xi} + \int_{\partial\Omega}\frac{\partial}{\partial n}\varphi\overline{\xi}$$

Let a function

$$f = \frac{1}{i\omega}(n_{j_1}\cdot E_{j_1} + n_{j_2}\cdot E_{j_2})$$

be defined on a set $D = \bigcup_{j_1\in I,\, j_2\in I, j_1\neq j_2}\partial V_{j_1}\cap\partial V_{j_2}$ consisting of boundaries between V_{j_1} and V_{j_2}. Then the equation for φ is obtained:

$$\int_\Omega \nabla\varphi\cdot\nabla\overline{\xi} + i\omega\int_\Omega\hat{\sigma}\mu\,\varphi\,\overline{\xi} = \int_D f\overline{\xi} + \int_{\partial\Omega}\left(\frac{\partial}{\partial n}\varphi\right)\overline{\xi} \tag{2.23}$$

Before we proceed, we have to consider boundary conditions for F. Rewriting (2.17) using (2.22) the following is obtained

$$E = -i\omega F + \nabla\left(\frac{\nabla \cdot F}{\hat{\sigma}\mu}\right) = -i\omega F + \nabla(i\omega\varphi) \tag{2.24}$$

The last equation together with the boundary condition (2.19) implies

$$n \cdot E = n \cdot \left(-i\omega\left(-\frac{1}{i\omega}\right)\right) E + n \cdot \nabla(i\omega\varphi) \quad \text{on } \partial\Omega$$

which is equivalent to

$$\frac{\partial}{\partial n}\varphi\Big|_{\partial\Omega} = 0 \tag{2.25}$$

On the other hand, because of (2.22), boundary condition (2.20) implies that

$$\varphi|_{\partial\Omega} = 0 \tag{2.26}$$

Returning back to (2.23), the term $\int_{\partial\Omega}(\frac{\partial}{\partial n}\varphi)\overline{\xi}$ vanishes if $\frac{\partial}{\partial n}\varphi = 0$ on $\partial\Omega$, as well as when $\varphi, \xi \in \mathcal{H}_0^1(\Omega)$, so $\xi = 0$ on $\partial\Omega$.

In the case of both Dirichlet ($\varphi|_{\partial\Omega} = 0$) and Neumann ($\frac{\partial}{\partial n}\varphi|_{\partial\Omega} = 0$) boundary conditions for φ, the weak equation for φ is:

$$\int_\Omega \nabla\varphi \cdot \nabla\overline{\xi} + i\omega \int_\Omega \hat{\sigma}\mu\,\varphi\,\overline{\xi} = \int_D f\overline{\xi} \tag{2.27}$$

To sum up, (2.17), (2.18), continuity of F, and boundary condition (2.19) imply that φ has to satisfy equation (2.27), for $\varphi, \xi \in \mathcal{H}^1(\Omega)$. Similarly, (2.17), (2.18), continuity of F, and boundary condition (2.20) imply that φ has to satisfy equation (2.27), for $\varphi, \xi \in \mathcal{H}_0^1(\Omega)$.

Existence and uniqueness of continuous F is equivalent to existence and uniqueness of the solution φ of equation (2.27) in $\mathcal{H}^1(\Omega)$ and $\mathcal{H}_0^1(\Omega)$ for (2.19) and (2.20), respectively.

Let us analyze the left-hand side of equation (2.27). It is a bilinear form $B(\varphi, \xi)$, which is bounded:

$$|B(\varphi,\xi)|^2 = \left|\int_\Omega \nabla\varphi\cdot\nabla\overline{\xi} + i\omega\int_\Omega \hat{\sigma}\mu\,\varphi\,\overline{\xi}\right|^2$$
$$\leq 2\left(\int_\Omega |\nabla\varphi|\,|\nabla\xi|\right)^2 + 2\omega^2\sigma_M^2\mu_M^2\left(\int_\Omega |\varphi|\,|\xi|\right)^2$$
$$\leq 2||\nabla\varphi||_0^2\,||\nabla\xi||_0^2 + 2\omega^2\sigma_M^2\mu_M^2||\varphi||_0^2\,||\xi||_0^2$$
$$\leq (2+2\omega^2\sigma_M^2\mu_M^2)\left(||\nabla\varphi||_0^2\,||\nabla\xi||_0^2 + ||\varphi||_0^2\,||\xi||_0^2\right)$$
$$\leq (2+2\omega^2\sigma_M^2\mu_M^2)\left(||\nabla\varphi||_0^2 + ||\varphi||_0^2)(||\nabla\xi||_0^2 + \,||\xi||_0^2\right)$$
$$= (2+2\omega^2\sigma_M^2\mu_M^2)||\varphi||_1^2||\xi||_1^2$$

For definition of $||.||_0, ||.||_1$ see the Appendix(Section 2.9). Moreover, this bilinear form is coercive:

$$|B(\xi,\xi)| = \left|\int_\Omega \nabla\xi\cdot\nabla\overline{\xi} + i\omega\int_\Omega \hat{\sigma}\mu\,\xi\,\overline{\xi}\right| = \left|\int_\Omega |\nabla\xi|^2 + i\omega\int_\Omega \hat{\sigma}\mu\,|\xi|^2\right|$$

The first term is purely real and the second term is purely imaginary ($\omega,\hat{\sigma},\mu\in\mathbb{R}$), so

$$|B(\xi,\xi)| \geq \frac{1}{\sqrt{2}}\left(\int_\Omega |\nabla\xi|^2 + \omega\int_\Omega \hat{\sigma}\mu\,|\xi|^2\right) \geq \frac{1}{\sqrt{2}}\left(\int_\Omega |\nabla\xi|^2 + \omega\sigma_m\mu_m\int_\Omega |\xi|^2\right)$$
$$\geq \frac{1}{\sqrt{2}}\min(1,\omega\sigma_m\mu_m)\left(\int_\Omega |\nabla\xi|^2 + \int_\Omega |\xi|^2\right) = \frac{1}{\sqrt{2}}\min(1,\omega\sigma_m\mu_m)||\xi||_1^2$$

To summarize, the left-hand side of the equation (2.27) is a bounded coercive bilinear form on $\mathcal{H}^1(\Omega)\times\mathcal{H}^1(\Omega)$ (and on $\mathcal{H}_0^1(\Omega)\times\mathcal{H}_0^1(\Omega)$), the right-hand side is a bounded linear functional on $\mathcal{H}^1(\Omega)$ (and also on $\mathcal{H}_0^1(\Omega)$), so from the Lax-Milgram theorem, there exists a unique $\varphi\in\mathcal{H}^1(\Omega)$ (or $\varphi\in\mathcal{H}_0^1(\Omega)$) satisfying equation (2.27). Hence there exists a unique F for boundary condition (2.19) as well as for boundary condition (2.20).

Let us now remove requirement that $\hat{\sigma}$ is real in assumption (2.16). Saying that $\hat{\sigma}=\sigma+i\omega\epsilon$ is real is equivalent to neglecting the term $i\omega\epsilon$. This is a simplifying assumption often used for low-frequency EM fields in a conducting medium, in particular for magnetotellurics. However, this assumption is not necessarily required for B to be coercive.

Indeed, consider $B(\xi,\xi)$, the term that has to be bounded from below, to prove coercivity of B. Substitution of $\sigma+i\omega\epsilon$ in place of $\hat{\sigma}$ gives:

$$|B(\xi,\xi)| \geq \left|\int_\Omega |\nabla\xi|^2 + i\omega\int_\Omega(\sigma+i\omega\epsilon)|\xi|^2\right| = \left|\int_\Omega |\nabla\xi|^2 - \int_\Omega \omega^2\epsilon|\xi|^2 + i\int_\Omega \omega\sigma|\xi|^2\right|$$

$$\geq \frac{1}{\sqrt{2}} \left(\left| \int_\Omega |\nabla \xi|^2 - \int_\Omega \omega^2 \epsilon |\xi|^2 \right| + \left| \int_\Omega \omega \sigma |\xi|^2 \right| \right) \tag{2.28}$$

- If there exists a positive constant γ such that $\sigma - \omega\epsilon \geq \gamma > 0$ in Ω, then B is coercive and the electric Schelkunoff potential F exists.

  Continuing with the calculation using this assumption, we derive the following estimate

$$|B(\xi,\xi)| \geq \frac{1}{\sqrt{2}} \left(\int_\Omega |\nabla \xi|^2 - \int_\Omega \omega^2 \epsilon |\xi|^2 + \int_\Omega \omega \sigma |\xi|^2 \right)$$
$$= \frac{1}{\sqrt{2}} \left(\int_\Omega |\nabla \xi|^2 + \int_\Omega \omega(\sigma - \omega\epsilon) |\xi|^2 \right) \geq \frac{1}{\sqrt{2}} \left(\int_\Omega |\nabla \xi|^2 + \omega\gamma \int_\Omega |\xi|^2 \right)$$
$$\geq \min \left(\frac{1}{\sqrt{2}}, \frac{\omega\gamma}{\sqrt{2}} \right) ||\xi||_1^2$$

  So the bilinear form B is coercive, and the hypothesis of Theorem 4 is true.

- Consider boundary condition (2.20). In this case, $\xi \in \mathcal{H}_0^1$. We will show that, if $c > \omega^2 \epsilon_M$, where c is a constant in Poincare inequality (2.52) for functions in $\mathcal{H}_0^1$, then B is coercive and the electric Schelkunoff potential exists even if $\sigma = 0$.

  Continuing with the calculation (2.28), let us drop the term containing σ

$$|B(\xi,\xi)| \geq \frac{1}{\sqrt{2}} \left(\left| \int_\Omega |\nabla \xi|^2 - \int_\Omega \omega^2 \epsilon |\xi|^2 \right| \right) \geq \frac{1}{\sqrt{2}} \left(\int_\Omega |\nabla \xi|^2 - \int_\Omega \omega^2 \epsilon_M |\xi|^2 \right)$$

  If $c > \omega^2 \epsilon_M$, then there is $\alpha \in (0,1)$ such that $c > \alpha c > \omega^2 \epsilon_M$, using Poincare inequality (2.52), one can rewrite the latter as

$$\frac{1}{\sqrt{2}} \left((1-\alpha) \int_\Omega |\nabla \xi|^2 + \alpha \int_\Omega |\nabla \xi|^2 - \int_\Omega \omega^2 \epsilon_M |\xi|^2 \right)$$
$$\geq \frac{1}{\sqrt{2}} \left((1-\alpha) \int_\Omega |\nabla \xi|^2 + \alpha c \int_\Omega |\xi|^2 - \int_\Omega \omega^2 \epsilon_M |\xi|^2 \right)$$
$$\geq \frac{1}{\sqrt{2}} \left((1-\alpha) \int_\Omega |\nabla \xi|^2 + (\alpha c - \omega^2 \epsilon_M) \int_\Omega |\xi|^2 \right)$$
$$\geq \min \left(\frac{1-\alpha}{\sqrt{2}}, \frac{\alpha c - \omega^2 \epsilon_M}{\sqrt{2}} \right) ||\xi||_1^2$$

  So even in the non-lossy medium case ($\sigma = 0$), if the frequency is small enough ($c > \omega^2 \epsilon_M$), B is coercive and unique electric Schelkunoff potential F exists for the case of boundary condition (2.20).

2.5 Difficulty in obtaining a weak form of the governing equation for the electric Schelkunoff potential

To be able to use the Finite Element Method for calculation of the EM field, a weak form of the governing equation satisfied by the electric Schelkunoff potential is needed.

In order to obtain a weak equation, one starts from Maxwell's equations (2.1). Dividing the first equation by $-i\omega\mu$, taking the curl and substituting into the second equation, one obtains

$$\nabla \times \frac{1}{-i\omega\mu}\nabla \times E - \hat{\sigma}E = J^{imp} \tag{2.29}$$

Next $-i\omega F + \nabla\left(\frac{\nabla\cdot F}{\hat{\sigma}\mu}\right)$ is substituted for E and the equation is multiplied by a test vector field A. The result is

$$\int_\Omega \left(\nabla \times \frac{1}{\mu}\nabla \times F\right)\cdot A - \int_\Omega \nabla\left(\frac{\nabla\cdot F}{\hat{\sigma}\mu}\right)\cdot(\hat{\sigma}A) + \int_\Omega i\omega\hat{\sigma}F\cdot A = \int_\Omega J^{imp}\cdot A$$

In order to integrate by parts the first term in the above equation, one uses continuity of the tangential component of $\frac{1}{\mu}\nabla\times F$, which is equivalent to continuity of tangential component of H and one needs tangential components of A to be continuous across $\partial V_{j_1}\cap\partial V_{j_2}$.

On the other hand, in order to integrate by parts the second term, one would use continuity of $\frac{\nabla\cdot F}{\hat{\sigma}\mu}$, and one needs normal components of $\hat{\sigma}A$ to be continuous. So if $\hat{\sigma}$ is discontinuous, so is the normal component of A. This is the essence of the problem in obtaining a proper weak form of the equation for F. A family of vector shape functions with continuous tangential components and normal components experiencing specific jumps is difficult to build. One may consider a mixed formulation involving scalar and vector potential (see Section 2.6), but that increases the number of degrees of freedom.

It turns out that, assuming that μ is constant, it is possible to obtain an equation involving only the vector potential, but for a Schekunoff potential representation of the magnetic field H. This idea is presented in Section 2.7.

2.6 Formulation with both scalar and vector potentials

If the original field is not divergence free, a potential equation involving both scalar and vector components must be considered. Although the number of degrees of freedom per point in space increases from 3 to 4, the obtained equation is valid for non-constant electromagnetic properties and the non-divergence-free field. Also the bilinear form of the equation remains coercive as $\omega \to 0$.

In [24], in the case of Lorentz gauge #2, the authors proved the uniqueness of Schelkunoff potential. The bilinear form

$$\begin{aligned}\mathcal{G}\left((F,\varphi),\,(A,\phi)\right) \quad &= \textstyle\int_\Omega [\nabla \times F \cdot \frac{1}{\mu}\nabla \times A + \nabla \cdot F \frac{1}{\mu}\nabla \cdot A - \omega^2 \epsilon F \cdot A \\ &\quad -i\omega\epsilon\nabla\cdot\phi F + \epsilon\nabla\varphi\cdot\nabla\phi - \omega^2\epsilon^2\mu\varphi\phi - i\omega\epsilon\nabla\cdot\varphi A]\end{aligned}$$

of the weak equation they propose (see (58) in [24]), for purely imaginary frequency $\omega = i\tilde{\omega},\ \tilde{\omega} > 0$, may be rewritten as

$$\mathcal{G} = \int_\Omega \frac{1}{\mu}(\nabla\cdot F + \mu\tilde{\omega}\epsilon\varphi)(\nabla\cdot A + \mu\tilde{\omega}\epsilon\phi) + \int_\Omega \epsilon(\nabla\varphi + \tilde{\omega}F)\cdot(\nabla\phi + \tilde{\omega}A) + \int_\Omega \frac{1}{\mu}(\nabla\times F)\cdot(\nabla\times A)$$

Using this form, we can prove boundedness and coercivity of $\mathcal{G}$ for $F,\ A \in \mathcal{H}_0(\nabla\times,\Omega) \cap \mathcal{H}(\nabla\cdot,\Omega),\ \varphi,\phi \in \mathcal{H}_0^1(\Omega)$. So from the Lax-Milgram theorem, there exists a unique solution to the equation for the Schelkunoff potential that is considered in [24]. The solution is a continuous electric Schelkunoff potential. This formulation may be easily adapted to a lossy medium, which is expressed in Theorem 6.

Theorem 6 *Let the assumptions of Theorem 4 be satisfied. The unique electric Schelkunoff potential F satisfying boundary condition (2.20), together with $\varphi = -\frac{\nabla\cdot F}{\hat{\sigma}\mu}$, satisfy the following equation*

$$\begin{aligned}&\textstyle\int_\Omega \frac{1}{\mu}(\nabla\times F)\cdot\overline{(\nabla\times A)} + \int_\Omega \frac{1}{\mu}(\nabla\cdot F + \mu\hat{\sigma}\varphi)\overline{(\nabla\cdot A + \mu\hat{\sigma}\phi)} \\ &\quad + \textstyle\int_\Omega \hat{\sigma}(i\omega F + \nabla\varphi)\cdot\overline{(i\omega A + \nabla\phi)} = \int_\Omega J^{imp}\cdot\overline{(i\omega A + \nabla\phi)}\end{aligned} \tag{2.30}$$

$$\forall A \in \mathcal{H}_0(\nabla\times)\cap\mathcal{H}(\nabla\cdot) \text{ and } \phi \in \mathcal{H}_0^1$$

$$F \in \mathcal{H}(\nabla\times)\cap\mathcal{H}(\nabla\cdot),\ n\times F = n\times(-i\omega E) \text{ on } \partial\Omega,\ \varphi \in \mathcal{H}_0^1$$

The bilinear form associated with the equation (2.30) is bounded and coercive with respect to the norm

$$||(A,\phi)||_B = \sqrt{||A||_0^2 + ||\nabla\times A||_0^2 + ||\nabla\cdot A||_0^2 + ||\nabla\phi||_0^2 + ||\phi||_0^2} \tag{2.31}$$

Hence if $J^{imp} \in L^2(\Omega)$, then the solution to this equation exists and is unique.

Remark 7

- *If the domain is a convex polygon, or if the domain has C^2 boundary, then one may use nodal shape functions to approximate both F and φ.*
- *In order to obtain the electric field, one has to calculate*

$$E = -i\omega F - \nabla\varphi$$

- *If one drops all the terms multiplied by ω, the resulting bilinear form remains coercive. To prove this, one has to use the Poincare inequality for $H_0(\nabla\times) \cap \mathcal{H}(\nabla\cdot)$ (see [29], Lemma 3.6). The proof of this result is easier than the proof of coercivity of the original bilinear form, so it is omitted.*

Proof : The fact that the pair (F, φ) satisfies the equation (2.30) is straightforward. As $\varphi = -\frac{\nabla\cdot F}{\hat{\sigma}\mu}$ and the boundary condition (2.20) is satisfied, $\varphi \in \mathcal{H}_0^1$. F is the electric Schelkunoff potential satisfying the hypothesis of Theorem 4 so F is continuous, and hence is a member of $\mathcal{H}(\nabla\times) \cap \mathcal{H}(\nabla\cdot)$. Moreover, on $\partial\Omega$

$$n \times F = n \times (-i\omega E - \nabla\varphi) = n \times (-i\omega E)$$

as $n \times \nabla\varphi = 0$ on $\partial\Omega$, which is a consequence of $\varphi \in \mathcal{H}_0^1(\Omega)$.

Moreover, if $\varphi = -\frac{\nabla\cdot F}{\hat{\sigma}\mu}$, then

$$\nabla \cdot F + \mu\hat{\sigma}\varphi = 0$$

so the middle term in equation (2.30) vanishes. If

$$E = -i\omega F - \nabla\varphi$$

is used, equation (2.30) simplifies to

$$\int_\Omega \frac{1}{\mu}(\nabla \times E) \cdot (\nabla \times \overline{A}) + i\omega \int_\Omega \hat{\sigma} E \cdot \left(\overline{A} - \frac{\nabla\varphi}{i\omega}\right) = -i\omega \int_\Omega J^{imp} \cdot \left(\overline{A} - \frac{\nabla\varphi}{i\omega}\right)$$

Since $A \in \mathcal{H}_0(\nabla\times) \cap \mathcal{H}(\nabla\cdot)$ and $\varphi \in \mathcal{H}_0^1$, then $\tilde{A} = \overline{A} - \frac{\nabla\varphi}{i\omega} \in \mathcal{H}_0(\nabla\times)$ and $\nabla \times \tilde{A} = \nabla \times A$, so it remains to show that for any $\tilde{A} \in \mathcal{H}_0(\nabla\times)$, the following equation is satisfied:

$$\int_\Omega \frac{1}{\mu}(\nabla \times E) \cdot (\nabla \times \tilde{A}) + i\omega \int_\Omega \hat{\sigma} E \cdot \tilde{A} = -i\omega \int_\Omega J^{imp} \cdot \tilde{A}$$

This equation is a standard equation satisfied by the electric field satisfying Maxwell's equations (2.1). It is satisfied for all $\tilde{A} \in \mathcal{H}_0(\nabla\times)$. We proved that (F, φ) satisfy equation (2.30).

Let us now focus on the proof of boundedness and coercivity of the bilinear form $B((F,\varphi),(A,\phi))$ defined as the left-hand side of the equation (2.30).

Boundedness of B is straightforward, as from Cauchy-Schwartz inequality, it follows that:

$$|B((F,\varphi),(A,\phi))| =$$

$$\int_\Omega \frac{1}{\mu}(\nabla\times F)\cdot\overline{(\nabla\times A)} + \int_\Omega \frac{1}{\mu}(\nabla\cdot F + \mu\hat{\sigma}\varphi)\overline{(\nabla\cdot A + \mu\hat{\sigma}\phi)} + \int_\Omega \hat{\sigma}(i\omega F + \nabla\varphi)\cdot\overline{(i\omega A + \nabla\phi)}$$

$$\leq \frac{1}{\mu_m}\int_\Omega |\nabla\times F|\,|\nabla\times A| + \int_\Omega \frac{1}{\mu_m}|\nabla\cdot F + \mu\hat{\sigma}\varphi|\,|\nabla\cdot A + \mu\hat{\sigma}\phi| + \int_\Omega \sigma_M |i\omega F + \nabla\varphi|\,|i\omega A + \nabla\phi|$$

$$\leq \frac{1}{\mu_m}||\nabla\times F||_0||\nabla\times A||_0 + \frac{1}{\mu_m}||\nabla\cdot F + \mu\hat{\sigma}\varphi||_0||\nabla\cdot A + \mu\hat{\sigma}\phi||_0 + \sigma_M||i\omega F + \nabla\varphi||_0||i\omega A + \nabla\phi||_0$$

$$\leq \frac{1}{\mu_m}||\nabla\times F||_0||\nabla\times A||_0 + \frac{1}{\mu_m}(||\nabla\cdot F||_0 + \mu_M\sigma_M||\varphi||_0)\,(||\nabla\cdot A||_0 + \mu_M\sigma_M||\phi||_0)$$

$$+\sigma_M(\omega||F||_0 + ||\nabla\varphi||_0)\,(\omega||A||_0 + ||\nabla\phi||_0)$$

$$\leq \max\left(\frac{1}{\mu_m}, \frac{\mu_M}{\mu_m}\sigma_M, \frac{\mu_M^2}{\mu_m}\sigma_M^2, \sigma_M, \sigma_M\omega, \sigma_M\omega^2\right)||(F,\varphi)||_B||(A,\phi)||_B$$

To prove coercivity, we have to prove that there exists a constant $\beta > 0$ such that for any $(A,\phi) \in (\mathcal{H}_0(\nabla\times,\Omega)\cap\mathcal{H}(\nabla\cdot,\Omega))\times\mathcal{H}_0^1(\Omega)$

$$B((A,\phi),(A,\phi)) \geq \beta||(A,\phi)||_B^2$$

It is enough to prove that it is not possible to have a sequence of (A_n,ϕ_n) such that

$$1 = ||(A_n,\phi_n)||_B^2 = ||A_n||_0^2 + ||\nabla\times A_n||_0^2 + ||\nabla\cdot A_n||_0^2 + ||\nabla\phi_n||_0^2 + ||\phi_n||_0^2 \qquad (2.32)$$

and

$$B((A_n,\phi_n),(A_n,\phi_n)) \xrightarrow[n\to\infty]{} 0 \qquad (2.33)$$

For proof by contradiction, assume that there is (A_n,ϕ_n) satisfying (2.32) and (2.33). Notice that

$$B((A_n,\phi_n),(A_n,\phi_n)) = \int_\Omega \frac{1}{\mu}|\nabla\times A_n|^2 + \int_\Omega \frac{1}{\mu}|\nabla\cdot A_n + \mu\hat{\sigma}\phi_n|^2 + \int_\Omega \hat{\sigma}|i\omega A_n + \nabla\phi_n|^2$$

All the terms in the sum are non-negative, so if the sum converges to 0, all of the terms converge to 0:

$$\int_\Omega \frac{1}{\mu}|\nabla \times A_n|^2 \xrightarrow[n\to\infty]{} 0 \tag{2.34}$$

$$\int_\Omega \frac{1}{\mu}|\nabla \cdot A_n + \mu\hat{\sigma}\phi_n|^2 \xrightarrow[n\to\infty]{} 0 \tag{2.35}$$

$$\int_\Omega \hat{\sigma}|i\omega A_n + \nabla\phi_n|^2 \xrightarrow[n\to\infty]{} 0 \tag{2.36}$$

In the following, we will prove that all the terms in $||(A_n, \phi_n)||_B^2$, the right-hand side of (2.32), converge to 0. Let us start with $||\nabla \times A_n||_0^2$:

$$||\nabla \times A_n||_0^2 = \int_\Omega |\nabla \times A_n|^2 \leq \mu_m \int_\Omega \frac{1}{\mu}|\nabla \times A_n|^2 \xrightarrow[n\to\infty]{} 0 \tag{2.37}$$

The last term converges to 0 from (2.34).

As $A_n \in \mathcal{H}_0(\nabla\times)$, there is Hodge decomposition of A_n (see [14], Appendix A). There exist unique $\psi_n \in \mathcal{H}_0^1$ and $A_n^\perp \in R(\nabla)^\perp$ such that

$$A_n = \nabla\psi_n + A_n^\perp$$

where $R(\nabla)^\perp$ is the space orthogonal to the range of the gradient, namely:

$$R(\nabla)^\perp = \left\{ A \in \mathcal{H}_0(\nabla\times) : \forall_{\xi\in\mathcal{H}_0^1} \int_\Omega \nabla\xi \cdot A = 0 \right\}$$

One can also say that this is a decomposition of A_n into a curl free part $\nabla\psi_n$ and a divergence free part $A_n^\perp$. Using this decomposition, we conclude that

$$||\nabla \times A_n||_0^2 = ||\nabla \times A_n^\perp||_0^2 \geq c_1||A_n^\perp||_0^2$$

Here c_1 depends on Ω only. The last inequality is a consequence of Poincare inequality for $\mathcal{H}(\nabla\times)$ (2.53) and the fact that $A_n^\perp \in R(\nabla)^\perp$. As a result, (2.37) implies that

$$||A_n^\perp||_0^2 \xrightarrow[n\to\infty]{} 0 \tag{2.38}$$

Moreover as n increases, the norm of the difference between $\nabla\phi_n$ and $-i\omega\nabla\psi_n$ goes to zero. Indeed,

$$\begin{aligned}
||\nabla\phi_n + i\omega\nabla\psi_n||_0^2 &= \textstyle\int_\Omega |\nabla\phi_n + i\omega\nabla\psi_n|^2 \\
&\leq \textstyle\int_\Omega 2\left[|\nabla\phi_n + i\omega\nabla\psi_n + i\omega A_n^\perp|^2 + |-i\omega A_n^\perp|^2\right] \\
&\leq \tfrac{2}{\sigma_m}\textstyle\int_\Omega \hat{\sigma}|\nabla\phi_n + i\omega(\nabla\psi_n + A_n^\perp)|^2 + 2\omega^2\int_\Omega |A_n^\perp|^2 \\
&= \tfrac{2}{\sigma_m}\textstyle\int_\Omega \hat{\sigma}|\nabla\phi_n + i\omega A_n|^2 + 2\omega^2||A_n^\perp||_0^2
\end{aligned}$$

The last two terms converge to 0 as a result of (2.36) and (2.38), respectively, hence

$$||\nabla\phi_n + i\omega\nabla\psi_n||_0^2 \xrightarrow[n\to\infty]{} 0 \tag{2.39}$$

The Poincare inequality for $\mathcal{H}_0^1$ (2.52) states that there exists a constant $c_2 > 0$ dependent on Ω only, such that

$$||\nabla\phi_n + i\omega\nabla\psi_n||_0 = ||\nabla(\phi_n + i\omega\psi_n)||_0 \geq c_2||\phi_n + i\omega\psi_n||_0$$

As a result

$$||\phi_n + i\omega\psi_n||_0 \xrightarrow[n\to\infty]{} 0 \tag{2.40}$$

Let us now work on the norms $||\phi_n||_0$ and $||\nabla\phi_n||_0$. Using the fact that $A_n^\perp \in R(\nabla)^\perp$ and integrating by parts, one can obtain

$$\int_\Omega |\nabla\psi_n|^2 = \int_\Omega \nabla\psi_n \cdot \nabla\overline{\psi_n} = \int_\Omega (A_n - A_n^\perp) \cdot \nabla\overline{\psi_n} = \int_\Omega A_n \cdot \nabla\overline{\psi_n} = -\int_\Omega \nabla \cdot A_n \, \overline{\psi_n}$$

Using this fact, it is possible to bound from above the following expression:

$$\begin{aligned}
&\frac{1}{\sqrt{2}}||\nabla\psi_n||_0^2 + \frac{\omega\sigma_m\mu_m}{\sqrt{2}}||\psi_n||_0^2 \leq \frac{1}{\sqrt{2}}\left[\int_\Omega |\nabla\psi_n|_0^2 + \int_\Omega \omega\hat{\sigma}\mu|\psi_n|^2\right] \\
&\leq \left|\int_\Omega |\nabla\psi_n|_0^2 + i\int_\Omega \omega\hat{\sigma}\mu|\psi_n|^2\right| = \left|\int_\Omega -\nabla\cdot A_n\,\overline{\psi_n} + i\int_\Omega \omega\hat{\sigma}\mu\psi_n\overline{\psi_n}\right| \\
&= \left|\int_\Omega -\nabla\cdot A_n\,\overline{\psi_n} - \phi_n\hat{\sigma}\mu\overline{\psi_n} + \phi_n\hat{\sigma}\mu\overline{\psi_n} + i\omega\hat{\sigma}\mu\psi_n\overline{\psi_n}\right| \\
&= \left|\int_\Omega -(\nabla\cdot A_n + \phi_n\hat{\sigma}\mu)\overline{\psi_n} + \int_\Omega \hat{\sigma}\mu(\phi_n + i\omega\psi_n)\overline{\psi_n}\right| \\
&\leq \int_\Omega |\nabla\cdot A_n + \phi_n\hat{\sigma}\mu|\,|\psi_n| + \sigma_M\mu_M\int_\Omega |\phi_n + i\omega\psi_n|\,|\psi_n|
\end{aligned}$$

From Cauchy-Schwarz inequality, the latter is less than

$$\begin{aligned}
&\sqrt{\int_\Omega |\nabla\cdot A_n + \hat{\sigma}\mu\phi_n|^2}||\psi_n||_0 + \sigma_M\mu_M||\phi_n + i\omega\psi_n||_0||\psi_n||_0 \\
&\leq \sqrt{\mu_M\int_\Omega \frac{1}{\mu}|\nabla\cdot A_n + \hat{\sigma}\mu\phi_n|^2}||\psi_n||_0 + \sigma_M\mu_M||\phi_n + i\omega\psi_n||_0||\psi_n||_0
\end{aligned}$$

This bound converges to 0 because of (2.35), (2.39) and the fact that the sequence of norms $||\psi_n||_0$ is bounded, which in turn is a consequence of (2.32) and (2.40). As a result

$$\frac{1}{\sqrt{2}}||\nabla\psi_n||_0^2 + \frac{\omega\sigma_m\mu_m}{\sqrt{2}}||\psi_n||_0^2 \xrightarrow[n\to\infty]{} 0$$

Hence

$$||\nabla\psi_n||_0^2 = \int_\Omega |\nabla\psi_n|^2 \xrightarrow[n\to\infty]{} 0 \tag{2.41}$$

$$||\ \psi_n||_0^2 = \int_\Omega |\psi_n|^2 \xrightarrow[n\to\infty]{} 0 \tag{2.42}$$

From (2.41) and (2.39), it follows that

$$||\nabla\phi_n||_0^2 \xrightarrow[n\to\infty]{} 0 \tag{2.43}$$

Similarly, (2.42) and (2.40) imply

$$||\phi_n||_0^2 \xrightarrow[n\to\infty]{} 0 \tag{2.44}$$

Let us consider $||\nabla \cdot A_n||_0$:

$$\begin{aligned} ||\nabla \cdot A_n||_0^2 = \int_\Omega |\nabla \cdot A_n|^2 &\le \int_\Omega 2\left[|\nabla \cdot A_n + \mu\hat{\sigma}\phi_n|^2 + |-\mu\hat{\sigma}\phi_n|^2\right] \\ &\le 2\mu_M \int_\Omega \frac{1}{\mu}|\nabla \cdot A_n + \mu\hat{\sigma}\phi_n|^2 + 2\mu_M\sigma_M||\phi_n||_0^2 \end{aligned}$$

The latter converges to 0 because of (2.35) and (2.44). Hence

$$||\nabla \cdot A_n||_0^2 \xrightarrow[n\to\infty]{} 0 \tag{2.45}$$

The last term to consider is $||A_n||_0$. It may be bounded as follows:

$$||A_n||_0 = ||\nabla\psi_n + A_n^\perp||_0 \le ||\nabla\psi_n||_0 + ||A_n^\perp||_0$$

The last two terms converge to 0, which was noted in (2.41) and (2.38).

All the terms of $||(A_n, \phi_n)||_B^2$ converge to 0 as $n \to \infty$, so it cannot be that $||(A_n, \phi_n)||_B^2 = 1$. Contradiction. Hence the bilinear form B is coercive.

If $J^{imp} \in L^2(\Omega)$, then the right-hand side of (2.30) is a bounded linear functional on $(\mathcal{H}_0(\nabla\times)\cap\mathcal{H}(\nabla\cdot))\times\mathcal{H}_0^1$ with the norm $||.||_B$, thus from the Lax-Milgram theorem, there exists a unique solution to equation (2.30).

The theorem above proves the existence and uniqueness of a continuous Schelkunoff potential by considering an equation satisfied by vector and scalar potentials. The main theorem of this paper, Theorem 4 shows that the existence and uniqueness of a continuous Schelkunoff potential is equivalent to a simpler equation (2.27) involving only the scalar potential. Also it is clearly shown how

the jumps of normal component of the electric field are represented by the jumps of the normal derivative of the scalar potential, allowing the vector potential to be continuous across the boundaries between regions with different properties. This vector-scalar formulation forms the basis for a general finite element simulation scheme for non-divergence-free EM fields.

2.7 A magnetic Schelkunoff potential

If the original field is divergence-free, a simpler weak equation involving only the vector potential may be obtained. This approach is presented for a magnetic Schelkunoff potential.

Instead of representing E field by (2.4), a similar representation for magnetic field H may be used, namely with a magnetic Schelkunoff potential G. This representation is mentioned in [21],

$$H = G - \nabla \left(\frac{\nabla \cdot G}{i\omega\hat{\sigma}\mu} \right) \tag{2.46}$$

Existence of this representation might be proven in a similar way as in Theorem 4. Alternatively, existence of potential G follows from Theorem 8 with $\kappa = \frac{1}{i\omega\hat{\sigma}\mu}$.

Although in a geophysical setting, it cannot be assumed that conductivity is constant, most of the rocks have magnetic permeability $\mu = \mu_0$. In this case, the magnetic field H is divergence free:

$$\nabla \cdot H = 0$$

In this situation, magnetic Schelkunoff potential satisfying $\nabla \cdot G = 0$ on $\partial\Omega$ coincides with the magnetic field:

$$G = H$$

If $\left(H - \nabla \left(\frac{\nabla \cdot H}{i\omega\hat{\sigma}\mu} \right) \right)$ is substituted in place of H in a standard curl-curl equation for magnetic field H, one obtains the equation presented below:

$$\int_\Omega \frac{1}{\hat{\sigma}} (\nabla \times H) \cdot (\nabla \times A) + \int_\Omega \frac{1}{\hat{\sigma}} (\nabla \cdot H)(\nabla \cdot A) + i\omega \int_\Omega \mu_0 H \cdot A = \int_\Omega \frac{1}{\hat{\sigma}} J^{imp} \cdot (\nabla \times A) \tag{2.47}$$

$$\forall\ A \in \mathcal{H}(\nabla\times) \cap \mathcal{H}(\nabla\cdot),\ n \times A|_{\partial\Omega} = 0$$

$$H \in \mathcal{H}(\nabla\times) \cap \mathcal{H}(\nabla\cdot), n \times H|_{\partial\Omega} = n \times \hat{H}|_{\partial\Omega}$$

where $n \times \hat{H}$ denote tangential boundary values for H.

The bilinear form of this equation for $\hat{\sigma} \in \mathbb{R}$ and $0 < \sigma_m < \hat{\sigma} < \sigma_M < \infty$ is coercive and bounded with respect to a norm

$$||A||_{DC} = \sqrt{||\nabla \times A||_0^2 + ||\nabla \cdot A||_0^2 + ||A||_0^2}$$

So the equation admits a unique solution, which is the magnetic field H.

The advantage of the equation (2.47) is that even if the term $i\omega \int_\Omega \mu_0 H \cdot A$ was not present, which happens when the frequency $w = 0$, the bilinear form would remain coercive, and as a result, the system matrix is well conditioned for small frequencies. If there is a jump in conductivity, the condition number of the system matrix increases, yet the situation is similar to a discontinuous coefficient in the Poisson equation. Even if there is a high contrast in conductivity, it should be sufficient to use standard vector multigrid preconditioners (see [25]) for the iterative solver to converge.

This kind of regularization has been studied in the literature (see [26, 27]) without introducing the notion of the Schelkunoff potential. Indeed, if the original field is divergence free, then the Schelkunoff potential with boundary condition (2.20) coincides with the original field. An interesting eigenvalue analysis for the equation with and without divergence term is presented in [30].

Caution is needed as $(\mathcal{H}^1)^3 \cap \mathcal{H}_0(\nabla\times)$ is not always dense in $\mathcal{H}(\nabla\cdot) \cap \mathcal{H}_0(\nabla\times)$ (see [26] or Appendix B in [14]). In geophysical applications a computational domain is usually a convex polygon (in magnetotellurics it is a cuboid). In this situation, $(\mathcal{H}^1)^3 \cap \mathcal{H}_0(\nabla\times)$ is dense in $\mathcal{H}(\nabla\cdot) \cap \mathcal{H}_0(\nabla\times)$, so the use of nodal shape functions leads to a convergent discretization. Numerical tests involving equation (2.47) are presented in Section 2.8.

2.8 Numerical results

In this section, the magnetic field for a plane-wave (magnetotelluric) source is calculated using equation (2.47) and compared with a field calculated by an independent integral equation code of [28].

The considered model is a conductive brick of resistivity $1\Omega m$ and dimensions 1km x 2km x 2km in the whole space of resistivity $100\Omega m$. The field is calculated 500m above the brick, along a line going in y-direction. Second-order nodal shape functions

are used for each component of the field. Hexahedral mesh is used. A sketch of the model and hexahedral mesh is presented in Figure 2.1.

The solution H is approximated by

$$H = \sum_{j=1}^{n} x_j N_j \tag{2.48}$$

where n is the number of degrees of freedom, N_j are shape functions. Inserting (2.48) into equation (2.47) gives

$$\int_\Omega \tfrac{1}{\hat{\sigma}} \left(\nabla \times \textstyle\sum_{j=1}^n x_j N_j\right) \cdot (\nabla \times N_k) + \int_\Omega \tfrac{1}{\hat{\sigma}} \left(\nabla \cdot \textstyle\sum_{j=1}^n x_j N_j\right) (\nabla \cdot N_k)$$
$$+ i\omega \int_\Omega \mu_0 \textstyle\sum_{j=1}^n x_j N_j \cdot N_k = \int_\Omega \tfrac{1}{\hat{\sigma}} J^{imp} (\nabla \times N_k)$$

which produces a linear system $Ax = b$ to be solved, where

$$A_{kj} = \int_\Omega \frac{1}{\hat{\sigma}} (\nabla \times N_j) \cdot (\nabla \times N_k) + \int_\Omega \frac{1}{\hat{\sigma}} (\nabla \cdot N_j)(\nabla \cdot N_k) + i\omega \int_\Omega \mu_0 N_j \cdot N_k$$
$$b_k = \int_\Omega \frac{1}{\hat{\sigma}} J^{imp} \cdot (\nabla \times N_k)$$

The total field of a plane wave in a whole space with a brick is decomposed $H_t = H_p + H_s$, $E_t = E_p + E_s$ into a primary electromagnetic field (H_p, E_p) and a secondary electromagnetic field (H_s, E_s). The primary field is a plane wave going in increasing z direction in whole space with the H field in y direction and the secondary field is the difference due to the presence of the brick. The code solves for the secondary field H_s, with $n \times H_s = 0$ on $\partial\Omega$. It is assumed that $\sigma = \sigma_t$ is the conductivity of a conducting brick in a whole space, with the source $J^{imp} = E_p \sigma_s$, where $\sigma_s = \sigma_t - \sigma_p$ is the difference between the conductivity of the whole space with the conducting brick and the conductivity of the whole space. Two frequencies were considered: 0.001Hz and 10Hz. The mesh consisted of 15x15x20 hexahedral elements and extended more than 20km from the brick. The inner part of the mesh is presented on the following figures. The linear system had $98,397$ unknowns. QMR with incomplete LU preconditioner converged to 10^{-7} of the starting residue in 28 iterations for the frequency 10Hz and in 54 iterations for 0.001Hz.

Figure 2.2 presents the ratio of the secondary field to the primary field. The fields calculated by Integral Equation code and FEM code that uses (2.47) are similar for both frequencies. The proposed method gives proper values of the magnetic field H.

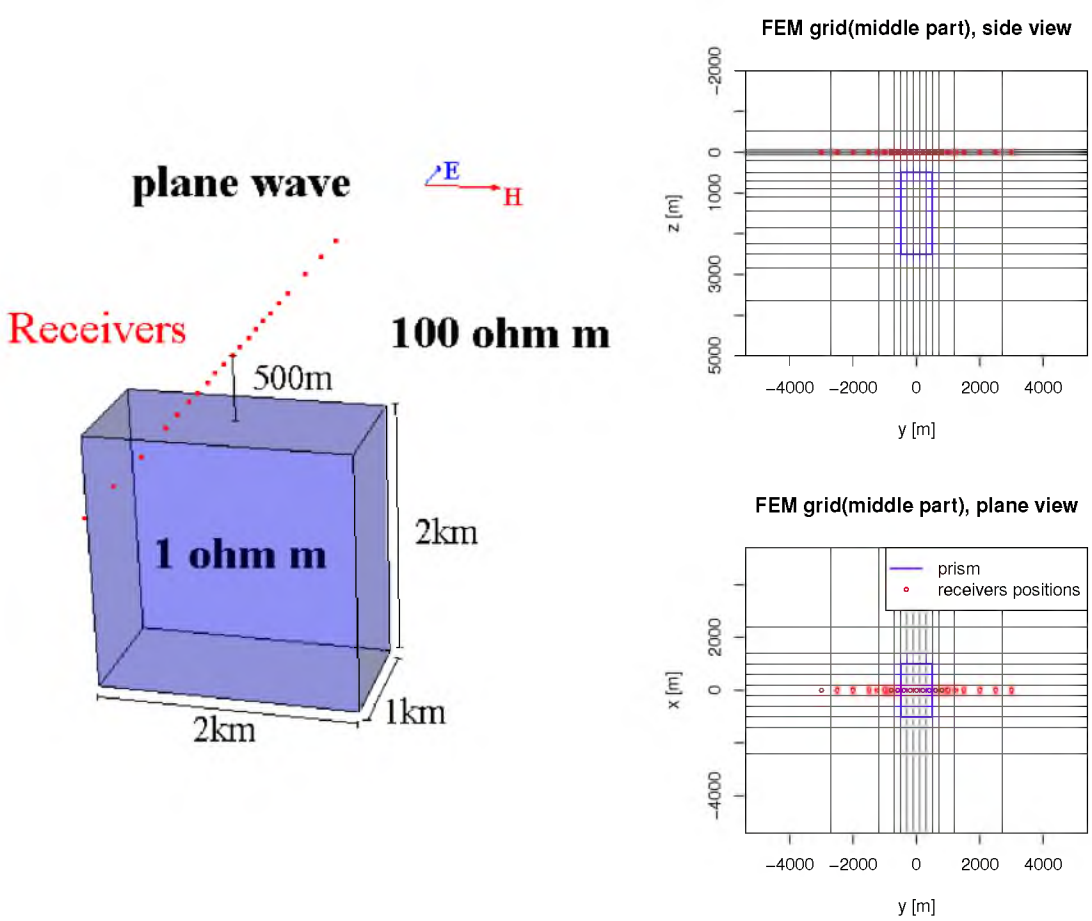

Figure 2.1. Sketch of a considered model for numerical simulation(left); Hexahedral mesh cross-sections(right)

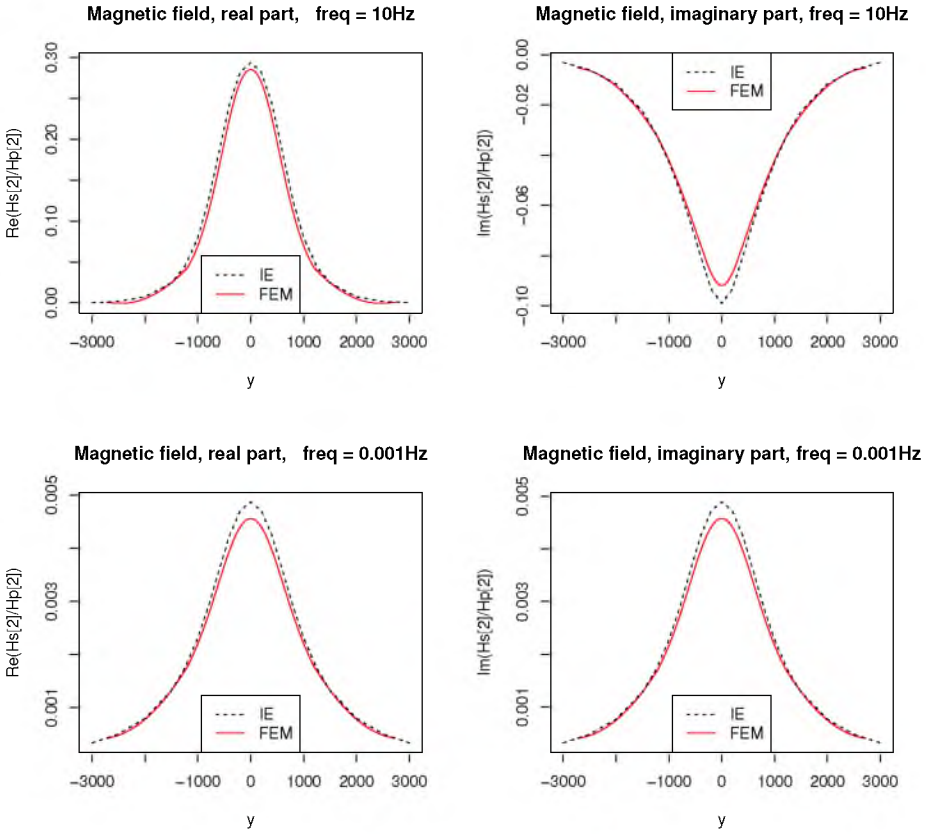

Figure 2.2. Ratio of the secondary field to the primary field for frequency 10Hz(top) and for frequency 0.001Hz(bottom)

Acknowledgments

We acknowledge the support of this work from the U.S. Dept. of Energy under contract DE-EE0002750 to PW. EC acknowledges the partial support of the U.S. National Science Foundation through grants ARC-0934721 and DMS-1413454.

2.9 Appendix

In this section, formal definitions and some intuition for the spaces used in the paper are presented.

$\mathcal{H}^1$ space, defined formally below, is a space of scalar functions in Ω that are square integrable, and for which it is possible to calculate gradient, and the gradient is square integrable as well. A useful intuition is to think about members of this family as being continuous. For example, if the domain is split into two subsets and the function experiences a jump on the boundary between those two subsets and otherwise is continuous, it is not a member of $\mathcal{H}^1$. This is a natural space for scalar potentials for which the gradient exists.

$\mathcal{H}(\nabla\times)$ is a space of vector fields, defined in Ω, which are square integrable, and for which it is possible to calculate the curl, and the curl is square integrable. It is useful to imagine the members to have continuous tangential components across any surface in Ω. Normal components do not have to be continuous. This is a natural space for force fields, electric field E and magnetic field H.

$\mathcal{H}(\nabla\cdot)$ is a space of vector fields, defined in Ω, which are square integrable, and for which it is possible to calculate the divergence, and the divergence is square integrable. Fields in this family will have continuous normal components across any surface in Ω. This is a natural space for fluxes. Total electric current J and magnetic induction B are members of this family.

$$\begin{array}{lcll} L^2 & = & L^2(\Omega) & = \left\{\varphi : \Omega \to \mathbb{C} : \int_\Omega |\varphi|^2 < \infty\right\} \\ \mathcal{H}^1 & = & \mathcal{H}^1(\Omega) & = \left\{\varphi : \Omega \to \mathbb{C} : \int_\Omega |\nabla\varphi|^2 + \int_\Omega |\varphi|^2 < \infty\right\} \\ \mathcal{H}(\nabla\times) & = & \mathcal{H}(\nabla\times,\, \Omega) & = \left\{A : \Omega \to \mathbb{C}^3 : \int_\Omega |\nabla \times A|^2 + \int_\Omega |A|^2 < \infty\right\} \\ \mathcal{H}(\nabla\cdot) & = & \mathcal{H}(\nabla\cdot,\, \Omega) & = \left\{A : \Omega \to \mathbb{C}^3 : \int_\Omega |\nabla \cdot A|^2 + \int_\Omega |A|^2 < \infty\right\} \end{array}$$

If homogeneous boundary conditions are assumed, a subscript "0" is added. For $\mathcal{H}_0^1$, $\mathcal{H}_0(\nabla\times)$, $\mathcal{H}_0(\nabla\cdot)$, the value of the function, tangential, or normal components of a vector field are fixed, respectively. If n is a vector normal to the boundary $\partial\Omega$, then

$$\begin{array}{lcll} \mathcal{H}_0^1 & = & \mathcal{H}_0^1(\Omega) & = \{\varphi \in \mathcal{H}^1(\Omega) : \varphi|_{\partial\Omega} = 0\} \\ \mathcal{H}_0(\nabla\times) & = & \mathcal{H}_0(\nabla\times, \Omega) & = \{A \in \mathcal{H}(\nabla\times, \Omega) : n \times A|_{\partial\Omega} = 0\} \\ \mathcal{H}_0(\nabla\cdot) & = & \mathcal{H}_0(\nabla\cdot, \Omega) & = \{A \in \mathcal{H}(\nabla\cdot, \Omega) : n \cdot A|_{\partial\Omega} = 0\} \end{array}$$

Additionally, norms defined below are used in the paper.

$$\begin{array}{lcl} ||\varphi||_0 & = & \sqrt{\int_\Omega |\varphi|^2} \\ ||\varphi||_1 & = & \sqrt{||\varphi||_0^2 + ||\nabla\varphi||_0^2} = \sqrt{\int_\Omega |\varphi|^2 + \int_\Omega |\nabla\varphi|^2} \end{array}$$

Three vector identities are used. For $K, L : \mathbb{R}^3 \to \mathbb{C}^3, u : \mathbb{R}^3 \to \mathbb{C}$, we have:

$$\nabla \times \nabla \times K = \nabla(\nabla \cdot K) - \nabla \cdot (\nabla K) \tag{2.49}$$

$$\int_\Omega (\nabla \times K) \cdot L = \int_\Omega K \cdot (\nabla \times L) + \int_{\partial\Omega} (n \times K) \cdot L \tag{2.50}$$

$$\int_\Omega \nabla u \cdot K = -\int_\Omega u \nabla \cdot K + \int_{\partial\Omega} u(K \cdot n) \tag{2.51}$$

Poincare inequalities (see Appendix A in [14]):

$$c||\varphi||_0 \leq ||\nabla\varphi||_0 \quad \text{for } \varphi \in \mathcal{H}_0^1 \tag{2.52}$$

$$c||A||_0 \leq ||\nabla \times A||_0 \quad \text{for } A \in \mathcal{H}_0(\nabla\times),\ A \in R(\nabla)^\perp \tag{2.53}$$

$$c||A||_0 \leq ||\nabla \cdot A||_0 \quad \text{for } A \in \mathcal{H}_0(\nabla\cdot),\ A \in R(\nabla\times)^\perp \tag{2.54}$$

The following theorem expresses the fact that a Schelkunoff type representation is valid not only for electric field E, magnetic field H, and for piecewise constant μ, σ, but is a general feature of members of $\mathcal{H}(\nabla\times)$ and any function κ that is bounded from above and below in Ω.

Theorem 8 *Let Ω be an open bounded subset of $\mathbb{R}^3$ with Lipschitz boundary. For any $K \in \mathcal{H}_0(\nabla\times, \Omega)$, and any $\kappa : \Omega \to \mathbb{C}$ such that $\kappa = \kappa_r + i\kappa_i$*

$$\forall x \in \Omega \quad \kappa_r(x) \geq 0, \quad 0 < \kappa_m \leq |\kappa(x)| \leq \kappa_M < \infty \tag{2.55}$$

there exists

$$F \in \mathcal{H}_0(\nabla\times, \Omega) \cap \mathcal{H}(\nabla\cdot, \Omega) \tag{2.56}$$

such that

$$\kappa\nabla \cdot F \in \mathcal{H}_0^1(\Omega) \tag{2.57}$$

and

$$K = F - \nabla(\kappa\nabla \cdot F) \tag{2.58}$$

Remark 9

- *One may also consider $K \in \mathcal{H}(\nabla\times)$ and $F \in \mathcal{H}(\nabla\times) \cap \mathcal{H}_0(\nabla\cdot)$. The proof is similar. Equation (2.59) has to be considered for $F, A \in \mathcal{H}_0(\nabla\cdot)$*
- *Notice that we no longer need the assumption that κ is piecewise constant or that the field K is piecewise divergence free.*

Proof : Consider $F \in \mathcal{H}(\nabla\cdot)$, a solution of the equation:

$$\forall\, A \in \mathcal{H}(\nabla\cdot) \qquad \int_\Omega F\cdot A + \int_\Omega \kappa(\nabla\cdot F)(\nabla\cdot A) = \int_\Omega K\cdot A \tag{2.59}$$

A calculation similar to the one in the proof of Theorem 4 shows that the left-hand side of (2.59) is a bounded and coercive bilinear form with respect to the norm

$$||A||_{\nabla\cdot} = \sqrt{\int_\Omega |A|^2 + \int_\Omega |\nabla\cdot A|^2}$$

Hence the problem has a unique solution from the Lax-Milgram theorem.

It remains to show that the solution F has all the desired properties. First, let us show that $\kappa\nabla\cdot F$ is a member of $\mathcal{H}_0^1(\Omega)$. It belongs to $L^2(\Omega)$, as $F \in \mathcal{H}(\nabla\cdot)$ and $|\kappa| \leq \kappa_M < \infty$. Let us consider $\nabla(\kappa\nabla\cdot F)$ as a distribution. Let us take any vector field with compact support in Ω, having derivatives of any order, $A \in (C_c^\infty(\Omega))^3$. Such A is also a member of $\mathcal{H}(\nabla\cdot)$, so it satisfies (2.59). Evaluating the value of the distribution $\nabla(\kappa\nabla\cdot F)$ at A, we obtain:

$$< \nabla(\kappa\nabla\cdot F), A >= - < \kappa\nabla\cdot F, \nabla\cdot A >= -\int_\Omega \kappa(\nabla\cdot F)(\nabla\cdot A) \overset{(2.59)}{=} \int_\Omega (F-K)\cdot A$$

Therefore, $\nabla(\kappa\nabla\cdot F) = F - K \in L^2(\Omega)$. We have shown that $\kappa\nabla\cdot F \in H^1(\Omega)$ and also that (2.58) is satisfied. Let us now show that a trace of $\kappa\nabla\cdot F$ on $\partial\Omega$ is 0. For any $A \in \mathcal{H}(\nabla\cdot)$, we have

$$\int_{\partial\Omega} (\kappa\nabla\cdot F)A\cdot n = \int_\Omega \nabla(\kappa\nabla\cdot F)\cdot A + \int_\Omega (\kappa\nabla\cdot F)(\nabla\cdot A) \overset{(2.58)}{=}$$
$$\int_\Omega (F-K)\cdot A + \int_\Omega (\kappa\nabla\cdot F)(\nabla\cdot A) \overset{(2.59)}{=} 0$$

So indeed, $(\kappa\nabla\cdot F) \in H_0^1$. This implies $\nabla(\kappa\nabla\cdot F) \in \mathcal{H}_0(\nabla\times)$, also $K \in \mathcal{H}_0(\nabla\times)$. From (2.58), one can conclude that $F = K + \nabla(\kappa\nabla\cdot F)$, so $F \in H_0(\nabla\times)$.

2.10 References

[1] W. Rodi and R. L. Mackie, "Nonlinear conjugate gradients algorithm for 2-D magnetotelluric inversion," *Geophysics*, vol. 66, no. 1, pp. 174–187, 2001.

[2] D. Avdeev and A. Avdeeva, "3D magnetotelluric inversion using a limited-memory quasi-Newton optimization," *Geophysics*, vol. 74, no. 3, F45–F57, 2009.

[3] C. G. Farquharson, D. W. Oldenburg, E. Haber, and R. Shekhtman, "An algorithm for the three-dimensional inversion of magnetotelluric data," in *72st Ann. Internat. Mtg., Soc. Expl. Geophys*, 2002, pp. 649–652.

[4] R.-U. Boerner, "Numerical modelling in geo-electromagnetics: advances and challenges," *Surv. Geophys.*, vol. 31, pp. 225–245, 2010.

[5] E. Haber, D. Oldenburg, and R. Shektman, "Inversion of time-domain three-dimensional data," *Geophys. J. Int.*, vol. 171, pp. 550–564, 2007.

[6] M. Commer and G. A. Newman, "New advances in three-dimensional controlled-source electromagnetic inversion," *Geophys. J. Int.*, vol. 172, pp. 513–535, 2008.

[7] E. S. Um, J. M. Harris, and D. L. Alumbaugh, "An iterative finite element time-domain method for simulating three-dimensional electromagnetic diffusion in the earth," *Geophys. J. Int.*, vol. 190, pp. 871–886, 2012.

[8] Y. Saad, *Iterative methods for sparse linear systems*. SIAM, 2003.

[9] R. L. Mackie, J. T. Smith, and T. R. Madden, "Three-dimensional electromagnetic modeling using finite difference equations: the magnetotelluric example," *Radio Science*, vol. 29, no. 4, pp. 923–935, 1994.

[10] J Smith, "Conservative modeling of 3-D electromagnetic fields, Part II: biconjugate gradient solution and an accelerator," *Geophysics*, vol. 61, no. 5, pp. 1319–1324, 1996.

[11] J. C. Nedelec, "Mixed finite elements in R^3," *Numer. Math.*, vol. 35, pp. 315–341, 1980.

[12] R. Hiptmair, "Finite elements in computational electromagnetism," *Acta Numerica*, vol. 11, pp. 237–339, 2002.

[13] A. Bermdez, R. Rodrguez, and P. Salgado, "Numerical analysis of electric field formulations of the eddy current model," *Numer. Math.*, vol. 102, pp. 181–201, 2005.

[14] P. B. Bochev and M. D. Gunzburger, *Least-Squares Finite Element Methods*. Springer New York, 2009.

[15] J. Xu, "The auxiliary space method and optimal multigrid preconditioning techniques for unstructured grids," *Computing*, vol. 56, no. 3, pp. 215–235, 1996.

[16] R. Hiptmair, "Multigrid method for Maxwell's equations," *SIAM Journal on Numerical Analysis*, vol. 36, no. 1, pp. 204–225, 1998.

[17] ——, "Analysis of multilevel methods for eddy current problems," *Mathematics of Computation*, vol. 72, no. 243, pp. 1281–1303, 2003.

[18] T. V. Kolev and P. S. Vassilevski, "Some experience with a H1-based auxiliary space AMG for H(curl) problems," *Lawrence Livermore National Laboratory, Technical report UCRL-TR-221841, Livermore, CA.*, 2006.

[19] D. Arnold, R. Falk, and R. Winther, "Multigrid in H(div) and H(curl)," *Numer. Math.*, vol. 85, no. 2, pp. 197–217, 2000.

[20] J. Xu, L. Chen, and R. H. Nochetto, *Optimal multilevel methods for H(grad), H(curl), and H(div) systems on graded and unstructured grids. In multi-scale, nonlinear and adaptive approximation.* Springer Berlin Heidelberg., 2009, pp. 599–659.

[21] S. Ward and G. Hohmann, *Electromagetic Theory for Geophysical Applications.* Electromagnetic Methods in Applied Geophysics-Theory Volume I, chapter 4., 1988.

[22] A. Bossavit, "On the Lorenz gauge," *COMPEL - the International Journal for Computation and Mathematics in Electrical and Electronic Engineering*, vol. 18, no. 3, pp. 323–336, 1999.

[23] E. S. Um, D. L. Alumbaugh, and J. M. Harris, "A Lorenz-gauged finite-element solution for transient CSEM modeling," in *SEG Annual Meeting, 17-22 October, Denver, Colorado*, 2010.

[24] W. E. Boyse, D. R. Lynch, K. D. Paulsen, and G. N. Minerbo, "Nodal-based finite-element modeling of Maxwell's equations," *IEEE Transactions on Antennas and Propagation*, vol. 40, no. 6, pp. 642–651, 1992.

[25] P. Vanek, J. Mandel, and M. Brezina, "Algebraic multigrid by smoothed aggregation for second and fourth order elliptic problems," *Computing*, vol. 56, no. 3, pp. 179–196, 1996.

[26] A. Dhia, C. Hazard, and S. Lohrengel, "A singular field method for the solution of Maxwell's equations in polyhedral domains," *SIAM J. Appl. Math.*, vol. 59, pp. 2028–2044. 1999.

[27] C. Amrouche, C. Bernardi, M. Dauge, and V. Girault, "Vector potentials in three-dimensional nonsmooth domains," *Mathematical Methods in the Applied Sciences*, vol. 21, pp. 823–864, 1998.

[28] P. Wannamaker, G. Hohmann, and W. SanFilipo, "Electromagnetic modeling of three-dimensional bodies in layered earths using integral equations," *Geophysics*, vol. 49, pp. 60–74, 1984.

[29] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier-Stokes Equations: Theory and Algorithms.* Springer Berlin Heidelberg, 1986, ISBN: 978-3-642-61623-5. DOI: 10.1007/978-3-642-61623-5.

[30] W. Shin and F. Shanhui, "Accelerated solution of the frequency-domain Maxwells equations by engineering the eigenvalue distribution of the operator," *Optics Express*, vol. 21, no. 19, pp. 22 578–22 595, 2013.

CHAPTER 3

3D MAGNETOTELLURIC INVERSION INCLUDING TOPOGRAPHY USING DEFORMED HEXAHEDRAL EDGE FINITE ELEMENTS AND DIRECT SOLVERS PARALLELIZED ON SMP COMPUTERS, PART I: FORWARD PROBLEM AND PARAMETER JACOBIAN[1]

Kordy M.[2,3], Wannamaker P.[3], Maris V.[3], and Cherkaev E.[2]

3.1 Abstract

We have developed an algorithm, which we call HexMT for 3D simulation and inversion of magnetotelluric (MT) responses using deformable hexahedral finite elements, that permits the incorporation of topography. Direct solvers parallelized on symmetric multiprocessor (SMP), single-chassis workstations with large RAM are used throughout, including the forward solution, the parameter Jacobian, and model parameter update. In Part I, the forward simulator and Jacobian calculations are presented. We use first-order edge elements to represent the secondary electric field (E), yielding accuracy O(h) for E and its curl (magnetic field). For very low frequencies or small material admittivities, the E-field requires divergence correction. With the help of Hodge decomposition, the correction may be applied in one step after

[1]Submitted to Geophysical Journal International in 2014

[2]Department of Mathematics, University of Utah

[3]Energy & Geoscience, University of Utah

the forward solution is calculated. This allows accurate E-field solutions in dielectric air. The system matrix factorization and source vector solutions are computed using the MUMPS library, which shows moderately good scalability through 12 processor cores but limited gains beyond that. The factored matrix is used to calculate the forward response as well as the Jacobian of EM field and MT responses using the reciprocity theorem. Comparison with other codes demonstrates accuracy of our forward calculations. We consider a popular conductive/resistive double brick structure and several topographic models. In particular, the ability of finite elements to represent smooth topographic slopes permits accurate simulation of refraction of electromagnetic waves normal to the slopes at high frequencies. Run time tests of the parallelized algorithm indicate that for meshes as large as 150x150x60 elements, the MT forward response and the Jacobian can be calculated in $\sim$ 2.5 hours per frequency. Together with an efficient inversion parameter step described in Part II, MT inversion problems of 200-300 stations are computable with total run times of several days on such workstations.

3.2 Introduction

Impressive progress has been made over the past several years in the simulation and inversion of three-dimensional (3D) diffusive electromagnetic (EM) responses for earth electrical resistivity structure. Most approaches have adopted finite difference or finite element numerical methods although the integral equations technique also has been utilized [see reviews by 1, 2]. An effective simulation and inversion algorithm needs to handle a large range of structural scales due to possibly complex resistivity distributions and the wide frequency bandwidth of survey techniques (e.g., potentially seven or more orders of magnitude in magnetotellurics). Furthermore, in many orogenic or resource settings, the earth's surface can show considerable topographic variation which will have its own EM response and introduces nonuniformity of receiver placement with respect to subsurface structure.

To include topography in earth resistivity models, we pursue the finite element method. Finite elements allow a relatively smooth representation of topography whereas the stair-step construction inherent with finite differences may introduce

spurious local electric field behavior [3]. Several authors have considered the choice between tetrahedral and hexahedral elements for this purpose [e.g., 4–6], with tetrahedra argued by some to allow a more arbitrary discretization of structure. However, we will show that much can be accomplished using hexahedral elements, and their simpler implementation in both the forward and inverse modules helps to keep computer resources manageable and may facilitate wider transfer of technique within the EM community. We solve for the electric field through the governing Helmholtz equation, so edge finite elements are used (lowest order type) [7]. These conforming elements allow field discontinuities normal to conductivity interfaces to be represented but preserve continuity of the tangential field component. Finite elements do not invoke material averaging procedures across cell boundaries as is done in staggered grid finite difference schemes [4] so there is no question about the placement of sharp interfaces.

The design factors cited above can put high demands upon mesh discretization and computing resources for larger data sets. Because of such demands, especially memory, iterative solutions have dominated the literature heretofore [8–11, and many others]. Since at least the work of [12], however, iterative forward solvers are known to become ill-conditioned and slow to converge if grid cell aspect ratios grow to be extreme. Moreover, iterative solutions for the Helmholtz equation require careful preconditioning and even so may sometimes fail to converge [also see 13]. They become expensive when many right-hand source vectors are needed, such as in controlled-source applications or the inversion approach we describe, as each source requires the work of a full simulation. Conditioning issues may apply as well to iteratively solving normal equations in the inversion parameter step (op. cit.).

Recent advances in computing power, especially emergence of less expensive many-core, symmetric multiprocessor (SMP) workstations with substantial RAM, have motivated us to implement direct solvers both for the forward model responses and for Gauss-Newton inversion parameter steps. This is intended to produce a practical 3D inversion code incorporating topography that can handle moderately large data sets on an affordable, single-box computer format. We find that accurate solutions for meshes with large element aspect ratios having run-times nearly independent of frequency are possible. The solution of hundreds of source vectors for the cost

of factoring the forward system matrix allows explicit calculation of the parameter Jacobian accurately and efficiently, as has been applied for some time with the 2D problem [e.g., 14–16].

We certainly are not the first to examine direct solutions for 3D problems. In [17], the author created a staggered-grid finite difference algorithm for simulating marine CSEM responses. The authors of [18, 19] used a direct solver in their finite difference H-field simulator for TDEM inversion. In [4], the authors utilized rectilinear edge finite elements in forward modeling of seafloor CSEM models. The author of [13] incorporated the solver of [17] to compute forward responses and parameter the Jacobian explicitly and create an inversion algorithm where the parameter step was estimated using a preconditioned conjugate gradient (PCG) scheme. In [6], the authors develop an unstructured mesh of tetrahedra with the forward problem solved directly and the parameter step computed via PCG or iterative quasi-Newton method.

In Part I of our contribution, we apply a direct solver to edge finite element (FE) equations of a deformed hexahedral mesh and verify that accurate responses are achieved for subsurface and topographic structure. Good responses are obtained also in the dielectric air portion of the model after applying a divergence correction. The parameter Jacobian is computed accurately and efficiently in the direct framework exploiting reciprocity. Moderately large meshes can be computed in what we believe are practical run times. In Part II, simulations are combined with MT data to form normal equations for a regularized inversion step. We investigate both model space and data space [20] formulations of the step and confirm that significantly larger parameter sets can be handled by the latter for typical MT data. We invert a well-known field data set to demonstrate algorithm performance in real-world settings. The algorithm, which we name HexMT, is parallelized for widely available, server-class SMP workstations.

3.3 Finite element formulation

For representing structure with topography, we use an FE mesh such as in Figure 3.1. Corners of elements at the air-earth interface (surface) are adjusted vertically to represent elevation changes. This is similar to the fashion of [21]. Sub- and suprajacent element layers are moved similarly but with steadily diminishing magnitude away from

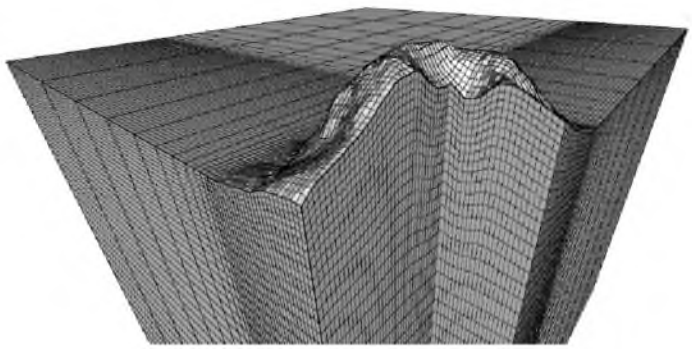

Figure 3.1. 3D view of an example hexahedral mesh with topography. Only the underground part of the mesh is shown. One sees increasingly high aspect ratio of elements approaching the boundary $\partial\Omega$.

the surface until upper and lower datum planes are reached. Beyond those planes, the element layers remain flat. The height and depth of these planes from the background air-earth interface typically are several times the maximal topographic model relief.

Formally, the spatial domain of Figure 3.1 is a cuboid Ω, whose top portion is air ($\sigma = 0$) and whose lower portion is earth's subsurface ($\sigma > 0$) which may exhibit topography in its central portion. We assume that the conductivity of the earth's subsurface may be an arbitrary three-dimensional (3D) isotropic function in the middle of the domain, while toward the distant domain boundaries, the conductivity becomes 1D with flat topography, i.e., changing only vertically. In the frequency domain with $e^{\mathrm{i}\omega t}$ time dependence and ω the angular frequency, the physical property variables are admittivity $\hat{\sigma} = \sigma + \mathrm{i}\omega\epsilon$ with electrical conductivity $\sigma \geq 0$, dielectric permittivity $\epsilon > 0$, and magnetic permeability $\mu > 0$.

Similar to numerous other authors [e.g., following 22], we define (E^p, H^p) as primary fields, which would be those within and over the 1D host, for use as an impressed source J^{imp}. Thus we denote

$$J^{\mathrm{imp}} = -(\hat{\sigma} - \hat{\sigma}^p)E^p \tag{3.1}$$

Secondary and primary fields are added to obtain total fields as:

$$E^t = E + E^p, \;\; H^t = H + H^p \tag{3.2}$$

One assumes that far from conductivity inhomogeneity, i.e., near $\partial\Omega$,

$$E^t \approx E^p, \;\; H^t \approx H^p \tag{3.3}$$

The secondary field E obeys the vector Helmholtz equation in the open spatial domain $\Omega \subset \mathbb{R}^3$:

$$\nabla \times \left(\frac{1}{\mu}\nabla \times E\right) + \mathrm{i}\omega\hat{\sigma}E = -\mathrm{i}\omega(\hat{\sigma} - \hat{\sigma}^p)E^p \tag{3.4}$$

[22]. As a basis for finite element formulation, we will consider a weak form of the equation (3.4):

$$\int_\Omega \frac{1}{\mu}\nabla \times E \cdot \nabla \times F + \mathrm{i}\omega \int_\Omega \hat{\sigma}E \cdot F = \int_\Omega J^{\mathrm{imp}} \cdot F \tag{3.5}$$

satisfied for all $F \in \mathcal{H}_0(\nabla\times,\Omega)$. The solution E should be a member of the same Sobolev space $\mathcal{H}_0(\nabla\times,\Omega)$, which is formally defined as

$$\mathcal{H}_0(\nabla\times,\Omega) = \left\{ F{:}\Omega \to \mathbb{C}^3 : \int_\Omega \left(|F|^2 + |\nabla \times F|^2\right) < \infty, \quad n \times F|_{\partial\Omega} = 0 \right\} \tag{3.6}$$

It is a space of complex valued vector fields that are square integrable with square integrable curl. Heuristically, one can think of the members of this space as having continuous tangential components across any surface going through Ω. $\mathcal{H}_0(\nabla\times,\Omega)$ is a natural space for the electric field E. The boundary condition $n \times E = 0$ is a natural consequence of (3.2) and (3.3). Equation (3.5) imposes weakly another condition: $\nabla \cdot E = 0$ on $\partial\Omega$, which is also a consequence of $E \approx 0$ close to $\partial\Omega$.

For numerical approximation, we choose first-order edge elements $\mathcal{H}_0^h(\nabla\times,\Omega)$ on a hexahedral mesh [see 7]. By construction $\mathcal{H}_0^h(\nabla\times,\Omega) \subset \mathcal{H}_0(\nabla\times,\Omega)$ and as the mesh element size $h \to 0$, $\mathcal{H}_0^h(\nabla\times,\Omega)$ approaches $\mathcal{H}_0(\nabla\times,\Omega)$. Therefore, this discretization is called "compatible". The tangential component of the members of $\mathcal{H}_0^h(\nabla\times,\Omega)$ are continuous across elements while the normal component may experience a jump. Degrees of freedom of the first-order edge elements are related to the integral of the E-field along an edge. Through Stokes theorem, an integration of E along the edges, around the face yields the flux of $\nabla \times E$ through the face. This shows that edge element discretization is compatible with the curl operator.

The electric field over Ω is represented as a linear combination of the edge shape functions N_i with coefficients ξ_i:

$$E = \sum_{i=1}^{n_e} \xi_i N_i \tag{3.7}$$

where $i = 1, \ldots, n_e$ are indices of the edges that do not lie on the boundary. By substituting this to equation (3.5) and using N_j as test functions, one obtains a linear system

$$A\xi = b \tag{3.8}$$

$$A_{i,j} = \left[\int_\Omega \frac{1}{\mu} \nabla \times N_i \cdot \nabla \times N_j + \mathrm{i}\omega \int_\Omega \hat{\sigma} N_i \cdot N_j \right] \tag{3.9}$$

$$b_i = \int_\Omega J^{\mathrm{imp}} \cdot N_i \tag{3.10}$$

The secondary magnetic field is calculated as

$$H = \frac{-\nabla \times E}{i\omega\mu} \tag{3.11}$$

This justifies the choice of first-order edge elements which have the same accuracy $O(h)$ for both the field and the curl.

Note that 1D host layer interfaces may project through individual deformed elements and as a result, J^{imp} is discontinuous within an element. The integration of terms in (3.9) and (3.10) is done using a quadrature integration of the form:

$$\sum_{i=1}^{n} f(u_i)v_i \tag{3.12}$$

where u_i are points in the reference element, which is a unit cube in our case and v_i are weights. If the integrand f is smooth in the element, which is true for (3.9) and for (3.10) if the 1D host layer interface does not project through an element, positions u_i and weights v_i are set according to Gaussian quadrature. Yet for (3.10), if a 1D conductivity layer interface splits the element, the integrated function is discontinuous and the integration is done by distributing u_i uniformly in the unit cube and setting all $v_i = \frac{1}{n}$. For accuracy of integration, n should have larger values than in the case of a smooth function. As will be seen, with sufficiently fine integration of the primary field over the element conductivity differences, we are able to achieve accurate responses.

In this paper, we consider the magnetotelluric (MT) source, namely that of a vertically propagating, planar EM wave. The total field components at specified surface locations and frequencies are interrelated through the tensor impedance Z and tipper K as:

$$\begin{bmatrix} E_x^t \\ E_y^t \\ H_z^t \end{bmatrix} = \begin{bmatrix} Z_{xx} & Z_{xy} \\ Z_{yz} & Z_{yy} \\ K_{zx} & K_{zy} \end{bmatrix} \begin{bmatrix} H_x^t \\ H_y^t \end{bmatrix} \tag{3.13}$$

where subscripts x, y, z denote components of a vector field. The equation (3.8) is solved twice for two polarizations (k = 1,2) of the source field E^p, typically in the x and then the y directions, to generate two equations in two unknowns for each row of the tensor (3.13). The impedance element equations are listed in, e.g., [23] and can be written analogously for the tipper.

A receiver can be positioned at an arbitrary location $\mathbf{r}$ with respect to element edges via appropriate interpolation. In general, let $\mathbf{r}$ be inside an element with edges $e_1, \ldots, e_{12}$. Then field E^k at location $\mathbf{r}$ is given by

$$E^k(\mathbf{r}) = \sum_{l=1}^{12} N_{e_l}(\mathbf{r})\xi_{e_l} = \begin{bmatrix} (w_x^E)^T\xi^k \\ (w_y^E)^T\xi^k \\ (w_z^E)^T\xi^k \end{bmatrix} \tag{3.14}$$

Here w_x^E, w_y^E, w_z^E contain interpolation vectors with at most 12 non-zero values corresponding to x, y, and z components of edge shape functions $N_{e_1}(\mathbf{r}), \ldots, N_{e_{12}}(\mathbf{r})$.

Similarly, the secondary magnetic field $H^k(\mathbf{r})$ for polarization k, calculated using (3.11) at location $\mathbf{r}$, is given by

$$H^k(\mathbf{r}) = \sum_{l=1}^{12} \frac{\nabla \times N_{e_l}(\mathbf{r})}{-i\omega\mu}\xi_{e_l} = \begin{bmatrix} (w_x^H)^T\xi^k \\ (w_y^H)^T\xi^k \\ (w_z^H)^T\xi^k \end{bmatrix} \tag{3.15}$$

This time, the only non-zero values of w_x^H, w_y^H, w_z^H are x, y, and z components of

$$\left(\frac{\nabla \times N_{e_1}(\mathbf{r})}{-i\omega\mu}, \ldots, \frac{\nabla \times N_{e_{12}}(\mathbf{r})}{-i\omega\mu}\right)$$

Total fields are obtained as in (3.2).

By convention and for inversion convenience, the location $\mathbf{r}$ of an MT receiver is at the earth's surface at an element face center. Because the tangential electric field is continuous across the surface, it is immaterial whether we approach the surface from within an element below or above the air-earth interface. E-fields normal to a surface would have to be evaluated on the side of interest. If magnetic permeability μ is the same above and below the surface, the magnetic field should be continuous as well. However, our edge element discretization allows for discontinuous tangential components of the magnetic field. Thus we use an average of the H-field just above and just below the surface. As a result, interpolation vectors w corresponding to the magnetic field may have up to 20 non-zero entries. The interpolation vectors w depend neither on the primary source fields nor on the conductivity model σ.

3.4 Divergence correction

It has been recognized for the first time by [24] that matrices formed from the numerical approximation of equation (3.4) suffer from a particular ill-conditioning.

The second term on the left side of (3.4) becomes very small at either low frequencies or small admittivities, so the solution becomes vulnerable to parasitic curl-free fields. These are manifest as erroneous divergences of current density within the earth model that require corrective steps. For example, consider a linear system (3.8) whose true solution is ξ, approximated by (3.7). Let the gradient of a potential field be added to the solution such that

$$\hat{E} = E + \nabla\tilde{\varphi} = \sum_{i=1}^{n_e} \hat{\xi} N_i \tag{3.16}$$

and let the values of $\nabla\tilde{\varphi}$ be of order 1. The residual r of equation (3.7) is defined by:

$$r = A\hat{\xi} - b = A\xi - b + A(\hat{\xi} - \xi) = A(\hat{\xi} - \xi) \tag{3.17}$$

The i-th component of the residual vector r is:

$$r_i = \sum_{j=1}^{n_e} A_{i,j}(\hat{\xi_j} - \xi_j)$$

which for air in particular reduces to

$$r_i = -\omega^2 \epsilon_0 \int_\Omega N_i \cdot \nabla\tilde{\varphi}$$

Thus the residual will be non-zero, but very small – of the order $\omega^2\epsilon_0$ for air. Even if we modify the field substantially by adding $\nabla\tilde{\varphi}$, there may be hardly any difference in the residual value. An eigenvalue analysis of ill-conditioning of equation (3.8) is presented in Appendix A (Section 3.11).

For iterative solutions to equation (3.8), the typical procedure for removing spurious curl-free fields (divergence correction) is to compute several solution iterations, estimate current divergences over the discretized model domain, calculate the curl-free fields arising from such divergences, and remove these fields from the full iterative solution at that stage [e.g., 23–27]. This is repeated numerous times until final convergence. In their direct solution, the authors of [13] augment equation (3.8) to explicitly enforce a divergence condition, resulting in an increase in matrix rank by four-thirds.

We present an alternative technique that achieves an efficient and accurate divergence correction for our FE method. Consider any domain Ω, with spatially changing

conductivity $\hat{\sigma}$, which includes both air and the subsurface. The space $H_0(\nabla\times,\Omega)$, defined in (3.6), may be decomposed into the null space of the curl and the space orthogonal to it [28]. Specifically:

$$H_0(\nabla\times) = R(\nabla) \oplus R(\nabla)^{\perp_{\hat{\sigma}}} \tag{3.18}$$

For every $F \in H_0(\nabla\times,\Omega)$, there is a unique decomposition:

$$F = \nabla\varphi_F + F_\perp, \quad \varphi_F \in H_0^1(\Omega),\ F_\perp \in R(\nabla)^{\perp_{\hat{\sigma}}} \tag{3.19}$$

where

$$\begin{array}{ll} \mathcal{H}_0^1(\Omega) = \{\varphi{:}\Omega \to \mathbb{C} & : \int_\Omega (|\varphi|^2 + |\nabla\varphi|^2) < \infty,\ \varphi|_{\partial\Omega} = 0\} \\ R(\nabla) = \{\nabla\varphi & : \varphi \in \mathcal{H}_0^1(\Omega)\} \\ R(\nabla)^{\perp_{\hat{\sigma}}} = \{K \in C_0 & : \int_\Omega \hat{\sigma} K \cdot \nabla\varphi = 0 \quad \forall\varphi \in \mathcal{H}_0^1(\Omega)\} \end{array}$$

For a discussion of Sobolev spaces $\mathcal{H}_0(\nabla\times)$ and $\mathcal{H}_0^1$ in our context, see [29, 30]. The decomposition (3.18) is called a Hodge decomposition of $\mathcal{H}_0(\nabla\times)$ and it exists when Ω includes both air and the earth's subsurface (see Appendix B in Section 3.12).

To visualize the subsequent derivations, consider Figure 3.2. Let the solution E to equation (3.5) be represented using the Hodge decomposition (3.19), namely

$$E = \nabla\varphi_E + E_\perp, \quad \varphi_E \in H_0^1(\Omega),\ E_\perp \in R(\nabla)^{\perp_{\hat{\sigma}}} \tag{3.20}$$

By setting $K = K_\perp \in R(\nabla)^{\perp_{\hat{\sigma}}}$ and then $K = \nabla\varphi$, one can show that (3.5) is equivalent to two uncoupled equations on $R(\nabla)^{\perp_{\hat{\sigma}}}$ and $R(\nabla)$, respectively:

$$\begin{array}{r} \int_\Omega \frac{1}{\mu}\nabla\times E_\perp\cdot\nabla\times K_\perp + \mathrm{i}\omega\int_\Omega \hat{\sigma} E_\perp\cdot K_\perp = \int_\Omega J^{\mathrm{imp}}\cdot K_\perp \\ \mathrm{i}\omega\int_\Omega \hat{\sigma}\nabla\varphi_E\cdot\nabla\varphi = \int_\Omega J^{\mathrm{imp}}\cdot\nabla\varphi \end{array} \tag{3.21}$$

The first equation is satisfied $\forall K_\perp \in R(\nabla)^{\perp_{\hat{\sigma}}}$, the second $\forall\varphi \in \mathcal{H}_0^1(\Omega)$. The second equation ensures that the component $\nabla\varphi_E$ is proper, so if we impose this equation, we may remove the error of the form $\nabla\tilde{\varphi}$. In a discrete case, we are dealing with $\mathcal{H}_0^h(\nabla\times,\Omega)$, which is the space of first-order edge elements. An important property of this space is that a Hodge decomposition similar to (3.18) exists [see 28]. The space $\mathcal{H}_0^1(\Omega)$ has to be replaced with $\mathcal{H}_0^{1,h}(\Omega)$ – the space spanned by first-order nodal shape functions on the same mesh.

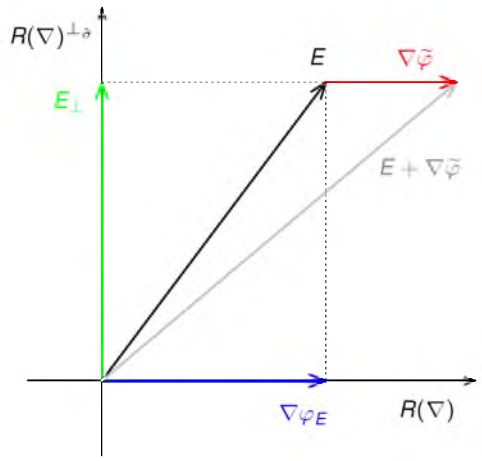

Figure 3.2. Hodge decomposition of the solution E, together with the added error of the form $\nabla\tilde{\varphi}$.

The correction is applied as follows. Let E be an approximation of the electric field given by (3.7). Solve Poisson equation for $\nabla\varphi_{\text{corr}} \in \mathcal{H}_0^{1,h}(\Omega)$, $\forall\varphi \in \mathcal{H}_0^{1,h}(\Omega)$:

$$\mathrm{i}\omega \int_\Omega \hat{\sigma}\nabla\varphi_{\text{corr}} \cdot \nabla\varphi = \int_\Omega (E - J^{\text{imp}})\cdot\nabla\varphi \tag{3.22}$$

The corrected electric field E_{corr} is

$$E_{\text{corr}} = E - \nabla\varphi_{\text{corr}} \tag{3.23}$$

The correction may be given further justification by considering the second equation in (3.21). Using the fact that $E_\perp = E - \nabla\varphi_E \in R(\nabla)^{\perp_{\hat{\sigma}}}$ and integrating by parts, we obtain:

$$\mathrm{i}\omega \int_\Omega (\nabla \cdot (\hat{\sigma}E))\varphi = \int_\Omega (\nabla \cdot J^{\text{imp}})\varphi \tag{3.24}$$

Thus we ensure that the divergence of electric current is proper weakly, on average, with $\varphi \in \mathcal{H}_0^{1,h}(\Omega)$ as a weight. The right-hand side of (3.22) may be viewed as excessive divergence of the electric current, which is removed when (3.23) is applied.

Divergence correction requires solving the Poisson equation (3.22), which we do using nodal-based finite elements. The divergence correction system matrix has three times less variables than the original system matrix, and at least four times less non-zeroes. In our experience, the divergence correction requires much less run-time than solving the original system (3.8); factorization phase is at least 8 times faster and solve phase is at least 5 times faster.

3.5 Field and MT response Jacobian

A primary goal in developing the FE simulator is to apply it to nonlinear inversion of MT field data. As described more fully in our companion paper, we examine both model and data space approaches to parameter updates under the Gauss-Newton framework [20]. For defining terms as related to FE simulation, the model space update equation is [e.g., 31, 32]:

$$[J^T B_d J + \lambda B_m](m_{n+1} - m_0) = J^T[d - F(m_n) - J(m_n - m_0)] \tag{3.25}$$

$F(m_k)$ is the MT response at iteration n using our finite element code, d is the vector of N_d observed MT data weighted against their estimated covariance matrix

B_d^{-1}, B_m^{-1} is a model covariance matrix which stabilizes or regularizes the N_m model parameter variations, m_0 is a reference model, and λ is a constant controlling the trade-off between data fit and model parameter stabilization.

Term J is the N_m by N_d matrix of the parameter Jacobian or derivatives [31] which specify the incremental change in the value of an MT response datum (in Z or K) to an incremental change in the value of a subsurface electrical conductivity parameter. First we focus on the derivatives of the secondary fields. There have been numerous ways to express this in the literature [e.g., 33]; here we basically generalize from the 2D approach of [15]. Recalling the interpolation vectors w, consider an entry σ_j of the FE mesh conductivity vector σ. The entry may correspond to a single element or a group of them. The derivative of a field value $w^T\xi^k$ with respect to σ_j may be evaluated as:

$$\begin{aligned}\frac{\partial(w^T\xi^k)}{\partial\sigma_j} = w^T\frac{\partial\xi^k}{\partial\sigma_j} = w^T\frac{\partial(A^{-1}b^k)}{\partial\sigma_j} &= w^T\left[\frac{\partial A^{-1}}{\partial\sigma_j}b^k + A^{-1}\frac{\partial b^k}{\partial\sigma_j}\right] \\ &= w^T\left[(-A^{-1}\frac{\partial A}{\partial\sigma_j}A^{-1})b^k + A^{-1}\frac{\partial b^k}{\partial\sigma_j}\right] \\ &= w^T\left[-A^{-1}\frac{\partial A}{\partial\sigma_j}(A^{-1}b^k) + A^{-1}\frac{\partial b^k}{\partial\sigma_j}\right] \\ &= w^T\left[-A^{-1}\frac{\partial A}{\partial\sigma_j}\xi^k + A^{-1}\frac{\partial b^k}{\partial\sigma_j}\right]\end{aligned}$$

for source polarization k. This reduces to:

$$\frac{\partial(w^T\xi^k)}{\partial\sigma_j} = w^T\left(A^{-1}\left[-\frac{\partial A}{\partial\sigma_j}\xi^k + \frac{\partial b^k}{\partial\sigma_j}\right]\right) \tag{3.26}$$

As written, in order to calculate the derivatives of the field values with respect to all $(\sigma_j)_{j=1}^{N_m}$, one would have to solve one linear equation for each polarization and for each σ_j, and then multiply by the proper w, to obtain the desired derivatives. That yields $2 \cdot N_m$ linear systems to solve, where N_m is the number of inversion voxels.

However, exploiting interchangeability of sources and receivers in reciprocity, (3.26) may be rewritten as

$$\begin{aligned}\frac{\partial(w^T\xi^k)}{\partial\sigma_j} &= \left(w^TA^{-1}\right)\left[-\frac{\partial A}{\partial\sigma_j}\xi^k + \frac{\partial b^k}{\partial\sigma_j}\right] \\ &= \left(A^{-T}w\right)^T\left[-\frac{\partial A}{\partial\sigma_j}\xi^k + \frac{\partial b^k}{\partial\sigma_j}\right]\end{aligned} \tag{3.27}$$

In this form, we solve one linear system for each field component. The method yields $5 \cdot N_{\text{rec}}$ linear systems to solve, where N_{rec} denotes the number of receivers. The matrix A, defined at (3.9), is symmetric, so $A^{-T} = A^{-1}$. To calculate $A^{-1}w$, we are

solving a linear system where the source w, defined at (3.14), (3.15) is distributed on the edges surrounding the receiver location [cf. 15].

The Jacobian for impedance Z and tipper K at each receiver follow by appling the chain rule to the equations for the impedance and tipper elements of (3.13) defined from applying the two source polarizations $k = 1, 2$. The individual impedance element derivatives are listed in [23] and the tipper element derivatives follow by analogy. For inversion implementation, derivatives are converted to be with respect to $\log_{10}$ resistivity [34].

3.6 Direct solver

Several attractive features of direct solutions were listed in the Introduction. Here we investigate the viability of 3D FE modeling and inversion performed on single-chassis, multicore, symmetric multiprocessing (SMP) computers typically used in server applications and which are relatively affordable. We were attracted to this platform at first for direct solution of the model-space, Gauss-Newton parameter step equation, which was parallelized using a matrix tiling approach under OpenMP compiler directives and showed good scalability across an 8-core workstation with 32 GB RAM [35]. Initially this tiling solution was applied also to the banded [4] FE matrix and showed good scalability across a newer 24-core workstation with 512 GB RAM [36]. However, solution time overall was slower than desired, for example, taking over 1 hour per frequency for a mesh with 85, 88, 50 cells in the x, y, z directions, respectively (85x 88y 50z) and two source vectors (i.e., no Jacobian).

Thus we have investigated the MUMPS library [37, 38], as have others [4, 17–19], to factorize $A = LL^T$. Athough this library is written for distributed memory architecture (using MPI), we show that it performs well on SMP machines. The matrix A in (3.8) is complex valued and symmetric (but not hermitian). The main idea is that a permutation of variables P is found and the matrix is replaced with PAP^T (permutation of both columns and rows of the matrix, so that the matrix remains symmetric). Then matrix L is found such that $PAP^T = LL^T$. The last step is the solve phase, i.e., solving the system (3.8), which may be written as

$$PAP^T(P\xi) = Pb \quad \text{or} \quad LL^T\tilde{\xi} = Pb, \quad \text{where} \quad \tilde{\xi} = P\xi$$

Permutation P is chosen to minimize the number of non-zero values in L matrix, and to allow for parallelization of the factorization and solve phases.

MUMPS allows use of third party ordering libraries; recommended choices are METIS [39] and SCOTCH [40]. We tested both together with their parallel versions ParMETIS and PT-SCOTCH. In the end, we have selected METIS to be most suitable based on the experience below.

A test problem for a hexahedral mesh with 53x 53y 38z cells was chosen, giving a system matrix A of $\sim 307,000$ variables. The test machine had 8 cores (two Intel X5355 Clovertown quad-core processors at 2.66 GHz) and the results are presented in Table 3.1. The analysis time is that needed to calculate the ordering matrix P, while factorization time is that needed to calculate L such that $PAP^T = LL^T$. Solve time is the time needed to solve (3.8) for 500 different vectors b, which would be needed to calculate the Jacobian matrix J in an MT problem with 100 receivers.

In this test, the parallel versions of the ordering libraries calculate matrix P faster, yet the ordering produced is not the same. In particular, ParMETIS gives an ordering with 30% more non-zeroes than METIS. The factorization time is nearly three times slower and the solve time is 1.5 times slower. Also one needs almost twice the memory to store L with ParMETIS, so we have stopped considering it. To chose among the remaining orderings, we consider a similar test on a 24-core workstation (four Intel E5-4610 Sandy Bridge hex-core processors at 2.4 GHz), the results of which appear in Table 3.2.

When 24 cores were used, the PT-SCOTCH ordering experienced some inconsistency. The number of non-zeroes is slightly larger than for SCOTCH and the factorization time is 50% bigger. Also the solve time is 30% greater. Our tests show that the parallel versions of orderings SCOTCH and METIS calculated ordering faster, but the ordering was of poorer quality. Comparing METIS and SCOTCH, because METIS has a shorter analysis time while the factorization and solve times are similar, we settle on the METIS library. However, results of different tests on other machines might vary.

Analysis time appears independent of the number of cores used. Additionally, the factorization and solve times increase more rapidly than the analysis time as

Table 3.1. Ordering library tests for MUMPS using both METIS and SCOTCH, and their parallel correspondants, on an 8-core workstation

ordering library	analysis time[s]	factorization time[s]	solve time for 500 rhs[s]	RAM memory used [GB]	number of non-zeroes in L
METIS	7.7	96.4	192.2	6.729	194,581,000
ParMETIS	3.2	283.0	299.0	11.898	252,859,250
SCOTCH	17.7	101.7	204.8	6.889	197,632,878
PT-SCOTCH	3.3	106.5	194.0	6.712	193,753,609

Table 3.2. Ordering library tests for MUMPS using METIS, SCOTCH, and PT-SCOTCH on a 24-core workstation

ordering library	analysis time[s]	factorization time[s]	solve time for 500 rhs[s]	RAM memory used [GB]	number of non-zeroes in L
METIS	7.6	20.5	34.9	8.181	192,585,549
SCOTCH	15.1	19.0	36.0	8.109	197,632,878
PT-SCOTCH	2.2	33.0	46.1	11.182	219,518,561

the problem gets larger. For example, for a model with mesh 85x 88y 46z, which gives around 1mln of columns in the matrix A, the analysis time, factorization time, and solve time for METIS are 20.5s, 180.8s, and 386s, respectively. Moreover, the solution phase time increases relative to factorization. Bigger problems also may mean more receivers in the survey, so more solution vectors (3.8) need to be reduced for a Gauss-Newton step.

Performance of MUMPS is increased considerably by using a BLAS library optimized for a given machine. Using the Intel MKL library provided a speedup versus a not optimized BLAS library of $2\times$. Additionally, the Intel library offered an option of using multithreaded functions for BLAS operations. Thus one can run MUMPS with some number of MPI processes and each process will use some number of threads. For example, on a 24-core machine running MUMPS with one MPI process and 24 threads, all parallelism is done by the BLAS library. If one uses 24 MPI processes and each process uses one thread, all the parallelism is done by MUMPS.

A summary of execution times with different numbers of processes and threads for METIS ordering on our 24-core workstation is presented in Table 3.3 using the mesh 85x 88y 46z. For factorization only, one sees that the best configuration is 6 MPI processes and 4 threads each. If solve time is considered, the fastest option is 24 MPI processes and 1 thread for each process. MUMPS requires more memory as more MPI processes are used.

MUMPS parallel speedup is about $7\times$ with 24 cores (see Figure 3.3). However not much is gained beyond 12 cores with a speedup of 6x. We can only speculate that memory or bus speeds have become limiting factors [cf. 13]. Nevertheless, we have found that for large models, MUMPS is at least $20\times$ and $5\times$ faster than our earlier tiling approach for factorization and solve phases, respectively. MUMPS also uses about two times less memory, allowing us to consider larger models with a given RAM.

3.7 Example forward calculations

The accuracy of our 3D FE forward code is tested against independent algorithms. These include a standard test of conductive and resistive heterogeneity under a flat

Table 3.3. Dependence of execution time of MUMPS with METIS ordering on the number of threads in BLAS, run on a 24-core workstation, for 85x 88y 46z mesh.

number of MPI processes	number of threads	analysis time[s]	factorization time[s]	solve time for 500 rhs[s]	RAM memory used [GB]	number of non-zeroes in L
24	1	20.5	180.8	176.4	40.5	998,715,481
12	2	19.9	154.5	191.5	38.5	998,715,481
6	4	19.3	150.5	267.5	33.3	998,715,481
2	12	19.4	186.5	422.8	28.3	998,715,481
1	24	19.3	208.2	886.2	22.4	1,092,567,331

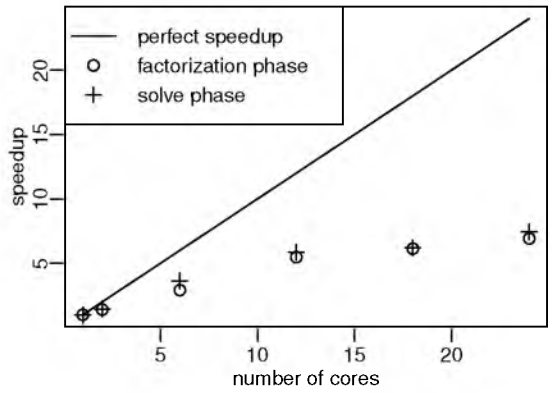

Figure 3.3. Speedup of MUMPS. Varying number of MPI processes, 1 thread per process.

surface, but we also focus on topography as a principal rationale for this work. Tests highlight the strength of the FE method in defining smooth, nonjagged topographic slopes.

3.7.1 Outcropping double brick model

First we consider the popular, outcropping double brick 3D model originally proposed in 2D by [41, 42] and included in the Commemi collection of trial models by [43]. The central portion of its finite element mesh appears in Figure 3.4. The mesh has 52x 53y 31z elements, out of which the two bodies consist of 20x 21y 8z elements. We used 10 layers for the air and 21 layers for the earth. Element sizes grow steadily away from the center of the domain to a total distance of 555km from the center. All calculations are done in double precision.

Complex tensor impedance Z and tipper K elements were calculated at the surface, over the cells centers, along a profile at $y = 16$km for frequencies of 0.001Hz and 0.1Hz. They are compared in Figure 3.5 with those computed using the Integral Equations code of [44], for which the body discretization coincides with that of the FE mesh. The agreement between the two codes clearly is very good, and compares favorably with the check against a finite difference approach in [45]. Comparison was similarly good for the profile at $y = 0$km (not shown) although Z_{xx}, Z_{yy}, K_{zy} are zero there.

The requirement for, and effectiveness of, the divergence correction described previously is demonstrated for a profile 2 km in the air over the center of the double brick model in Figure 3.6. On the left is the electric field in the x-direction across the sides of the body at the low frequency of 0.001 Hz. It consists mainly of numerical noise due to spurious curl-free electric fields. Nevertheless, as seen on the right side, the divergence correction is able to remove the error leaving a response which is a smooth, upward-continued version of a surface response (cf. Z_{xy} in Figure 3.5). Thus we are able to model accurate E-fields in the air with our FE method as would be desired under efforts to create airborne MT platforms [e.g., 46].

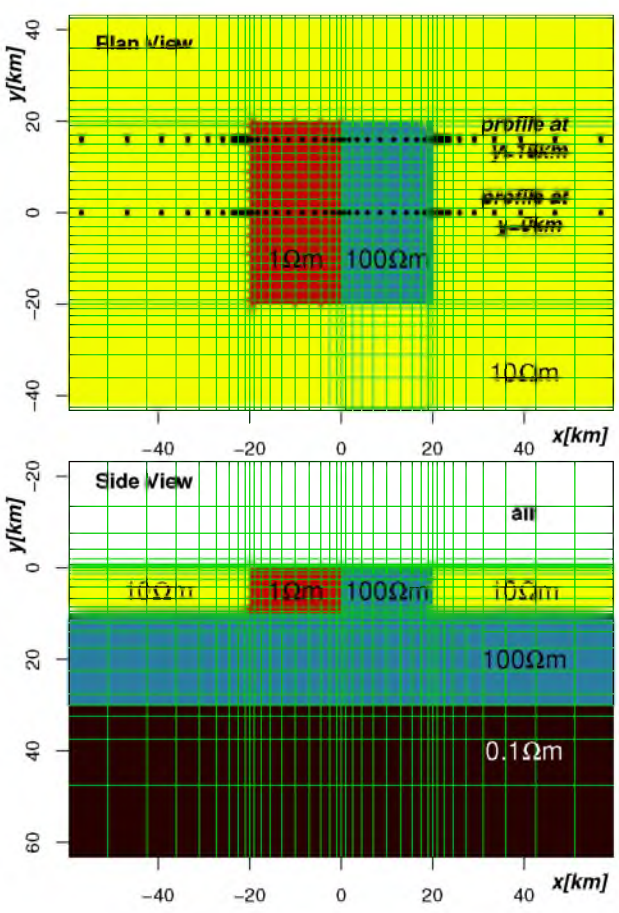

Figure 3.4. Outcropping double brick resistivity model, together with the mesh used. Element boundaries are drawn as solid green lines.

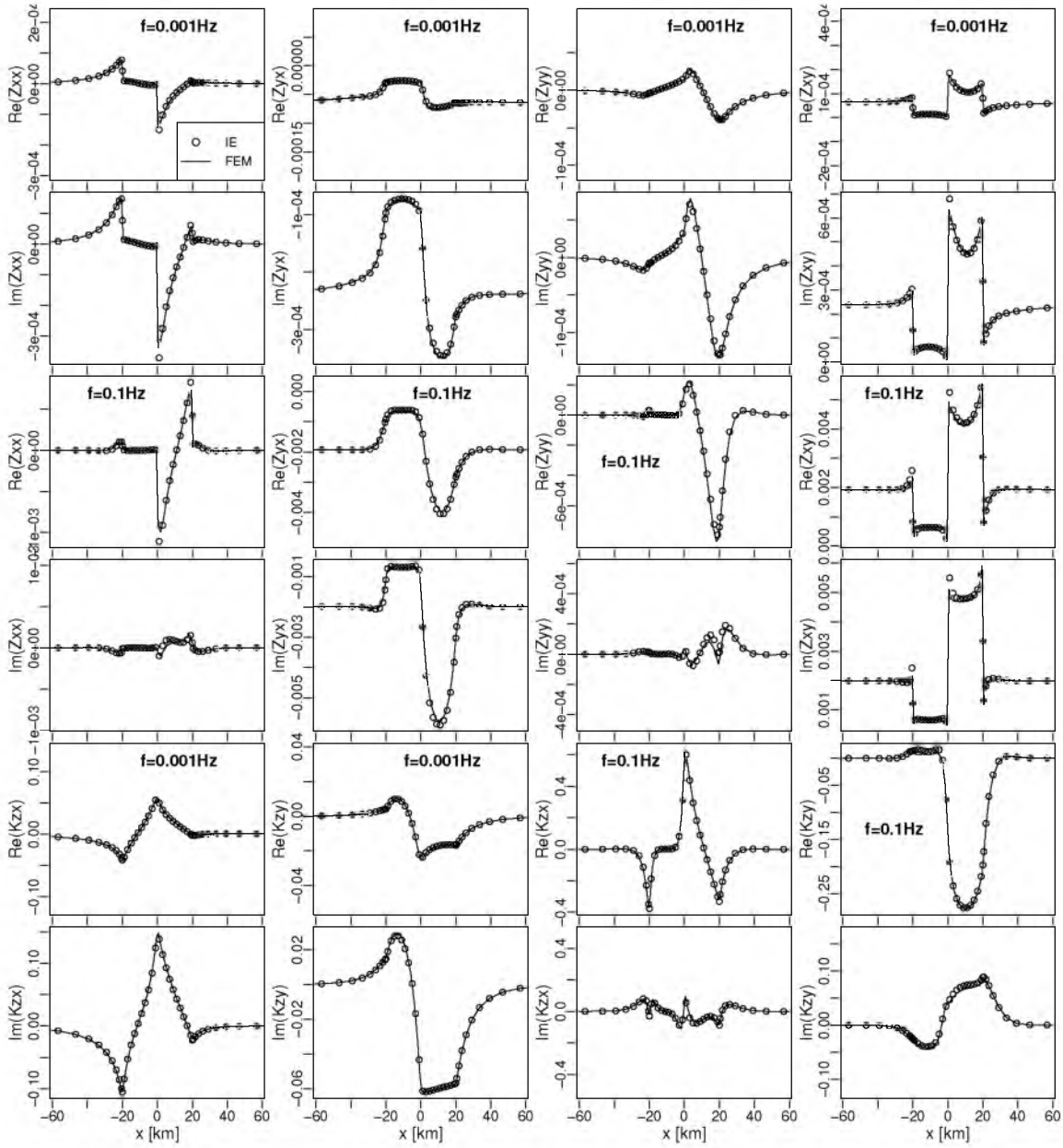

Figure 3.5. Forward MT response of a double brick model for profile at $y = 16$km for comparison with Integral Equation code response. Frequencies are 0.001Hz and 0.1Hz

3.7.2 2D valley and hill

Because topographic simulation and inversion is a principal motivation for this work, we present several accuracy checks here. First, we compare fields over an elongate 3D valley with those of the 2D valley model of [47] computed with their nodal FE code. The valley is 450m deep, 500m wide at the bottom, and 3km wide at the top in a host of resistivity $\rho = 100\Omega$m (3D cross-section in Figure 3.7). In 3D, infinite strike is approximated with a 30km length (Figure 3.8). The entire 3D mesh consisted of 39x 41y 30z elements while the valley portion was covered by 21 elements across the y direction. The mesh extended to 6km above the ground, 11km below the ground and laterally 26km and 14km from the valley in x and y directions, respectively.

A coarse and a finer 2D discretization are considered. The coarse valley is made up of 20 layers of elements each 22.5m thick. The finer valley is made up of 40 layers of elements each 11.25m thick. Element dimensions grow steadily away from the center to a total distance of >20 km to the sides and depth. Note that the 2D mesh is rectilinear such that slopes must be made up of triangles rather than deformed quadrilaterals [47]. The E- and H-fields across the valley center normalized by the primary fields are plotted in Figure 3.9. The responses of the 3D and 2D codes are in close agreement.

For the hill model, we consider the high frequency of 1000Hz to test whether the 3D code can accurately simulate refraction of the EM fields normal to the slope, as was done in 2D in [47]. The small skin depth ($\sim$ 160m) requires a finer mesh closer to the hill surface, although the lateral limits do not have to be so far (Figure 3.10). The mesh extends 3km above the highest point, 1.3km below the base and laterally 5km and 4km from the hill in x and y directions. The 2D hill has the same dimensions and discretization as the valley for both coarse and fine versions. We compute the E- and H- fields parallel to the slope of the hill, and use those values to calculate apparent resistivity ρ_a.

The high-frequency ρ_a should approach the true resistivity of the ground (100Ωm) at high frequency because the total EM fields ideally become purely parallel to the slope. We present ρ_a and phase for TE and TM modes in Figure 3.11. Note that

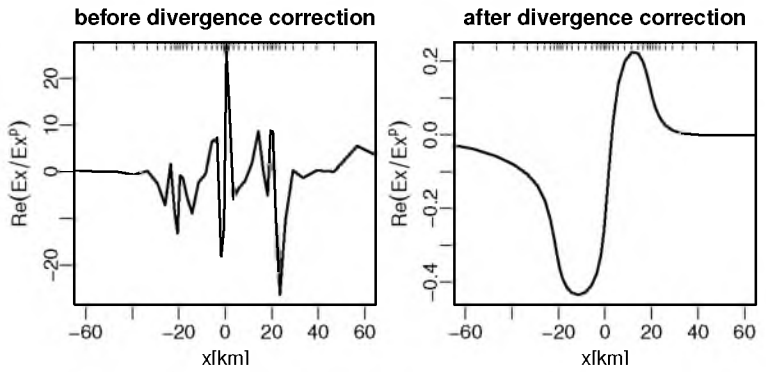

Figure 3.6. Real component of electric field E_x at a height of 2 km for the double brick model calculated before and after divergence correction for a frequency of 0.001 Hz. Y-axis scales are different for figure on the right and on the left; the field before correction is 10× larger overall in magnitude. Tick marks along top plot border show calculation point locations.

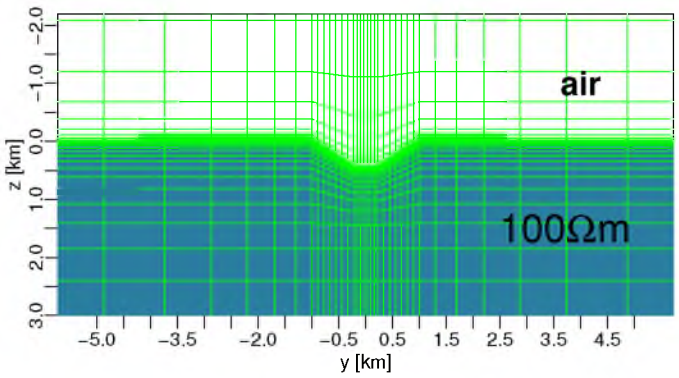

Figure 3.7. YZ cross-section of a 2D valley, together with the central part of the 3D FEM mesh. Element boundaries are drawn as solid green lines.

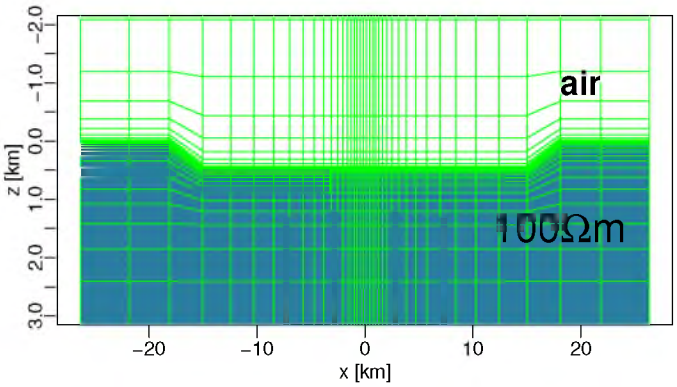

Figure 3.8. XZ cross-section of the 2D valley, together with the 3D FEM mesh. 5× vertical exaggeration.

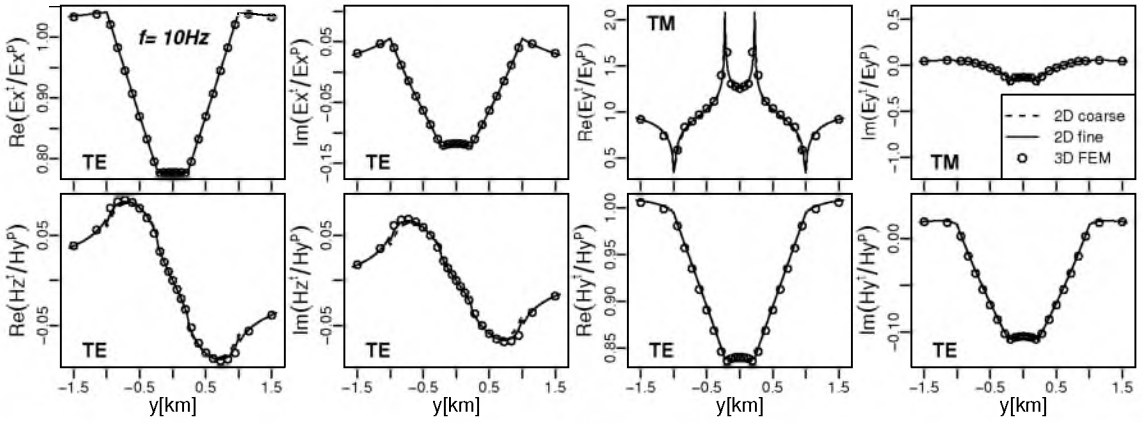

Figure 3.9. Normalized EM fields along the profile across the 2D valley, for $x = 0$km.

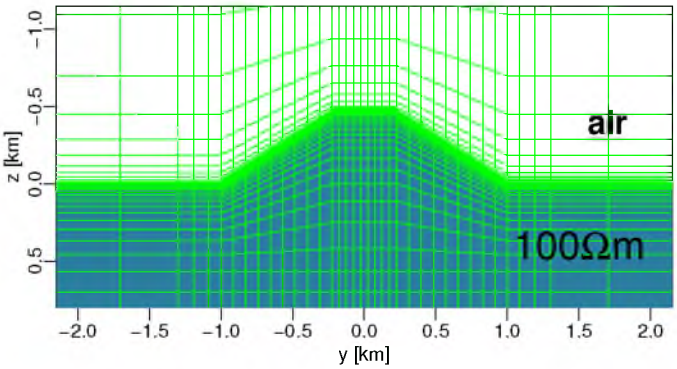

Figure 3.10. YZ cross-section of a 2D hill, together with the central part of 3D FEM mesh.

for both fine mesh 2D code and 3D code results, away from breaks in slope, ρ_a is close to 100Ωm and phase is close to 45°. In fact, the 3D phase results look the most accurate, which may reflect a greater ease for layers of hexahedral elements to simulate essentially 1D fields than for triangles, although the 2D results are converging with finer discretization.

3.7.3 3D trapezoidal hill

The final test model is the 3D hill model considered in [21]. It has the same dimensions as the previous 2D hill, but is square in cross-section (see Figure 3.12). It is 0.45km high, 0.5km at the hilltop, 2km wide at the base with resistivity of 100Ωm. It is calculated for 2Hz, and the MT response is compared to that of [21]. Two grids were considered, the finer grid being 97x, 97y, and 50z while the coarser grid is 27x, 27y, 24z. The ρ_a and phase along a profile across the center of the hill is presented in Figure 3.13. The MT response calculated in [21] and the field calculated by our FE code appear very similar.

3.7.4 Jacobian test calculations

Next, we test the calculation of MT response Jacobian as they are essential for inversion purposes. We consider derivatives with respect to $\log_{10}$ resistivity model $m = (m_j)_{j=1}^{N_m}$, where in principle, each m_j could be parsed as finely as a single finite element. We consider the coarse 3D hill mesh with receivers over the centers of surface element faces at $y = 0$km. For the test parameter, we use two adjacent finite elements on the facing hill slope (Figure 3.14). The Jacobian is calculated using reciprocity as described previously and it is compared with a symmetric difference approximation of the derivative, i.e.,

$$\frac{\partial(Z,K)}{\partial m_j}(m) \approx \frac{(Z,K)(m+e_j h) - (Z,K)(m-e_j h)}{2h}$$

where e_j is a vector with only one non-zero entry at the jth position, which is equal to 1. In Figure 3.15, we present the result of calculation for frequencies of 100Hz and 0.001Hz for Zxy, Zyx, Kzy and for the inversion voxel marked in the model figure. We used $h = 0.05$. A wide range of frequencies and various locations of the

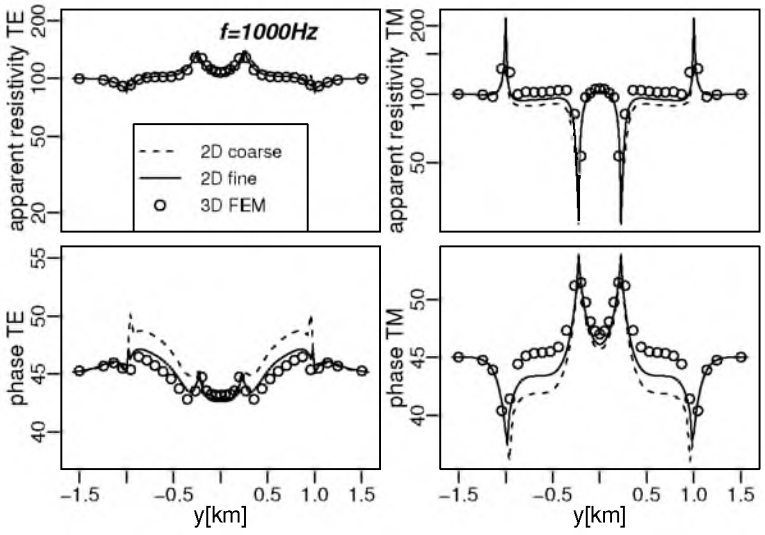

Figure 3.11. The ρ_a and phase responses from the 2D code and 3D codes for the 2D hill at 1000 Hz.

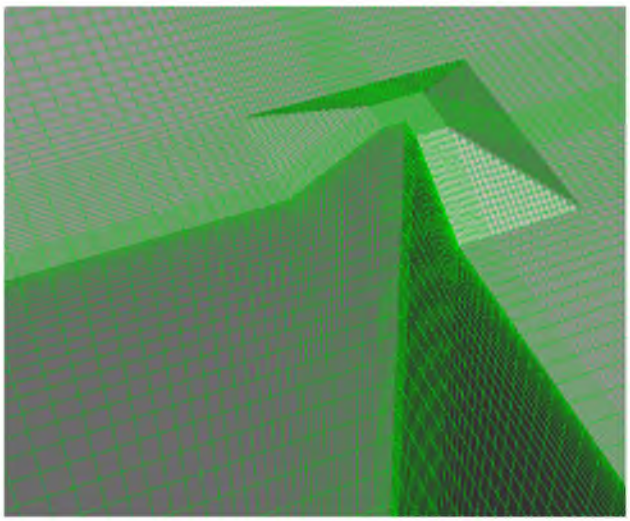

Figure 3.12. Central part of the finer mesh for the 3D hill model.

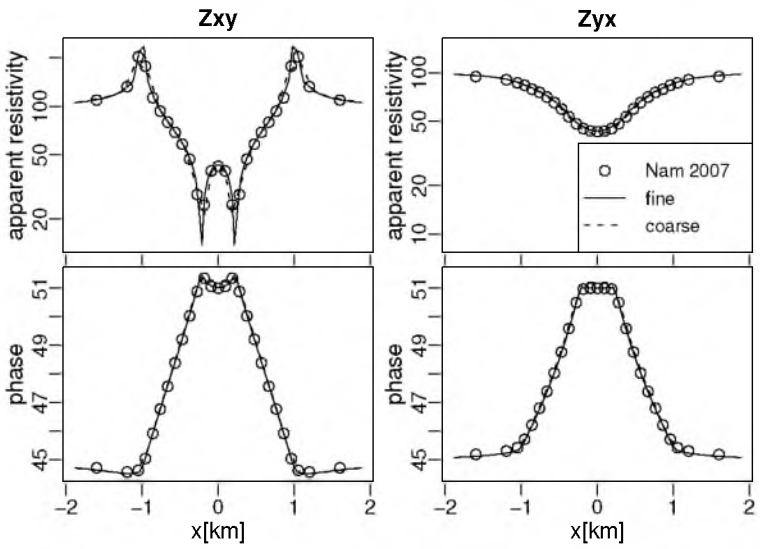

Figure 3.13. MT responses of the 3D hill model along a profile across the hill compared with the result of [21]. The results of Nam have been discretized from their plots.

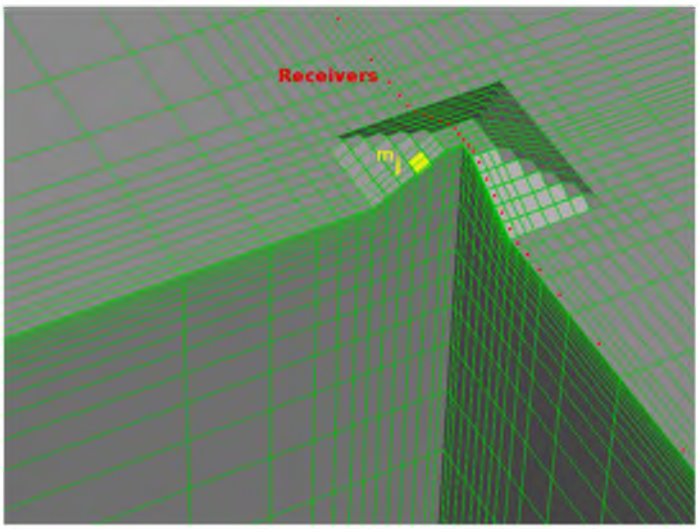

Figure 3.14. Central part of the coarse mesh of 3D hill, together with the location of receivers and the chosen inversion voxel m_j

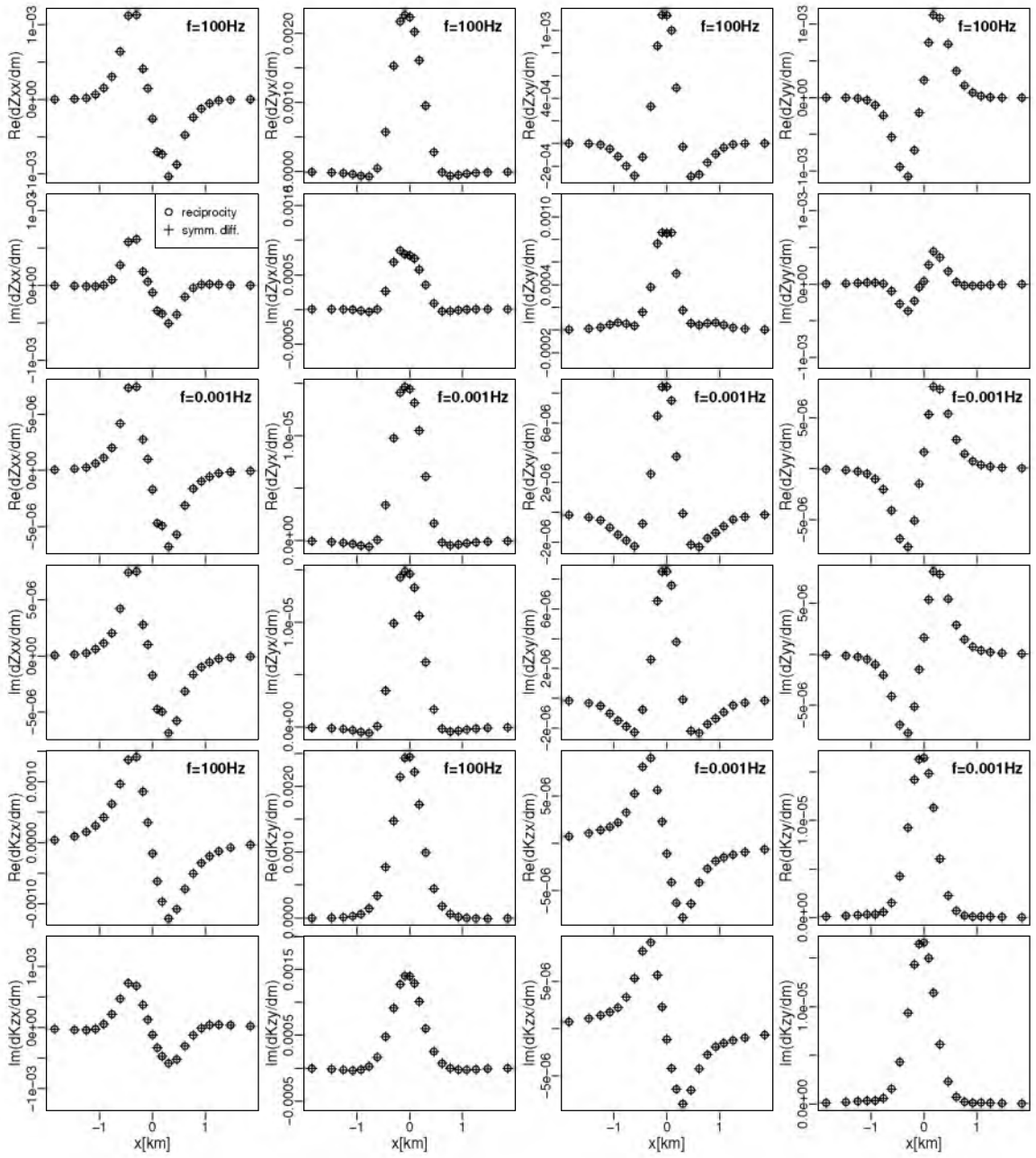

Figure 3.15. Comparison of a Jacobian of Z, K calculated using reciprocity and symmetric difference

inversion voxel have been tried and in all cases the values of the Jacobian showed precise agreement.

3.8 Example run times

In Tables 3.4 and 3.5, we present run times related to solving the equation (3.8) with the MUMPS library. Recall that MUMPS finds a permutation matrix P (analysis phase), then calculates matrix L such that $P^T AP = LL^T$ (factorization phase). MUMPS then uses L to solve a linear system (3.8) for numerous rhs vectors b. Times in the tables correspond to work done for a single frequency. In order to calculate full MT Jacobian, for each receiver location, one needs to solve 5 linear systems (3.8). For example, 500rhs in Table 3.5 would correspond to a survey with 100 receivers. As expected, run time increase is geometric with respect to number of unknowns. With data space parameter step formulation, as discussed in Part II, the inversion run time will be dominated by the forward problem and Jacobian. For the largest test mesh and assuming each element can be a parameter, a 400 site survey (20 x 20) could be inverted using a mesh with five columns of parameters per site in both x and y directions, leaving >20 columns of padding to far distances outside the survey domain.

3.9 Conclusions

Finite elements provide a flexible and accurate means of simulating EM responses of 3D resistivity structure beneath topographic variations. Hexahedral elements provide a straightforward means of representing earth surface slopes, are compatible with the Helmholtz governing equation as discretization increases, and generate FE system matrices of simple structure. In particular, discretization requirements for topography at high frequencies are modest compared to those for traditional rectilinear meshes because layers of elements can lie parallel to the earths surface. By invoking an efficient current divergence correction, accurate E-field results may be obtained at very low frequencies and small admittivities, even those of dielectric air. Because we utilize a secondary field approach, it should be straightforward to generalize to finite source problems. As will be shown in Part II, hexahedral elements also provide a simple path to regularized inversion, for example, by direct mapping

Table 3.4. MUMPS analysis and factorization phase times for factoring matrix A in (3.8) for various meshes (in hr:min:sec).

mesh	number of unknowns in A	number of non-zeroes in A	number of non-zeroes in L	analysis time	factorization time	RAM memory used [GB]
30x 30y 25z	62,785	1,008,965	21,614,954	00 : 00 : 01	00 : 00 : 02	1.27
50x 50y 35z	250,635	4,111,215	146,622,950	00 : 00 : 04	00 : 00 : 13	6.9
75x 75y 45z	734,820	12,179,490	647,221,887	00 : 00 : 14	00 : 01 : 55	29
100x 100y 50y	1,460,250	24,316,190	1,679,493,542	00 : 00 : 32	00 : 06 : 08	70
125x 125y 55z	2,519,680	42,085,390	3,487,912,888	00 : 00 : 59	00 : 17 : 12	147
150x 150y 60z	3,969,360	66,443,340	6,643,228,266	00 : 01 : 36	00 : 53 : 10	273

Table 3.5. MUMPS solution phase time for the linear system (3.8), for various meshes and numbers of rhs vectors b.

mesh	100rhs	500rhs	1000rhs	1500rhs	2000rhs
30x 30y 25z	00 : 00 : 01	00 : 00 : 07	00 : 00 : 14	00 : 00 : 21	00 : 00 : 28
50x 50y 35z	00 : 00 : 06	00 : 00 : 32	00 : 01 : 04	00 : 01 : 36	00 : 02 : 08
75x 75y 45z	00 : 00 : 27	00 : 02 : 16	00 : 04 : 32	00 : 06 : 48	00 : 09 : 04
100x 100y 50y	00 : 01 : 16	00 : 06 : 21	00 : 12 : 43	00 : 19 : 04	00 : 25 : 26
125x 125y 55z	00 : 02 : 36	00 : 13 : 01	00 : 26 : 01	00 : 39 : 02	00 : 52 : 03
150x 150y 60z	00 : 04 : 36	00 : 23 : 02	00 : 46 : 04	01 : 09 : 06	01 : 32 : 08

of triaxial parameter roughness damping into deformed coordinates.

Efficient and affordable parallel computing solutions have emerged that are putting direct solutions to fairly large 3D EM simulation problems within reach of an increasing number of users. These include a powerful public-domain library for direct solutions (MUMPS) that is seeing increased community use. Because the factorization provided by direct solutions allows economical solution of large numbers of source vectors, explicit and accurate values of the parameter Jacobian can be obtained. Technological advances also include single-box, server-class workstations with numerous cores and substantial RAM that provide relatively affordable computing. Parallelization of the direct solver MUMPS on multicore SMP computers was moderately good but reached communication limits beyond ~ 12 cores. Further improvements might be achieved with faster memory performace. Parallelization also could be increased with distributed computing using multiple SMP. Nevertheless, direct simulations including the Jacobian can be done on a single workstation for meshes with one million elements in 1-2 hours.

3.10 Acknowledgements

We acknowledge the support of this work from the U.S. Dept. of Energy under contract DE-EE0002750 to PW. EC acknowledges the partial support of the U.S. National Science Foundation through grants ARC-0934721 and DMS-1413454. The University of Utah/EGI funded acquisition of the 24-core workstation.

3.11 Appendix A

To explain the ill-conditioning related to spurious curl-free E-fields, let us analyze eigenvalues of the system matrix A. A good approximation of those eigenvalues are eigenvalues of the operator

$$\mathcal{L}(F) = \nabla \times \left(\frac{1}{\mu} \nabla \times F \right) + \mathrm{i}\omega\hat{\sigma} F \tag{3.28}$$

$\mathcal{L}$ should be defined on some suitable finite dimensional space, dependent on the mesh size h. First let us consider infinite dimensional space of vector fields $F \in \mathcal{H}_0(\nabla\times, \Omega)$, with the additional assumption that $\nabla \times \frac{1}{\mu} \nabla \times F$ exists and is square integrable. Let

the domain be a cube $\Omega = [0, M]^3$ with $\hat{\sigma}, \mu = \text{const}$. It is straightforward to verify that the eigenvectors of $\mathcal{L}$ are of the form:

$$\begin{bmatrix} C_x \cos\left(\pi \frac{k_x}{M} x\right) \sin\left(\pi \frac{k_y}{M} y\right) \sin\left(\pi \frac{k_z}{M} z\right) \\ C_y \sin\left(\pi \frac{k_x}{M} x\right) \cos\left(\pi \frac{k_y}{M} y\right) \sin\left(\pi \frac{k_z}{M} z\right) \\ C_z \sin\left(\pi \frac{k_x}{M} x\right) \sin\left(\pi \frac{k_y}{M} y\right) \cos\left(\pi \frac{k_z}{M} z\right) \end{bmatrix} \tag{3.29}$$

for $k_x, k_y, k_z \in \mathbb{N}$ where

$$\begin{bmatrix} C_x \\ C_y \\ C_z \end{bmatrix} = \begin{bmatrix} k_x \\ k_y \\ k_z \end{bmatrix}, \quad \begin{bmatrix} -k_z \\ 0 \\ k_x \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} 0 \\ k_z \\ -k_y \end{bmatrix} \tag{3.30}$$

The corresponding eigenvalues are:

$$\lambda = \quad \mathrm{i}\omega\hat{\sigma}, \quad \frac{\pi^2|k|^2}{\mu M^2} + \mathrm{i}\omega\hat{\sigma}, \quad \frac{\pi^2|k|^2}{\mu M^2} + \mathrm{i}\omega\hat{\sigma} \tag{3.31}$$

For an infinite dimensional space, we have $3 < |k|^2 < \infty$, yet for a discretization with spatial mesh parameter h, it would be:

$$3 < |k|^2 < O\left(\left(\frac{M}{h}\right)^2\right) \tag{3.32}$$

Let us look at those eigenvalues for some practical setting for magnetotellurics. Let $M = 10\text{km}$, $\frac{M}{h} = 50$. The quantity $\frac{\pi^2|k|^2}{\mu M^2}$ is in the interval $[2{\cdot}10^{-1}, 2{\cdot}10^{-2}]$. The values of the first eigenvalue $i\omega\hat{\sigma}$ for conductivity corresponding to the earth subsurface ($\hat{\sigma} = 0.01\frac{\text{S}}{\text{m}}$) and air ($\hat{\sigma} = i\omega\epsilon_0$) are presented in Table 3.6. The corresponding condition numbers of the system matrix, defined as $\text{cond}(A) = \frac{\max(|\lambda|)}{\min(|\lambda|)}$, are presented in Table 3.7. One can see that if a conductivity corresponding to the earth is used, the condition number increases as the frequency decreases, yet remains at a reasonable level of 10^{+5} even for frequency as small as 0.01Hz. If conductivity of the air is used, the situation is different. As the frequency decreases, the condition number increases quadratically and reaches very large value of of 10^{+18} for the frequency 0.01Hz.

In magnetotellurics, the domain contains both earth and air. Because of the presence of the air, the matrix is ill-conditioned. Nevertheless, the calculated electric field, approximated by solving (3.8) using a direct solver, and using (3.7) has improper values only in the air. The electric field below the earth's surface does not suffer from

Table 3.6. Values of $|\omega\hat{\sigma}|$ for different σ and ω. Unit of $|\omega\hat{\sigma}|$ is $\frac{\mathrm{S\,Hz}}{\mathrm{m}}$

ω	$2\pi\,100\mathrm{Hz}$	$2\pi\,1\mathrm{Hz}$	$2\pi\,0.01\mathrm{Hz}$
earth: $\hat{\sigma} = 0.01\frac{\mathrm{S}}{\mathrm{m}}$	6.3	$6.3 \cdot 10^{-2}$	$6.3 \cdot 10^{-4}$
air: $\hat{\sigma} = \mathrm{i}\omega\epsilon_0$	$5.6 \cdot 10^{-5}$	$5.6 \cdot 10^{-11}$	$5.6 \cdot 10^{-17}$

Table 3.7. Condition number of the system matrix A as a function of frequency ω and $\hat{\sigma}$

ω	$2\pi\,100\mathrm{Hz}$	$2\pi\,1\mathrm{Hz}$	$2\pi\,0.01\mathrm{Hz}$
earth: $\hat{\sigma} = 0.01\frac{\mathrm{S}}{\mathrm{m}}$	$3.1 \cdot 10^{+1}$	$3.1 \cdot 10^{+3}$	$3.1 \cdot 10^{+5}$
air: $\hat{\sigma} = \mathrm{i}\omega^2\epsilon_0$	$3.5 \cdot 10^{+6}$	$3.5 \cdot 10^{+12}$	$3.5 \cdot 10^{+18}$

numerical instability. It is also worth mentioning, that the magnetic field, calculated using the curl of electric field as in (3.11), has proper values in all of the domain. It shows that the error added to the electric field in the air is curl-free.

The condition number of the matrix gets this large because of the smallness of the eigenvalue $\lambda = \mathrm{i}\omega\hat{\sigma}$, corresponding to the first eigenvector in (3.30). Notice that this eigenvector is curl-free, whereas the other two are not. Moreover, the first eigenvector is equal to $\nabla\varphi$, where

$$\varphi = \sin\left(\pi\frac{k_x}{M}x\right)\sin\left(\pi\frac{k_y}{M}y\right)\sin\left(\pi\frac{k_z}{M}z\right) \tag{3.33}$$

and $\varphi|_{\partial\Omega} = 0$. This is not a coincidence.

3.12 Appendix B

Theorem 10 (Hodge decomposition for $\mathcal{H}_0(\nabla\times,\Omega)$) *If*

$$F \in \mathcal{H}_0(\nabla\times,\Omega) \tag{3.34}$$

$$Re(\hat{\sigma}) \geq 0, \quad Im(\hat{\sigma}) \geq 0, \quad 0 < \sigma_m \leq |\hat{\sigma}| \leq \sigma_M \tag{3.35}$$

then there exist unique $\varphi_F \in \mathcal{H}_0^1(\Omega)$, $F_\perp \in R(\nabla)^{\perp_{\hat{\sigma}}}$ *such that*

$$F = \nabla\varphi_F + F_\perp \tag{3.36}$$

Proof : φ_F, $F_\perp$ satisfy (3.36) if and only if $\varphi_F \in \mathcal{H}_0^1(\Omega)$ satisfies

$$\int_\Omega \hat{\sigma}\nabla\varphi_F\nabla\varphi = \int_\Omega F\nabla\varphi \qquad \forall\varphi \in \mathcal{H}_0^1(\Omega) \tag{3.37}$$

and $F_\perp$ is defined as $F_\perp = F - \varphi_F$.

Existence and uniqueness of solution to (3.37) follows from the Lax-Milgram theorem if only F is square integrable (which is true for $F \in \mathcal{H}_0(\nabla\times)$) and the left-hand side of (3.37) is a bounded, coercive bilinear form. It is true, because of Poincare inequality for $\mathcal{H}_0^1(\Omega)$ if only $\hat{\sigma}$ is in the first quadrant of the complex plane $\mathbb{C}$ and is bounded from 0 and ∞. This is what we assume in (3.35).

Remark 11 *The important conclusion is that the decomposition exists in our situation, when the domain* Ω *includes both the air and the earth's subsurface. In the earth* $\hat{\sigma} = \sigma + \mathrm{i}\omega\epsilon_0$ *and in the air* $\hat{\sigma} = \mathrm{i}\omega\epsilon_0$. *If we assume that* σ *is bounded, assumption (3.35) is satisfied as* $\sigma \geq 0$ *and* $\epsilon_0 > 0$.

3.13 References

[1] R.-U. Boerner, "Numerical modelling in geo-electromagnetics: advances and challenges," *Surv. Geophys.*, vol. 31, pp. 225–245, 2010.

[2] M. E. Everett, "Theoretical developments in electromagnetic induction geophysics with selected applications in the near-surface," *Surv. Geophys.*, vol. 33, pp. 29–63, 2012.

[3] J. Liu, M. Brio, and J. Moloney, "Overlapping Yee FDTD method on nonorthogonal grids," English, *J. Sci. Comput.*, vol. 39, no. 1, pp. 129–143, 2009, ISSN: 0885-7474. DOI: 10.1007/s10915-008-9253-1. [Online]. Available: http://dx.doi.org/10.1007/s10915-008-9253-1.

[4] N. V. daSilva, J. V. Morgan, L. MacGregor, and M. Warner, "A finite element multi-frontal method for 3D CSEM modeling in the frequency domain," *Geophysics*, vol. 77, E101–E115, 2012.

[5] P. G. Lelievre and C. G. Farquharson, "Gradient and smoothness regularization operators for geophysical inversion on unstructured meshes," *Geophys. J. Int.*, vol. 195, pp. 330–341, 2013.

[6] C. Schwarzbach and E. Haber, "Finite element based inversion for time-harmonic electromagnetic problems," *Geophys. J. Int.*, vol. 193, pp. 615–634, 2013.

[7] J.-C. Nédélec, "A new family of mixed finite elements in R^3," *Numer. Math.*, vol. 50, no. 1, pp. 57–81, 1986, ISSN: 0029-599X. DOI: 10.1007/BF01389668. [Online]. Available: http://dx.doi.org/10.1007/BF01389668.

[8] E. Haber, D. Oldenburg, and R. Shektman, "Inversion of time-domain three-dimensional data," *Geophys. J. Int.*, vol. 171, pp. 550–564, 2007.

[9] R.-U. Boerner, O. Ernst, and K. Spitzer, "Fast 3D simulations of transient electromagnetic fields by model reduction in the frequency domain using Krylov subspace projection," *Geophys. J. Int.*, vol. 173, pp. 766–780, 2008.

[10] M. Commer and G. A. Newman, "New advances in three-dimensional controlled-source electromagnetic inversion," *Geophys. J. Int.*, vol. 172, pp. 513–535, 2008.

[11] E. S. Um, J. M. Harris, and D. L. Alumbaugh, "An iterative finite element time-domain method for simulating three-dimensional electromagnetic diffusion in the earth," *Geophys. J. Int.*, vol. 190, pp. 871–886, 2012.

[12] D. F. Pridmore, G. W. Hohmann, S. H. Ward, and W. R. Sill, "An investigation of finite element modeling for electrical and electromagnetic data in three dimensions," *Geophysics*, vol. 46, pp. 1009–1024, 1981.

[13] A. V. Grayver, R. Streich, and O. Ritter, "Three-dimensional parallel distributed inversion of CSEM data using a direct forward solver," *Geophys. J. Int.*, vol. 193, pp. 1432–1446, 2013.

[14] C. deGroot Hedlin and S. Constable, "Occams inversion to generate smooth two-dimensional models from magnetotelluric data," *Geophysics*, vol. 55, pp. 1613–1624, 1990.

[15] P. P. deLugao and P. E. Wannamaker, "Calculating the two-dimensional magnetotelluric jacobian in finite elements using reciprocity," *Geophys. J. Int.*, vol. 127, pp. 806–810, 1996.

[16] K. Key and S. Constable, "Coast effect distortion of marine magnetotelluric data: insights from a pilot study offshore northeastern Japan," *Phys. Earth Planet. Inter.*, vol. 184, pp. 194–207, 2011.

[17] R. Streich, "3D finite-difference frequency-domain modeling of controlled-source electromagnetic data: direct solution and optimization for high accuracy," *Geophysics*, vol. 74, F95–F105, 2009.

[18] D. W. Oldenburg, E. Haber, and R. Shekhtman, "Forward modeling and inversion of multi-source TEM data," in *78th Ann. Internat. Mtg, SEG, las Vegas, Exp. Abstr.*, 2008, pp. 559–563.

[19] ——, "Three dimensional inversion of multi-source time domain electromagnetic data," *Geophysics*, vol. 78, no. 1, E47–E57, 2013.

[20] W. Siripunvaraporn, G. Egbert, Y. Lenbury, and M. Uyeshima, "Three-dimensional magnetotelluric inversion: data-space method," *Phys. Earth Planet. Inter.*, vol. 150, pp. 3–14, 2005.

[21] M. J. Nam, H. J. Kim, Y. Song, T. J. Lee, J.-S. Son, and J. H. Suh, "Three-dimensional magnetotelluric modeling including surface topography," *Geophys. Prospect.*, vol. 55, pp. 277–287, 2007.

[22] G. W. Hohmann, "Numerical modeling for electromagnetic methods of geophysics," in *Electromagnetic methods in applied geophysics*, M. N. Nabighian, Ed., Soc. Expl. Geophys., 1988, pp. 313–363.

[23] G. A. Newman and D. L. Alumbaugh, "Three-dimensional magnetotelluric inversion using non-linear conjugate gradients," *Geophys. J. Int*, vol. 140, pp. 410–424, 2000.

[24] J Smith, "Conservative modeling of 3-D electromagnetic fields, Part II: biconjugate gradient solution and an accelerator," *Geophysics*, vol. 61, no. 5, pp. 1319–1324, 1996.

[25] Y. Sasaki, "Full 3-D inversion of electromagnetic data on PC," *J. Appl. Geophys.*, vol. 46, pp. 45–54, 2001.

[26] W. Siripunvaraporn, G. Egbert, and Y. Lenbury, "Numerical accuracy of magnetotelluric modeling: a comparison of finite difference approximations," *Earth Planets Space*, vol. 54, pp. 721–725, 2002.

[27] C. G. Farquharson and M. Meinsopust, "Three-dimensional finite-element modeling of magnetotelluric data with a divergence correction," *J. Appl. Geophys.*, vol. 75, pp. 699–710, 2011.

[28] P. B. Bochev and M. D. Gunzburger, *Least-Squares Finite Element Methods*. Springer New York, 2009.

[29] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier-Stokes Equations: Theory and Algorithms*. Springer Berlin Heidelberg, 1986, ISBN: 978-3-642-61623-5. DOI: 10.1007/978-3-642-61623-5.

[30] R. Adams and J. Fournier, *Sobolev Spaces*. Elsevier Science, 2003, ISBN: 9780080541297. [Online]. Available: http://books.google.com/books?id=R5A65Koh-EoC.

[31] A. Tarantola, *Inverse Problems Theory, Methods for Data Fitting and Model Parameter Estimation*. Elsevier Netherlands, 1987.

[32] R. L. Mackie, B. R. Bennett, and T. R. Madden, "Long-period magnetotelluric measurements near the central California coast: a land-locked view of the conductivity structure under the Pacific Ocean," *Geophys. J. Int.*, vol. 95, pp. 181–194, 1988.

[33] P. R. McGillivray, O. D. W., R. G. Ellis, and T. M. Habashy, "Calculation of sensitivities for the frequency-domain electromagnetic problem," *Geophys. J. Int.*, vol. 116, pp. 1–4, 1994.

[34] G. W. Hohmann and A. P. Raiche, "Inversion of controlled-source electromagnetic data," in *Electromagnetic methods in applied geophysics*, M. N. Nabighian, Ed., Tulsa, OK: Soc. Expl. Geophys., 1988, pp. 469–503.

[35] V. Maris and P. E. Wannamaker, "Parallelizing a 3D finite difference MT inversion algorithm on a multicore PC using OpenMP," *Computers & Geosciences*, doi: 10.1016/j.cageo.2010. 03.001, 5 pp. 2010.

[36] M. Kordy, V. Maris, P. Wannamaker, and E. Cherkaev, "3D edge finite element solution for scattered electric field using a direct solver parallelized on an SMP workstation," in *5th International Symposium on Three-Dimensional Electromagnetics, Sapporo, May 7-9*, 2013, p. 4.

[37] P. R. Amestoy, I. S. Duff, J. Koster, and J.-Y. L'Excellent, "A fully asynchronous multifrontal solver using distributed dynamic scheduling," *SIAM J. Matrix Anal. and Appl.*, vol. 23, no. 1, pp. 15–41, 2001.

[38] P. R. Amestoy, A. Guermouche, J.-Y. L'Excellent, and S. Pralet, "Hybrid scheduling for the parallel solution of linear systems," *Parallel Comput.*, vol. 32, no. 2, pp. 136–156, 2006.

[39] G. Karypis, "Multi-constraint mesh partitioning for contact/impact computations," in *Proc. SC2003, Phoenix, AZ*, ACM, 2003, ISBN: 1-58113-695-1.

[40] C. Chevalier and F. Pellegrini, "PT-SCOTCH: a tool for efficient parallel graph ordering," *Parallel Comput.*, vol. 34, pp. 318–331, 2008.

[41] J. T. Weaver, B. V. LeQuang, and G. Fischer, "A comparison of analytic and numerical results for a two-dimensional control model in electromagnetic

induction - I. B-polarization calculations," *Geophys. J. Roy. Astron. Soc.*, vol. 82, pp. 263–277. 1985.

[42] ——, "A comparison of analytic and numerical results for a two-dimensional control model in electromagnetic induction - II. E-polarization calculations," *Geophys. J. Roy. Astron. Soc.*, vol. 87, pp. 917–948. 1986.

[43] M. S. Zhdanov, I. M. Varentsov, J. T. Weaver, N. G. Golubev, and V. A. Krylov, "Methods for modelling electromagnetic fields: results from COMMEMI - the international project on the comparison of modelling methods for electromagnetic induction," *J. Appl. Geophys.*, vol. 37, pp. 133–271. 1997.

[44] P. E. Wannamaker, "Advances in three-dimensional magnetotelluric modeling using integral equations," *Geophysics*, vol. 56, pp. 1716–1728, 1991.

[45] R. L. Mackie, T. R. Madden, and P. E. Wannamaker, "Three-dimensional magnetotelluric modeling using difference equations – theory and comparisons to integral equation solutions," *Geophysics*, vol. 58, pp. 215–226, 1993.

[46] J. Macnae, "Electric field measurements in air," in *80th Ann. Internat. Mtg, SEG, Denver, Exp. Abstr.*, 2010, pp. 1773–1777.

[47] P. Wannamaker, J. Stodt, and L. Rijo, "Two-dimensional topographic responses in magnetotelluric modeling using finite elements," *Geophysics*, vol. 51, pp. 2131–2144, 1986.

CHAPTER 4

3D MAGNETOTELLURIC INVERSION INCLUDING TOPOGRAPHY USING DEFORMED HEXAHEDRAL EDGE FINITE ELEMENTS AND DIRECT SOLVERS PARALLELIZED ON SMP COMPUTERS, PART II: DIRECT DATA SPACE INVERSE SOLUTION[1]

Kordy M.[23], Wannamaker P.[3], Maris V.[3], Cherkaev E.[2], and Hill G.[4]

4.1 Abstract

Following the creation described in Part I of a deformable edge finite element simulator for 3D magnetotelluric (MT) responses using direct solvers, in Part II, we develop an algorithm named HexMT for 3D regularized inversion of MT data including topography. Direct solvers parallelized on large-RAM, symmetric multiprocessor (SMP) workstations are used also for the Gauss-Newton parameter step estimate. By exploiting the data space approach, the computational cost of the parameter step becomes much less in both time and computer memory than the cost of the forward simulation. In order to regularize using the second norm of the gradient,

---

[1]Submitted to Geophysical Journal International in 2014

[2]Department of Mathematics, University of Utah

[3]Energy & Geoscience, University of Utah

[4]GNS Science, Lower Hutt,Wellington, New Zealand; Australian Crustal Research Centre, Monash University, Melbourne, Australia

we factor the matrix related to the regularization term and apply its inverse to the Jacobian, which is done using the MUMPS library. For dense matrix multiplication and factorization related to the parameter step, we use the PLASMA library which shows very good scalability across processor cores. A synthetic test inversion using a simple hill model shows that including topography can be important; in this case, depression of the electric field by the hill can cause false conductors at depth or mask the presence of resistive structure. With a simple model of two buried bricks, a uniform spatial weighting for the norm of model smoothing recovered more accurate locations for the tomographic images compared to weightings which were a function of parameter Jacobian. We implement joint inversion for static distortion matrices tested using the Dublin secret model 2, for which we are able to reduce nRMS to 1.1 while avoiding oscillatory convergence. Finally, we test the code on field data by inverting full impedance and tipper MT responses collected around Mount St. Helens in the Cascade volcanic chain. Among several prominent structures, the north-south trending, eruption-controlling shear zone is clearly imaged in the inversion.

4.2 Introduction

In Part I [1], we have shown that moderately large 3D magnetotelluric models including topography can be simulated accurately in practical run times using a direct solver on a single-box, server-class multicore workstation with large RAM. The deformable mesh approach allows us to avoid expending many rows of cells to define just the topography as is done with finite differences, and which even then may not escape local electric field distortion [e.g., 2, 3]. The public-domain solver library MUMPS is effective on this platform, showing moderately good scalability across at least 12 cores that appears better than scalability for distributed clusters [cf. 4]. For a mesh of $150x$ $150y$ $60z$ elements, 2000 source vectors (corresponding to 400 MT sites) could be solved in just under twice the time required for factorization, with total time for both under 2.5 hours. Meshes comparable to that could simulate site arrays of similar size to the Earthscope MT Transportable Array of the U.S. Pacific Northwest using this parallelized direct solver [5].

Here in Part II, we also use direct solvers exclusively to create a 3D regularized

inversion algorithm for MT data including topography, which we name HexMT. Due to its good convergence properties, we pursue a Gauss-Newton formulation for the nonlinear, iterative parameter update, as have others [4, 6–8]. The number of parameters usually is significantly greater than the number of data for tomographic-style, regularized inversion. As noted by [9], inverse formulations using fewer parameters than data may suffer from a dependence of solution upon parameterization. One may also expect some lack of fit to data to occur if parameters are not defined optimally. On the other hand, tomographic inversions for MT data sets of a few hundred sites may require a number of parameters of order one million [e.g., 5]. Direct solution of the parameter step matrix in the traditional model-space definition [e.g., 6, 10], even using parallelization across multicore [11], is not practical for that scale of parameterization. As a result, researchers have tended to retain iterative solvers for the parameter step solution whether cast as Gauss-Newton or otherwise [e.g., 4, 12–14].

An alternative is to investigate the data space formulation for solving the Gauss-Newton parameter step [9, 15, 16]. This approach reduces the size of the parameter step matrix from $N_m \times N_m$ to $N_d \times N_d$ (m = model parameters, d = data), while the solutions in theory are identical. Consider an MT survey of 400 sites with 20 frequencies (four per decade say) and twelve data per frequency (four complex impedance and two complex tipper elements). The total data set size would be 96,000. As we show, this turns out to be a very manageable size of matrix to invert using direct solvers, particularly as parallelized across multicore SMP computers. Matrices twice this size in fact are not impractical, allowing data sets of more sites, greater bandwidth, or finer frequency sampling, with a fairly arbitrary number of model parameters.

This paper sets out with a brief overview of both model- and data space approaches to solving the Gauss-Newton step. Attention is paid to the mechanics of solving stably the step equation for a model gradient regularization functional. Run time and scalability of the step solver is investigated for multicore using different sized trial models. At this point, it appears that parameter step solution time will remain significantly smaller than forward simulation run time across all models with moderately fine parameter discretization. The inversion code is tested on several

models. These include a simple conductive brick below a hill to show the strength of effect that topography can have on inversion models assuming a flat surface. Subsequently, we examine a multiprism test model used as a community standard [17] and experiment with various regularization weighting schemes. Finally, we invert an extensive MT data set acquired over the volcano Mount St. Helens [18] to show performance for a model where parameter number approaches one million.

4.3 Forward problem

The forward problem is described in detail in [1], touched on briefly here to define terms. We consider the magnetotellurics (MT) problem in a domain Ω that includes the air and earth's subsurface. The earth's surface is allowed to have topography. In order to calculate the MT response due to an arbitrary 3D conductivity structure σ, we consider a hexahedral edge finite element discretization of the equation for the secondary electric field E:

$$\int_\Omega \frac{1}{\mu} \nabla \times E \cdot \nabla \times F + \mathrm{i}\omega \int_\Omega \hat{\sigma} E \cdot F = \int_\Omega -\mathrm{i}\omega(\hat{\sigma} - \hat{\sigma}^p) E^p \cdot F \tag{4.1}$$

for $E, F \in \mathcal{H}_0(\nabla\times)$, where ω is angular frequency, $\epsilon > 0$ is dielectric permittivity, μ is magnetic permeability, $\hat{\sigma} = \sigma + \mathrm{i}\omega\epsilon$, and $\hat{\sigma}^p = \sigma_p + \mathrm{i}\omega\epsilon$. E^p is the primary electric field, which is that of a plane wave traveling downwards in primary conductivity structure σ_p, the conductivity of a 1D earth. We assume that $\sigma \approx \sigma_p$ close to the domain boundaries. The solution space is defined below:

$$\begin{aligned} \mathcal{H}_0(\nabla\times, \Omega) = \big\{ F{:}\Omega \to \mathbb{C}^3 : \textstyle\int_\Omega (|F|^2 + |\nabla \times F|^2) < \infty, \\ n \times F|_{\partial\Omega} = 0 \big\} \end{aligned} \tag{4.2}$$

The approximate solution to equation (4.1) is obtained using edge elements. Secondary magnetic field H is calculated as

$$H = \frac{-\nabla \times E}{i\omega} \tag{4.3}$$

The total field E^t, H^t is a sum of secondary and primary fields:

$$E^t = E + E^p, \quad H^t = H + H^p \tag{4.4}$$

The MT response is obtained by finding Z, K such that

$$\begin{bmatrix} E_x^t \\ E_y^t \\ H_z^t \end{bmatrix} = \begin{bmatrix} Z_{xx} & Z_{xy} \\ Z_{yz} & Z_{yy} \\ K_{zx} & K_{zy} \end{bmatrix} \begin{bmatrix} H_x^t \\ H_y^t \end{bmatrix} \tag{4.5}$$

is satisfied no matter what is the polarization of the primary (E^p, H^p) plane wave.

4.4 Gauss-Newton inversion procedure

For defining inversion terminology, we consider again Figure 3.1 where layers of hexahedral elements are deformed vertically to represent topography. This was efficient for the forward problem, but also will be for inversion. Although elements below the earth surface could be grouped to form a parameter, for maximal flexibility, we usually consider each element as being a possible parameter linked through regularization in a tomographic inversion.

4.4.1 Description of the method

As is usual, the portion of the model domain Ω below the air-earth interface is split into N_m model parameters, which are disjoint regions with constant resistivity. Let $m = (m_1, \ldots, m_{N_m})$ be the vector of parameter $\log_{10}$ resistivity values. We work with $\log_{10}$ resistivity as this ensures that resistivity remains positive during inversion and makes the square norms of the step matrix columns more nearly equal in magnitude [19]. There are N_d data points collected, denoted as $d = (d_1, \ldots, d_{N_d})$. As individual data values, we consider the real and imaginary parts of all entries in Z, K for all N_{rec} receivers, namely $N_d = 12N_{\text{rec}}$, and for all frequencies. Let $e_1, \ldots, e_{N_d}$ be the vector of measurements errors, for which standard deviations s_i are known. By $F(m) \in \mathbb{R}^{N_d}$ we denote the response of the current model m, calculated by the forward code.

The inversion procedure seeks a model m such that

$$|F_i(m) - d_i| \approx s_i, \quad i = 1, \ldots, N_d \tag{4.6}$$

together with the constraint that some measure of roughness of the model m is limited. The roughness will be measured by

$$||m - m_0||^2_{B_m} = (m - m_0)^T B_m (m - m_0) \tag{4.7}$$

where m_0 is a reference model and B_m is a symmetric non-negative definite matrix, so that $||.||_{B_m}$ is a seminorm. Often B_m is such that $||m - m_0||_{B_m} = ||\nabla(m - m_0)||_{L_2}$, where ∇ denotes spatial gradient (in all three directions) of the $\log_{10}$ resistivity model. In the deformed mesh geometry we implement, the three directions in general are not

purely perpendicular; one remains vertical while the other two lie along the variably deformed layer of elements.

Specifically, in the inversion, we seek a model m that minimizes the functional

$$W(m) = (F(m) - d)^T B_d (F(m) - d) + \lambda (m - m_0)^T B_m (m - m_0) \tag{4.8}$$

for some suitable value of $\lambda > 0$, where B_d is a diagonal matrix with $\frac{1}{s_i^2}$ as entries.

The Gauss-Newton procedure is an iterative one that seeks a minimizer of (4.8). It starts with an initial guess m_1. Given a current model m_n, the Gauss-Newton scheme approximates the response $F(m)$ around m_n by its linear part:

$$F(m) \approx F(m_n) + J(m - m_n) \tag{4.9}$$

where J is a $N_d \times N_m$ matrix of derivatives of F

$$J_{i,j} = \frac{\partial F_i}{\partial m_j}(m),\ i = 1, \ldots, N_d,\ j = 1, \ldots, N_m \tag{4.10}$$

whose computation we have described in Part I. If (4.9) is used, the functional (4.8) becomes quadratic and the minimizer m_{n+1} satisfies a linear equation:

$$[J^T B_d J + \lambda B_m](m_{n+1} - m_0) = J^T B_d \hat{d} \tag{4.11}$$

where $\hat{d} = d - F(m_n) - J(m_n - m_0)$.

The matrix in the equation (4.11) is dense, symmetric positive definite, and has dimension $N_m \times N_m$. This is the traditional model-space parameter step formulation. The time to solve it using Cholesky decomposition is $O(\frac{1}{3}N_m^3)$. This cubical growth eventually makes direct solution of the model-space Gauss-Newton scheme impractical for arbitrarily large model size N_m.

The data space method [9, 15, 16] replaces equation (4.11) with a linear equation having only N_d unknowns. For the moment, we assume that B_m is invertible (which implies that B_m is positive definite and $||.||_{B_m}$ is a norm), the treatment of which we will revisit shortly. When (4.11) is left-multiplied by B_m^{-1}, we obtain:

$$B_m^{-1}[J^T B_d J + \lambda B_m](m_{n+1} - m_0) = B_m^{-1} J^T B_d \hat{d}$$

$$B_m^{-1} J^T (B_d J)(m_{n+1} - m_0) + \lambda (m_{n+1} - m_0) = B_m^{-1} J^T B_d \hat{d}$$

$$m_{n+1} - m_0 = B_m^{-1} J^T \frac{1}{\lambda} B_d [\hat{d} - J(m_{n+1} - m_0)]$$

This proves that

$$m_{n+1} - m_0 = B_m^{-1} J^T \beta \tag{4.12}$$

for some $\beta \in \mathbb{R}^{N_d}$. When (4.12) is plugged into (4.11) and the equation is left-multiplied by B_m^{-1}, we obtain an equation equivalent to (4.11):

$$B_m^{-1} J^T [B_d J B_m^{-1} J^T + \lambda I]\beta = B_m^{-1} J^T B_d \hat{d} \tag{4.13}$$

This equation will be satisfied if β satisfies

$$[B_d J B_m^{-1} J^T + \lambda I]\beta = B_d \hat{d} \tag{4.14}$$

which is equivalent to

$$[J B_m^{-1} J^T + \lambda B_d^{-1}]\beta = \hat{d} \tag{4.15}$$

The latter equation has a unique solution as $J B_m^{-1} J^T$ is symmetric non-negative definite and B_d^{-1} is symmetric positive definite. The data space Gauss-Newton method finds β, the solution to (4.15), and uses (4.12) to calculate a model update m_{n+1}.

4.4.2 Computational considerations

In the data space method, one has to invert B_m and apply it to J^T in order to calculate $J B_m^{-1} J^T$. In [16], matrix B_m^{-1} was denoted C_m and called the model covariance matrix. It was not defined explicitly as a result of inverting a norm matrix B_m, but was treated as a natural matrix to consider for regularization.

The choice of a proper regularization functional $||.||_{B_m}$ is important, as minimizing the functional W in (4.8) is equivalent to minimizing the data misfit

$$(F(m) - d)^T B_d (F(m) - d) \tag{4.16}$$

subject to a condition on the model norm

$$||m - m_0||_{B_m} \leq \delta \tag{4.17}$$

where $\delta > 0$ depends on the choice of λ. The regularization functional we consider is L_2 norm of the gradient of the model,

$$||m - m_0||_{B_m} = ||\nabla(m - m_0)||_{L_2} \tag{4.18}$$

In order to use this functional in data space method, matrix B_m must be inverted. However, if (4.18) is used, B_m is non-negative definite and thus singular. To make it positive definite, we add a small number $\epsilon > 0$ to its diagonal before inverting. The functional is negligibly modified as then

$$||v||_{B_m} = \sqrt{||\nabla v||^2_{L_2} + \epsilon \sum_{i=1}^{N_m} v_i^2} \tag{4.19}$$

To estimate the cost of calculation of $B_m^{-1} J^T$, we exploit the coincidence that the matrix B_m has a similar non-zero pattern as a scalar Poisson equation over the inversion voxel grid. Even if each inversion voxel consists of one element, the number of inversion voxels will be similar to the number of interior nodes in the earth's subsurface. Thus the non-zero pattern in B_m should be similar to those of the matrix used for the divergence correction described in Part I and the number of variables should be less. Even though the number of linear systems to be solved is $12 \times N_{\text{rec}}$, and for divergence correction it was $5 \times N_{\text{rec}}$, here all variables are real valued. Thus if the solution library MUMPS is used, the time of factorization of B_m and applying B_m^{-1} to J^T is expected to be less than the cost of applying the divergence correction, which takes a fraction of time of the forward problem. One concludes that calculation of $B_m^{-1} J^T$ will not add significant execution time to the inversion process no matter how large the model is, as long as the direct solver is used for forward modeling.

For matrix multiplications like $J^T B_d J$ and the Cholesky factorization needed to solve equations (4.11) and (4.15), we use the PLASMA library [20, 21]. PLASMA is a linear algebra library for dense matrices, parallelized for shared memory machines (like the SMP unit we use). It uses a matrix tiling approach [cf. 11, 22, 23], which reduces the time of transporting the matrix entries from RAM to CPU. The scalability of Cholesky factorization and matrix multiplication using PLASMA is presented in Figure 4.1. The speedup using 24 cores is ∼17 and ∼19 for Cholesky factorization and matrix multiplication, respectively.

To assemble the model-space Gauss-Newton matrix (4.11) , one has to evaluate $J^T B_d J$, which has numerical complexity $O(N_m^2 N_d)$. Solving the matrix as noted previously has complexity $O(\frac{1}{3} N_m^3)$. For the data space method on the other hand, to assemble the matrix equation (4.15), one has to evaluate $J(B_m^{-1} J^T)$, which has

numerical complexity $O(N_m N_d^2)$. Solving the equation using Cholesky decomposition has complexity $O(\frac{1}{3}N_d^3)$. As typically $N_d < N_m$, the computational cost associated with data space method is less. The difference becomes more pronounced for larger MT surveys.

Example computation times are presented in Table 4.1. The models that are considered in the tests are listed in Table 4.2. The time to solve the model-space Gauss-Newton equation (4.11) increases rapidly with the model size and quickly gets impractical, reaching over 27 hours for the largest model considered. On the other hand, the time to solve the data space equation (4.15) remains short, less than one minute for all the models considered. In the case of the data space method, more time consuming than solving equation (4.15) is evaluating $J(B_m^{-1}J^T)$, which takes 1 hour for the largest model considered. Nevertheless, the corresponding evaluation of $J^T B_d J$ for model-space Gauss-Newton takes longer, more than seven times longer for model 5. Comparable to the time of calculation of $J(B_m^{-1}J^T)$ is the time of evaluation of $B_m^{-1}J^T$. However, as we expected, this time is less than that of applying the divergence correction (more than two times less), which in turn is almost 10 times less than the total time spent solving forward problem. Thus the advantage of the data space Gauss-Newton approach over the model-space version for this application is clear.

Concerning RAM requirements, for model-space GN, one needs to store the matrix J, which is $N_d \times N_m$ as well as the matrix in the equation (4.11), which is a symmetric $N_m \times N_m$ matrix. In the data space version, one needs to store matrix J and the matrix $B_m^{-1}J^T$ of the same size, but depending on implementation, those may be saved on hard disk and parts of them may be read into RAM memory when needed. Also, one needs to store the matrix of equation (4.15), which is a symmetric $N_d \times N_d$ matrix. One can see that as typically $N_d < N_m$ the memory requirements are smaller for data space Gauss-Newton than for its model-space cousin.

In Table 4.3, we present the RAM memory requirements for the models considered. The matrix J is $N_d \times N_m$ in size, but largely can be stored on hard disk and parts accessed as needed. For the largest model considered, the model-space GN step requires 413GB, whereas the data space step requires as little as 4.6GB of RAM

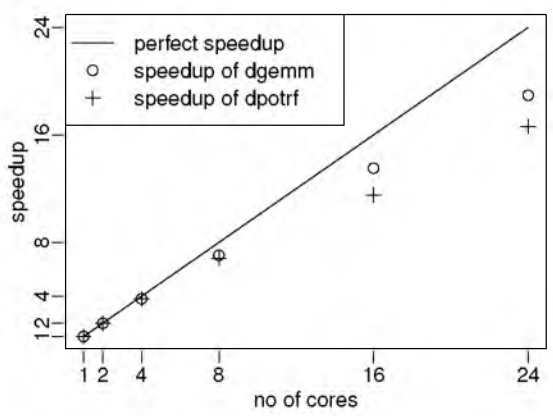

Figure 4.1. Speedup of PLASMA library for Cholesky factorization (dpotrf) and matrix multiplication (dgemm).

Table 4.1. Run times in format hh:mm:ss for models listed in Table 4.2. "FP" denotes the total time of calculation of the forward problem (response $F(m)$ and Jacobian J) using MUMPS. "FP DC" is the time spent on divergence correction using MUMPS(fraction of "FP"). "GN $J^T B_d J$" denotes the time spent on evaluation of $J^T B_d J$ using PLASMA. "GN" solve denotes time needed to solve the Gauss-Newton equation (4.11) using PLASMA once the matrix $J^T B_d J + \lambda B_m$ has been assembled. "DS $B_m^{-1} J^T$" denotes the time spent to calculate $B_m^{-1} J^T$ using MUMPS. "DS JJ^T" denotes the time spent to evaluate $J(B_m^{-1} J^T)$ using PLASMA. "DS solve" denotes the time spent to solve equation (4.15) using PLASMA, once the matrix $JB_m^{-1} J^T + \lambda B_d^{-1}$ has been assembled. * denotes the estimated time, the calculation hasn't been done due to insufficient RAM memory. The calculations have been done on 24-core workstation (four Intel E5-4610 Sandy Bridge hex-core processors at 2.4 GHz).

ID	FP total	FP DC	GN $J^T B_d J$	GN solve	DS $B_m^{-1} J^T$	DS JJ^T	DS solve
1	00:08:35	00:01:34	00:00:39	00:00:27	00:00:33	00:00:24	00:00:02
2	03:11:13	00:11:27	00:14:33	01:03:39	00:04:00	00:04:23	00:00:10
3	04:16:10	00:15:43	00:21:51	05:10:34	00:04:03	00:03:44	00:00:02
4	09:19:08	01:10:57	02:16:54	05:10:20	00:11:31	00:14:54	00:00:43
5	17:18:00	02:12:53	07:11:27*	27:12:21*	01:00:30	01:02:24	00:00:43

Table 4.2. Statistics of test models used for run times testing. The brick under a hill model with topography has been used.

ID	FEM grid	inversion grid	edges no	recv no	freq no	N_d	N_m
1	$41x$ $41y$ $30z$	$39x$ $39y$ $19z$	143,120	81	13	12,636	28,899
2	$65x$ $65y$ $45z$	$63x$ $63y$ $29z$	550,400	121	15	21,780	115,101
3	$81x$ $81y$ $47z$	$79x$ $79y$ $31z$	896,960	64	15	11,520	193,471
4	$81x$ $81y$ $47z$	$79x$ $79y$ $31z$	896,960	196	15	35,280	193,471
5	$101x$ $101y$ $50z$	$99x$ $99y$ $34z$	1,489,800	196	15	35,280	333,234

Table 4.3. RAM memory needed for matrices related to the Gauss-Newton step using model-space and data space for models listed in Table 4.2. "FP" denotes RAM needed to calculate the forward problem. "J" denotes the memory needed to store the matrix J of size $N_m \times N_d$. "GN" denotes the memory needed to store the (symmetric) matrix $J^T B_d J + \lambda B_m$ of size $N_m \times N_m$. "DS" denotes the memory needed to store the (symmetric) matrix $JB_m^{-1}J^T + \lambda B_d^{-1}$ of size $N_d \times N_d$

ID	N_m	N_d	FP	J	GN	DS
1	28899	12636	3.4GB	2.7GB	3.1GB	0.6GB
2	115101	21780	21.6GB	18.7GB	49.4GB	1.8GB
3	193471	11520	38.8GB	16.6GB	139.4GB	0.5GB
4	193471	35280	38.0GB	50.9GB	139.4GB	4.6GB
5	333234	35280	76.1GB	87.6GB	413.7GB	4.6GB

memory depending upon treatment of J.

4.4.3 Regularization norm weight

Up to this point, we have not specified details of the entries of matrix B_m, other than that it is a finite difference representation of spatial gradients in the model parameter vector m. Several investigators have explored whether entries of B_m should also be weighted according to influence (Jacobian) of their corresponding parameter [e.g., 24, 25]. Here we present three different regularization functionals, the performance of which will be compared with numerical tests.

Consider the infinite-dimensional problem and its response $F(m)$ given a spatially varying $\log_{10}$ resistivity model $m = m(\mathbf{r})$. Also consider the derivative $S(\mathbf{r})$ of F with respect to m satisfying:

$$F(m + \delta m) \approx F(m) + \int_\Omega S(\mathbf{r})\delta m(\mathbf{r})d\mathbf{r} \tag{4.20}$$

for a small change in model δm. The quantity $||S(\mathbf{r})||_2$ measures the sensitivity of the response F to the change of the conductivity at the point $\mathbf{r}$.

As the regularization functional $(m-m_0)^T B_m(m-m_0)$, we will consider L_2 norms of the gradient of m with a weight $\nu(\mathbf{r}) > 0$ defined as follows:

$$||\nabla(m - m_0)||^2_{L_2(\nu)} = \int_\Omega |\nabla(m - m_0)|^2\nu(\mathbf{r})d\mathbf{r} \tag{4.21}$$

Further, we consider three possible values for ν:

$$\begin{aligned} \nu(\mathbf{r}) &= ||S(\mathbf{r})||_2 \\ \nu(\mathbf{r}) &= 1 \\ \nu(\mathbf{r}) &= \tfrac{1}{||S(\mathbf{r})||_2} \end{aligned} \tag{4.22}$$

Norm $||\nabla(m - m_0)||_{L_2(1)}$ uses no information about the influence of the inversion voxel on the data. If norm $||\nabla(m-m_0)||_{L_2(||S||_2)}$ is used for regularization, smoothing is suppressed for parameter regions with low sensitivity, allowing them to show additional structure [cf. 24]. Norm $||\nabla(m - m_0)||_{L_2\left(\frac{1}{||S||_2}\right)}$ will suppress regions with low sensitivity, using the reasoning that if we cannot detect the properties of a region well, we will make it similar to its surroundings. This is similar to the approach of [25], although they make a rigorous evaluation of the parameter resolution matrix with is computationally intractible for the larger problems we consider here.

Norm $||m||_{L_2(\nu)}$ has the property that if the inversion mesh is changed in such a way that the model $m(\mathbf{r})$ remains the same, the norm remains the same. For example, if one decides to split a given voxel V_j into two voxels V_{j_1}, V_{j_2} and sets $m_{j_1} = m_{j_2} = m_j$, the corresponding function $m(\nu)$ remains the same and so does $||m||_{L_2(\nu)}$. It is a desired property especially in the case of our hexahedral mesh, when the elements and as a result inversion voxels vary in size and shape; some of them may be nearly cubic in shape while others may be long and thin.

More details on how these norms are approximated are provided in Appendix A (Section 4.9).

4.5 Synthetic inversion examples

In this section, we present results of the inversion of synthetic MT data to evaluate algorithm performance under controlled conditions. As a measurement error, we will use the value

$$\begin{aligned} e(Z_{ij}) &= \max\left\{3.5\%\tfrac{|Z_{xy}-Z_{yx}|}{2}\right\} && ,i,j=x,y \\ e(K_{zj}) &= 0.03 && ,j=x,y \end{aligned} \tag{4.23}$$

We use $Z_{xy}-Z_{yx}$ because it is a rotational invariant and shows relative stability to data noise [see 26]. As the measurement error s_j (noted in (4.6)) for real or imaginary part of Z and K, we take the above value e. The data used in the inversion are calculated by the forward problem for the true conductivity model, with Gaussian noise having zero mean and standard deviation s added to the real and imaginary parts of Z and K.

To assess goodness of fit of a model response to the data, we use the normalized root mean square (nRMS), defined as:

$$\mathrm{nRMS}(m) = \sqrt{\frac{1}{N_d}\sum_{j=1}^{N_d}\left(\frac{d_j - F_j(m)}{s_j}\right)^2} \tag{4.24}$$

where d is the vector of our synthetic data, F is a vector of response of the model m, and s is the measurement error vector.

In the inversion process, the parameter λ in (4.8) that is used to obtain model m_{j+1} is set as:

$$\lambda = \mathrm{nRMS}(m_j)\cdot\kappa \tag{4.25}$$

Parameter κ is set to some initial value $\kappa_0 > 0$ at the beginning of the inversion. If at some iteration the nRMS does not decrease by more than 5%, parameter κ is decreased by a factor of two. As a result in our inversions, the parameter λ steadily decreases and the model acquires increasing amoounts of structure. The scaling with respect to nRMS is consistent with experience in Constable et al. (1987) where the optimal λ decreased as iterations proceeded and misfit improved. However, we to not sweep through a series of λ values at each iteration due to computational expense. Procedures for determining λ warrant more investigation.

4.5.1 Brick under a hill

Our first model is a brick under a hill in 100Ωm background. The hill has dimensions 2000m and 4000m in x and y directions at the bottom and 500m and 1000m at the top. The hill is 450m high. The object is placed below the hill with the top and the bottom of the object 650m and 1600m, respectively, below the top of the hill and its XY cross-section is a square $[-328\text{m}, 328\text{m}] \times [-700\text{m}, 700\text{m}]$. We consider a conductive (5Ωm) and resistive (2000Ωm) object as well as no object at all. We compare the inversion that has the mesh conforming to the topography as in Figure 4.2 to the mesh without topography(flat surface). Both meshes have the same location of voxels in x and y directions and the same x and y coordinates of receivers. The only difference is the elevation of layers close to the earth surface.

We generated the data using a different grid than the one used for inversion. The forward code grid consisted of 95 cells in x direction, 95 cells in y direction, and 50 cells in z direction ($95x$, $95y$, $50z$) and extended to 18km from the grid center in x and y directions, 5.6km above the earth's surface (air layer), and 12.5km below the surface. The inversion grid was $41x$, $41y$, and $30z$. It extended 14km and 15km from the center of the grid in x and y directions, respectively. There were 106 receivers. The location of a receiver is always at the center of the face of an element lying on the earth's surface, thus the location of the forward code receivers is slightly different than the inversion receivers. The inversion grid, together with the location of the brick and receivers in 3 of the 4 quadrants, is presented in cutaway view in Figure 4.2.

The data consisted of the impedance Z and the tipper K for 13 frequencies between

1Hz and 1000Hz distributed evenly in $\log_{10}$ space. We added Gaussian error with standard deviation (4.23) to the forward data. The initial value of κ_0 started at the same value for all inversions. The starting and reference (a priori) models were set to 50Ωm uniformly. The regularization functional used was $||\nabla m||_{L_2(1)}$.

Iteration history is presented in Figure 4.3. The regularization norm $||m_j - m_0||_{B_m}$ increases as the inversion proceeds and λ decreases in the effort to decrease nRMS. One can see that inversion with topography is able to achieve nRMS close to 1 in less than 3 iterations, whereas the inversion without topography is struggling to decrease nRMS below 1.6 even though the model norm is larger than in the case of inversion with topography.

We have plotted cross-sections of selected models for 6 inversions in Figures 4.4 and 4.5 for comparison. In all cases, the inversion with topography is able to recover a smoothed version of the original object (or no object in the no brick case). Inversion without topography puts a more conductive object below the ground to make-up for the absence of a hill. This occurs because the electric field is reduced by the hill as background electric current only partially flows upward into that volume [see TM mode results in 27]. Even for the resistive brick forward data, the inversion without topography returns a (somewhat) conductive object. The inversion also creates an oscillatory region above the object (more apparent on XZ cross-section) that resembles the shape of a hill. These results emphasize the importance of including the topography in the inversion of MT data.

4.5.2 Simple two brick model

Our next synthetic model consists simply of two buried and separated bricks, one conductive (2Ωm) and one resistive (1000Ωm), in a 40Ωm half-space (Figure 4.6). With this model, we examine the effect of inversion regularization weights on model characteristics.

The forward mesh consists of $58x$ $58y$ $45z$ elements. In the XY plane, the central 33×33 elements are square with sides = 0.333km. Further from the center, the element sizes grow gradually and extend 130km from the center of the grid. In the Z direction, there were 34 elements below the surface and 11 elements in the air. The mesh extends to 100km above the surface and 140km below the surface.

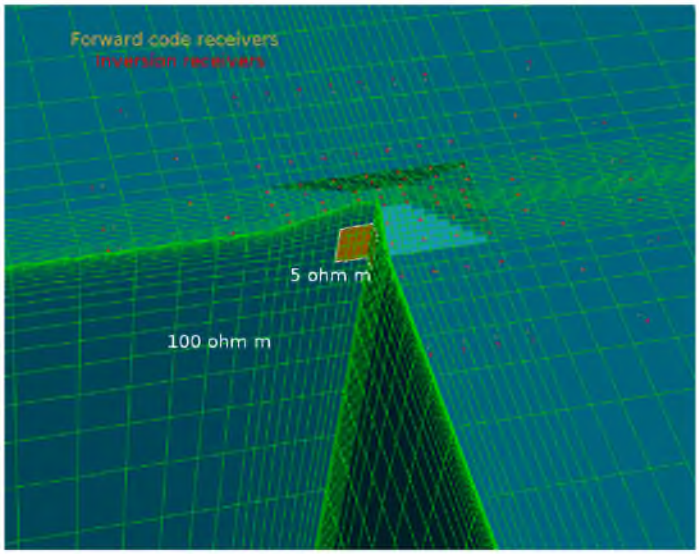

Figure 4.2. Central part of the inversion grid together with the receiver locations in 3 quadrants. Conductive brick is shown below the hill.

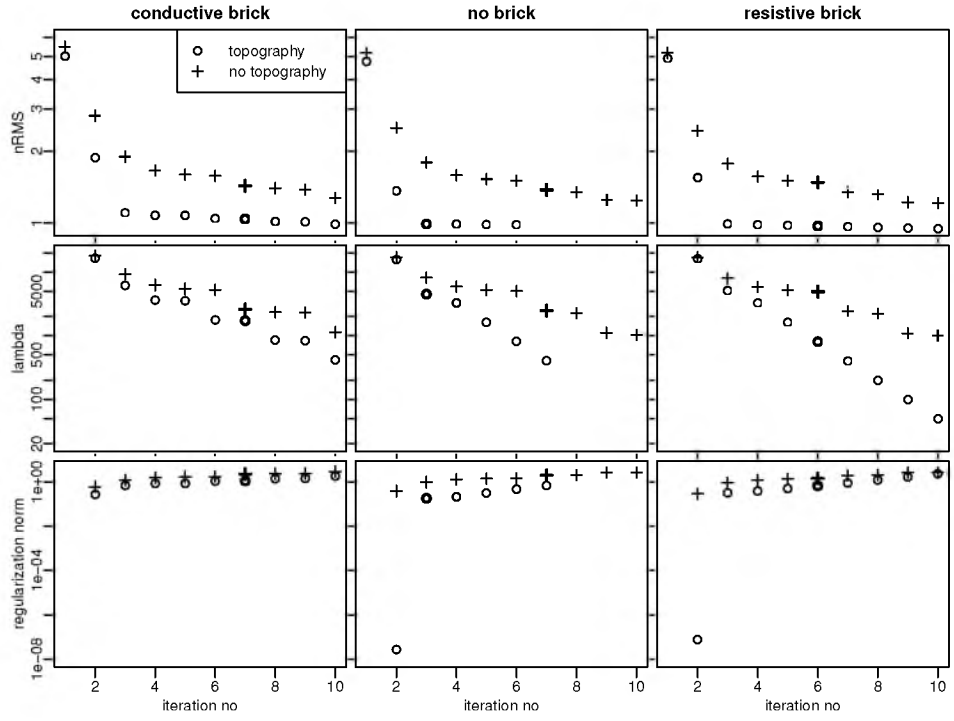

Figure 4.3. Inversion iteration history for model of bricks under a hill. nRMS(m_j),λ used to obtain model m_j, $||m_j - m_0||_{B_m}$ as a function of iteration number j. The models plotted in Figures 4.4 and 4.5 are denoted by bold symbols.

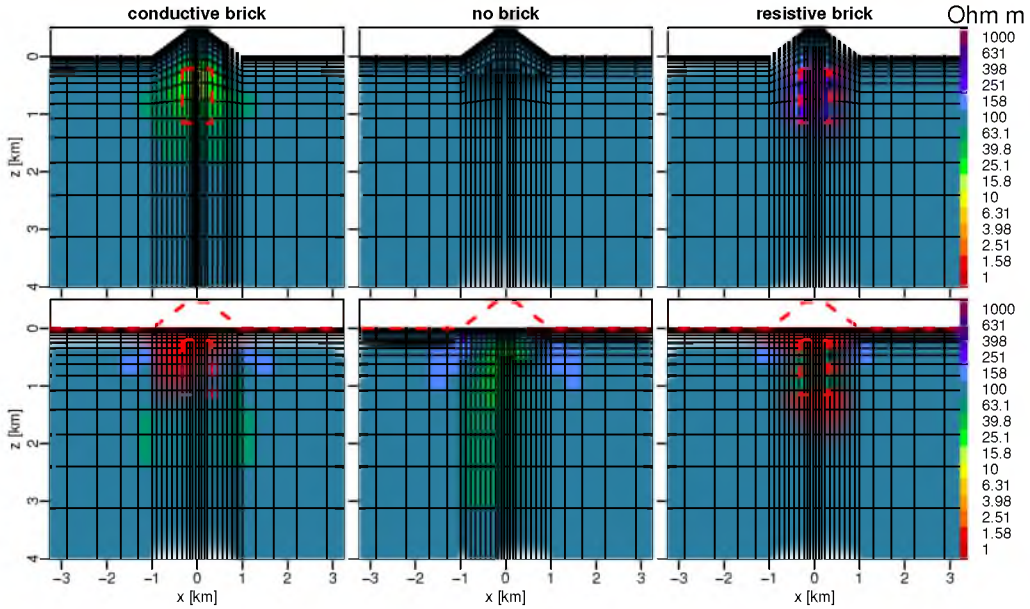

Figure 4.4. Inversion results for bricks under a hill along XZ cross-section at $y = 0$km. Top row shows inversion with topography, bottom row the inversion without topography.

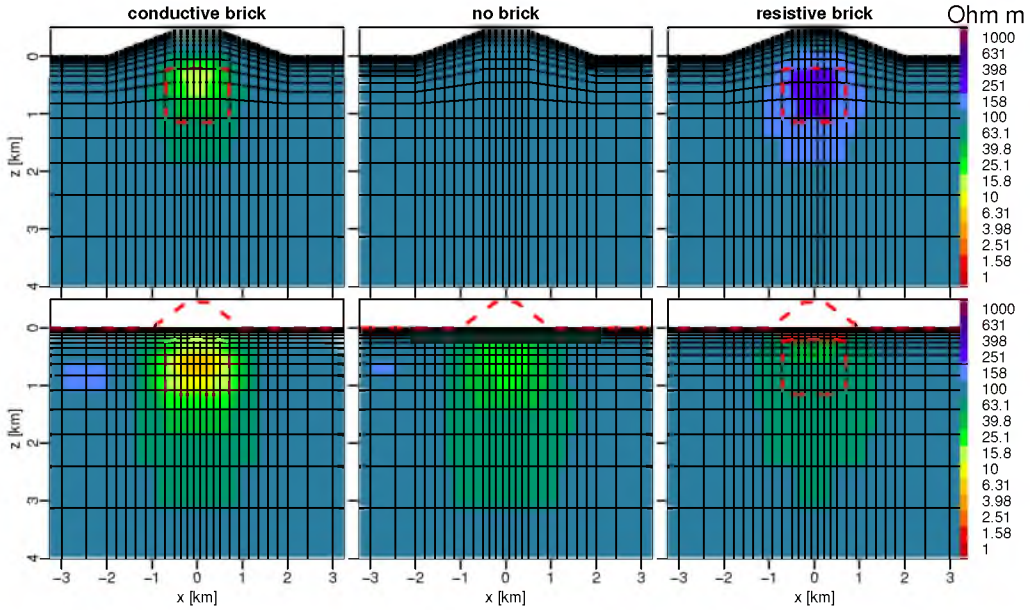

Figure 4.5. Inversion results for bricks under a hill along YZ cross-section at $x = 0$km. Top row shows inversion with topography, bottom row the inversion without topography.

There are 10×10 receivers evenly distributed in XY plane, separated by 10km. The forward response (impedance Z and tipper K) was generated for 31 frequencies evenly distributed in log space between 0.01Hz and 1000Hz, which gives 6 frequencies per decade. We added Gaussian error with standard deviation (4.23) to the forward data.

The inversion mesh consists of $41x$ $41y$ $41z$ elements. In the XY plane, the central 24 by 24 elements are square with sides = 0.5km. Further from the center, the element sizes grow gradually and extend 135km from the center of the grid. In the Z direction, there were 31 elements below the surface and 10 elements in the air. The mesh extends to 110km above the surface and 140km below the surface. Thus the forward and inversion meshes differ in discretization but have the same locations for the receivers, which are at the center of elements faces in both cases. The inversion mesh is presented in Figure 4.7.

For this model, we conducted inversions using different regularization functionals. We used (4.30), (4.27) and (4.32) that give regularization functionals resembling $||\nabla m||^2_{L_2(||S||_2)}$, $||\nabla m||^2_{L_2(1)}$ and $||\nabla m||^2_{L_2\left(\frac{1}{||S||_2}\right)}$, respectively. The value of $||S||_2$ was confined to change within a factor of 10^4. More precisely, two values were found S_1 and S_2, such that $\frac{S_2}{S_1} = 10^4$, and the value of $||S||_2$ was truncated if it lies outside the interval $[S_1, S_2]$. Additionally, weights ν have been multiplied by a normalization constant, so that the average ν over the central part of the domain is the same in all cases. This allows us to use the same initial value of λ. The nRMS, λ, and the regularization norm as a function of the iteration number are presented in Figure 4.8. One can see that the nRMS values as a function of iteration number are almost the same for the different regularization schemes, and thus the amount of regularization is similar for all weightings.

The models calculated by the different inversion schemes at iteration 6 are presented in Figure 4.9, with weights ν of $L_2(\rho)$ norms used for regularization to obtain those models plotted in Figure 4.10. Generally speaking, the weight $\nu = ||S||_2$ decreases with depth and the weight $\nu = \frac{1}{||S||_2}$ increases with depth. Thus the effect of using $L_2(||S||_2)$ is to prolong the depth extent of the formed image to minimize the value of the regularization norm. Similarly for $L_2\left(\frac{1}{||S||_2}\right)$, the recovered objects tend to be compressed toward the surface for comparable reasons. In the case of $L_2\left(\frac{1}{||S||_2}\right)$,

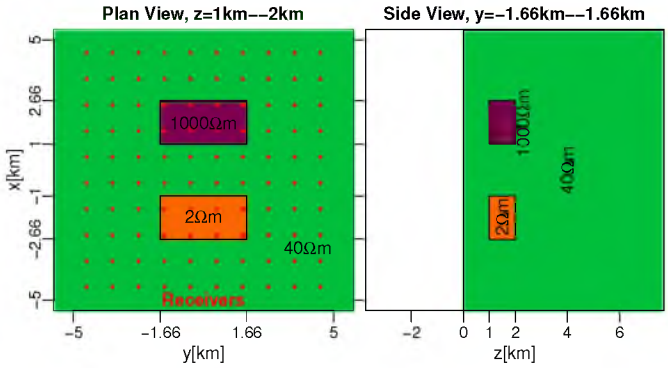

Figure 4.6. Sketch of two bricks model.

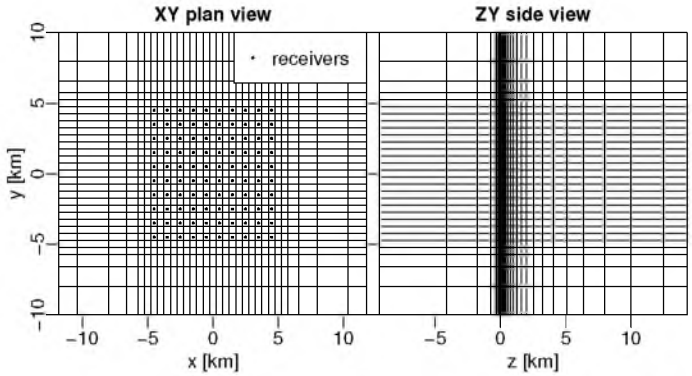

Figure 4.7. Cross-sections of the inversion grid for the two bricks model. Central part of the grid is shown.

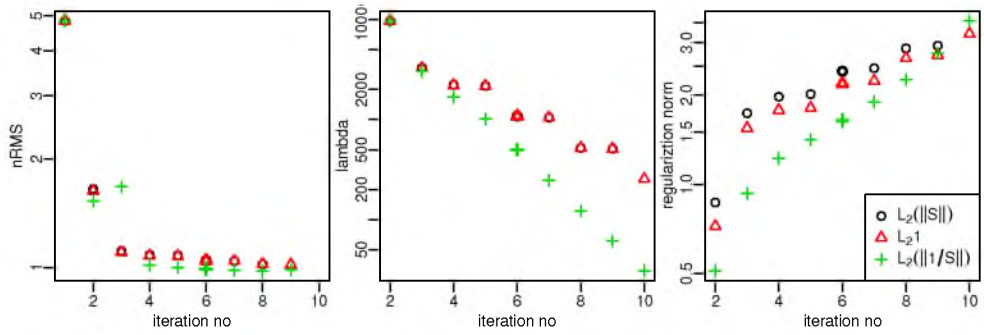

Figure 4.8. Two bricks model iteration history. nRMS(m_j), λ used to obtain model m_j, $||m_j - m_0||_{B_m}$ as a function of iteration number j. The model number 6, plotted on Figure 4.9 is denoted by a bold symbol.

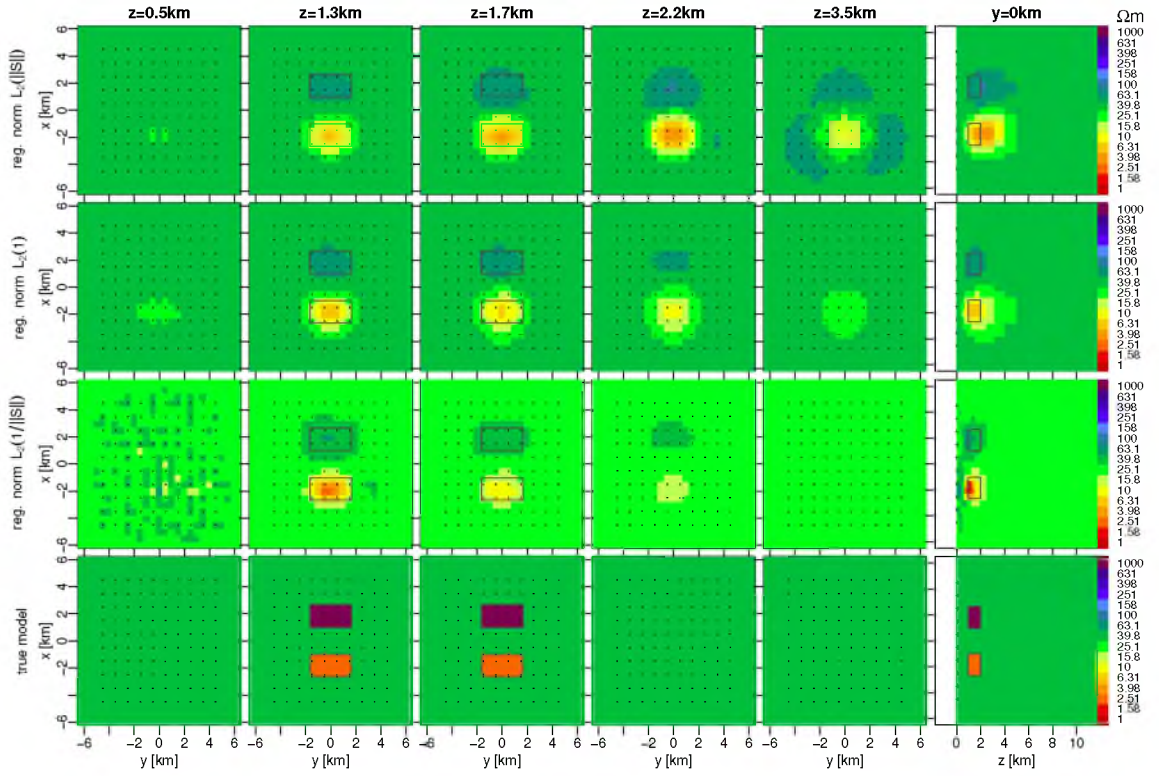

Figure 4.9. Models calculated by inversions of synthetic responses of two bricks using regularization functionals $||\nabla m||^2_{L_2(||S||_2)}$, $||\nabla m||^2_{L_2(1)}$ and $||\nabla m||^2_{L_2\left(\frac{1}{||S||_2}\right)}$. In each case, the model obtained at iteration 6 is plotted.

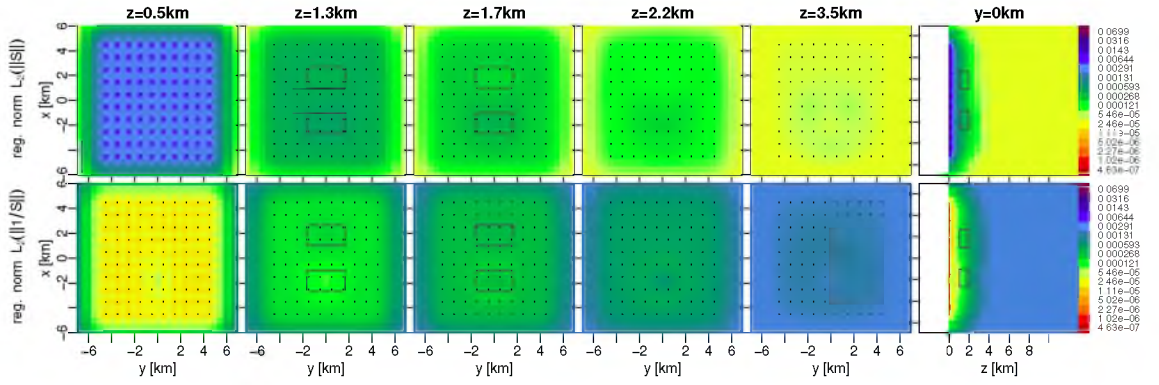

Figure 4.10. Weight ν of $L_2(\nu)$ norm used to obtain model number 6 for regularization schemes used for two bricks model. For $L_2(1)$ regularization, the weight is constant so is not plotted.

significant resistivity oscillations are apparent at shallow depths; one of their effects is to drive the background resistivity toward $25\Omega m$ rather than the true $40\Omega m$ because of the voxel-scale heterogeneity formed under the receivers. Nevertheless, the nRMS values are all very close, underscoring the nonuniqueness inherent in this ill-posed inversion problem. We observed no systematic difference in the fit of the final models across the frequency range for the three regularizations. Results for other models might differ, however. Further challenges in establishing appropriate regularization may be expected for more complex settings.

4.5.3 DSM2 model

The Dublin MT Modeling and Inversion workshops have provided model results for the EM community to test newly developed simulation and imaging codes [see 17]. Here we consider inversion of the MT responses of the Dublin Secret Model 2 (DSM2) presented in Figure 4.11. It is a flat-earth model with two contacting, shallow bricks in a four-layer earth. There are 144 MT receivers arranged in a uniform grid 12x12 with 7km spacing.

The forward data, supplied by the workshop organizers, consist of the impedance tensor Z values only (no tipper) for 30 frequencies between 0.016s and 10000s evenly distributed in $\log_{10}$. Random galvanic distortion was applied to the responses by the organizers as described in [17]. Gaussian noise of 5% of the maximal impedance value also had been added to the distorted data set. This supplied error bound was treated as a standard deviation and was used for both real and imaginary parts of Z. The data from all sites and frequencies were used in our inversion.

The applied static distortion provides an opportunity for us to implement and test recent distortion removal procedures [28–30]. Specifically, we follow the approach of Avdeeva et al., summarized in Appendix B (Section 4.10). Initially, an inversion model is sought without distortion correction. This model is used as a initial guess to estimate a new, more stable model plus the static distortion matrices of the impedance Z. We invert the data using the $L_2(1)$ regularization functional.

We considered coarse and fine inversion meshes. The coarse mesh has two columns of parameters per site in the central portion of the model whereas the fine mesh has five columns of parameters per site. The purpose of the latter mesh is to test whether

a fine discretization allows formation of a small-scale shallow structure which can simulate the impedance galvanic distortion without having to solve explicitly for correction factors [cf., e.g., 5].

Specifically, the coarse(fine) mesh consisted of $45x\ 45y\ 41z(78x\ 78y\ 50z)$ elements. In the XY plane, the central 23 by 23(58 by 58) elements are squares with sides = 3.5km(1.4km). Further from the center, the element sizes grow gradually and extend 600km from the center of the grid. In the Z direction, there were 31(38) elements below the surface and 10(12) elements in the air. The mesh extends to 300km above the surface and 700km below the surface. The central part of the coarse mesh is presented in Figure 4.12.

The inverted models are presented in Figure 4.13. Inverting only for $\log_{10}$ resistivity on the coarse mesh with no distortion correction yields a model with nRMS of 4.2, with little further improvement by relaxing the regularization factor (see Figure 4.14). Subsequently, inverting also for the distortion matrices obtains a model with nRMS of 1.1. The latter model achieves generally smoother resistivity structure with values closer to the true values, especially in the deeper structure, than does the former model. For the coarse model, there is some scatter in the norm of distortion matrices versus iteration. This presumably is a result of small regularization ($\tau = 0.01$). Further investigation is warranted as to when, and to what degree of regularization, distortion should be estimated through the iteration history.

When the fine mesh inversion for $\log_{10}$ resistivity only is considered, the resulting model has nRMS of 2.2, significantly less than the similar model obtained on a coarse mesh. The fine mesh inversion is able to represent some of the distortion by small-scale variability of $\log_{10}$ resistivity in the vicinity of the receivers, at shallow depths. Nevertheless, the fine mesh inversion for $\log_{10}$ resistivity including the distortion estimation provides a smoother model with a smaller nRMS of 1.1 (see Figure 4.14). Here we see smoother behavior in the estimated distortion versus iteration. From this result, we suggest that distortion matrices should be considered in tensor impedance inversion even for fine discretizations. However, we also advocate that fine discretization be used to the extent practical to ensure that nongalvanic variations at the highest frequencies are accommodated by the smallest-scale mesh structure.

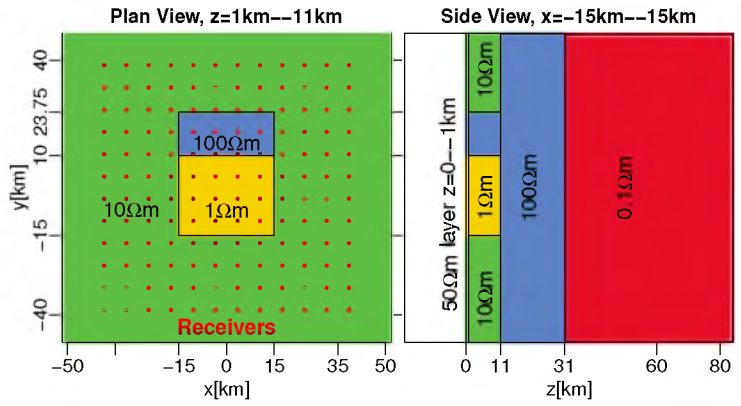

Figure 4.11. Sketch of DSM2 model

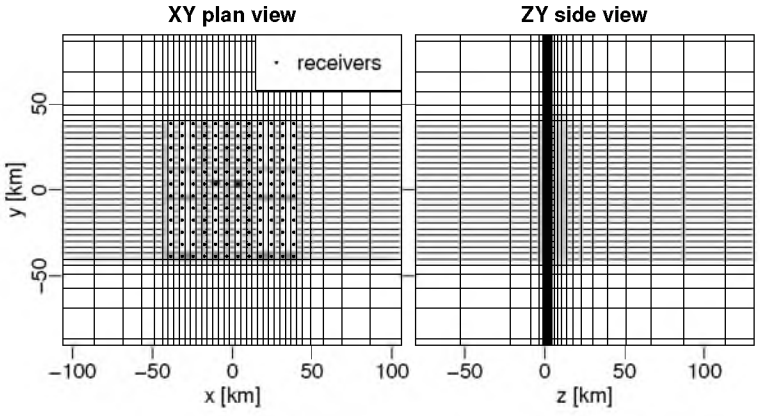

Figure 4.12. Cross-sections of the coarse inversion grid for DSM2 model.

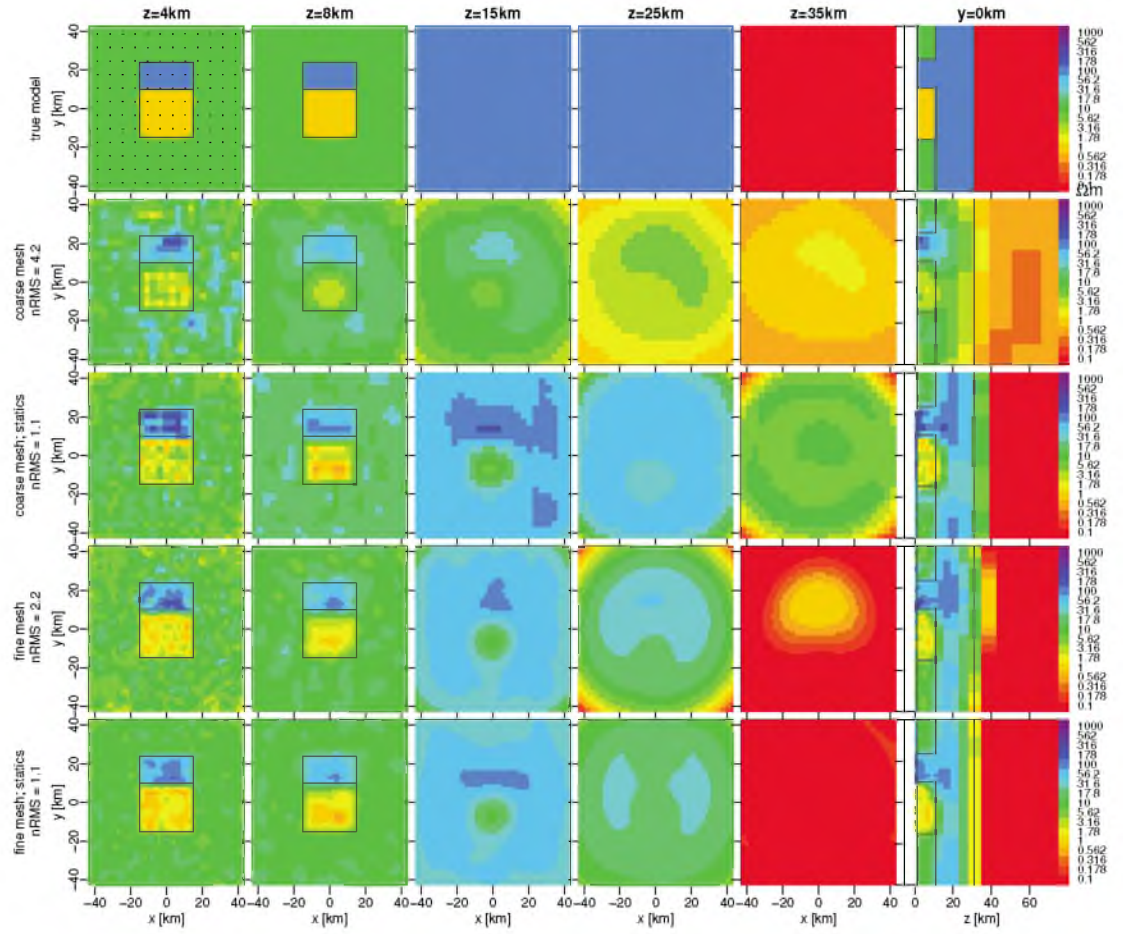

Figure 4.13. Inversion models obtained for DSM2. Top two rows consider a coarse mesh. Row three and four show inversions on a fine mesh. Rows second and fourth, denoted by 'statics', show results of inversion for $\log_{10}$ resistivity and the static distortion matrices. The last row shows the true model.

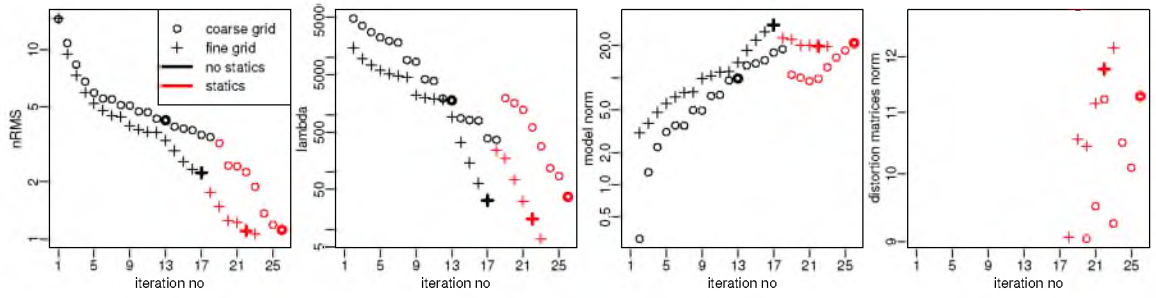

Figure 4.14. Iteration history for DSM2 model, for coarse and fine meshes. Initial inversion without static distortion is shown in black. Subsequent inversion with distortion matrix estimation is plotted in red. Bold symbols denote models shown in Figure 4.13.

4.6 Field inversion examples

4.6.1 Mount St. Helens

Finally, we examine the MT field data set collected by the authors of [18] from the north-central Cascade volcanic environment in Washington State, USA, to demonstrate the ability of our solution to handle moderately large models with topography. There are 82 soundings primarily clustered over the recently-active Mount St. Helens volcano, but with 14 of the sites extending in a nearly E-W profile past the north side of Mount Adams (Figure 4.15). This gives us the opportunity also to compare 3D inversion of profile data [e.g., 32] with 2D inversion results. We invert the complex tensor impedance Z and tipper K for 20 frequencies log-uniformly distributed from 100 through 0.0018 Hz.

The mesh, presented in Figure 4.16, consists of $111x$ $167y$ $62z$ elements in total. This requires storage of 500 GB, which fills the capacity of our particular workstation. Over the large central section including the two volcanoes, horizontal dimensions of the elements were in the 500× 600 to 500×1000 m range typically. Around this region, the element sizes grew gradually, covering a total area of 375 km× 425 km. In the z direction, there are 50 elements below the ground and 12 elements in the air. The elements at the earth's surface have a thickness of 80 m (at mesh edge) and grow gradually to reach an elevation of 250 km above the surface and a depth of 220 km below the ground. We did not attempt to include the Pacific Ocean nearly 200 km to the west, as that distance is significantly larger than the depth range of interest here ($<$ 100 km). A rim of one element around the side edges and bottom of the mesh was excluded from the inversion and fixed to be a 1D (flat) initial model. Thus there are $109\times165\times49 = 881265$ inversion parameters in the Mount St. Helens model. However, in data space formulation, the rank of the step matrix is only 19680. This may seem like a large model to handle 82 MT sites, but that is a result of particular site distribution. In principle, many more MT sites could be placed in this model mesh with additional computational cost only being more Jacobian source reductions and a larger (though still modest) data space step matrix.

The inversion was run without distortion matrix estimation for 11 iterations, with iteration history shown in Figure 4.17. Data error floors as given in equation (23) were

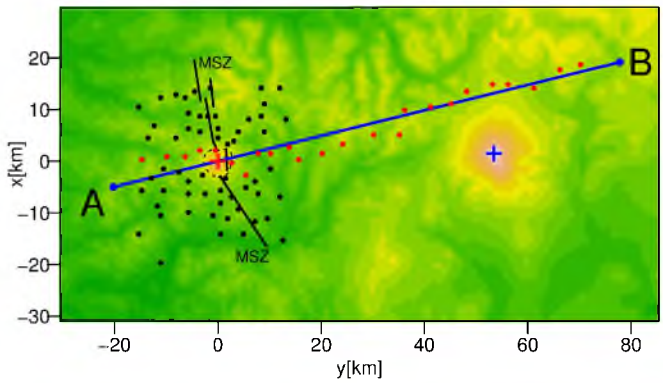

Figure 4.15. Mount St. Helens inversion model. Elevation map of the central part of the domain. Coordinate $(0, 0)$ corresponds to the location of Mount St. Helens, marked by a red cross. Blue cross denotes Mount Adams. Blue line denotes profile A B used in Figure 4.18. MT receiver locations are marked by black and red dots. Red dots denote receivers used in 2D inversion of [18]. Mount St. Helens shear zone (MSZ) after [31].

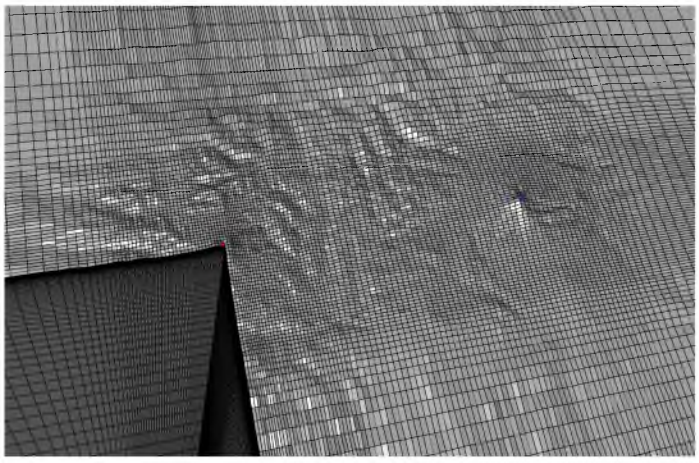

Figure 4.16. Central part of the mesh for the Mount St. Helens inversion model. Blue and red crosses denote Mount Adams and Mount St. Helens, respectively.

adopted. The starting model was a 100 ohm-m half-space, the same as considered in [18], and the starting nRMS was ∼11.5. Run time on the 24-core workstation was ∼30 hours per iteration, which was by far dominated by the forward and Jacobian calculations over the 20 frequencies. Model 11 has nRMS of 1.2, which is considered a good fit so that distortion correction should yield little improvement and was not carried out.

Model cross-section and plan views are presented in Figures 4.18 and 4.19, and can be compared to the original results of [18]. The cross-section overall bears a close ressemblance to the 2D inversion of Hill et al., which emphasized the nominal TM mode (relative to profile orientation) of data. Steep low resistivity is seen in the middle crust directly under Mount St. Helens, presumably related to recent eruptive processes, of which more will be discussed shortly. This gives way at depths $>$ 20 km to broad, quasi-horizontal low resistivity between the two volcanoes, which we attribute to lower crustal magmatic underplating and high temperature fluid release. Shallow, very low resistivity overlies the deep crustal conductor approaching Mount Adams which may reflect in part the presence of graphitic metasediments associate with a suture between the Siletz terrane and former North American margin [the southern Washington Cascades conductor or SWCC or 33], although this interpretation is nonunique and not without controversy [18, 34]. A large resistive body extending to $>$ 15 km depth lies between the Mount St. Helens and Mount Adams and could be correlated with earlier Western Cascades intrusive rocks [see 35].

The steep low resistivity directly under Mount St. Helens in Figure 4.18 is similar to that in the flat-earth 3D inversion model of [18] although the most anomalous portion does not extend to quite as shallow a depth as that in Hill's. This may in part be explained by the conical ediface of the volcano inducing additional depression of the electric field as discussed with Figure 4.4. A second, somewhat lesser conductor in the 5-9 km depth range appears just west of the first one, which is more subtly expressed in the model of Hill et al. In plan view at 7 km depth (Figure 4.19), we see that this steep conductor is strongly linear in a nearly N-S direction and is associated with the Mount St. Helens shear zone passing through the west flank of the volcanic

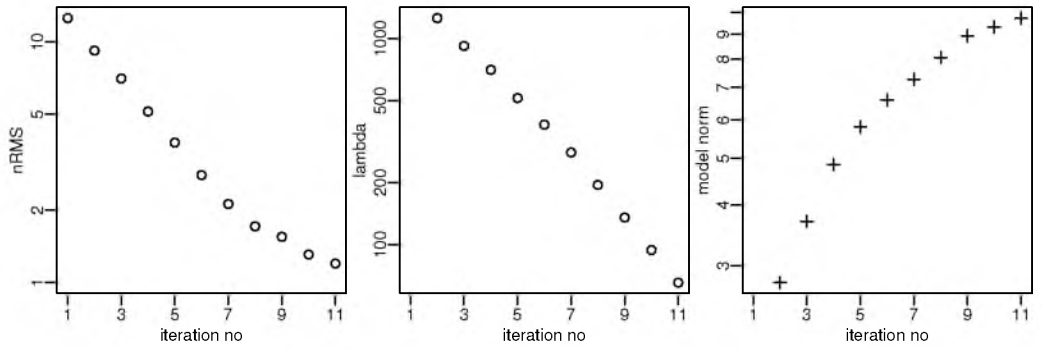

Figure 4.17. Values of nRMS, λ, and model norm as a function of iteration number for the Mount St. Helens inversion.

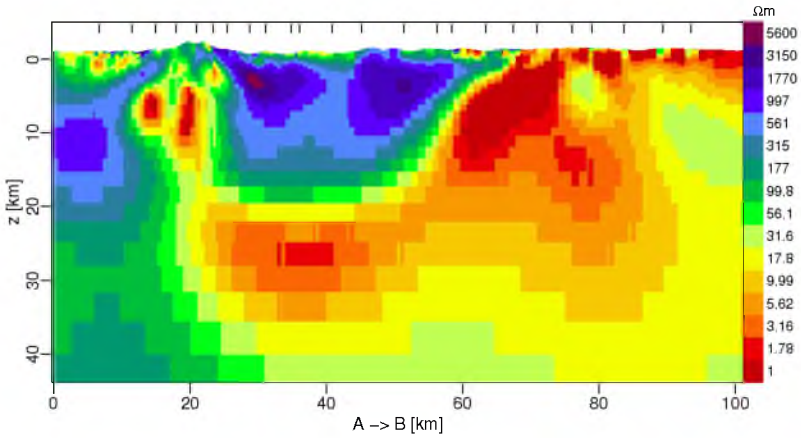

Figure 4.18. Cross-section of Mount St. Helens inversion at iteration 11 along profile A–B marked on Figure 4.15. Black ticks at the top denote the locations of receivers denoted red on Figure 4.15 that were used in 2D inversion of [18]

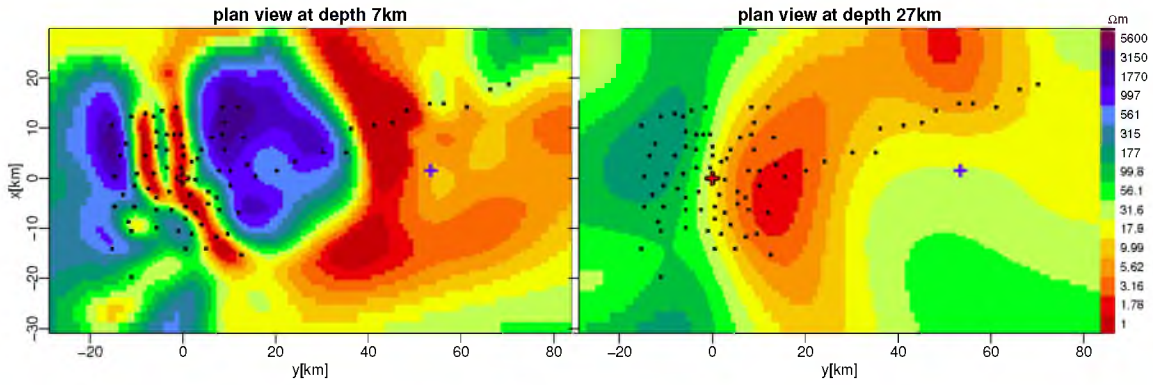

Figure 4.19. Plan views of Mount St. Helens inversion model 11. Receivers locations marked by black dots, Mount St. Helens by red cross, and Mount Adams by blue cross.

edifice [31, 36]. Clear representation of this structure in our model we believe may be due to inclusion of the tipper elements in the inversion, as the tipper shows a subtle reversal on the west flank of the volcano [18, also see our Supplemental Material Section]. The second, subsidiary conductor flanks the shear zone nearby to the west.

The large resistor east of Mount St. Helens confines the large conductor further east to be in the Mount Adams area, providing better resolution than prior 3D images based just on regional tipper data [34]. The NNW-SSE limits of the resistor cannot be considered as well-resolved, however, without site coverage. At lower crustal depths (27 km in Figure 4.19), resistivity under Mount St. Helens decreases from west to east as in [18] and the low is somewhat elongate toward the south-southeast. On the other hand, low deep crustal resistivity under Mount Adams expands to the north. It is tempting to assign this geometry to an offset in lower crustal magmatic underplating associated with the E-W offset in the Cascade volcanic chain at this latitude. However, such conjecture should await better resistivity structural constraints from further 3D MT coverage both north and south of the current data set.

4.7 Conclusions

As other researchers are finding as well, direct solutions to various aspects of the diffusive EM inversion problem are becoming increasingly practical. Here we have shown that direct solvers can effectively handle the Gauss-Newton step for inverse problems approaching one million parameters with parallelization on multicore SMP workstations and large RAM if the step is formulated in data space. Thus, with the forward problem and the Jacobian computed directly using MUMPS, our entire inversion process is done now with direct solutions. In this case, the limiting computational cost both in run time and memory is the forward problem (including the Jacobian). Finite element models of order $150\times150\times60$ elements fill a workstation with 0.5 TB RAM but such meshes can, for example, fit large MT data sets of 400 sites with five columns of parameters per site in both x and y directions with padding of 25 expanding element columns around the mesh edges. We have not experienced system conditioning problems due to high element aspect ratios with our direct solutions.

Single-box SMP workstation capabilities continue to progress, with platforms holding up to 4 TB RAM available at the time of this writing. Although scalability of MUMPS performance on multicore appears to be better than that across distibuted cluster systems [4, cf.], finite scalability at present is an impediment to exploiting machines with larger numbers of cores. We suspect that this will be helped with improvements in memory speed and latency. Otherwise, run times of significantly larger models than that for Mount St. Helens may be prohibitive. One option could be to construct a distributed cluster whose nodes were large-RAM multicore machines such as we employed herein each devoted to a different response frequency, although at considerably greater hardware investment. Finally, we find that the deformable hexahedral mesh framework lends a predictability to mesh design and performance of libraries such as MUMPS that counters concerns that the geometries of simulation with such a mesh may not be as arbitrary as is possible with assemblies of tetrahedra.

4.8 Acknowledgements

We acknowledge the support of this work from the U.S. Dept. of Energy under contract DE-EE0002750 to PW. EC acknowledges the partial support of the U.S. National Science Foundation through grants ARC-0934721 and DMS-1413454.

4.9 Appendix A: Approximation of regularization norms

Here, we present how we approximate norms $||\nabla(m - m_0)||_{L_2(1)}$,$||\nabla(m - m_0)||_{L_2(||S||_2)}$,$||\nabla(m-m_0)||_{L_2\left(\frac{1}{||S||_2}\right)}$. For simplicity, we will write m instead of $m-m_0$.

First we will consider norms $||m||_{L_2(1)}$, $||m||_{L_2(||S||_2)}$, $||m||_{L_2\left(\frac{1}{||S||_2}\right)}$. To approximate them, we will take norm of the form:

$$||m||^2_{B_m} = m^T B_m m \tag{4.26}$$

with appropriate matrix B_m, where $m = (m)_{j=1}^{N_m}$ is a vector of $\log_{10}$ resistivities of inversion voxels and it corresponds to a function $m = m(\mathbf{r})$, where $\mathbf{r}$ is a location in space. $m(\mathbf{r})$ is piecewise constant and $m(\mathbf{r}) = m_j$ whenever $\mathbf{r}$ is in the space occupied by the inversion voxel V_j.

If one takes B_m to be a diagonal matrix with entries w_j equal to volumes of inversion voxels:

$$w_j = \#V_j = \int_{V_j} d\mathbf{r} \tag{4.27}$$

then one obtains a model norm $||m||_{B_m}$ that is equal to the $L_2(1)$ norm of the model $m(\mathbf{r})$:

$$\begin{aligned} ||m||^2_{B_m} &= \sum_{j=1}^{N_m} m_j^2 B_m(j,j) = \sum_{j=1}^{N_m} m_j^2 \#V_j \\ &= \sum_{j=1}^{N_m} \int_{V_j} m(\mathbf{r})^2 d\mathbf{r} = \int_{\Omega} m(\mathbf{r})^2 d\mathbf{r} \\ &= ||m||^2_{L_2(1)} \end{aligned} \tag{4.28}$$

Consider the derivative S of infinite dimensional problem defined at (4.20). Assuming that the discretization of the domain is fine enough so that the finite dimensional approximation of the problem is close to the infinite dimensional problem, using F for finite dimensional response, one could write that the j-th column of Jacobian matrix J is:

$$J_{\cdot j} = \frac{\partial F}{\partial m_j} = \int_{V_j} S(\mathbf{r}) d\mathbf{r} \tag{4.29}$$

where V_j is a j-th inversion voxel. If we assume that the inversion voxel V_j is small enough that $S(\mathbf{r}) \approx S_j = \text{const}$ for $\mathbf{r} \in V_j$, then a sensitivity of inversion voxel V_j is obtained:

$$\begin{aligned} w_j &= \sqrt{\sum_{i=1}^{N_m} J_{ij}^2} = \sqrt{J_{\cdot j}^T J_{\cdot j}} = ||J_{\cdot j}||_2 = \left|\left| \int_{V_j} S(\mathbf{r}) d\mathbf{r} \right|\right|_2 \\ &\approx ||\#V_j S_j||_2 = \#V_j ||S_j||_2 = \int_{V_j} ||S_j||_2 d\mathbf{r} \\ &\approx \int_{V_j} ||S(\mathbf{r})||_2 d\mathbf{r} \end{aligned} \tag{4.30}$$

If we define B_m to be a diagonal matrix with w_j as entries, then

$$\begin{aligned} ||m||^2_{B_m} &= \sum_{j=1}^{N_m} m_j^2 w_j \approx \sum_{j=1}^{N_m} m_j^2 \int_{V_j} ||S(\mathbf{r})||_2 d\mathbf{r} \\ &= \int_{\Omega} ||m(\mathbf{r})||_2^2 ||S(\mathbf{r})||_2 d\mathbf{r} \\ &= ||m||^2_{L_{2(||S||_2)}} \end{aligned} \tag{4.31}$$

The regularization norm is approximately equal to the weighted L_2 norm with $||S||_2$ as a weight. Notice also that to calculate $w_j = ||J_{\cdot j}||_2$, one does not need to know the voxel volume, only Jacobian matrix J is used. This regularization was considered in [24] (see equation (3.89)).

The third weight we consider is defined as

$$w_j = \frac{(\#V_j)^2}{\sqrt{J_{.j}^T J_{.j}}} \approx \frac{(\#V_j)^2}{\#V_j ||S_j||_2} = \#V_j \frac{1}{||S_j||_2} \tag{4.32}$$

The corresponding norm of the model is approximately

$$\begin{aligned} ||m||_{B_m}^2 &= \sum_{j=1}^{N_m} m_j^2 w_j \approx \sum_{j=1}^{N_m} m_j^2 \#V_j \tfrac{1}{||S_j||_2} \\ &\approx \sum_{j=1}^{N_m} m_j^2 \int_{V_j} \tfrac{1}{||S(\mathbf{r})||_2} d\mathbf{r} \\ &= \int_\Omega ||m(\mathbf{r})||_2^2 \tfrac{1}{||S(\mathbf{r})||_2} d\mathbf{r} \\ &= ||m||_{L_2\left(\frac{1}{||S||_2}\right)}^2 \end{aligned} \tag{4.33}$$

The norm is approximately equal to the weighted L_2 norm with $\frac{1}{||S||_2}$ as a weight. This norm will suppress regions with low sensitivity, using the reasoning that if we cannot detect the properties of a region well, we will make it similar to its surroundings. This is similar to the approach of [25].

To get an approximation of a norm of the model gradient, rather than of the model, so a norm that resembles $||\nabla m||_{L_2(\nu)}$ rather than $||m||_{L_2(\nu)}$, we will do as follows. Assume that the inversion voxel consists of one finite element. Air layers as well as one layer of elements close to the boundary are not used in the inversion. As a result, the inversion voxels can be addressed using three indices $i_x = 1, \ldots, n_x$, $i_y = 1, \ldots, n_y$, $i_z = 1, \ldots, n_z$, where the total number of inversion voxels is $N_m = n_x n_y n_z$. Matrix B_m is such that

$$\begin{aligned} &||m||_{B_m}^2 = \\ &\sum_{i_x=2}^{n_x} \sum_{i_y=1}^{n_y} \sum_{i_z=1}^{n_z} \tilde{w}_{i_x,i_y,i_z}^x \left(\frac{m_{i_x,i_y,i_z} - m_{i_x-1,i_y,i_z}}{x_{i_x,i_y,i_z} - x_{i_x-1,i_y,i_z}} \right)^2 \\ &+ \sum_{i_x=1}^{n_x} \sum_{i_y=2}^{n_y} \sum_{i_z=1}^{n_z} \tilde{w}_{i_x,i_y,i_z}^y \left(\frac{m_{i_x,i_y,i_z} - m_{i_x,i_y-1,i_z}}{y_{i_x,i_y,i_z} - y_{i_x,i_y-1,i_z}} \right)^2 \\ &+ \sum_{i_x=1}^{n_x} \sum_{i_y=1}^{n_y} \sum_{i_z=2}^{n_z} \tilde{w}_{i_x,i_y,i_z}^z \left(\frac{m_{i_x,i_y,i_z} - m_{i_x,i_y,i_z-1}}{z_{i_x,i_y,i_z} - z_{i_x,i_y,i_z-1}} \right)^2 \end{aligned}$$

where

$$\begin{aligned} \tilde{w}_{i_x,i_y,i_z}^x &= \tfrac{w_{i_x-1,i_y,i_z} + w_{i_x,i_y,i_z}}{2} \\ \tilde{w}_{i_x,i_y,i_z}^y &= \tfrac{w_{i_x,i_y-1,i_z} + w_{i_x,i_y,i_z}}{2} \\ \tilde{w}_{i_x,i_y,i_z}^z &= \tfrac{w_{i_x,i_y,i_z-1j} + w_{i_x,i_y,i_z}}{2} \end{aligned}$$

and

$$\left[x_{i_x,i_y,i_z}, y_{i_x,i_y,i_z}, z_{i_x,i_y,i_z} \right]$$

is the location of the center of mass of the inversion voxel denoted by i_x, i_y, i_z.

Using the procedure described above with w given by (4.30), (4.27), or (4.32), one gets norms of model m resembling $||\nabla m||_{L_2(||S||_2)}$, $||\nabla m||_{L_2(1)}$, and $||\nabla m||_{L_2\left(\frac{1}{||S||_2}\right)}$, respectively. Those norms are used for regularization and the inversion results are compared.

4.10 Appendix B: Inversion for static distortion matrices

We present the inversion for the impedance static distortion matrices similar to the approach of [30]. Shallow conductivity structure causes a static distortion of the impedance such that

$$Z_k^{\text{obs}}(\omega) = C_k Z_k(\omega) \tag{4.34}$$

where Z_k is the impedance without the shallow conductivity structure and Z_k^{obs} is the impedance with the shallow conductivity structure. Matrix $C_k \in \mathbb{R}^{2\times 2}$ is real valued and not dependent on frequency, yet different for each receiver $k = 1, \ldots, N_{\text{rec}}$ [see 30].

In the inversion procedure, apart from calculating the unknown model $m = (m_j)_{j=1}^{N_m}$ of $\log_{10}$ resistivities, we invert also for real valued matrices $C = (C_k)_{k=1}^{N_{\text{rec}}}$, one for each receiver location.

The forward problem response $F(m)$, defined by (4.5), is modified by applying (4.34) to obtain $F(m, C)$. The regularized functional to be minimized changes from (4.8) by adding squares of Frobenius norms $||.||_F$ of the difference between distortion matrices C_k and identity matrix I, yielding:

$$\begin{aligned} \tilde{W}(m, C) = (F(m, C) - d)^T B_d (F(m, C) - d) + \\ \lambda (m - m_0)^T B_m (m - m_0) + \tau \sum_{k=1}^{N_{\text{rec}}} ||C_k - I||_F^2 \end{aligned} \tag{4.35}$$

Notice that if we define $\tilde{N}_m = N_m + 4N_{\text{rec}}$,

$$\begin{aligned} \tilde{m} = (\tilde{m}_k)_{k=1}^{\tilde{N}_m} = \; & (m_1, \ldots, m_{N_m}, \\ & C_{1,xx}, C_{1,yx}, C_{1,yx}, C_{1,yy}, \ldots, \\ & C_{N_{\text{rec}},xx}, C_{N_{\text{rec}},yx}, C_{N_{\text{rec}},yx}, C_{N_{\text{rec}},yy}) \\ \tilde{m}_0 = (\tilde{m}_{k,0})_{k=1}^{\tilde{N}_m} = \; & (m_{1,0}, \ldots, m_{N_m,0}, \\ & 1, 0, 0, 1, \ldots, \\ & 1, 0, 0, 1) \\ \tilde{B}_m = \; & \begin{bmatrix} B_m & 0 \\ 0 & \frac{\tau}{\lambda} I \end{bmatrix} \end{aligned}$$

Then (4.35) may be written in the form similar to (4.8):

$$\tilde{W}(\tilde{m}) = (F(\tilde{m}) - d)^T B_d(\tilde{F}(m) - d) + \lambda(\tilde{m} - \tilde{m}_0)^T \tilde{B}_m(\tilde{m} - \tilde{m}_0)$$

Jacobian $\tilde{J}$ of the forward response $F(\tilde{m})$ may be easily obtained from J using chain rule. As a result, one can apply Gauss-Newton and data space Gauss-Newton procedure similarly to the case of inversion for m only.

Similarly to [30], we use $\tau = 0.01$. Notice that this value of τ is very small, giving almost no regularization for distortion matrices term in (4.35). Yet it is enough to obtain good models, if only the starting model is not far from the true model. It is our experience so far that using the starting model that was obtained in the inversion without the distortion matrix yields good results. On the other hand, if one starts from half space that is far from the true model, the iteration may not converge to a plausible model. In this case, we have seen the matrices C converge to 0.

4.11 Supplementary materials

We present here additional materials related to the Mount St. Helens inversion. In Figure 4.20, we show the induction vectors, where Parkinson convention is used (real and imaginary part of $-K$ is plotted), so the vectors points towards the conductors. In Figure 4.21, we present the final nRMS as a function of receiver location and MT response component.

In Table 4.4, we list the values of nRMS for all frequencies for the starting and the final models. For the half-space starting model, the initial nRMS values at the upper frequencies are about one-half those of the higher frequencies. This carries through to the relative nRMS values for the final model. We expect that if a starting model was used where the initial nRMS was more evenly distributed, the final nRMS would be more even as well. Such a model might be a depth profile obtained through 1D inversion of the invariant Zxy-Zyx integrated over the survey area [see e.g., 35].

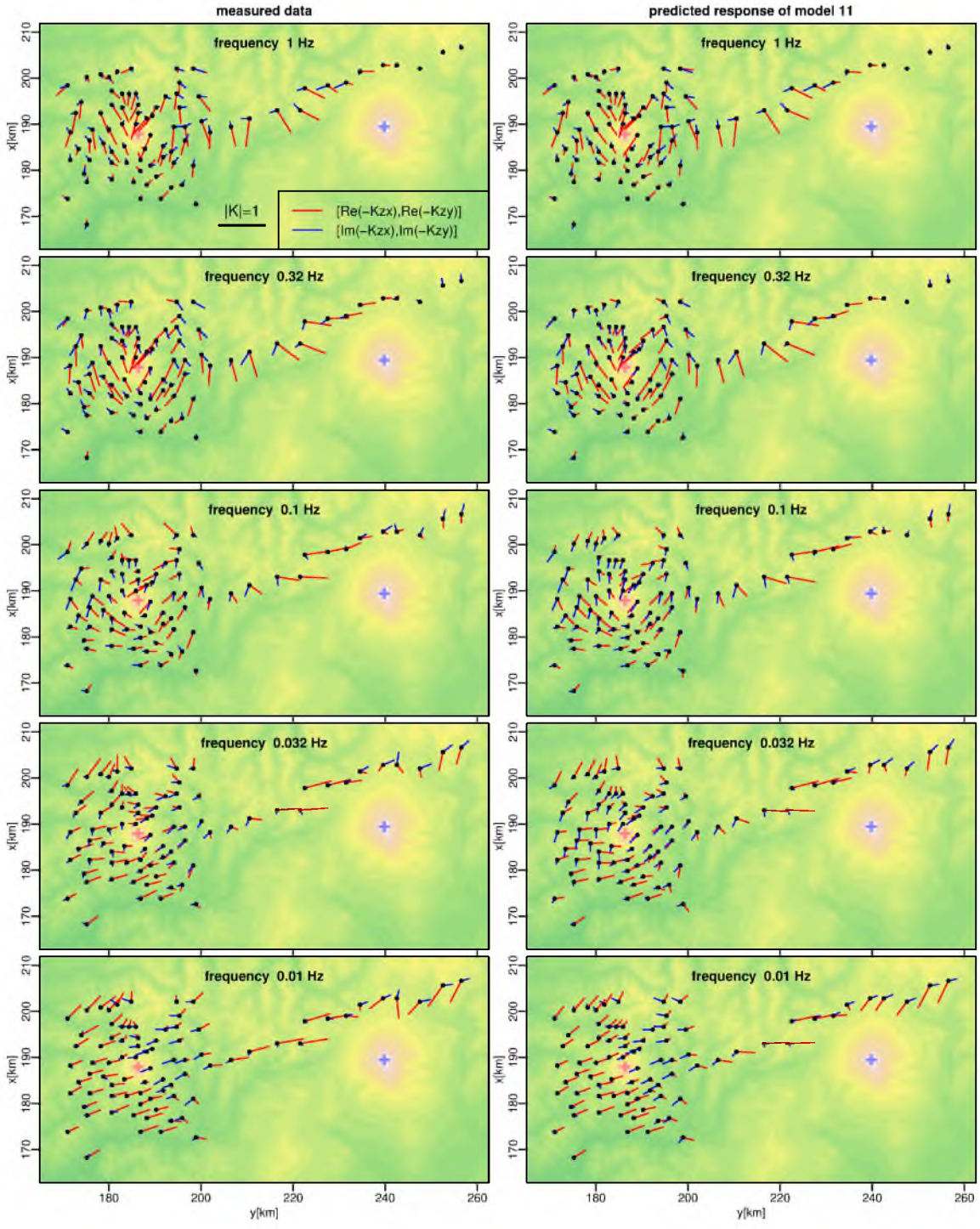

Figure 4.20. Real and imaginary induction vectors for the measured data and the prediction by the model obtained in the Mount St. Helens inversion for five frequencies between 1Hz and 0.01Hz. Parkinson convention is used.

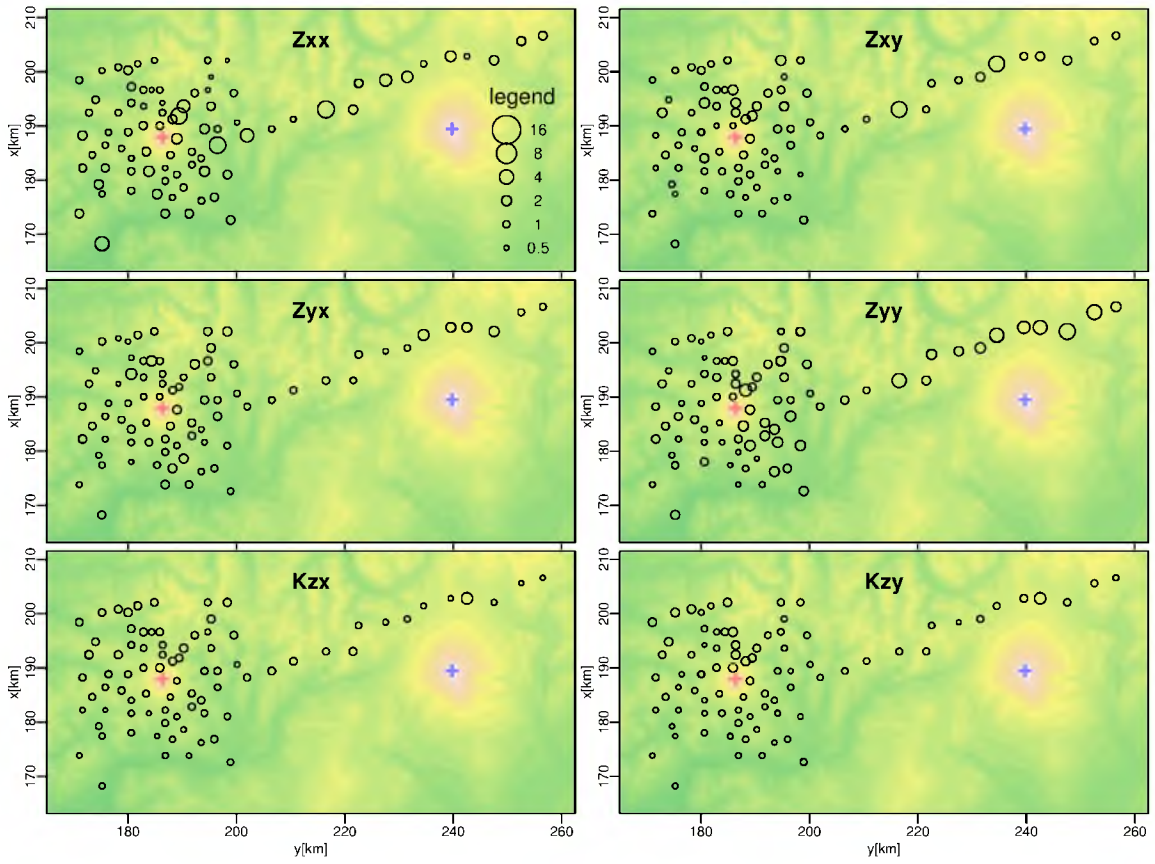

Figure 4.21. nRMS for each component of the MT response and for each receiver for the final model of Mount St. Helens inversion.

Table 4.4. Table of nRMS as a function of frequency for the starting model and the final model for Mount St. Helens inversion.

Frequency [Hz]	Model 1	Model 11
100	8.1	0.85
56	8.2	0.7
32	8.3	0.7
18	8.7	0.79
10	9.2	0.94
5.6	9.4	0.97
3.2	9.3	1.02
1.8	9.7	1.33
1	10.2	1
0.56	10.8	1.15
0.32	11.8	1.07
0.18	12.4	1.32
0.1	12.4	1.2
0.056	12.9	1.22
0.032	13.8	1.38
0.018	15.7	1.53
0.01	17.9	1.49
0.0056	19	1.47
0.0032	17.3	1.62
0.0018	16.4	1.57

4.12 References

[1] M. Kordy, P. Wannamaker, V. Maris, and E. Cherkaev, "Three-dimensional magnetotelluric inversion including topography using deformed hexahedral edge finite elements and direct solvers parallelized on SMP computers, Part I: forward problem and parameter jacobians," *submitted to Geophys. J. Int.*, 2014.

[2] J. Liu, M. Brio, and J. Moloney, "Overlapping Yee FDTD method on nonorthogonal grids," English, *J. Sci. Comput.*, vol. 39, no. 1, pp. 129–143, 2009, ISSN: 0885-7474. DOI: 10.1007/s10915-008-9253-1. [Online]. Available: http://dx.doi.org/10.1007/s10915-008-9253-1.

[3] M. A. Stark, S. W., H. S., and M. D. Watts, "Distortion effects on magnetotelluric sounding data investigated by 3d modeling of high-resolution topography," *Geothermal Resources Council Trans.*, vol. 37, pp. 521–527, 2013.

[4] A. V. Grayver, R. Streich, and O. Ritter, "Three-dimensional parallel distributed inversion of CSEM data using a direct forward solver," *Geophys. J. Int.*, vol. 193, pp. 1432–1446, 2013.

[5] N. M. Meqbel, G. D. Egbert, P. E. Wannamaker, A. Kelbert, and A. Schultz, "Deep electrical resistivity structure of the northwestern U.S. derived from 3-D inversion of USArray magnetotelluric data," *Earth and Planetary Science Letters*, dx.doi.org/10.1016/j.epsl.2013.12.026, 2014.

[6] C. deGroot Hedlin and S. Constable, "Occams inversion to generate smooth two-dimensional models from magnetotelluric data," *Geophysics*, vol. 55, pp. 1613–1624, 1990.

[7] K. Key and S. Constable, "Coast effect distortion of marine magnetotelluric data: insights from a pilot study offshore northeastern Japan," *Phys. Earth Planet. Inter.*, vol. 184, pp. 194–207, 2011.

[8] D. W. Oldenburg, E. Haber, and R. Shekhtman, "Three dimensional inversion of multi-source time domain electromagnetic data," *Geophysics*, vol. 78, no. 1, E47–E57, 2013.

[9] W. Siripunvaraporn, G. Egbert, Y. Lenbury, and M. Uyeshima, "Three-dimensional magnetotelluric inversion: data-space method," *Phys. Earth Planet. Inter.*, vol. 150, pp. 3–14, 2005.

[10] Y. Sasaki, "Full 3-D inversion of electromagnetic data on PC," *J. Appl. Geophys.*, vol. 46, pp. 45–54, 2001.

[11] V. Maris and P. E. Wannamaker, "Parallelizing a 3D finite difference MT inversion algorithm on a multicore PC using OpenMP," *Computers & Geosciences*, doi: 10.1016/j.cageo.2010. 03.001, 5 pp. 2010.

[12] M. Commer and G. A. Newman, "New advances in three-dimensional controlled-source electromagnetic inversion," *Geophys. J. Int.*, vol. 172, pp. 513–535, 2008.

[13] M. S. Zhdanov, L. Wan, A. Gribenko, M. Cuma, K. Key, and S. Constable, "Large-scale 3D inversion of marine magnetotelluric data: case study from the Gemini prospect, Gulf of Mexico," *Geophysics*, vol. 76, F77–F87, 2011.

[14] C. Schwarzbach and E. Haber, "Finite element based inversion for time-harmonic electromagnetic problems," *Geophys. J. Int.*, vol. 193, pp. 615–634, 2013.

[15] R. L. Parker, *Geophysical Inverse Theory*. Princeton University Press, Princeton, NJ, 386 pp, 1994.

[16] W. Siripunvaraporn and G. Egbert, "An efficient data-subspace inversion method for 2-D magnetotelluric data," *Geophysics*, vol. 65, pp. 791–803, 2000.

[17] M. P. Miensopust, P. Queralt, A. G. Jones, and the 3D Modelers, "Magnetotelluric 3-D inversiona review of two successful workshops on forward and inversion code testing and comparison," *Geophys. J. Int.*, vol. 193, pp. 1216–1238, 2013.

[18] G. J. Hill, T. G. Caldwell, W. Heise, D. G. Chertkoff, H. M. Bibby, M. K. Burgess, J. P. Cull, and R. A. F. Cass, "Distribution of melt beneath Mount St Helens and Mount Adams inferred from magnetotelluric data," *Nature Geosci.*, vol. 2, pp. 785–789, 2009.

[19] G. W. Hohmann and A. P. Raiche, "Inversion of controlled-source electromagnetic data," in *Electromagnetic methods in applied geophysics*, M. N. Nabighian, Ed., Tulsa, OK: Soc. Expl. Geophys., 1988, pp. 469–503.

[20] M. Baboulin, D. Becker, and J. Dongarra, "A parallel tiled solver for dense symmetric indefinite systems on multicore architectures," *University of Tennessee Computer Science Technical Report, ICL-UT-11-07*, 2011.

[21] J. Kurzak, P. Luszczek, A. YarKhan, M. Faverge, J. Langou, H. Bouwmeester, and J. Dongarra, "Multithreading in the PLASMA library," *Multi and Many-Core Processing: Architecture, Programming, Algorithms, & Applications, Ahmed, M., Ammar, R., Rajasekaran, S. eds. Taylor & Francis*, 2013.

[22] M. Baboulin, L. Giraud, and S. Gratton, "A parallel distributed solver for large dense symmetric systems: applications to geodesy and electromagnetism problems," *International Journal of High Performance Computing Applications*, vol. 19, no. 4, pp. 353–363, 2005.

[23] M. Kordy, V. Maris, P. Wannamaker, and E. Cherkaev, "3D edge finite element solution for scattered electric field using a direct solver parallelized on an SMP workstation," in *5th International Symposium on Three-Dimensional Electromagnetics, Sapporo, May 7-9*, 2013, p. 4.

[24] M. S Zhdanov, *Geophysical inverse theory and regularization problems*. Elsevier, 2002.

[25] M. J. Yi, J. H. Kim, and S. H. Chung, "Enhancing the resolving power of least-squares inversion with active constraint balancing," *Geophysics*, vol. 68, no. 3, pp. 931–941, 2003.

[26] R. W. Groom and K. Bahr, "Corrections for near surface effects: decomposition of the magnetotelluric impedance tensor and scaling corrections for regional resistivities: a tutorial," *Surveys in Geophysics*, vol. 13, no. 4-5, pp. 341–379, 1992.

[27] P. Wannamaker, J. Stodt, and L. Rijo, "Two-dimensional topographic responses in magnetotelluric modeling using finite elements," *Geophysics*, vol. 51, pp. 2131–2144, 1986.

[28] M. Miensopust, "Multidimensional magnetotellurics: a 2d case study and a 3d approach to simultaneously invert for resistivity structure and distortion parameters," *PhD thesis*, 2010.

[29] A. Avdeeva, M. M., and A. D., "Three-dimensional joint inversion of magnetotelluric tensor data and full distortion matrix," *Extended abstract, 21st Biennial Workshop on EM Induction in the Earth, Darwin, Australia, July 25-21*, p. 4, 2012.

[30] A. Avdeeva, M. Moorkamp, D. Avdeev, M. Jegen, and M. Miensopust, "Three-dimensional inversion of magnetotelluric impedance tensor data and full distortion matrix," *Geophysical Journal International*, in press, 2014.

[31] A. M. F. Lagmay, B. van Wyk de Vries, N. Kerle, and D. M. Pyle, "Volcano instability induced by strike-slip faulting," *Bull. Volcanol.*, vol. 62, pp. 331–346. 2000.

[32] W. Siripunvaraporn, G. Egbert, and U. M., "Interpretation of two-dimensional magnetotelluric profile data with three-dimensional inversion: synthetic examples," *Geophys. J. Int.*, doi: 10.1111/j.1365–246X.2005. 02527.x. 2005.

[33] W. D. Stanley, S. Y. Johnson, A. Y. Qamar, C. S. Weaver, and J. M. Williams, "Tectonics and seismicity of the southwest Washington Cascade Range," *Bull. Seis. Soc. Amer.*, vol. 86, pp. 1–18, 1996.

[34] P. G. D. Egbert and J. R. Booker, "Imaging crustal structure in southwestern washington with small magnetometer arrays," *J. Geophys. Res.*, vol. 98, pp. 15,967–15,985, 1993.

[35] P. E. Wannamaker, R. L. Evans, P. A. Bedrosian, M. J. Unsworth, V. Maris, and R. S. McGary, "Segmentation of plate coupling, fate of subduction fluids, and modes of arc magmatism in Cascadia, inferred from magnetotelluric resistivity," *Geochemistry, Geophysics, Geosystems, in revision.*, 2014.

[36] C. Weaver, W. C. Grant, and J. E. Shemeta, "Local crustal extension at Mount St. Helens, Washington," *J. Geophys. Res.*, vol. 92, 10, pp. 170–10,178. 1987.

CHAPTER 5

FORWARD AND INVERSE MULTIPLE FREQUENCY PROBLEM FOR MAXWELL'S EQUATIONS USING ADAPTIVE MODEL ORDER REDUCTION[1]

Kordy M.[2,3], Cherkaev E.[2], and Wannamaker P.[3]

5.1 Abstract

This work develops a model order reduction method for numerical solution of forward and inverse multifrequency eddy current problem. Using Helmholz decomposition, we extend previously developed technique to the case when the operator has a non-empty null space. In the case of finite element discretization of Maxwell's equations with edge elements, the discrete Helmholtz decomposition is accomplished by solving a Poisson equation on nodal elements of the same grid. Exploiting analyticity of the electromagnetic field, we use Pade interpolation in the complex frequency plane; this allows us to approximate the forward solution as well as the frequency-dependent Jacobian in the inversion procedure. To adaptively choose interpolating frequencies, we propose to minimize the maximal approximation error. We discuss several error estimates and propose a fast method of calculating the residual across a range of frequencies. The efficiency of the developed approach is demonstrated by applying it to the forward and inverse magnetotelluric problem, which is a

[2]Department of Mathematics, University of Utah

[3]Energy & Geoscience, University of Utah

geophysical electromagnetic remote sensing method used in mineral, geothermal, and groundwater exploration. Numerical tests show excellent performance of the proposed methods characterized by a significant reduction of computational time without loss of accuracy.

5.2 Introduction

Model order reduction (MOR) is a powerful technique to reduce dimensionality of a problem. It has become popular recently and has been used in a variety of contexts [1–7]. This work is inspired mainly by the work of [8, 9], where the authors consider adaptive choice of shifts for approximation of the transfer function using rational Krylov subspaces, with application to a time-domain electromagnetic geophysical forward problem. Their work is followed by the application to the inverse problem [10, 11], which considers a regularization through a small admissible set of conductivity models.

Here we develop a different approach. We consider the Gauss-Newton method for the minimization of the inversion functional and we use model order reduction through rational Krylov subspaces only to speedup the calculation of the transfer function used in the forward problem as well as a transfer function used in the calculation of the Jacobian. The application of interest is magnetotellurics (MT), which is a frequency domain electromagnetic remote-sensing geophysical method used in mineral, geothermal, ad groundwater exploration. In this case, the transfer function $h(s) = (A + sI)^{-1}b$ for the forward problem has a complex valued right-hand side (rhs) b dependent on frequency. In the case of the calculation of the Jacobian, the rhs b is not dependent on frequency and is real valued.

The rational Krylov subspaces are build from the values of the transfer function, evaluated at a number of points, called here interpolating shifts. In our application the transfer function values are needed in a purely imaginary interval, so we consider interpolating shifts in the same imaginary interval. In the case of real valued rhs b, we consider also real shifts, following the suggestion of [9].

The interpolating shifts are chosen to minimize the maximal error of interpolation and as the true error is unknown, an error indicator is needed. We consider error indicator suggested by [9] and compare it with the residual suggested as the error

indicator by [12], which turns out to be superior in our numerical tests. We propose a fast way to calculate the residual at multiple frequencies, which makes it a practical error indicator.

A matrix A, arising in discretization of the variational problem has a nonempty nullspace. This is not a problem in the case of $b \perp \text{null}(A)$. Yet if $b \not\perp \text{null}(A)$, we propose to decompose b into the part in $\text{null}(A)$ and the part orthogonal to it. We calculate the part of the transfer function h that is in the null space exactly and use model order reduction only for the part orthogonal to $\text{null}(A)$. This procedure is particularly useful in the case of rhs b dependent on frequency (the forward problem case), making the approximation better by two orders of magnitude. In our application, the decomposition of rhs b is possible using Helmholtz decomposition on a discrete level, which is valid for mimetic finite element approximation through edge elements [13, 14].

We present a simple theorem of a lucky failure of the model order reduction algorithm for the case of rhs b not dependent on frequency. For the case of b dependent on the frequency, we prove that a failure will almost always be lucky, if only b is an analytic function of the frequency.

In our application, the transfer function is of the form $\tilde{h}(s) = (\tilde{A} + s\tilde{B})^{-1}\tilde{b}$, where $\tilde{B}$ is not the identity matrix. We show that it is related with the case of $\tilde{B} = I$ through a simple scaling using $\tilde{B}^{\frac{1}{2}}$, that need not be calculated in the approximation procedure, but is useful to simplify the analysis.

In the numerical tests, the speedup of applying model order reduction gets better when the number of frequencies considered in MT survey increases. For 30 frequencies, the speedup is 2 times for the forward problem and 4 times for the Jacobian.

The paper is organized as follows. In Section 5.3, we present the magnetotelluric formulation of the forward and inverse problem explaining how approximation of the transfer function may be used to speedup the calculations. Next we show the theory of approximation of the transfer function using rational Krylov subspaces for the case of b dependent and not dependent on frequency. Then we propose the idea for the treatment of $\text{null}(A)$.

In Section 5.4, we present the error indicator functions and algorithms based on

them. We also give details of the numerical implementation. In particular, we propose a fast way to calculate the residual. In Section 5.5, we show results of numerical tests for a 3D magnetotelluric model with non-constant conductivity structure and with a hill and a valley in topography.

5.3 Theory

5.3.1 Numerical formulation of magnetotelluric problem

The forward and inverse magnetotelluric problem is described in detail in [15, 16]. We consider a domain Ω that includes the air and earth's subsurface. The earth's surface is allowed to have topography. In order to calculate the MT response due to an arbitrary 3D conductivity structure $\sigma > 0$, we consider edge finite element discretization of the equation for the secondary electric field E. Though the numerical tests presented are done using lowest order edge hexahedral discretization, all the methods may be applied to tetrahedral mesh. Higher order edge elements may be used as well.

Define the solution space for the unknown electric field

$$\mathcal{H}_0(\nabla\times,\Omega) = \left\{F{:}\Omega \to \mathbb{C}^3 : \textstyle\int_\Omega \left(|F|^2 + |\nabla \times F|^2\right) < \infty,\ n \times F|_{\partial\Omega} = 0\right\} \tag{5.1}$$

Consider Maxwell's equation in frequency domain for low frequency, where the term $\mathrm{i}\omega\epsilon$, related to displacement current, is neglected. Assuming further that E is a secondary field, the equation for E is

$$\int_\Omega \frac{1}{\mu}\nabla \times E \cdot \nabla \times F + \mathrm{i}\omega \int_\Omega \sigma E \cdot F = \int_\Omega -\mathrm{i}\omega(\sigma - \sigma^p)E^p \cdot F \tag{5.2}$$

for $E, F \in \mathcal{H}_0(\nabla\times)$, where the source term depends on E^p, the primary electric field which is a plane wave traveling in a primary conductivity structure σ_p, the conductivity of a 1D earth. We assume that $\sigma \approx \sigma_p$ close to the domain boundaries. We denote angular frequency by ω and magnetic permeability by μ. Most of the methods presented may be adapted to the case when the term $\mathrm{i}\omega\epsilon$ is present.

The electric field over Ω is represented as a linear combination of the edge shape functions S_i with coefficients ξ_i:

$$E = \sum_{i=1}^{N} \xi_i S_i \tag{5.3}$$

where $i = 1, \ldots, N$ are indices of the edges that do not lie on the boundary. By substituting this to equation (5.2) and using S_j as test functions, one obtains a linear system

$$(\tilde{A} + \mathrm{i}\omega\tilde{B})\xi = g \tag{5.4}$$

$$\tilde{A}_{i,j} = \int_\Omega \frac{1}{\mu}\nabla \times S_i \cdot \nabla \times S_j, \quad \tilde{B}_{i,j} = \int_\Omega \sigma S_i \cdot S_j \tag{5.5}$$

$$g_i = g_i(\omega, \sigma) = \int_\Omega -\mathrm{i}\omega(\sigma - \sigma^p)E^p \cdot S_i \tag{5.6}$$

Secondary magnetic field H is calculated as

$$H = \frac{-\nabla \times E}{i\omega} \tag{5.7}$$

The total field E^t, H^t is a sum of secondary and primary fields:

$$E^t = E + E^p, \quad H^t = H + H^p \tag{5.8}$$

The MT response is obtained by finding impedance Z and tipper K such that

$$\begin{bmatrix} E_x^t \\ E_y^t \\ H_z^t \end{bmatrix} = \begin{bmatrix} Z_{xx} & Z_{xy} \\ Z_{yz} & Z_{yy} \\ K_{zx} & K_{zy} \end{bmatrix} \begin{bmatrix} H_x^t \\ H_y^t \end{bmatrix} \tag{5.9}$$

is satisfied no matter what is the polarization of the primary (E^p, H^p) plane wave.

A receiver can be positioned at an arbitrary location $\mathbf{r}$ with respect to element edges via appropriate interpolation. In general, let $\mathbf{r}$ be inside an element with edges $e_1, \ldots, e_{12}$. Then field E at location $\mathbf{r}$ is given by

$$E(\mathbf{r}) = \sum_{l=1}^{12} S_{e_l}(\mathbf{r})\xi_{e_l} = \begin{bmatrix} (v_x^E)^T\xi \\ (v_y^E)^T\xi \\ (v_z^E)^T\xi \end{bmatrix} \tag{5.10}$$

Here v_x^E, v_y^E, v_z^E contain interpolation vectors with at most 12 non-zero values corresponding to x, y, and z components of edge shape functions $S_{e_1}(\mathbf{r}), \ldots, S_{e_{12}}(\mathbf{r})$.

Similarly, the secondary magnetic field $H(\mathbf{r})$, calculated using (5.7) at location $\mathbf{r}$, is given by

$$H(\mathbf{r}) = \sum_{l=1}^{12} \frac{\nabla \times S_{e_l}(\mathbf{r})}{-i\omega\mu} \xi_{e_l} = \frac{1}{i\omega} \begin{bmatrix} (v_x^H)^T \xi \\ (v_y^H)^T \xi \\ (v_z^H)^T \xi \end{bmatrix} \tag{5.11}$$

This time the only non-zero values of v_x^H, v_y^H, v_z^H are x, y, and z components of

$$\left(\frac{\nabla \times S_{e_1}(\mathbf{r})}{-\mu}, \ldots, \frac{\nabla \times S_{e_{12}}(\mathbf{r})}{-\mu} \right)$$

As a result, each component of secondary electric and magnetic fields E, H at a specific receiver location may be represented using

$$v^T \xi = v^T \left[(\tilde{A} + \mathrm{i}\omega\tilde{B})^{-1} g \right] \tag{5.12}$$

where v is a real valued vector, $g = g(\omega)$ is complex valued, and one should write $\frac{g}{\mathrm{i}\omega}$ in place of g in the case of H. The calculation is done in such a way that the quantity in square brackets is evaluated first and then it is multiplied by v^T. This gives us the values of electric and magnetic secondary fields at a receiver location, which are enough to calculate the MT response at this location.

In the inversion of the MT data, one seeks a conductivity model that fits the measured data. One of the ways to find it is to use Gauss-Newton algorithm (see for example [16]) to minimize the functional that consists of data misfit and the regularization term. In order to use this algorithm, one needs the Jacobian J of the response functional. Let the domain be split into N_m inversion cells C_j, which form a partition of the subsurface part of the domain (the domain excluding the air). We assume that each cell consists of a number of finite elements and the conductivity σ is constant in each cell. Let σ_j denote the value of the conductivity in an inversion cell C_j. The conductivity in the whole earth's subsurface is given by the vector of conductivities $(\sigma_j)_{j=1}^{N_m}$. Each entry of Jacobian matrix J consists of values of the derivative of the magnetotelluric response Z, K at some receiver location $\mathbf{r}$ with respect to σ_j, for some $\mathbf{r}$ and some $j = 1, \ldots, N_m$. The response is a function of secondary electromagnetic field E, H, so its derivative with respect to σ_j may be calculated using chain rule, if the derivatives

$$\frac{\partial E}{\partial \sigma_j}, \quad \frac{\partial H}{\partial \sigma_j}$$

are found. In order to find the latter, one has to evaluate the derivative of expression (5.12):

$$\begin{aligned}
&\frac{\partial}{\partial\sigma_j}\left(v^T\left((\tilde{A}+\mathrm{i}\omega\tilde{B})^{-1}g\right)\right) = v^T\frac{\partial\left((\tilde{A}+\mathrm{i}\omega\tilde{B})^{-1}g\right)}{\partial\sigma_j} = \\
&v^T\left(\frac{\partial(\tilde{A}+\mathrm{i}\omega\tilde{B})^{-1}}{\partial\sigma_j}g + (\tilde{A}+\mathrm{i}\omega\tilde{B})^{-1}\frac{\partial g}{\partial\sigma_j}\right) = \\
&v^T\left(-(\tilde{A}+\mathrm{i}\omega\tilde{B})^{-1}\frac{\partial(\tilde{A}+\mathrm{i}\omega\tilde{B})}{\partial\sigma_j}(\tilde{A}+\mathrm{i}\omega\tilde{B})^{-1}g + (\tilde{A}+\mathrm{i}\omega\tilde{B})^{-1}\frac{\partial g}{\partial\sigma_j}\right) = \\
&v^T\left(-(\tilde{A}+\mathrm{i}\omega\tilde{B})^{-1}\frac{\partial\mathrm{i}\omega\tilde{B}}{\partial\sigma_j}\xi + (\tilde{A}+\mathrm{i}\omega\tilde{B})^{-1}\frac{\partial g}{\partial\sigma_j}\right) = v^T\left((\tilde{A}+\mathrm{i}\omega\tilde{B})^{-1}\left(-\frac{\partial\mathrm{i}\omega\tilde{B}}{\partial\sigma_j}\xi + \frac{\partial g}{\partial\sigma_j}\right)\right) = \\
&\left(v^T(\tilde{A}+\mathrm{i}\omega\tilde{B})^{-1}\right)\left(-\frac{\partial\mathrm{i}\omega\tilde{B}}{\partial\sigma_j}\xi + \frac{\partial g}{\partial\sigma_j}\right) = \left[(\tilde{A}+\mathrm{i}\omega\tilde{B})^{-1}v\right]^T\left(-\frac{\partial\mathrm{i}\omega\tilde{B}}{\partial\sigma_j}\xi + \frac{\partial g}{\partial\sigma_j}\right)
\end{aligned} \tag{5.13}$$

where ξ and g should be replaced with $\frac{\xi}{\mathrm{i}\omega}$ and $\frac{g}{\mathrm{i}\omega}$ in the case of H. The calculation of the quantity above is done in such a way that the expression in square brackets is evaluated first and then it is multiplied by $\left(-\frac{\partial\mathrm{i}\omega\tilde{B}}{\partial\sigma_j}\xi + \frac{\partial g}{\partial\sigma_j}\right)$ for each σ_j. Notice that one has to find ξ before the multiplication is done and that both the matrix $\frac{\partial\mathrm{i}\omega\tilde{B}}{\partial\sigma_j}$ and the vector $\frac{\partial g}{\partial\sigma_j}$ are sparse.

Magnetotelluric response Z, K is a smooth function of frequency, so it may be efficiently interpolated between frequencies. Such an interpolation has been considered [17]. In Figure 5.1, we present values of Z_{xx} in the complex plane when the frequency is changed. One can see that Z_{xx} is a quite complicated function of frequency; piecewise linear or high order polynomial interpolation is not accurate enough. Model order reduction interpolation, which is the topic of this paper, is much more appropriate.

5.3.2 Model order reduction

Values of the E, H fields (5.12) and the Jacobian (5.13) may be approximated in such a way that the vectors in square brackets in (5.12), (5.13) are approximated first, and then they are multiplied by the remaining part of expressions (5.12), (5.13). Thus in both cases, we are interested in approximation of the expression

$$\tilde{h}(s) = (\tilde{A}+s\tilde{B})^{-1}\tilde{b} \tag{5.14}$$

Here $s = \mathrm{i}\omega$ for some chosen finite number of frequencies ω at which the MT measurements are taken. Those frequencies are usually log-uniformly distributed in an interval $[\omega_{\min}, \omega_{\max}]$. In the case of (5.12), $\tilde{b} = g(\omega)$, so it is dependent on ω

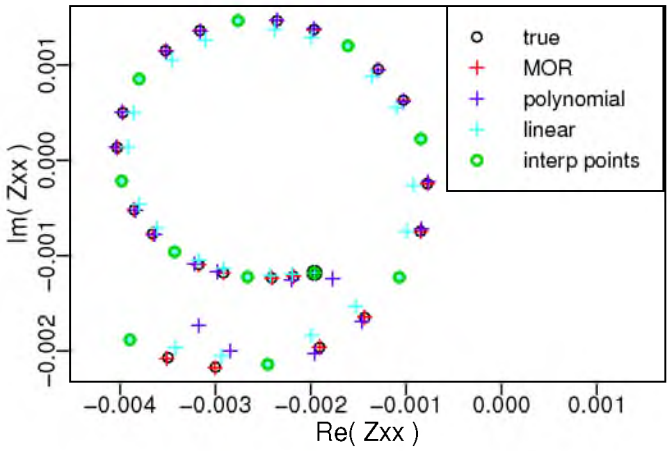

Figure 5.1. Z_{xx} in the complex plane for a model considered in Section 5.5 for 31 frequencies log-uniformly distributed in the interval [1Hz, 1000Hz]. True values are shown together with high order polynomial, piecewise linear, and model order reduction interpolation. Every third value (shown in green) is used as an interpolation point.

and complex valued. In the case of (5.13), $\tilde{b} = v$, so it is not dependent on ω and real valued. Matrix $\tilde{A}$, defined at (5.5), is real valued symmetric, non-negative definite, with a significant null space. We assume that the conductivity $\sigma > 0$, so matrix $\tilde{B}$, defined at (5.5), is real valued, symmetric positive definite.

We will consider approximation of (5.14) using the model order reduction method [1–11]. Some of its theory is presented next.

Let us start with the equation satisfied by $\tilde{h}$:

$$(\tilde{A} + s\tilde{B})\tilde{h} = \tilde{b} \tag{5.15}$$

Consider $\tilde{V}$, which is a $N \times n$ matrix whose columns span the space

$$\text{colsp}(\tilde{V}) = \text{span}\left\{(\tilde{A} + s_1\tilde{B})^{-1}\tilde{b}, (\tilde{A} + s_2\tilde{B})^{-1}\tilde{b}, \ldots, (\tilde{A} + s_n\tilde{B})^{-1}\tilde{b}\right\} \tag{5.16}$$

for some complex values $s_1, \ldots, s_n$, which satisfy

$$s_i \neq s_j \text{ if } i \neq j \tag{5.17}$$

As $\tilde{A}$ is non-negative definite and $\tilde{B}$ is positive definite, eigenvalues of $\tilde{B}^{-\frac{1}{2}}\tilde{A}\tilde{B}^{-\frac{1}{2}}$ are in $[0, \infty)$. Thus, in order for the equation (5.14) to have a solution, we assume that

$$s,\ s_j \notin (-\infty, 0] \tag{5.18}$$

If we consider approximation of the solution of (5.15) by a vector in $\text{colsp}(\tilde{V})$, namely $\tilde{h}_{\tilde{V}} = \tilde{V}\beta$, for $\beta \in \mathbb{C}^n$ and if we make the residual orthogonal to $\text{colsp}(\tilde{V})$, then we get an equation for β:

$$\tilde{V}^*(\tilde{A} + s\tilde{B})(\tilde{V}\beta) = \tilde{V}^*\tilde{b} \tag{5.19}$$

and if the equation has a unique solution (see Theorem 12), then we obtain the approximation $h_{\tilde{V}}(s) = \tilde{V}\beta$ to $h(s)$:

$$\tilde{h}_{\tilde{V}}(s) = \tilde{V}\left(\tilde{V}^*(\tilde{A} + s\tilde{B})\tilde{V}\right)^{-1}\tilde{V}^*\tilde{b} \tag{5.20}$$

We assume that $n << N$, so equation (5.19) is much easier to solve than equation (5.15). In a simplest case $s_j = \mathrm{i}\omega_j$, for $\omega_j \in [\omega_{\min}, \omega_{\max}]$. In this case, $\omega_1, \ldots, \omega_n$ may be called interpolating frequencies as the following theorem holds:

Theorem 12 *If the matrix $\tilde{V}$ satisfying (5.16) is full rank, then for $s \in \mathbb{C}$, the solution to (5.19) exists and is unique. Moreover,*

$$\tilde{h}_{\tilde{V}}(s_j) = \tilde{h}(s_j), \quad j = 1, \ldots, n \tag{5.21}$$

Proof : First we show that the matrix in equation (5.19) is not singular. Indeed

$$\tilde{V}^*(\tilde{A} + s\tilde{B})\tilde{V} = \tilde{V}^*\tilde{A}\tilde{V} + \tilde{V}^*s\tilde{B}\tilde{V} = A_1 + sB_1$$

As $\tilde{V}$ is full rank and $\tilde{A}$ is symmetric non-negative definite, A_1 is hermitian, non-negative definite. Similarly, as B is symmetric positive definite, B_1 is hermitian positive definite. Thus the matrix may be written as:

$$A_1 + sB_1 = B_1^{\frac{1}{2}} \left(B_1^{-\frac{1}{2}} A_1 B_1^{-\frac{1}{2}} + sI \right) B_1^{\frac{1}{2}}$$

with $B_1^{\frac{1}{2}}$ symmetric, invertible. $B_1^{-\frac{1}{2}} A_1 B_1^{-\frac{1}{2}}$ is hermitian, non-negative definite, so it has real, non-negative eigenvalues $a_1, \ldots, a_n$. As a result, the eigenvalues of $B_1^{-\frac{1}{2}} A_1 B_1^{-\frac{1}{2}} + sI$ are $a_1 + s, \ldots, a_n + s$. With the assumption (5.18) that s is not real, non-positive, none of the eigenvalues $a_i + s$ can be equal to zero, thus matrix $B_1^{-\frac{1}{2}} A_1 B_1^{-\frac{1}{2}} + sI$ is invertible and so is $A_1 + sB_1$ as a product of invertible matrices. We have proven that (5.19) has a unique solution.

Next notice that because of (5.16), for each j, there is β_j such that $\tilde{V}\beta_j = (\tilde{A} + s_j\tilde{B})^{-1}\tilde{b}$. Thus β_j satisfies (5.19) for $s = s_j$. This implies

$$\tilde{h}_{\tilde{V}}(s_j) = \tilde{V}\beta_j = (\tilde{A} + s_j\tilde{B})^{-1}\tilde{b} = \tilde{h}(s_j) \tag{5.22}$$

The theorem above is valid for both the approximation of E, H when $\tilde{b}$ is dependent on ω and the approximation of the Jacobian, when $\tilde{b}$ is constant.

5.3.3 Relationship with $(A + sI)^{-1}b$

Let us relate our problem to the situation when $\tilde{B}$ is an identity matrix I. If we define

$$A = \tilde{B}^{-\frac{1}{2}}\tilde{A}\tilde{B}^{-\frac{1}{2}}, \quad b = \tilde{B}^{-\frac{1}{2}}\tilde{b}, \quad h(s) = (A + sI)^{-1}b \tag{5.23}$$

we can rewrite $\tilde{h}(s)$ as

$$\tilde{h}(s) = (\tilde{A} + s\tilde{B})^{-1}\tilde{b} = \tilde{B}^{-\frac{1}{2}}(A + sI)^{-1}b = \tilde{B}^{-\frac{1}{2}}h(s) \tag{5.24}$$

Moreover, if we define the matrix V as $V = \tilde{B}^{\frac{1}{2}}\tilde{V}$, then

$$\text{colsp}(V) = \text{span}\left\{(A + s_1 I)^{-1}b, (A + s_2 I)^{-1}b, \ldots, (A + s_n I)^{-1}b\right\} \tag{5.25}$$

Notice that particular form of $\tilde{V}$ does not matter, as long as (5.16) is satisfied. So to make the presentation easier, we will assume that columns of V are othogonal. With this assumption we obtain

$$\tilde{V}^*\tilde{A}\tilde{V} = V^*AV, \quad \tilde{V}^*\tilde{B}\tilde{V} = V^*V = I \tag{5.26}$$

and

$$\tilde{h}_{\tilde{V}}(s) = \tilde{V}\left(\tilde{V}^*(\tilde{A} + s\tilde{B})\tilde{V}\right)^{-1}\tilde{V}^*\tilde{b} = \tilde{B}^{-\frac{1}{2}}V\left(V^*(A + sI)V\right)^{-1}V^*b = \tilde{B}^{-\frac{1}{2}}h_V(s) \tag{5.27}$$

Combining (5.24) and (5.27) together allows us to relate the error of approximation of $\tilde{h}(s)$ with the error of approximation $h(s)$:

$$\tilde{h}(s) - \tilde{h}_{\tilde{V}}(s) = \tilde{B}^{-\frac{1}{2}}(h(s) - h_V(s)) \tag{5.28}$$

Consider the diagonalization of A:

$$A = U\Lambda U^* \tag{5.29}$$

where Λ is a diagonal matrix with eigenvalues λ_k as entries and columns of U are eigenvectors u_k. With this notation, it is easy to write $h(s)$ as a vector valued function with each entry being a rational function of s:

$$h(s) = (A + sI)^{-1}b = \sum_{k=1}^{N}\frac{u_k(u_k^*b)}{\lambda_k + s} \tag{5.30}$$

For $s = \mathrm{i}\omega$ with $b = b(i\omega)$ being analytic, $h(\mathrm{i}\omega)$ is an analytic function of ω. This rational (or analytic) function is approximated by $h_V(s)$:

$$h_V(s) = V\left(V^*AV + sI\right)^{-1}V^*b = \sum_{j=1}^{n}\frac{V\gamma_j(\gamma_j^*V^*b)}{\hat{\lambda}_j + s} \tag{5.31}$$

where $\hat{\lambda}_j$ and γ_j are eigenvectors and eigenvalues of V^*AV. Notice that for an analytic function, Laurent series may exist outside the Taylor series circle of convergence. Thus

it makes more sense to approximate an analytic function by a rational function than to approximate it by a polynomial. Extensive theory of rational approximation to an analytic function may be found in [18].

Intuitively, the approximation (5.31) will be good if $\tilde{\lambda}_j$, $V\gamma_j$ $j = 1, \ldots, n$ approximates λ_k, u_k, $k = 1, \ldots, N$ for k such that $u_k^* b \neq 0$. Yet the quality of this approximation depends on the choice of V, which in turn depends on the choice of interpolating shifts s_j. We will discuss algorithms that are able to choose s_j to adapt to the part of the spectrum of A for which $u_k^* b \neq 0$.

5.3.4 The case of $\tilde{b}$ not dependent on frequency ω

Next, we consider the case of approximating of the Jacobian. We will assume that $\tilde{b}$ is not dependent on frequency ω. At some places, we will also use the fact that $\tilde{b}$ is real valued. This implies that b, defined at (5.23), is not dependent on ω and real valued.

Using the idea that for any vector z, and any s_1, s_2

$$A(A + s_1 I)^{-1}(A + s_2 I)^{-1} z + s_1 (A + s_1 I)^{-1}(A + s_2 I)^{-1} z = (A + s_2 I)^{-1} z \qquad (5.32)$$

and using assumption (5.17), it is easy to show that the space (5.25) may be written as a Krylov subspace (see also [8])

$$\mathrm{colsp}(V) = \mathrm{span}\left\{q, Aq, \ldots, A^{n-1}q\right\}, \quad q = \left(\prod_{j=1}^{n} (A + s_j I)^{-1}\right) b \qquad (5.33)$$

With this notation we can formulate and prove the following theorem:

Theorem 13 *If (5.17) is satisfied and the number of distinct eigenvalues λ_k such that $u_k^* b \neq 0$ is greater than or equal to n, then V is full rank.*

Proof : Assume that V is not full rank and take a linear combination of columns of V and assume that it is 0. Using the representation (5.33), there is a polynomial p of degree $\leq n - 1$ such that

$$0 = p(A)q$$

Using diagonalization (5.29) and the definition of q, we obtain:

$$0 = p(A)q = Up(\Lambda)U^* q = Up(\Lambda)\left(\prod_{j=1}^{n} (\Lambda + s_j I)^{-1}\right) U^* b$$

From the latter, as columns of U are linearly independent, we can conclude that for each $k = 1, \ldots, N$ we have

$$0 = \frac{p(\lambda_k)}{\prod_{j=1}^{n}(\lambda_k + s_j)} u_k^* b$$

so for each $k = 1, \ldots, N$

$$0 = \frac{p(\lambda_k)}{\prod_{j=1}^{n}(\lambda_k + s_j)} \quad \text{or} \quad 0 = u_k^* b$$

Given the assumption that the number of distinct λ_k such that $u_k^* b \neq 0$ is greater or equal than n, we obtain that the polynomial p has at least n distinct roots. As p has degree not greater than $n-1$, it has to be that

$$p = 0$$

This proves that V is full rank. Notice that the assumption in Theorem 13 is true in most practical applications. If the number of distinct eigenvalues λ_k is greater than or equal to n, the assumption not satisfied is equivalent for b to be in one of $\binom{N}{n-1}$ subspaces of $\mathbb{C}^N$. If b is chosen randomly, the probability of that happening is 0.

Moreover, failure to satisfy the assumption of Theorem 13 is a lucky failure as the following theorem is true.

Theorem 14 *If $s_1, \ldots, s_n$ satisfy (5.17) and the space $\mathbf{U}(b)$ defined as*

$$\mathbf{U}(b) = span\{u_k : u_k^* b \neq 0\} \tag{5.34}$$

has dimension n, then the approximation $h_V(s)$ to $h(s)$ is exact for any s.

Proof : Given (5.30), for any s, $h(s) \in \mathbf{U}(b)$, so using definition of V (5.25), $\text{colsp}(V) \subset \mathbf{U}(b)$. Moreover using Theorem 13,

$$\dim(\text{colsp}(V)) = \text{rank}(V) = n = \dim(\mathbf{U}(b))$$

thus

$$\text{colsp}(V) = \mathbf{U}(b)$$

This implies that for any s, $h(s) \in \text{colsp}(V)$, so there exists β_s such that $h(s) = V\beta_s$ given uniqueness of solution to (5.19) (see Theorem (12)). Hence

$$h_V(s) = h(s)$$

Next, we will follow [8] to present an interesting interpretation of the approximation $h_V(s)$ to $h(s)$, which allows us to derive the error of the approximation.

Notice, that all vectors $x \in \text{colsp}(V)$, using (5.33) may be written as

$$x = Ug(\Lambda)U^*b, \quad g \in \mathcal{V} \tag{5.35}$$

where $\mathcal{V}$ is a space of rational functions, defined as follows

$$\mathcal{V} = \left\{ \frac{p(\lambda)}{\prod_{j=1}^{n}(\lambda + s_j)} \ : \ p \text{ is a polynomial of degree } \leq n-1 \right\} \tag{5.36}$$

One may write an equation for h_V as

$$(\alpha V)^*(A + sI)h_v(s) = (\alpha V)^* b \tag{5.37}$$

for all $\alpha \in \mathbb{R}^n$. From (5.35), one can conclude that there is $f \in \mathcal{V}$ such that

$$Uf(\Lambda)U^*b = h_V(s) = V\left(V^*(A+sI)V\right)^{-1}V^*b$$

Similarly, each $\alpha V \in \text{colsp}(V)$, so it may be represented as $Ug(\Lambda)U^*b$ for some $g \in \mathcal{V}$. Using those representations, one can rewrite (5.37) as

$$(Ug(\Lambda)U^*b)^*(A+sI)(Uf(\Lambda)U^*b) = (Ug(\Lambda)U^*b)^*b$$

for all $g \in \mathcal{V}$. The latter may be rewritten as

$$\begin{array}{l} b^*U\overline{g(\Lambda)}(\Lambda + sI)f(\Lambda)U^*b = b^*U\overline{g(\Lambda)}U^*b \\ \sum_{k=1}^{N}\overline{g(\lambda_k)}(\lambda_k + s)f(\lambda_k)|u_k^*b|^2 = \sum_{k=1}^{N}\overline{g(\lambda_k)}|u_k^*b|^2 \end{array}$$

And that may be rewritten as

$$< (\lambda + s)f - 1, g >_{\mu} = 0 \quad \forall g \in \mathcal{V} \tag{5.38}$$

where the measure μ is defined on the spectrum of A (which is a subset of $[0, \infty)$) as a linear combination of delta measures:

$$\mu = \sum_{k=1}^{N} |u_k^*b|^2 \delta_{(\lambda = \lambda_k)} \tag{5.39}$$

Theorem 15 *If the number of distinct eigenvalues λ_k such that $u_k^*b \neq 0$ is greater than or equal to n, then $f \in \mathcal{V}$ satisfying (5.38) approximates $\frac{1}{\lambda+s}$ with a relative error*

$$\frac{f(\lambda, s) - \frac{1}{\lambda+s}}{\frac{1}{\lambda+s}} = (\lambda + s)f(\lambda, s) - 1 = -\prod_{j=1}^{n} \frac{(s - s_j)(\lambda - \hat{\lambda}_j)}{(\lambda + s_j)(s + \hat{\lambda}_j)} \tag{5.40}$$

where $\hat{\lambda}_j$, $j = 1, \ldots, n$ are eigenvalues of V^*AV.

The proof, shorter than the one in [8], is given in the Appendix (Section 5.7).

5.3.5 The case of $\tilde{b}$ dependent on frequency ω

In the case of approximation of the forward MT response, $\tilde{b}(\mathrm{i}\omega) = g(\omega)$ where $g(\omega)$ is defined at (5.6), so that $\tilde{b}$ depends on the value of shift s. In this case, Theorem 12 as well as the formulas (5.30) and (5.31) are valid.

We will investigate now what it means that at some point in the iterations, V is not full rank. Theorem 13 and Theorem 14 say that in the case of $\tilde{b}$ not dependent on ω, V becoming not full rank is a lucky failure. It turns out that we can get a result almost as strong in the case of $\tilde{b}(s)$, if the dependence on s is analytic.

Let I be a closed, bounded, connected set in the complex plane. In can be in particular the interval in which we will calculate h and in which we will choose interpolating shifts s_j. We can think of a purely imaginary interval $I = \{\mathrm{i}\omega : \ \omega \in [\omega_{\min}, \omega_{\max}]\}$. Crucial for further investigation is the representation of $h(s)$ by (5.30).

Define functions

$$d_k(s) = \frac{u_k^* b(s)}{\lambda_k + s}, \quad k = 1, \ldots, N \tag{5.41}$$

for $s \in I$

Using the above definition, (5.29) and the definition of V, we can rewrite the condition of V being full rank:

$$\operatorname{rank}(V) = n \Leftrightarrow \operatorname{rank} \begin{bmatrix} d_1(s_1) & \cdots & d_1(s_n) \\ \vdots & \ddots & \vdots \\ d_N(s_1) & \cdots & d_N(s_n) \end{bmatrix} = n \tag{5.42}$$

Let us assume that all the entries of $b(s)$ are analytic functions on I. Let us also assume that the singularities of h (which are $s = -\lambda_k$) are not in I. In this case, h as well as $d_k(s)$ are analytic on I.

Theorem 16 *Assume that there are at least n functions $d_{k_1}, \ldots, d_{k_n}$ among $(d_k)_{k=1}^N$ that are linearly independent (domain of those functions is assumed to be I). Then for all $j = 1, \ldots, n$ the following is true. If*

$$\mathit{rank} \begin{bmatrix} d_{k_1}(s_1) & \cdots & d_{k_1}(s_{j-1}) \\ \vdots & \ddots & \vdots \\ d_{k_n}(s_1) & \cdots & d_{k_n}(s_{j-1}) \end{bmatrix} = j - 1 \tag{5.43}$$

then there is at most a finite number of points $s \in I$ for which

$$rank \begin{bmatrix} d_{k_1}(s_1) & \cdots & d_{k_1}(s_{j-1}) & d_{k_1}(s) \\ \vdots & \ddots & \vdots & \\ d_{k_n}(s_1) & \cdots & d_{k_n}(s_{j-1}) & d_{k_n}(s) \end{bmatrix} < j \tag{5.44}$$

Remark 17 *The theorem above tells us that if there are n linearly independent functions among d_k, then at each iteration, it is highly unlikely for V not be full rank. At each step, almost any choice of $s = s_j$ in I is good (any apart from at most a finite number of points).*

Before we prove Theorem 16, let us formulate another theorem, similar to 14, that tells us that failure to satisfy assumption of Theorem 16 is a lucky failure.

Theorem 18 *If there are exactly n linearly independent functions among d_k and V is full rank, then the approximation $h_V(s)$ to $h(s)$ is exact for any s.*

In the proof, we will use the following lemma

Lemma 19 *Assume that a function f is analytic on a bounded, closed, connected set I. If there are infinitely many points $s_j \in I$ for which $f(s_j) = 0$, then $f(s) = 0$ for all $s \in I$.*

Proof : As I is bounded and closed, it is compact, thus there is a subsequence of points $(s_{j_l})_{l=1}^{\infty}$ convergent to $s_0 \in I$. We use the fact that if an analytic function has an accumulation point of zeros, then it is equal to zero everywhere on the connected component containing the accumulation point. In our case, I is connected, so we obtain that $f(s) = 0$ for all $s \in I$.

Proof : (of Theorem 16)

Assume that (5.44) is not satisfied.

If $j = 1$, it means that there is an infinite number of points $s \in I$ for which $d_{k_1}(s) = \ldots = d_{k_n}(s) = 0$. From Lemma 19, we obtain that $d_{k_1}(s) = \ldots = d_{k_n}(s) = 0$ for all $s \in I$, which contradicts $d_{k_1}, \ldots, d_{k_n}$ being linearly independent.

Consider $j > 1$. As (5.43) is true, there is a submatrix with j rows of the matrix in (5.43) which has rank $j - 1$. To simplify the notation, assume that the first j rows form a matrix with rank $j - 1$, in other words, we have

$$\text{rank}\begin{bmatrix} d_{k_1}(s_1) & \cdots & d_{k_1}(s_{j-1}) \\ \vdots & \ddots & \vdots \\ d_{k_j}(s_1) & \cdots & d_{k_j}(s_{j-1}) \end{bmatrix} = j-1 \tag{5.45}$$

Moreover, (5.44) not satisfied implies

$$f(s) = \det\begin{bmatrix} d_{k_1}(s_1) & \cdots & d_{k_1}(s_{j-1}) & d_{k_1}(s) \\ \vdots & \ddots & \vdots & \vdots \\ d_{k_j}(s_1) & \cdots & d_{k_j}(s_{j-1}) & d_{k_j}(s) \end{bmatrix} = 0 \tag{5.46}$$

for infinitely many points s in I. The determinant $f(s)$ is analytic on I as it is a sum of products of analytic functions, so from Lemma 19, (5.46) is satisfied for all $s \in I$. Given (5.45), it has to be that for all $s \in I$, the last column of matrix in (5.46) has to be a linear combination of the rest of the columns. In other words, there exist functions $\beta_1(s), \ldots, \beta_k(s),\ s \in I$ (the coefficients of the combination) such that

$$d_{k_l}(s) = \sum_{i=1}^{j-1} \beta_i(s) d_{k_l}(s_l) \text{ for } l = 1, \ldots, j, \quad s \in I \tag{5.47}$$

So if one thinks about the linear subspaces of functions defined on I, we have

$$\text{span}(d_{k_1}, \ldots, d_{k_j}) \subset \text{span}(\beta_1, \ldots, \beta_{j-1}) \tag{5.48}$$

This contradicts with the assumption of $d_{k_1}, \ldots, d_{k_j}$ being linearly independent.

Proof : (of Theorem 18) If there are no more than n linearly independent functions among d_k, then for any points $s_1, \ldots, s_n, s \in I$ we have:

$$\text{rank}\begin{bmatrix} d_1(s_1) & \cdots & d_1(s_n) & d_1(s) \\ \vdots & \ddots & \vdots & \vdots \\ d_N(s_1) & \cdots & d_N(s_n) & d_N(s) \end{bmatrix} \leq n \tag{5.49}$$

With the assumption that V is full rank, which implies (5.42), we obtain that the rank in (5.49) is exactly n and the last column of the matrix in (5.49) is a linear combination of the first n columns:

$$\begin{bmatrix} d_1(s) \\ \vdots \\ d_N(s) \end{bmatrix} = \sum_{j=1}^{n} \alpha_j \begin{bmatrix} d_1(s_j) \\ \vdots \\ d_N(s_j) \end{bmatrix}, \text{ for } s \in I$$

Using (5.41) and (5.29), the previous relationship may be rewritten as

$$(A + sI)^{-1} b(s) = \sum_{j=1}^{n} \alpha_j (A + s_j I)^{-1} b(s_j), \text{ for } s \in I$$

which, from the definition of h_V implies

$$h_V(s) = h(s), \text{ for } s \in I \tag{5.50}$$

From representations (5.30) and (5.31), for b having analytic coefficients, both h and h_V are analytic in I. As a result, (5.50) is satisfied wherever b and h may be analytically extended from I. For example, if b has coefficients which are entire functions, then (5.50) is satisfied in all of the domain of h.

To draw a connection between assumptions of Theorems (16), (18), valid for the case of b dependent on frequency and Theorems (13), (14) formulated for the case of b not dependent on frequency, we notice that if b is not dependent on frequency, $u_{k_j}^* b \neq 0$ for $j = 1, \ldots, n$ and $\{\lambda_{k_j}\}_{j=1}^n$ are n distinct values, then $(d_{k_j})_{j=1}^n$, defined in (5.41) are linearly independent. Also, $(d_{k_j})_{j=1}^n$ are linearly independent if $u_{k_j}^* b$ are entire functions and $u_{k_j}^* b(-\lambda_{k_j}) \neq 0$, as then each d_{k_j} has a simple pole at $s = -\lambda_k$ and is elsewhere analytic.

If one goes back to the considered application, then $\tilde{b}(\mathrm{i}\omega) = g(\omega)$, where g is defined at (5.6) using E^p. E^p is a plane wave going downwards and if we consider the electric field normalized to 1 at the earth's surface, then the dependence of E^p on the frequency at any location in the domain Ω is analytic. This implies that the coefficients of g are analytic functions of the frequency ω.

Yet if one thinks about E^p being a plane wave with a fixed frequency ω, there is a question of magnitude of the wave as well as the way the magnitude changes as the frequency ω changes. For example, one could choose $E^p(\omega)$ to be equal to 1 at the top of the domain, at the earth's surface or according to some other recipe. The following theorem says that the relative error of approximation will be the same in all cases.

Theorem 20 *Let g be any scalar function defined in I, such that $g(s) \neq 0$ for $s \in I$. Take any vector valued function $b(s)$, $s \in I$ and define*

$$\breve{b}(s) = b(s)g(s)$$

Consider $h(s)$ and $\breve{h}(s)$ satisfying

$$(A + sI)h(s) = b(s)$$
$$(A + sI)\breve{h}(s) = \breve{b}(s)$$

Take any distinct interpolating shifts $(s_j)_{j=1}^n$, *let* V *and* $\breve{V}$ *be defined as*

$$colsp(V) = span\{h(s_j) : \ j = 1, \ldots, n\}$$
$$colsp(\breve{V}) = span\{\breve{h}(s_j) : \ j = 1, \ldots, n\}$$

Model order reduction approximation is defined in a natural way as

$$h_V(s) = V\alpha, \ \textit{where } V^*(A + sI)V\alpha = V^*b(s)$$
$$\breve{h}_{\breve{V}}(s) = \breve{V}\breve{\alpha}, \ \textit{where } \breve{V}^*(A + sI)\breve{V}\breve{\alpha} = \breve{V}^*\breve{b}(s)$$

Then the relative errors of approximation are the same

$$\frac{||\breve{h}_{\breve{V}}(s) - \breve{h}(s)||}{||\breve{h}(s)||} = \frac{||h_V(s) - h(s)||}{||h(s)||} \tag{5.51}$$

Proof : Using the definition of $h(s)$ and $\breve{h}(s)$, we obtain immediately that

$$\breve{h}(s) = h(s)g(s) \tag{5.52}$$

This, together with the definition of V and $\breve{V}$ implies

$$\text{colsp}\{V\} = \text{colsp}\{\breve{V}\} \tag{5.53}$$

which results in

$$\breve{h}_{\breve{V}}(s) = h_V(s)g(s) \tag{5.54}$$

The hypothesis (5.51) is an immediate consequence.

Remark 21 *The quantities of interest in MT are impedance* Z *and tipper* K, *which are defined in (5.9). Using a reasoning similar to the one in the proof, one can show that if in the formulas for numerical approximation of electric (5.10) and magnetic (5.11) fields, we use approximation* $\breve{h}_{\breve{V}}(\mathrm{i}\omega)$ *instead of* ξ, *it does not matter if we use* $h_V(\mathrm{i}\omega)$ *of* $\breve{h}_{\breve{V}}(\mathrm{i}\omega)$, *the value of* Z, K *will be exactly the same.*

5.3.6 Treatment of the null space of A

Existing algorithms that consider approximation of $\tilde{h}(s)$ for real shifts $s \in [\lambda_{\min}, \lambda_{\max}]$ require $\lambda_{\min} > 0$ [9], which means that the matrix A has a trivial null space, so is positive definite. If $\lambda_{\min} = 0$, the error indicator might have the maximum at $s = 0$, for which the equation (5.15) is not solvable. For example, for algorithm AR(defined in Section 5.4), if at some point one of the Ritz values $\hat{\lambda}_j = 0$, the error

indicator (5.75) approaches ∞ as s approaches 0. We present a solution to this problem, which requires us to decompose b into the part lying on the null space of A and the part orthogonal to it.

Take any matrix $\tilde{K}$, whose columns are the basis of null space of $\tilde{A}$:

$$\text{colsp}(\tilde{K}) = \text{null}(\tilde{A}) \tag{5.55}$$

Define

$$K = \tilde{B}^{\frac{1}{2}}\tilde{K} \tag{5.56}$$

with this definition, we have

$$\text{colsp}(K) = \text{null}(A) \tag{5.57}$$

Take a matrix W, whose columns are an orthonormal basis the range of A. Matrix A is hermitian, so its range is orthogonal to its null space, thus:

$$[K\ W]^*[K\ W] = \begin{bmatrix} K^*K & 0 \\ 0 & I \end{bmatrix} \tag{5.58}$$

Consider a representation of h in the bases of columns of K and W:

$$h = [K\ W]\begin{bmatrix} \alpha_K \\ \alpha_W \end{bmatrix} = K\alpha_K + W\alpha_W = h_K + h_W \tag{5.59}$$

Using this decomposition, one can rewrite the equation

$$(A + sI)h = b \tag{5.60}$$

as

$$\begin{cases} W^*(A + sI)(K\alpha_K + W\alpha_W) = W^*b \\ K^*(A + sI)(K\alpha_K + W\alpha_W) = K^*b \end{cases}$$

which is equivalent to two uncoupled equations for α_w and α_K:

$$\begin{cases} (W^*AW + sI)\alpha_W &= W^*b \\ s(K^*K)\alpha_K &= K^*b \end{cases} \tag{5.61}$$

If b is orthogonal to the null space of A, in which case, $K^*b = 0$, then $\alpha_K = 0$ and thus:

$$h(s) = (A + sI)^{-1}b = W(W^*AW + sI)^{-1}W^*b \tag{5.62}$$

In this case, one could modify the matrix eigenvalues on the null space of A in an arbitrary way, and h would be the same. This explains that if $b \perp \text{null}(A)$, then

it is enough to consider $[\lambda_{\min}, \lambda_{\max}]$ to be the spectral interval of W^*AW, which is the effective spectral interval of A (spectral interval of A, disregarding the null space). Notice that even more is true; it is enough to consider the interval containing eigenvalues λ_k of A for which $u_k^* b \neq 0$ which are the support of the measure μ defined at (5.39).

Let us focus on the situation when b is not orthogonal to the null space of A. One can solve the second equation in (5.61) obtaining

$$h_K(s) = K\alpha_K = \frac{1}{s} K(K^*K)^{-1}K^*b \tag{5.63}$$

And then finding h_W may be done equivalently by solving the original equation (5.60) with a modified right-hand side, consisting only of the component b_W in the range of A:

$$(A + sI)h_W = b_W \tag{5.64}$$

where

$$b_W = b - K(K^*K)^{-1}K^*b \tag{5.65}$$

We propose to calculate $h_K(s)$ exactly and to use model order reduction techniques only to approximate $h_W(s)$, the solution of equation (5.64). In this case, one can use algorithms AR, ARR, and NARR with the effective spectral interval $[\lambda_{\min}, \lambda_{\max}]$ of A.

Notice that in the case of b not dependent on s, in order to get $h_K(s)$ for all s, one needs to calculate $K(K^*K)^{-1}K^*b$ once only.

Although AI, AIR, and NAIR algorithms (defined in Section 5.4) might be used without this decomposition, the technique described above is likely to improve the approximation of $h(s)$ firstly because one can obtain h_K exactly, and secondly because the location of interpolation frequencies used for (5.64) will need to adapt to μ with a smaller support (not containing 0) than in the case of equation (5.60).

Representation of the null space of A, allowing for a sparse K^*K, is not possible in all situations. Yet in the case of edge element discretization of the equation (5.2) for electric field E, the null space of $\tilde{A}$, which is the null space of the curl operator $\nabla\times$, is the range of the gradient operator ∇ acting on the space $\mathcal{H}_0^{1,h}(\Omega)$ of nodal shape functions on the same mesh (see Appendix B in [14]). Let φ be a scalar field

defined at vertices inside the domain Ω ($\varphi = 0$ on $\partial\Omega$), $\varphi \in \mathcal{H}_0^{1,h}(\Omega)$. $\nabla\varphi$ is defined at edges of Ω. Let edge e point from vertex v_1 to vertex v_2, then the operator ∇ acts on φ in such a way, that

$$(\nabla\varphi)(e) = \varphi(v_2) - \varphi(v_1)$$

Let N_v be the number of vertices inside Ω. If $(\psi_j)_{j=1}^{N_v}$ are nodal shape functions, then we define $\tilde{K}$ as a matrix with entries $1, -1$ such that for $\varphi = \sum_{j=1}^{N_v} \xi_j \psi_k$

$$((\nabla\varphi)(e))_{k=1}^{N} = \tilde{K}\xi \tag{5.66}$$

Notice that, using (5.56)

$$K^*K = \tilde{K}^*\tilde{B}\tilde{K} = \left[\int_\Omega \sigma \nabla\psi_i \cdot \nabla\psi_j\right]_{i,j=1}^{N_v} \tag{5.67}$$

Thus finding $(K^*K)^{-1}K^*b = (\tilde{K}^*\tilde{B}\tilde{K})^{-1}\tilde{K}^*\tilde{b}$ requires us to solve a Poisson equation with σ as a coefficient (compare with the divergence correction [15]). The matrix (5.67) is real valued and also $\tilde{b} \in \mathbb{R}$ in our situation. Moreover, the number of vertices N_v is more than three times less than the number of edges N, so finding $(K^*K)^{-1}K^*b$ is more than 10 times faster than solving evaluating $\tilde{h}(s)$ for one value of s.

We write the resulting procedure of approximating $\tilde{h}(s)$ using original quantities of interest (not scaled by $\tilde{B}^{\frac{1}{2}}$). We use the fact that

$$\tilde{h}(s) = \tilde{h}_{\tilde{W}}(s) + \tilde{h}_{\tilde{K}}(s) = (\tilde{A} + s\tilde{B})^{-1}\tilde{b}_{\tilde{W}} + \frac{1}{s}(\tilde{K}^*\tilde{B}\tilde{K})^{-1}\tilde{K}^*\tilde{b} \tag{5.68}$$

where

$$\tilde{b}_{\tilde{W}} = \tilde{B}^{-\frac{1}{2}}b_W = \tilde{b} - \tilde{K}(\tilde{K}^*\tilde{B}\tilde{K})^{-1}\tilde{K}^*\tilde{b} \tag{5.69}$$

We calculate $\tilde{h}_{\tilde{K}}(s)$ exactly and observe that it may be found at the expense of one solve of Poisson equation per s (in the case of b not dependent on s one solve of Poisson equation for all s). We further approximate $\tilde{h}_{\tilde{W}}(s)$ by $\tilde{h}_{\tilde{W},\tilde{V}}(s)$ using model order reduction techniques. The final approximation to $\tilde{h}(s)$ is given by

$$\tilde{h}(s) \approx \tilde{h}_{\tilde{W},\tilde{V}}(s) + \tilde{h}_{\tilde{K}}(s) \tag{5.70}$$

5.4 Algorithms

5.4.1 The case of $\tilde{b}$ not dependent on frequency ω

In this Section, we present the algorithms for choosing the interpolating shifts s_j. We will consider the values of the shifts in an purely imaginary interval $\{s = \mathrm{i}\omega : \omega \in [\omega_{\min}, \omega_{\max}]\}$, or in a purely real interval $[\lambda_{\min}, \lambda_{\max}]$. Both intervals are presented in Figure 5.2. Considering that we need to evaluate the transfer function in a purely imaginary interval, the best choice of shifts would be such that minimizes the maximum relative error of approximation:

$$\min_{s_1,\ldots,s_n \in I} \max_{\omega \in [\omega_{\min}, \omega_{\max}]} \frac{||\tilde{h}(\mathrm{i}\omega) - \tilde{h}_{\tilde{V}}(\mathrm{i}\omega)||_2}{||\tilde{h}(\mathrm{i}\omega)||_2} \tag{5.71}$$

Such an approach is not pursued in this paper for two reasons. Firstly, because the true relative error of approximation is not known, we have to use an error indicator $e_{(s_1,\ldots,s_n)}(\mathrm{i}\omega)$. Secondly, if (5.71) was used, the values $(s_1, \ldots, s_n)$ would most likely not appear among $(s_1, \ldots, s_n, s_{n+1})$.

Because of that we consider an approach in which at each iteration, given $(s_1, \ldots, s_n)$, we will add one value s_{n+1} in order to form $(s_1, \ldots, s_n, s_{n+1})$. We use the fact that at each interpolation shift, the error of approximation is 0, thus the value of the next interpolation shift s_{n+1} is chosen as the maximum of the error indicator function in the interval I:

$$s_{n+1} = \mathrm{argmax}_{s \in I}\ e_{(s_1,\ldots,s_n)}(s) \tag{5.72}$$

Next, we present two error indicator functions and the algorithm based on them. Let us start with the idea of [9]. The relative error (5.40) is related to the residual:

$$\begin{aligned} R = (A + sI)h_V(s) - b \ &= (A + sI)(h_V(s) - h(s)) = U((\Lambda + s)f(\Lambda, s) - 1)U^*b \\ &= -\left[\prod_{j=1}^n \tfrac{s - s_j}{s + \hat{\lambda}_j}\right]\left[U\left(\prod_{j=1}^n (\Lambda - \hat{\lambda}_j I)\prod_{j=1}^n (\Lambda + s_j I)^{-1}\right)U^*b\right] \end{aligned} \tag{5.73}$$

The quantity above is split into two parts. The first part is a scalar, the second is a vector. Once the interpolation shifts s_j are fixed, the Ritz values $\hat{\lambda}_j$ minimize the norm of the second part:

$$\left|\left| U\left(\prod_{j=1}^n (\Lambda - \hat{\lambda}_j I)\prod_{j=1}^n (\Lambda + s_j I)^{-1}\right)U^* \right|\right|_2 \tag{5.74}$$

For the proof of this statement, see the Theorem 22 in the Appendix (Section 5.7). This property of eigenvalues $\hat{\lambda}_j$ is a basis of the adaptive choice of interpolating shifts

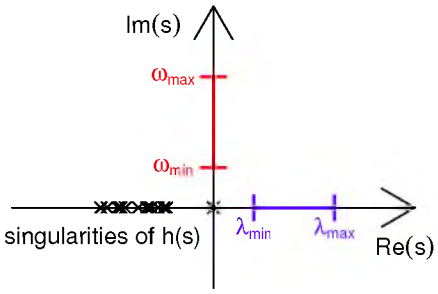

Figure 5.2. Complex plane – the domain of h. The interval over which the values of h are needed is shown in red. Interval of effective spectrum of A shown in blue.

s_j in [9]. Authors of [9] consider s real valued and in the interval $[\lambda_{\min}, \lambda_{\max}]$ where $\lambda_{\min}, \lambda_{\max} > 0$ are the smallest and the largest eigenvalues of A. The algorithm of [9] is presented below:

ALGORITHM AR (ADAPTIVE CHOICE OF REAL SHIFTS):

1. Set $n = 2$, choose $s_1 = \lambda_{\min}$, $s_2 = \lambda_{\max}$ and set $\tilde{V}_{1:2}$ in such a way that

$$\mathrm{colsp}(\tilde{V}_{1:2}) = \mathrm{span}\{(\tilde{A} + s_1\tilde{B})^{-1}\tilde{b}, (\tilde{A} + s_2\tilde{B})^{-1}\tilde{b}\}$$

2. Find s_{n+1} as a maximizer of

$$\left| \prod_{j=1}^{n} \frac{s - s_j}{s + \hat{\lambda}_j} \right| \tag{5.75}$$

over $s \in [\lambda_{\min}, \lambda_{\max}]$, where $\hat{\lambda}_1, \ldots, \hat{\lambda}_n$ are eigenvalues of $\tilde{V}^*\tilde{A}\tilde{V}$.

3. Set $\tilde{V}_{1:(n+1)}$ in such a way that

$$\mathrm{colsp}(\tilde{V}_{1:(n+1)}) = \mathrm{colsp}(\tilde{V}_{1:n}) \oplus \mathrm{span}\{(\tilde{A} + s_{n+1}\tilde{B})^{-1}\tilde{b}\}$$

4. If exit criteria met(approximation good enough), stop

5. Set $n = n + 1$ and jump to 2

The details of update of $\ddot{V}$ at 3 are presented in Section 5.4.2. Notice that according to (5.26) $\ddot{V}^*\ddot{A}\ddot{V} = V^*AV$, so those two matrices have the same eigenvalues.

This approach is valid for A symmetric positive definite or for b orthogonal to the null space of A (in this case, $\lambda_{\min}$ is the smallest of non-zero eigenvalues). In the current paper, we suggest what may be done in the case when b is not orthogonal to the null space of A (see Section 5.3.6). To use this algorithm, one needs estimates of $\lambda_{\min}$ and $\lambda_{\max}$, which might not be easy if A has a nontrivial null-space. The advantage of this approach is that at each iteration, to enlarge V, one needs to solve the linear system (5.15) for real s. In this case, the matrix $\tilde{A} + s\tilde{B}$ is real, symmetric positive definite, and solution of the linear system with iterative solvers is faster than for the complex case. One matrix vector multiplication is 4 times faster. And the total number of iterations is less as methods for hermitian matrices (like Conjugate Gradient) may be used.

Next we propose a similar algorithm, but for purely imaginary values of $s = \mathrm{i}\omega$, where the frequency is in an interval of interest $\omega \in [\omega_{\min}, \omega_{\max}]$. This is a natural approach for MT, as one needs the response for a number of frequencies log-uniformly distributed in an interval. Also if $\tilde{b} \in \mathbb{R}$, $\omega \in \mathbb{R}$ then

$$\overline{\tilde{h}(\mathrm{i}\omega)} = \overline{(\tilde{A} + \mathrm{i}\omega\tilde{B})^{-1}\tilde{b}} = (\tilde{A} - \mathrm{i}\omega\tilde{B})^{-1}\tilde{b} = \tilde{h}(-\mathrm{i}\omega) \tag{5.76}$$

so if one calculates $\tilde{h}$ at $s = \mathrm{i}\omega_j$, then one simultaneously gets the value of $\tilde{h}$ at $s = -\mathrm{i}\omega_j$. Hence with one interpolating frequency ω_j, the dimension of $\mathrm{colsp}(\tilde{V})$ may be increased by 2, using $\tilde{h}_{\tilde{V}}(\mathrm{i}\omega)$ and $\overline{\tilde{h}_{\tilde{V}}(\mathrm{i}\omega)}$. In this setting when at each iteration we add two shifts $s = \pm\omega_j$, the squared norm of the first part of the residual (5.73) evaluated at $s = \mathrm{i}\omega$ is

$$\left|\prod_{j=1}^{2n} \frac{s - s_j}{s + \hat{\lambda}_j}\right|^2 = \left|\frac{\prod_{j=1}^{n}(\mathrm{i}\omega - \mathrm{i}\omega_j)(\mathrm{i}\omega + \omega_j)}{\prod_{j=1}^{2n}(\mathrm{i}\omega + \hat{\lambda}_j)}\right|^2 = \frac{\prod_{j=1}^{n}(\omega^2 - \omega_j^2)^2}{\prod_{j=1}^{2n}(\omega^2 + \hat{\lambda}_j^2)} \tag{5.77}$$

where $\hat{\lambda}_j \in \mathbb{R}$, $j = 1, \ldots, 2n$ are eigenvalues of $V^*AV = \tilde{V}^*\tilde{A}\tilde{V}$. The resulting algorithm is presented below:

Algorithm AI (Adaptive choice of Imaginary shifts):

1. Set $n = 2$, choose $\omega_1 = \omega_{\min}$, $\omega_2 = \omega_{\max}$, and set $\tilde{V}_{1:4}$ in such a way that

   $$\mathrm{colsp}(\tilde{V}_{1:4}) = \mathrm{span}\{(\tilde{A}+\mathrm{i}\omega_1\tilde{B})^{-1}\tilde{b}, (\tilde{A}-\mathrm{i}\omega_1\tilde{B})^{-1}\tilde{b}, (\tilde{A}+\mathrm{i}\omega_2\tilde{B})^{-1}\tilde{b}, (\tilde{A}-\mathrm{i}\omega_2\tilde{B})^{-1}\tilde{b}\}$$

2. Find ω_{n+1} as a maximizer of

   $$\frac{\prod_{j=1}^{n}(\omega^2 - \omega_j^2)^2}{\prod_{j=1}^{2n}(\omega^2 + \hat{\lambda}_j^2)} \tag{5.78}$$

   over $w \in [\omega_{\min}, \omega_{\max}]$, where $\hat{\lambda}_1, \ldots, \hat{\lambda}_{2n}$ are eigenvalues of $\tilde{V}^*\tilde{A}\tilde{V}$.

3. Set $\tilde{V}_{1:2(n+1)}$ in such a way that

   $$\mathrm{colsp}(\tilde{V}_{1:2(n+1)}) = \mathrm{colsp}(\tilde{V}_{1:2n}) \oplus \mathrm{span}\{(\tilde{A} + \mathrm{i}\omega_{n+1}\tilde{B})^{-1}\tilde{b}, (\tilde{A} - \mathrm{i}\omega_{n+1}\tilde{B})^{-1}\tilde{b}\}$$

4. If exit criteria met(approximation good enough), stop

5. Set $n = n + 1$ and jump to 2

In this algorithm, the update 3 may be done in such a way that $\tilde{V}$ is a real matrix, which saves some computational time (see Section 5.4.2).

The advantage over the AR algorithm is that one does not need to estimate $\lambda_{\min}, \lambda_{\max}$ and the approximation is tuned for the interval of interest $[\omega_{\min}, \omega_{\max}]$.

Both AR and AI algorithms adapt to the spectral distribution μ. The error indicators (5.75), (5.78) depend on Ritz values $\hat{\lambda}_j$, which in turn depend on μ. Both AR and AI algorithms choose new values of shifts in an attempt to decrease the maximum norm of the residual (5.73). Yet they do it by considering a part of the norm of the residual as the error indicator. One could also consider the norm of the residual itself as the error indicator [12]. The norm of the residual may be approximated as

$$||(A+sI)h_V(s)-b||_2 = ||\tilde{B}^{-\frac{1}{2}}((\tilde{A}+s\tilde{B})\tilde{h}_{\tilde{V}}(s)-\tilde{b})||_2 \approx ||\tilde{B}_d^{-\frac{1}{2}}((\tilde{A}+s\tilde{B})\tilde{h}_{\tilde{V}}(s)-\tilde{b})||_2 \tag{5.79}$$

where $\tilde{B}_d$ approximates $\tilde{B}$ (in numerical tests, we use the diagonal part of $\tilde{B}$). This gives rise to the following two algorithms:

Algorithm ARR (Adaptive choice of Real shifts; norm of the Residual as the error indicator):

The algorithm is the same as AR, apart from step 2, which is replaced with

2. Find s_{n+1} as a maximizer of residual

$$||\tilde{B}_d^{-\frac{1}{2}}((\tilde{A}+s\tilde{B})\tilde{h}_{\tilde{V}}(s)-\tilde{b})||_2 \tag{5.80}$$

over $s \in [\lambda_{\min}, \lambda_{\max}]$.

Algorithm AIR (Adaptive choice of Imaginary shifts; norm of the Residual as the error indicator):

The algorithm is the same as AI, apart from step 2, which is replaced with

2. Find ω_{n+1} as a maximizer of residual

$$||\tilde{B}_d^{-\frac{1}{2}}((\tilde{A}+\mathrm{i}\omega\tilde{B})\tilde{h}_{\tilde{V}}(\mathrm{i}\omega)-\tilde{b})||_2 \tag{5.81}$$

over $\omega \in [\omega_{\min}, \omega_{\max}]$.

In Section 5.4.3, we present a fast method of calculating the residual over an interval, allowing for practical implementation of algorithms ARR and AIR.

The previous methods used the approach where at each s, the approximation $h_V(s)$ to $h(s)$ is obtained as $Uf(\Lambda, s)U^*$, where $f \in \mathcal{V}$ is such that (5.38) is satisfied. The adaptive choice of shifts was such that the maximum of the residual that may be written as

$$||(\lambda + s)f - 1||_\mu \tag{5.82}$$

was minimized for $s \in [\lambda_{\min}, \lambda_{\max}]$(or $s = \mathrm{i}\omega,\ \omega \in [\omega_{\min}, \omega_{\max}]$). There is a natural question: For s fixed, why not choose $f \in \mathcal{V}$ such that the residual (5.82) is minimal? The approximation $\tilde{h}_V^{\mathrm{norm}}(s)$ to $\tilde{h}(s)$ is given by

$$\tilde{h}_V^{\mathrm{norm}}(s) = \tilde{V}\beta, \text{where } \beta \text{ minimizes } ||\tilde{B}^{-\frac{1}{2}}((\tilde{A} + s\tilde{B})\tilde{V}\beta - \tilde{b})||_2 \tag{5.83}$$

In practice, we will approximate the above by

$$\tilde{h}_V^{\mathrm{norm}}(s) \approx \tilde{V}\beta', \text{where } \beta' \text{ minimizes } ||\tilde{B}_d^{-\frac{1}{2}}((\tilde{A} + s\tilde{B})\tilde{V}\beta' - \tilde{b})||_2 \tag{5.84}$$

To find β', one needs to solve the normal equation

$$(\tilde{A} + s\tilde{B})^*(\tilde{B}_d^*)^{-\frac{1}{2}}\tilde{B}_d^{-\frac{1}{2}}(\tilde{A} + s\tilde{B})\tilde{V}\beta' = (\tilde{A} + s\tilde{B})^*(\tilde{B}_d^*)^{-\frac{1}{2}}\tilde{b} \tag{5.85}$$

This gives rise to two algorithms:

Algorithm NARR (Normal equation; Adaptive choice of Real shifts; norm of the Residual as the error indicator):

The algorithm is the same as AR, apart from step 2, which is replaced with

2. Find s_{n+1} as a maximizer of residual

$$||\tilde{B}_d^{-\frac{1}{2}}((\tilde{A} + s\tilde{B})\tilde{h}_{\tilde{V}}^{\mathrm{norm}}(s) - \tilde{b})||_2 \tag{5.86}$$

over $s \in [\lambda_{\min}, \lambda_{\max}]$, where $\tilde{h}_{\tilde{V}}^{\mathrm{norm}}(s)$ is given by the approximation (5.84).

Also $\tilde{h}_{\tilde{V}}^{\mathrm{norm}}(s)$ instead of $\tilde{h}_{\tilde{V}}(s)$ is used as the approximation to $\tilde{h}(s)$.

Algorithm NAIR (Normal equation; Adaptive choice of Imaginary shifts; norm of the Residual as the error indicator):

The algorithm is the same as AI, apart from step 2, which is replaced with

2. Find ω_{n+1} as a maximizer of residual

$$||\tilde{B}_d^{-\frac{1}{2}}((\tilde{A}+\mathrm{i}\omega\tilde{B})\tilde{h}_{\tilde{V}}^{\mathrm{norm}}(\mathrm{i}\omega)-\tilde{b})||_2 \tag{5.87}$$

over $\omega \in [\omega_{\min}, \omega_{\max}]$, where $\tilde{h}_{\tilde{V}}^{\mathrm{norm}}(\mathrm{i}\omega)$ is given by the approximation (5.84) for $s = \mathrm{i}\omega$.

Also $\tilde{h}_{\tilde{V}}^{\mathrm{norm}}(s)$ instead of $\tilde{h}_{\tilde{V}}(s)$ is used as the approximation to $\tilde{h}(s)$.

Details of solving (5.85) for β' and finding maximum of the residuals (5.86), (5.87) are discussed in Section 5.4.3.

5.4.2 Update of matrix $\tilde{V}$

Let v_i denote the columns of $\tilde{V}$. We update V in such a way that the columns of $\tilde{V}$ are orthonormal. If we need to add a vector u to the column space $\tilde{V}$, we orthogonalize it according to the algorithm:

$$\begin{aligned} &v_{n+1} = u \\ &\text{for } (j = 1 : n)\{ \\ &\quad v_{n+1} = v_{n+1} - (v_{n+1}, v_j)v_j \\ &\} \\ &v_{n+1} = \tfrac{v_{n+1}}{||v_{n+1}||_2} \end{aligned} \tag{5.88}$$

In the case of AR, ARR, and NARR algorithms, the vector to be added is

$$u = \tilde{h}(s) = (\tilde{A}+s\tilde{B})^{-1}\tilde{b}$$

for some real s. As a result, $u \in \mathbb{R}$ and the matrix V is real.

In the case of AI, AIR, and NAIR algorithms when $\tilde{b} \in \mathbb{R}$, for some ω instead of adding $\tilde{h}(\mathrm{i}\omega)$ and its complex conjugate $\overline{\tilde{h}(\mathrm{i}\omega)} = \tilde{h}(-\mathrm{i}\omega)$, we add its real and imaginary parts $\mathrm{Re}(\tilde{h}(\mathrm{i}\omega))$, $\mathrm{Im}(\tilde{h}(\mathrm{i}\omega))$. The latter two vectors span the same two-dimensional space as $\tilde{h}(\mathrm{i}\omega)$, $\overline{\tilde{h}(\mathrm{i}\omega)}$, yet they are real valued and as a result, the matrix V is real valued. This allows for computational savings whenever columns of $\tilde{V}$ have to be multiplied by matrices $\tilde{A}, \tilde{B}$ or when inner products have to be calculated (see also Section 5.4.3).

5.4.3 Fast evaluation of the residual

We assume that all of the error indicators have a single maximum in the interval between two interpolating shifts (or interpolating frequencies). For AR and AI

algorithms, this is quite obvious. For the other algorithms, it might not be so easy to prove, yet this is what we observe in numerical tests. In order to maximize the error indicator, one needs to find a single maximum in each of the $n-1$ subintervals and then choose the largest of them for a global maximum. Finding a maximum in one of the subintervals requires a number of evaluations of the error indicator. Because of that it is important for those evaluations to be fast.

Using the notation:

$$\begin{aligned} \acute{A} &= \acute{B}_d^{-\frac{1}{2}} \tilde{A} \\ \acute{B} &= \tilde{B}_d^{-\frac{1}{2}} \tilde{B} \\ \acute{b} &= \tilde{B}_d^{-\frac{1}{2}} \tilde{b} \\ h_{\tilde{V}} &= \tilde{V}\beta \end{aligned} \tag{5.89}$$

the residual (5.80) (and similarly (5.86)) may be rewritten as

$$\begin{aligned} ||\acute{B}_d^{-\frac{1}{2}}((\tilde{A}+s\tilde{B})\tilde{h}_{\tilde{V}}(s)-\tilde{b})||_2^2 = ||\acute{b}||_2^2 - 2\mathrm{Re}\left(\beta^*([(\acute{A}\tilde{V})^*\acute{b}]+s[(\acute{B}\tilde{V})^*\acute{b}])\right) + \\ \beta^*\left([(\acute{A}\tilde{V})^*(\acute{A}\tilde{V})]+s[(\acute{B}\tilde{V})^*(\acute{A}\tilde{V})]+\overline{s}[(\acute{A}\tilde{V})^*(\acute{B}\tilde{V})]+|s|^2[(\acute{B}\tilde{V})^*(\acute{B}\tilde{V})]\right)\beta \\ =: ||\acute{b}||_2^2 - 2\mathrm{Re}\left(\beta^*(q_1+sq_2)\right)+\beta^*\left(Q_1+sQ_2+\overline{s}Q_3+|s|^2Q_4\right)\beta \end{aligned} \tag{5.90}$$

For ARR and AIR, β is found as a solution to (5.19), which may be rewritten as

$$([\tilde{V}^*\tilde{A}\tilde{V}]+s[\tilde{V}^*\tilde{B}\tilde{V}])\beta = \tilde{V}^*\tilde{b} \tag{5.91}$$

In the case of NARR and NAIR, β is a minimizer of (5.90). In order to make the evaluation of residual norm fast, we suggest to precompute matrices $Q_1,\ldots,Q_4$ and vectors q_1, q_2 and to use them in order to evaluate the residual norm at a given value of s. The matrices are of size $m\times m$ ($m=n$ for ARR, NARR, $m=2n$ for AIR, NAIR) and the vectors of length m. With this approach, the numerical cost of evaluation of the residual norm at each s is of the order of $O(m^3)$. And as will be seen in the numerical Section, we consider m to be not more than 30, so the cost of evaluation at a one s is much less than the cost of a single sparse matrix vector multiplication which has numerical complexity of order $O(N)$, where N is the number of edges in the domain. For example, if $N=300,000$, the average number of non-zeros per row of $\tilde{A}+s\tilde{B}$ is 72(hexahedral mesh), $m=30$, then the cost of evaluation of the error indicator at each frequency is about 2200 times less than the cost of a single complex matrix-vector multiplication. If there were 29 subintervals, we could easily evaluate the error indicator 75 times in each subinterval. This takes into account only the

numerical complexity. In practice, the residual evaluation for each frequency will be even faster as all the data will fit into the cache memory.

Also at each iteration of one of the algorithms, one has to update matrices in square brackets of (5.90) and (5.91). The numerical cost is $2(5m+2)N$ double precision floating point operations (flops) (for ARR, AIR, matrices in (5.90) and (5.91) needed) or $2(3m+2)N$ flops (for NARR, NAIR, only matrices in (5.90) needed). For $m = 30$ this is less than one sparse complex matrix-vector multiplication. Notice also that for AR or AI, the algorithms that do not use the residual as the indicator, one needs to solve equation (5.91) once, for a chosen shift s_j. So one needs the matrices $\ddot{V}^*\dot{A}\ddot{V}$, $\tilde{V}^*\dot{B}\ddot{V}$, update of which has numerical cost of $2mN$ flops. Also orthogonalization described at (5.4.2) requires us to calculate m inner products each time a vector is added, which has numerical cost of $2mN$ flops. All of this discussion shows that if the residual is used as the error indicator, it will not add a significant computation time.

The procedure described above is not able to calculate the relative residual which is less than the square root of the machine precision. This is because, simplifying a little, to calculate the square of the residual $||x-y||_2^2$, we are adding three numbers of similar magnitude:

$$||x-y||_2^2 = (x-y, x-y) = (x,x) - 2(x,y) + (y,y) \tag{5.92}$$

As a result, $||x-y||_2^2$ cannot be calculated if it is less than $\xi||x||_2^2$, where ξ is the machine precision. The residual is a square root of (5.92), so may not be evaluated if it is less than $\sqrt{\xi}||x||_2$. As a consequence, if double precision is used, the residual used as an error indicator if evaluated this way stops working when the relative residual of the approximation approaches 10^{-7} (c.f. Figure 5.3).

Approximate solution with the relative residual of 10^{-7} is considered satisfactory in many applications. If a better approximation is needed, one can use quadruple precision, which is only 4 times slower than double precision on modern machines [19] or double-double precision [20] if quadruple precision on the machine used is slow. Higher precision is needed to evaluate the residual for each frequency as well as to form matrices in square brackets in (5.90), (5.91), for which one has to evaluate a number of inner products. We checked though that for calculation of $\dot{A}\ddot{V}$, $\dot{B}\ddot{V}$ double

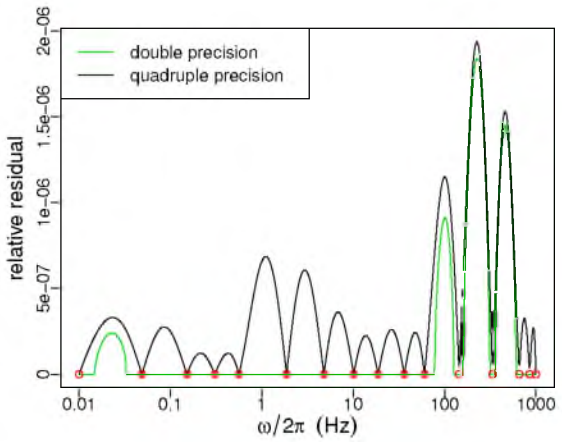

Figure 5.3. Relative residual calculated at the iteration $n = 16$ of AIR for Jacobian of E_x. The figure shows that the calculation in double precision is not sufficiently accurate. Values equal to zero indicate that the numerical value of the squared residual is non-positive.

precision is sufficient. To evaluate the inner products, the input vectors may be stored in double precision, only the calculations have to be done in quadruple precision. As a result reading from RAM is the same, only the calculations at the CPU are slower than in double precision. Example code in Fortran showing precisely what we mean is presented in the Appendix (Section 5.7).

5.4.4 The case of $\tilde{b}$ dependent on frequency ω

In the case of the forward problem approximation, when $\tilde{b}$ is complex valued and depends on the frequency, we consider only one algorithm, similar to AIR. As $\tilde{b}$ depends on the value of shift, it makes more sense to approximate it in the interval of interest $\{\mathrm{i}\omega : \omega \in [\omega_{\min}, \max]\}$.

Also this time, it is more natural to use the relative residual as the error indicator, instead of the residual. This will make our algorithm not dependent on the strength of the primary plane wave E^p (compare with Theorem 20). The evaluation of the residual is more expensive than in the case of $\tilde{b}$ not dependent on shift s. In order to find β, the solution of equation (5.91), one has to evaluate $\tilde{V}^*\tilde{b}(s)$, which is a cost of order $O(nN)$ for each value of s. To calculate the residual of the approximation, we cannot use (5.90) with precomputed matrices $Q_1, \ldots, Q_4$ and vectors q_1, q_2 as they depend on s through $\tilde{b}$. Thus we propose to evaluate the relative residual by calculating

$$\frac{||\tilde{B}_d^{-\frac{1}{2}}(([\tilde{A}\tilde{V}] + s[\tilde{B}\tilde{V}])\beta - \tilde{b}(s))||_2}{||\tilde{B}_d^{-\frac{1}{2}}\tilde{b}(s)||_2} \tag{5.93}$$

for precomputed $[\tilde{A}\tilde{V}]$, $[\tilde{B}\tilde{V}]$. The most expensive part is evaluation of $([\tilde{A}\tilde{V}] + s[\tilde{B}\tilde{V}])\beta$, which has numerical complexity $O(nN)$. This tells us that the numerical cost of evaluation of the error indicator (5.93) is of order $O(nN)$, which is more computationally expensive than in the case of $\tilde{b}$ not dependent on s.

Also, as was mentioned before, in MT, there is a number of frequencies of interest $\breve{\omega}_1, \ldots, \breve{w}_m$ log-uniformly distributed in an interval. We propose an algorithm that considers the values of interpolating shifts only in the set $\{\mathrm{i}\breve{\omega}_1, \ldots, \mathrm{i}\breve{\omega}_m\}$, for m small. If $m = 30$, the cost of evaluation of the error indicator at each of those frequencies is comparable with one matrix vector multiplication.

Moreover, as only a fixed number of frequencies is considered, we can apply the null space treatment. As a first step, one evaluates the null space part $\tilde{h}_{\tilde{K}}(\mathrm{i}\tilde{\omega}_1), \ldots, \tilde{h}_{\tilde{K}}(\mathrm{i}\omega_m)$, and then constructs the model order reduction approximation to (5.64).

The algorithm is presented below

ALGORITHM AIRD (ADAPTIVE CHOICE OF IMAGINARY SHIFTS; NORM OF THE RESIDUAL AS THE ERROR INDICATOR; $\tilde{b}$ DEPENDENT ON ω):

The algorithm is the same as AI, apart from step 2, which is replaced with

2. Find ω_{n+1} as a maximizer of the (scaled) relative residual

$$\frac{||\tilde{B}_d^{-\frac{1}{2}}((\tilde{A}+\mathrm{i}\omega\tilde{B})\tilde{h}_{\tilde{V}}(\mathrm{i}\omega)-\tilde{b}(\mathrm{i}\omega)))||_2}{||\tilde{B}_d^{-\frac{1}{2}}\tilde{b}(\mathrm{i}\omega)||_2} \tag{5.94}$$

over $\omega \in \{\omega_1, \ldots, \omega_m\}$.

5.5 Numerical results

Although the proposed methods are designed primarily for iterative solvers, in order to test the algorithms, we use a forward solver of [15]. We consider a model presented in Figure 5.4. There is a 3d hill and a 3d valley. Below the surface, there is a conductive (1Ωm) brick in 50Ωm background. The conductive object is placed at $[-700\mathrm{m}, 700\mathrm{m}] \times [-328\mathrm{m}.3, 328.3\mathrm{m}]$ in XY plane and extends from on average 450m to 1153m depth. The air is approximated by $10^7\Omega$m and the term $\mathrm{i}\omega\epsilon$ is dropped completely in the whole domain. In the numerical test, we will consider magnetotelluric response Z, K at one receiver location at the bottom of the valley, marked blue in Figure 5.4. We are interested in the response $\tilde{h}(\mathrm{i}\omega)$ for a range of frequencies considered in Magnetotellurics $\frac{\omega}{2\pi} \in [0.01\mathrm{Hz}, 1000\mathrm{Hz}]$. YZ cross-section plotted in Figure 5.5 shows the location of the object. The hexahedral mesh consists of 31, 31, and 25 elements in x, y, and z directions, respectively and it extends to 45km from the center in x and y directions. In z direction, it extends to 32km above the surface and 47km deep.

5.5.1 The case of $\tilde{b}$ not dependent on frequency ω

First, let us consider the approximation of the Jacobian, the case of $\tilde{b} = v$ for v not dependent of frequency. Vector v has been defined at (5.10) and (5.11) for the electric

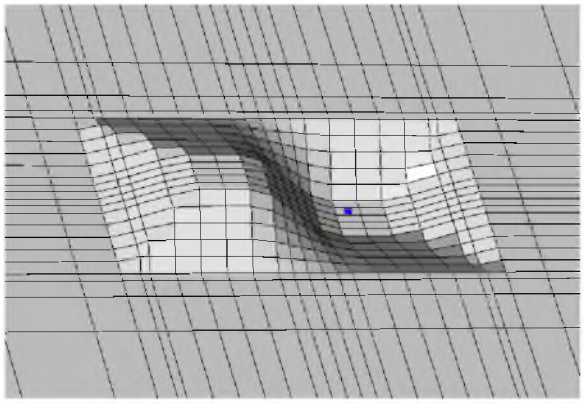

Figure 5.4. Central part of the surface mesh, together with the location of the receiver in blue

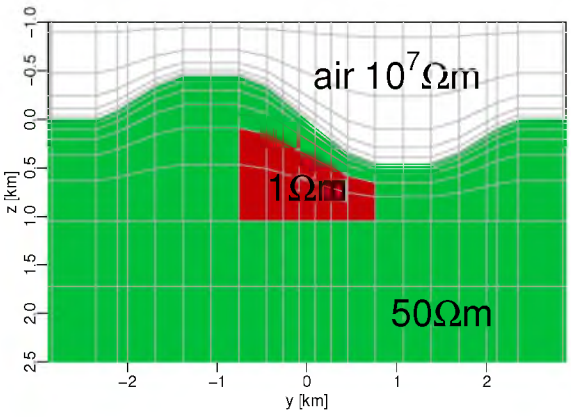

Figure 5.5. Central part of the YZ cross-section at x=0m

field E and the magnetic field H, respectively. We consider v for E corresponding to x and y directions and H corresponding to x, y, and z directions. For the magnetic field H, vector b is orthogonal to the null space of the matrix A (see Appendix in Section 5.7 for a proof), yet in the case of the electric field E, the decomposition described in Section 5.3.6 will be used.

First, consider the case of the magnetic field H. To compare the quality of approximation, we calculate the maximum relative error of approximation

$$\max_{\omega\in[\omega_{\min},\omega_{\max}]} \frac{||\tilde{h}_{\tilde{V}}(\mathrm{i}\omega)-\tilde{h}(\mathrm{i}\omega)||_2}{||\tilde{h}(\mathrm{i}\omega)||_2} \tag{5.95}$$

The max relative residual is defined as

$$\max_{\omega\in[\omega_{\min},\omega_{\max}]} \frac{||(A+\mathrm{i}wI)h_V(\mathrm{i}\omega)-b||_2}{||b||_2} \approx \max_{\omega\in[\omega_{\min},\omega_{\max}]} \frac{||\tilde{B}_d^{-\frac{1}{2}}\left[(\tilde{A}+\mathrm{i}\omega\tilde{B})\tilde{h}_{\tilde{V}}(\mathrm{i}\omega)-\tilde{b}\right]||_2}{||\tilde{B}_d^{-\frac{1}{2}}\tilde{b}||_2} \tag{5.96}$$

In Figure 5.6, we compare the strategies that evaluate the error indicator over the purely imaginary interval shown in red in Figure 5.2. AIR and NAIR strategies decrease the maximal relative error (5.95) with a similar speed, reaching the value of 10^{-10} in about 30 iterations. The strategy AI, using an error indicator, which is a part of the residual, based on the idea of [9] does poorly in comparison. To understand the reason for that, in Figure 5.7 we plot the relative error of approximation and the error indicator as a function of frequency $\omega \in [\omega_{\min}, \omega_{\max}]$ for two consecutive iterations for all three strategies. One can see that all of the strategies tend to focus on the high frequency part putting more interpolation shifts there. AI strategy does it to the biggest extent and thus fails to reduce the error of approximation at low frequencies.

Let us now consider strategies that use real valued shifts: AR, ARR, NARR. In this case, we need an estimate of minimum $\lambda_{\min}$ and maximum $\lambda_{\max}$ eigenvalues from effective spectrum of A. We do not discuss the ways to do that, but for the purpose of comparison with the imaginary shifts strategies, we use minimum and maximum of the spectrum of $\tilde{V}^*\tilde{A}\tilde{V}$, where $\tilde{V}$ is a matrix obtained by applying the strategy AI. As a result, our estimates may give a slightly narrower interval. But it focuses only on the important part of the spectrum, the one related to $U(b)$, defined in (14).

Max relative error and max relative residual as a function of iteration n is presented in Figure 5.8. One can see that the strategy of [9], AR does more poorly than the

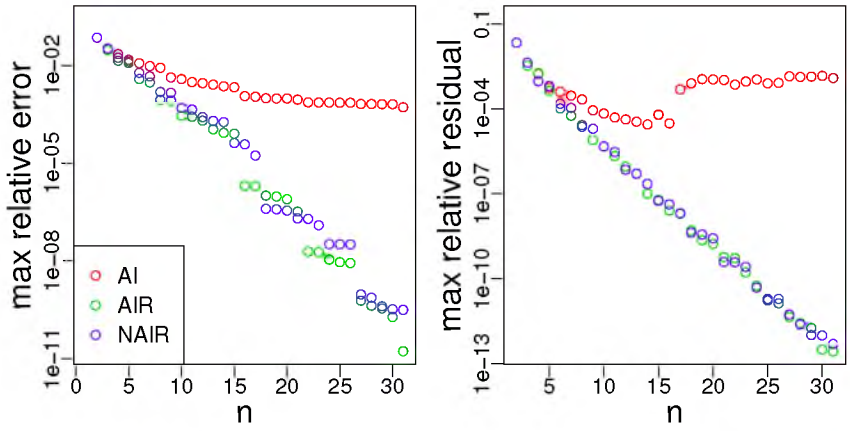

Figure 5.6. Relative error(left) and the relative residual(right) as a function of iteration number n for strategies that choose imaginary shifts. The case of a Jacobian of z component of the magnetic field H.

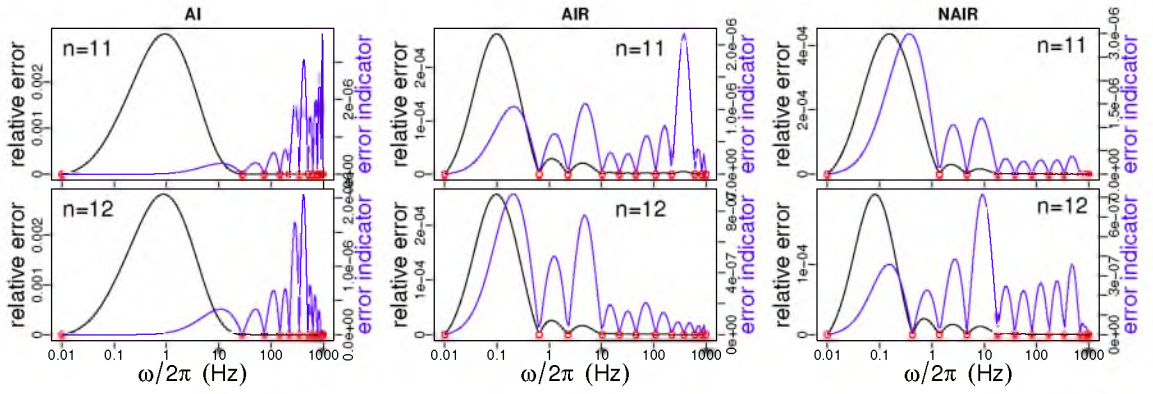

Figure 5.7. The true relative error(black) and the error indicator(blue) as a function of the frequency for two consecutive iterations(n=11,12). For three strategies: AI, AIR, NAIR. The case of a Jacobian of z component of the magnetic field H.

strategies based on residual as the error indicator: ARR, NARR. In Figure 5.9, we present the errors as a function of frequency. One can see that AR focuses more on the high frequency end. This explains why it reduces the max relative error of approximation slower than the other strategies.

All of the real shifts strategies seem to perform worse than the strategies based on imaginary interpolating shifts: AI, AIR, NAIR. One has to remember, though, that one iteration of a real shift strategy needs a solution of equation (5.15) for $s \in \mathbb{R}$ and as $\tilde{b} \in \mathbb{R}^N$, everything is real valued, so the solution time is less. One real matrix vector multiplication is 4 times faster than a complex matrix vector multiplication. Moreover, the real system matrix is hermitian, related to a minimization of a quadratic functional, whereas for complex s, we have a saddle point problem, so the number of iterations needed to solve (5.15) for real s should be less than for complex s. In Figure 5.10, we plot max relative error and max relative residual for all of the strategies as a function of time, for which one unit is the cost of solving (5.15) once with complex s. We assume that the solution for real s is four times faster. Comparing this way, real shift strategies appear to be two times faster in reducing the max relative error of approximation.

To assess the importance of the null space treatment, let us consider the case of calculation of the Jacobian for the electric field E, the case when b is not orthogonal to the null space of A. In this case, the estimation is done according to (5.70). We consider the maximum relative error of approximation of the part orthogonal to the null space of A, using model order reduction techniques:

$$\max_{\omega \in [\omega_{\min}, \omega_{\max}]} \frac{||\tilde{h}_{\tilde{W},\tilde{V}}(\mathrm{i}\omega) - \tilde{h}_{\tilde{W}}(\mathrm{i}\omega)||_2}{||\tilde{h}_{\tilde{W}}(\mathrm{i}\omega)||_2} \tag{5.97}$$

We also consider the maximum of total relative error of approximation

$$\max_{\omega \in [\omega_{\min}, \omega_{\max}]} \frac{||[\tilde{h}_{\tilde{W},\tilde{V}}(\mathrm{i}\omega) + \tilde{h}_{\tilde{K}}(\mathrm{i}\omega)] - \tilde{h}(\mathrm{i}\omega)||_2}{||\tilde{h}(\mathrm{i}\omega)||_2} \tag{5.98}$$

Notice that in the case of a Jacobian of the magnetic field H, $\tilde{h}_{\tilde{K}}(\mathrm{i}\omega) = 0$, so (5.97) coincides with (5.98) and is equal to (5.95).

In Figure 5.11, we present the max relative error and max relative residual as a function of iteration number for strategies selecting imaginary shifts: AI, AIR, NAIR.

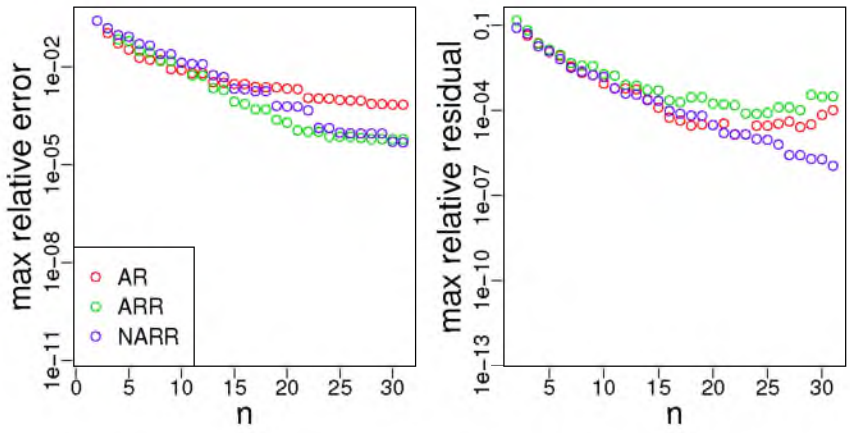

Figure 5.8. Relative error(left) and the relative residual(right) as a function of iteration number n for strategies that choose real shifts. The case of a Jacobian of z component of the magnetic field H.

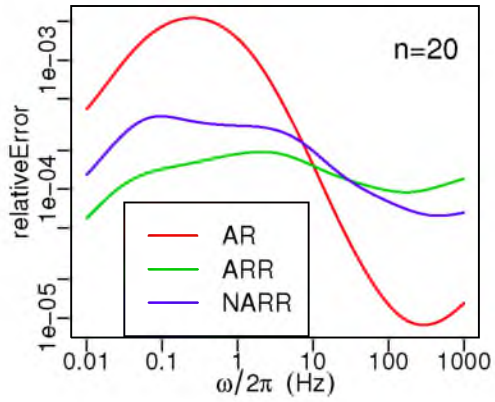

Figure 5.9. The true relative error as a function of the frequency for one iteration(n=20). For three strategies: AR, ARR, NARR. The case of a Jacobian of z component of the magnetic field H.

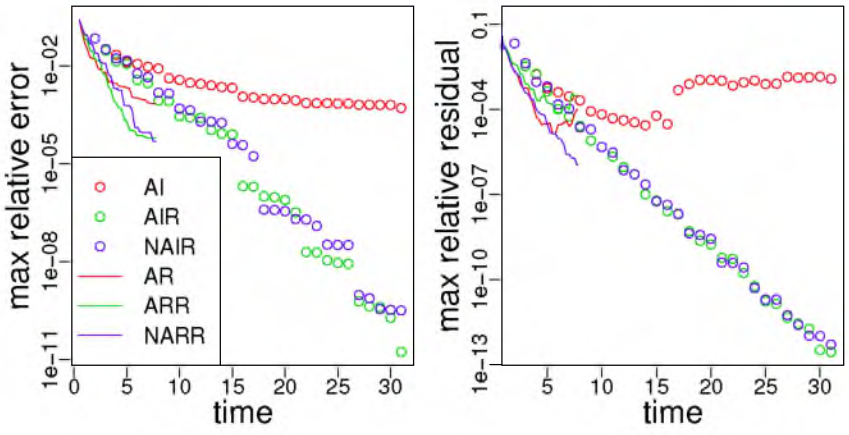

Figure 5.10. Relative error(left) and the relative residual(right) as a function of iteration number n for all strategies. The case of a Jacobian of z component of the magnetic field H. One unit of time is the cost of solving (5.15) once with complex s.

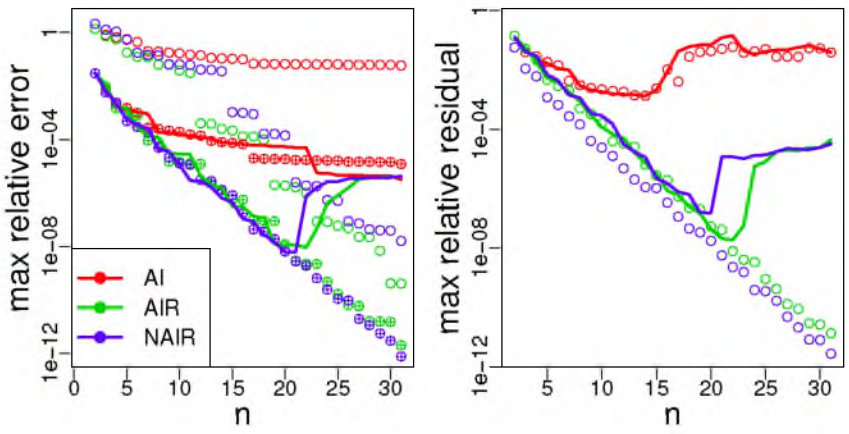

Figure 5.11. Relative error(left) and the relative residual(right) as a function of iteration number n for strategies choosing imaginary shifts. The case of a Jacobian of x component of the electric field E. $\bigcirc$ denotes the error (5.97) of approximation of the part orthogonal to the null space of A, $\oplus$ corresponds to the total error (5.98), the solid line corresponds to using model order reduction for the original b, without any null space treatment. The residuals of approximation of h_W and h do not differ substantially, so only one is plotted.

If one looks at the error of model order reduction (5.97) (denoted in the figure by $\bigcirc$), the comparison between strategies is similar to the case of the Jacobian for the magnetic field H. AIR and NAIR perform better than AI. The total error (5.98) (denoted in the figure by $\oplus$) is around three orders of magnitude less. Again AI performs worse than AIR, NAIR.

We compared those results with a case of not considering the null space treatment at all. In this case, the model order reduction is used for the equation with the original vector b. The max relative error (5.95) and the corresponding relative residual are denoted by a solid line on 5.11. One can see that the strategies without the null space treatment are less stable numerically. They diverge beyond $n = 20$. What is interesting though, the quality of approximation is almost the same as the quality of approximation with the null space treatment for the first 20 iterations. This is a little surprising at first. Recollect that for a given s, the rational approximation f satisfies (5.38). And for the original b, measure μ, defined at (5.39), contains a significant portion of mass at point 0, so f has to be good estimator of $\frac{1}{\lambda+s}$ also at $\lambda = 0$. It might though be that the difference between null space treatment and no null space treatment is not that large because $\lambda = 0$ is only one point. Hence $h(s)$ differ from $h_W(s)$ only by adding a single vector multiplied by $\frac{1}{s}$ (c.f. Section 5.3.6). So if one has a space $\mathrm{colsp}(V)$ that is good for approximation of $h_W(s)$, it is enough to add one vector to be able to approximate $h(s)$ with the same quality.

This result encourages us to consider the strategies without the null space treatment, with real shifts in $s_j \in [\lambda_{\min}, \lambda_{\max}]$, where $\lambda_{\min}, \lambda_{\max}$ are bounds of the essential spectrum. In Figure 5.12, we present the max relative error and max relative residual as a function of iteration number n. One can see that the strategies without the null space treatment perform similarly to the strategies with the null space treatment.

Additionally, although ARR and NARR are better than AR in decreasing the error (5.97) of approximating $\tilde{h}_{\tilde{W}}$, when the total error (5.98) is considered, strategy AR seems to perform better. It seems to be a result of a two errors canceling each other, which we describe below.

First notice that the error of approximation of $\tilde{h}_{\tilde{W}(s)}$, (5.97) and the total error (5.98) differ by a multiplicative constant, dependent on frequency:

$$\frac{||[\tilde{h}_{\bar{W},\bar{V}}(\mathrm{i}\omega)+\tilde{h}_{\bar{K}}(\mathrm{i}\omega)]-\tilde{h}(\mathrm{i}\omega)||_2}{||\tilde{h}(\mathrm{i}\omega)||_2}=\frac{||\tilde{h}_{\bar{W},\bar{V}}(\mathrm{i}\omega)-\tilde{h}_{\bar{W}}(\mathrm{i}\omega)||_2}{||\tilde{h}_{\bar{W}}(\mathrm{i}\omega)||_2}C(\mathrm{i}\omega)$$

for

$$C(\mathrm{i}\omega)=\frac{||\tilde{h}_{\bar{W}}(\mathrm{i}\omega)||_2}{||\tilde{h}(\mathrm{i}\omega)||_2}=\frac{||\tilde{h}_{\bar{W}}(\mathrm{i}\omega)||_2}{||\tilde{h}_{\bar{W}}(\mathrm{i}\omega)+\frac{1}{\mathrm{i}\omega}(\tilde{K}^*\tilde{B}\tilde{K})^{-1}\tilde{K}^*\tilde{b}||_2} \tag{5.99}$$

The value of C as a function of frequency is plotted in Figure 5.13. One can see that it spans almost 5 orders of magnitude. If we are interested in values of $\tilde{h}(\mathrm{i}\omega)$, and we use model order reduction for approximation of $\tilde{h}_{\bar{W}}(\mathrm{i}\omega)$, the ideal error indicator function should focus on high frequency more. AR strategy focuses on high frequency (see Figure 5.9), thus although the error of approximating $\tilde{h}_{\bar{W}}$ is bigger for it than for ARR of NARR, the total error is less for AR.

The conclusion one might draw is that in order to profit from null space treatment, one has to deal with the factor $C(\mathrm{i}\omega)$ in a proper way.

To calculate the speedup of calculation of Jacobian, one has to do the following. Assume that the iterative solver is used to solve the linear systems, and that the solver is such that the speed of reducing the log of residual does not depend on the value of the relative residual, i.e., the same number of iterations needed to reduce the relative residual from 10^{-1} to 10^{-2} is the same as the number of iterations needed to reduce the relative error from 10^{-5} to 10^{-6}. It is true for solvers that use multigrid techniques, divergence correction, and if the mesh does not have high aspect ratio elements [21, 22]. Assume further that the approximation $\tilde{h}_{\bar{V}}(\mathrm{i}\omega)$ is used as $\tilde{h}(\mathrm{i}\omega)$ if it has sufficiently small residual and if not, it is used as a starting guess for the iterative solver. Assume that the cost of applying the model reduction techniques is negligible compared with the cost of calculating $(\tilde{A}+\mathrm{i}\omega\tilde{B})^{-1}\tilde{b}$ for one ω using an iterative solver. If we are interested in finding solutions that have relative residuals no more than 10^{-6} for 30 frequencies, in our numerical test, the speedup of using AIR or NAIR versus not using model order reduction is 4 times. The speedup is higher if more frequencies are needed. For example, for 60 frequencies, the speedup would be 8 times. The speedup may be different for different geometries of the model and for finer discretization.

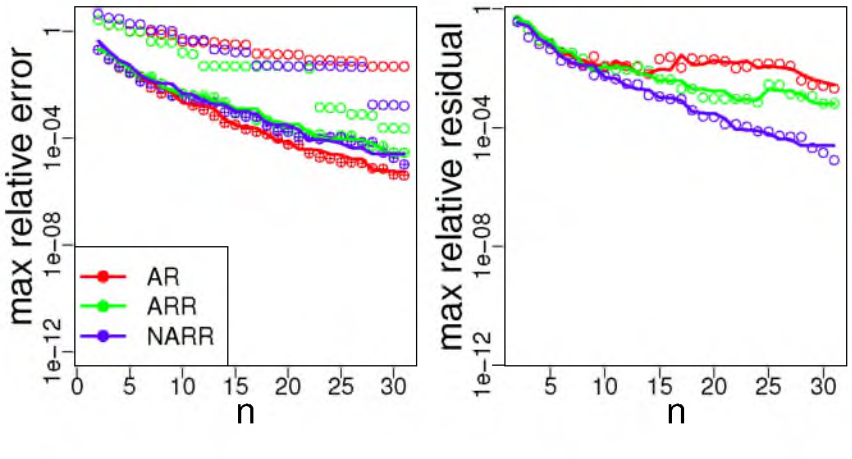

Figure 5.12. Relative error(left) and the relative residual(right) as a function of iteration number n for strategies choosing real shifts. The case of a Jacobian of x component of the electric field E. $\bigcirc$ denotes the error (5.97) of approximation of the part orthogonal to the null space of A, $\oplus$ corresponds to the total error (5.98), the solid line corresponds to using model order reduction to the original b, without any null space treatment. The residuals of approximation of h_W and h do not differ substantially, so only one is plotted.

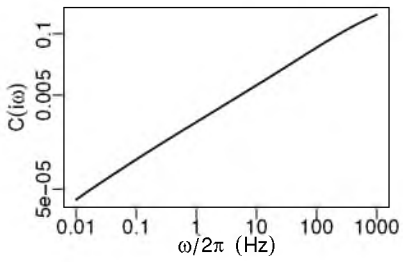

Figure 5.13. The ratio of relative errors $C(\mathrm{i}\omega)$ defined at (5.99). The case of a Jacobian of x component of the electric field E.

5.5.2 The case of b dependent on frequency ω

Let us consider the approximation of the forward MT response, so the case of $\tilde{b}(\mathrm{i}\omega) = g(\omega)$ where $g(\omega)$, defined at (5.6) depends on the primary plane wave field. Usually in MT, one considers two cases: E field purely in x direction(with H purely in y direction) and E field purely in y direction(with H purely in x direction).

As the source in equation (5.2), the primary plane wave electric field multiplied by the conductivity difference $(\sigma - \sigma^p)E^p$, is not divergence free, the vector $b(\mathrm{i}\omega)$ is not orthogonal to the null space of A. Moreover, its part lying in the null space $b_K(\mathrm{i}\omega)$ depends on ω. Thus we can expect the null space treatment to have a bigger impact on the quality of approximation than in the case of $\tilde{b}$ not dependent on ω.

We consider algorithm AIRD for 31 frequencies of interest log-uniformly distributed in $[\omega_{\min}, \omega_{\max}]$. In Figure 5.14, we present maximum over the whole interval $[\omega_{\min}, \omega_{\max}]$ of the relative error and relative residual as a function of iteration number. This time, the null space treatment decreases the relative error of approximation by nearly two orders of magnitude, allowing us to decrease the maximum relative error of approximation to 10^{-7} in 25 iterations. For n larger than 25, we can see a stagnation caused by the limitation of the choice of interpolation frequencies to $\{\breve{\omega}_1, \ldots, \breve{\omega}_m\}$

In Figure 5.15, we present a similar plot, but this time, the maximum of the relative error and the relative residual is taken only over the frequencies of interest $\breve{\omega}_1, \ldots, \breve{\omega}_m$. The main difference with the previous figure is seen for large n.

We calculate the time savings from using the AIRD algorithm, with the same assumptions as in the case of Jacobian. If we are interested in finding a solution that has relative residual no more than 10^{-6} for 30 frequencies, the speedup AIRD versus not using model order reduction is 2 times. So the speedup in the forward problem case is two times less than in the case of the Jacobian. There may be two reasons for that. One is that this time $\tilde{b}$ depends on ω, the other is that we limit the values of frequencies to $\breve{\omega}_1, \ldots, \breve{\omega}_{30}$, thus the location of interpolation points is not optimal.

Although the speedup for using model order reduction techniques is less for the forward problem case, the majority of calculation at each iteration of the Gauss-Newton method is due to calculation of the Jacobian. If the same mesh is used for all of the receivers, one needs two linear solves to obtain two ξ (one for each plane

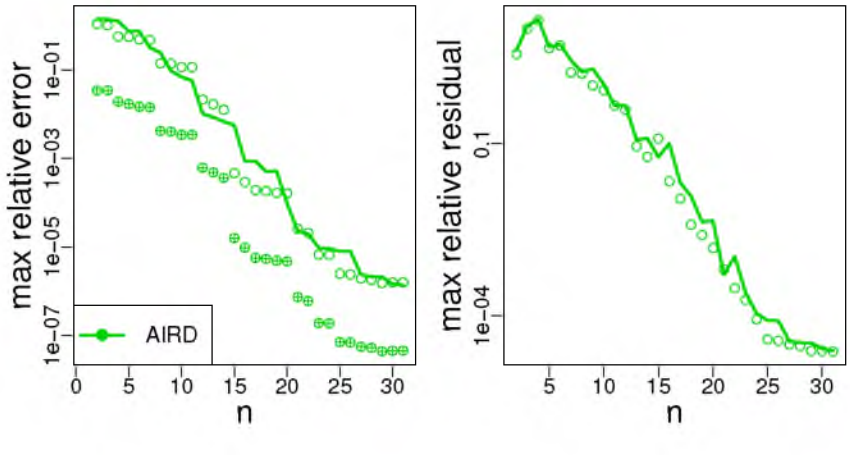

Figure 5.14. Relative error(left) and the relative residual(right) as a function of iteration number n for strategies choosing imaginary shifts. The case of the forward problem for source plane wave with E purely in x direction. $\bigcirc$ denotes the error (5.97) of approximation of the part orthogonal to the null space of A, $\oplus$ corresponds to the total error (5.98), the solid line corresponds to using model order reduction to the original b, without any null space treatment. The residuals of approximation of h_W and h do not differ substantially, so only one is plotted.

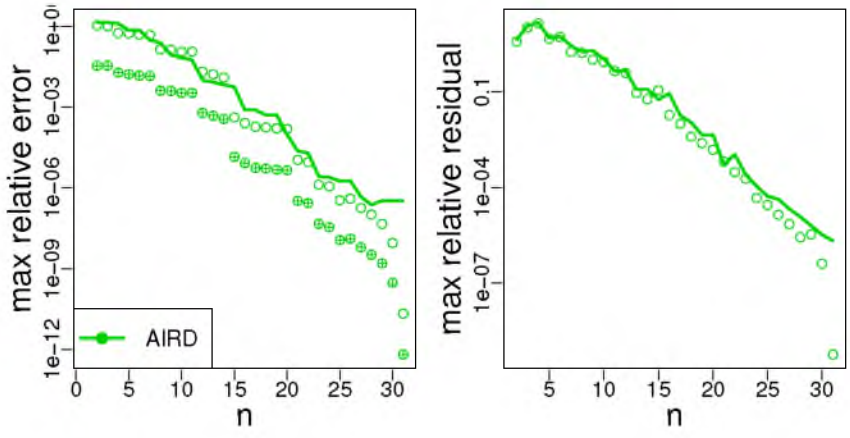

Figure 5.15. Same Figure as 5.14, except for the max error and max residual being calculated only among the trial frequencies $\breve{\omega}_1, \ldots, \breve{\omega}_{31}$

wave polarization) that are needed in the forward response calculation (5.12) and subsequently as much as 5 times the number of receivers linear solves to obtain the Jacobian in (5.13). With 100 MT receivers in the domain, the ratio of workload is 500 to 2. In this case, one does not need to consider model order reduction for the forward problem at all.

Obviously, one could only be interested in forward modeling and then the speedup of the forward problem matters, but this speedup might be of interest also in inversion using the Gauss-Newton method. The method we propose for the calculation of the Jacobian considers each receiver separately, so one could use a different mesh for each receiver (local meshes). In this case, one needs two forward problem solves for each receiver, and then the workload is 5 to 2 so the speedup of the forward solve matters much more.

5.6 Acknowledgments

We acknowledge the support of this work from the U.S. Dept. of Energy under contract DE-EE0002750 to PW. EC acknowledges the partial support of the U.S. National Science Foundation through grants ARC-0934721 and DMS-1413454.

5.7 Appendix

Proof : (proof of Theorem 15) Let $\hat{\lambda}_j$, γ_j be eigenvalues and eigenvectors of V^*AV. As V^*AV is hermitian, non-negative definite

$$\hat{\lambda}_j \geq 0, \quad j = 1, \ldots, n \tag{5.100}$$

and γ_j might be chosen to be orthonormal. Using representation (5.35), define $z_j \in \mathcal{V}$ such that

$$V\gamma_j = Uz_j(\Lambda)U^*b \tag{5.101}$$

functions z_j are μ orthonormal eigenvectors of operator of multiplication by λ in $\mathcal{V}$:

$$\delta(i=j) = \gamma_i^*\gamma_j = \gamma_i^*V^*V\gamma_j = (V\gamma_i)^*(V\gamma_j) = (Uz_i(\Lambda)U^*b)^*Uz_j(\Lambda)U^*b =< z_j, z_i >_\mu \tag{5.102}$$

$$\hat{\lambda}_j < z_j, z_i >_\mu = \hat{\lambda}_j\gamma_i^*\gamma_j = \gamma_i^*(\hat{\lambda}_j\gamma_j) = \gamma_i^*(V^*AV\gamma_j) = (V\gamma_i)^*A(V\gamma_j) =< \lambda z_j, z_i >_\mu \tag{5.103}$$

Using the latter and (5.38), one can calculate the coefficients of f in eigendirections z_j to obtain:

$$f(\lambda, s) = \sum_{j=1}^{n} \frac{< 1, z_j >_\mu}{\hat{\lambda}_j + s} z_j(\lambda) \tag{5.104}$$

Using the definition (5.36) of $\mathcal{V}$, let p_j be a polynomial of degree at most $n-1$ such that

$$z_j(\lambda) = \frac{p_j(\lambda)}{\prod_{i=1}^{n}(\lambda + s_i)} \tag{5.105}$$

This allows us to rewrite (5.104) as

$$f(\lambda, s) = \sum_{j=1}^{n} \left(\frac{< 1, z_j >_\mu}{\hat{\lambda}_j + s} \frac{p_j(\lambda)}{\prod_{i=1}^{n}(\lambda + s_i)} \right) = \frac{p(\lambda, s)}{\prod_{j=1}^{n}(\hat{\lambda}_j + s)(\lambda + s_j)} \tag{5.106}$$

for some p, which is a polynomial of the order at most $n-1$ with respect to λ as well as with respect to s. Define

$$\check{f}(\lambda, s) = \frac{1}{\lambda + s} \left[1 - \prod_{j=1}^{n} \frac{(s - s_j)(\lambda - \hat{\lambda}_j)}{(\lambda + s_j)(s + \hat{\lambda}_j)} \right] = \frac{\prod_{j=1}^{n}(\lambda + s_j)(s + \hat{\lambda}_j) - \prod_{j=1}^{n}(s - s_j)(\lambda - \hat{\lambda}_j)}{(\lambda + s)\prod_{j=1}^{n}(\lambda + s_j)(s + \hat{\lambda}_j)}$$

following the idea of [23], we notice that if we plug in $-s$ for λ, then the numerator is 0. Thus the polynomial in the numerator is divisible by $(\lambda + s)$. This allows us to write $\check{f}$ as

$$\check{f}(\lambda, s) = \frac{\check{p}(\lambda, s)}{\prod_{j=1}^{n}(\hat{\lambda}_j + s)(\lambda + s_j)}$$

where $\check{p}$ is a polynomial of order at most $n-1$ of s and of λ. In order to prove Theorem 15, it remains to show that $p = \check{p}$.

Define

$$h'_V(s) = U\check{f}(\Lambda, s)U^*b \tag{5.107}$$

Using (5.21) and (5.28), and the definition of $\check{f}$, we have

$$\begin{aligned}
h_V(s) &= h'_V(s_j), && \text{for } s = s_j,\ j = 1, \ldots, n \\
Uf(\Lambda, s)U^*b &= U\check{f}(\Lambda, s)U^*b, && \text{for } s = s_j,\ j = 1, \ldots, n \\
f(\Lambda, s)U^*b &= \check{f}(\Lambda, s)U^*b, && \text{for } s = s_j,\ j = 1, \ldots, n \\
f(\lambda_k, s)u_k^*b &= \check{f}(\lambda_k, s)u_k^*b, && \text{for } s = s_j,\ j = 1, \ldots, n,\quad k = 1, \ldots, N
\end{aligned}$$

Let us now fix λ_k. Assuming $u_k^*b \neq 0$, we have

$$\frac{p(\lambda_k, s)}{\prod_{i=1}^{n}(\hat{\lambda}_i + s)(\lambda_k + s_i)} = \frac{\check{p}(\lambda_k, s)}{\prod_{i=1}^{n}(\hat{\lambda}_i + s)(\lambda_k + s_i)}, \quad \text{for } s = s_j,\ j = 1, \ldots, n$$

As $\lambda_k, \hat{\lambda}_i \geq 0$, with the assumption (5.18), we can conclude that

$$\begin{array}{ll} p(\lambda_k, s) = \check{p}(\lambda_k, s), & \text{for } s = s_j,\ j = 1, \ldots, n \\ p(\lambda_k, s) - \check{p}(\lambda_k, s) = 0, & \text{for } s = s_j,\ j = 1, \ldots, n \end{array}$$

$p(\lambda_k, s) - \check{p}(\lambda_k, s)$ is a polynomial of degree at most $n-1$. As it has n distinct roots, it has to be equal to 0. We have obtained:

$$u_k^* b \neq 0 \ \Rightarrow \ \forall s \ p(\lambda_k, s) = \check{p}(\lambda_k, s)$$

Now, fix s. $p(\lambda, s) - \check{p}(\lambda, s)$ is a polynomial of λ of degree at most $n-1$. With the assumption that the number of distinct λ_k such that $u_k^* b \neq 0$ is greater or equal n, this polynomial has at least n distinct roots, so it is equal to 0. We have obtained that

$$p(\lambda, s) = \check{p}(\lambda, s), \quad \forall \lambda, s$$

which implies

$$f(\lambda, s) = \check{f}(\lambda, s), \quad \forall \lambda, s$$

which in turn implies (5.40).

Theorem 22 *For any values $s_j \notin (-\infty, 0]$, the eigenvalues $\hat{\lambda}_j$ of V^*AV minimize the norm (5.74).*

Proof : The square of the norm may be rewritten as

$$\begin{array}{l} \left|\left| U \left(\prod_{j=1}^n (\Lambda - \hat{\lambda}_j I) \prod_{j=1}^n (\Lambda + s_j I)^{-1} \right) U^* b \right|\right|_2^2 = \int_0^\infty \left| \prod_{j=1}^n (\lambda - \hat{\lambda}_j) \right|^2 \left| \prod_{j=1}^n (\lambda - s_j) \right|^{-2} d\mu(\lambda) \\ = \int_0^\infty \prod_{j=1}^n (\lambda - \hat{\lambda}_j)^2 d\tau(\lambda) = ||\tilde{\pi}_n||_\tau^2 \end{array} \tag{5.108}$$

where the measure τ is defined as

$$d\tau(\lambda) = \left| \prod_{j=1}^n (\lambda - s_j) \right|^{-2} d\mu(\lambda) \tag{5.109}$$

and the polynomial $\tilde{\pi}_n$ is

$$\tilde{\pi}_n(\lambda) = \prod_{j=1}^n (\lambda - \hat{\lambda}_j) \tag{5.110}$$

The expression (5.103) for eigenvalues $\hat{\lambda}_j$ and eigenvectors z_j, defined at (5.101) may be rewritten using the measure τ as

$$< (\lambda - \hat{\lambda}_j) p_j, p_i >_\tau = 0 \tag{5.111}$$

for all $i, j = 1, \ldots, n$. Polynomial p_j has been defined at (5.105). Consider a polynomial $\pi_n(\lambda)$ of order n, which is τ orthogonal to all the polynomials of degree at most $n-1$. If the measure τ has support consisting of at least n points (which is true if the number of distinct eigenvalues λ_k for which $u_k^* b \neq 0$ is at least n), then π_n has n distinct zeros. This is a standard result of the theory of orthogonal polynomials, the proof is not given here. According to (5.111), we have

$$\pi_n(\lambda) = C_j(\lambda - \hat{\lambda}_j)p_j$$

where C_j is a constant. This shows that $\hat{\lambda}_j$ for $j = 1, \ldots, n$ are roots of π_n. Thus

$$\pi_n(\lambda) = C\prod_{j=1}^{n}(\lambda - \hat{\lambda}_j) = C\tilde{\pi}_n(\lambda) \tag{5.112}$$

for some constant C.

If one varies the values $\hat{\lambda}$, the quantity (5.108) yields $||p(\lambda)||_\tau^2$ for some polynomial p of degree n and the coefficient next to λ^n equal to 1. If p is divided by π_n, one obtains

$$p(\lambda) = \pi_n(\lambda) + r(\lambda) \tag{5.113}$$

where the degree of r is at most $n-1$. Because of (5.112), π_n is τ orthogonal to r and thus

$$||p||_\tau^2 = ||\pi_n||_\tau^2 + ||r||_\tau^2$$

so

$$||p||_\tau^2 \geq ||\pi_n||_\tau^2$$

Example code for high precision inner product follows. Calculations are done in quadruple precision, yet input vectors are in double precision.

```
subroutine HighPrecisionInnerProduct(N,x,y,res)
  implicit none
  integer ,intent(in)  :: N
  real(8) ,intent(in)  :: x(N),y(N)  !! input in double precision
  real(16),intent(out) :: res        !! output in quadruple precision
```

```
  ! ! local variables:
  integer :: i
  res = 0
  do i = 1,N
     res = res + real(y(i),16)*real(x(i),16) !! quadruple prec.
  enddo
endsubroutine HighPrecisionInnerProduct
```

Theorem 23 *If* $b = \tilde{B}^{\frac{1}{2}} v$ *for* $v = v_x^H, v_y^H$ *or* v_z^H, *defined at (5.11), then*

$$b \perp null(A)$$

Proof : We present a proof for lowest order edge elements on hexahedral grid. A proof for tetrahedral grid is similar.

We have to prove that

$$0 = K^* b = \tilde{K}^* v$$

for $\tilde{K}$ defined (5.66). Consider an element, plotted in Figure 5.16, that contains the location $\mathbf{r}$ of the receiver. As $\tilde{K}$ corresponds to discrete gradient ∇ operator, $\tilde{K}^*$ corresponds to negative discrete divergence $-\nabla\cdot$ operator. Each row of $\tilde{K}$ corresponds to a vertex, and each entry of $\tilde{K}^* v$ corresponds to a value of sink at the vertex. The values v_e are added for all edges connected to the vertex, with the sign related to the orientation towards the vertex. If we consider the vertex j in Figure 5.16, only edges e_1, e_2, e_3 need to be considered as $v_e = 0$ for other edges and

$$\begin{bmatrix} \tilde{K}^* v_x \\ \tilde{K}^* v_y \\ \tilde{K}^* v_z \end{bmatrix}_j = \sum_e (\tilde{K}_{j,e})^* (\nabla \times S_e)(r) = (\nabla \times S_{e_1})(r) + (\nabla \times S_{e_2})(r) + (\nabla \times S_{e_3})(r) = 0$$

The above is true, as $\nabla \times S_{e_1} + \nabla \times S_{e_2} + \nabla \times S_{e_3} = 0$. To explain the latter, notice that for each edge e, $\nabla \times S_e$ is a member of $\mathcal{H}_0^h(\nabla\cdot)$, which have degrees of freedom being fluxes through faces. For face f_1 in Figure 5.16, using Stokes theorem, we obtain

$$(\nabla \times S_{e_1} + \nabla \times S_{e_2} + \nabla \times S_{e_3})_{f_1} = (\nabla \times S_{e_1} + \nabla \times S_{e_2})_{f_1} = 1 - 1 = 0$$

Similarly for f_2, f_3. As all the coefficients of vector field $\nabla \times S_{e_1} + \nabla \times S_{e_2} + \nabla \times S_{e_3}$ are 0, it is identically zero.

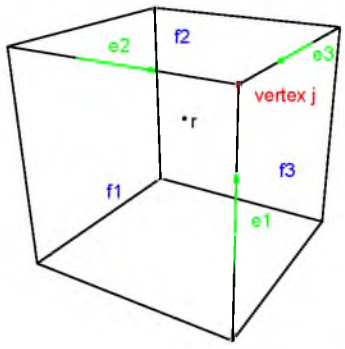

Figure 5.16. Plot showing a hexahedral element used in the proof of Theorem 23

5.8 References

[1] M. Rewienski and J. White, "A trajectory piecewise-linear approach to model order reduction and fast simulation of nonlinear circuits and micromachined devices," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 22, no. 2, pp. 155–170, 2003.

[2] X.-D. T. Sheldon and H. Lei, *Advanced Model Order Reduction Techniques in VLSI Design*. Cambridge University Press, 2007.

[3] J. S. Han, E. B. Rudnyi, and J. G. Korvink, "Efficient optimization of transient dynamic problems in mems devices using model order reduction," *Journal of Micromechanics and Microengineering*, vol. 15, no. 4, p. 822, 2005.

[4] Z. Bai, "Krylov subspace techniques for reduced-order modeling of large-scale dynamical systems," *Appl. Numer. Math.*, vol. 43, pp. 9–44. 2002.

[5] R. W. Freund, "Model reduction methods based on Krylov subspaces," *Acta Numer.*, vol. 12, pp. 267–319. 2003.

[6] E. J. Grimme, "Krylov projection methods for model reduction," PhD thesis, Citeseer, 1997.

[7] S. Gugercin, A. Antoulas, and C. Beattie, "A rational Krylov iteration for optimal H2 model reduction," in *Proceedings of the 17th International Symposium on Mathematical Theory of Networks and Systems, Kyoto, Japan*, 2006, pp. 1665–1667.

[8] L. Knizhnerman, V. Druskin, and M. Zaslavsky, "On optimal convergence rate of the rational Krylov subspace reduction for electromagnetic problems in unbounded domains," *SIAM Journal on Numerical Analysis*, vol. 47, no. 2, pp. 953–971, 2009.

[9] V. Druskin, C. Lieberman, and M. Zaslavsky, "On adaptive choice of shifts in rational Krylov subspace reduction of evolutionary problems," *SIAM Journal on Scientific Computing*, vol. 32, no. 5, pp. 2485–2496, 2010.

[10] V. Druskin, V. Simoncini, and M. Zaslavsky, "Solution of the time-domain inverse resistivity prolem in the model reduction framework part I: one-dimensional problem with SISO data," *SIAM Journal on Scientific Computing*, vol. 35, no. 3, A1621–A1640, 2013.

[11] M. Zaslavsky, V. Druskin, A. Abubakar, T. Habashy, and Simoncini, "Large-scale Gauss-Newton inversion of transient CSEM data using the model order reduction framework," *Geophysics*, vol. 78, no. 4, E161–E171, 2013.

[12] J. F. Villena and L. M. Silveira, "ARMS - automatic residue-minimization based sampling for multi-point modeling techniques," in *DAC 09 Proceedings of the 46th Annual Design Automation Conference*, 2009, pp. 951–956.

[13] J.-C. Nédélec, "A new family of mixed finite elements in R^3," *Numer. Math.*, vol. 50, no. 1, pp. 57–81, 1986, ISSN: 0029-599X. DOI: 10.1007/BF01389668. [Online]. Available: http://dx.doi.org/10.1007/BF01389668.

[14] P. B. Bochev and M. D. Gunzburger, *Least-Squares Finite Element Methods*. Springer New York, 2009.

[15] M. Kordy, P. Wannamaker, V. Maris, and E. Cherkaev, "Three-dimensional magnetotelluric inversion including topography using deformed hexahedral edge finite elements and direct solvers parallelized on SMP computers, Part I: forward problem and parameter jacobians," *submitted to Geophys. J. Int.*, 2014.

[16] M. Kordy, P. Wannamaker, V. Maris, E. Cherkaev, and G. J. Hill, "Three-dimensional magnetotelluric inversion including topography using deformed hexahedral edge finite elements and direct solvers parallelized on SMP computers, Part II: Direct data-space inverse solution," *submitted to Geophys. J. Int.*, 2014.

[17] W. Siripunvaraporn, G. Egbert, Y. Lenbury, and M. Uyeshima, "Three-dimensional magnetotelluric inversion: data-space method," *Phys. Earth Planet. Inter.*, vol. 150, pp. 3–14, 2005.

[18] G. A. Baker Jr and P. Graves-Morris, *Padé approximants*. Encyclopedia of Mathematics and its Applications., 1996.

[19] H. Hasegawa, "Utilizing the quadruple-precision floating-point arithmetic operation for the Krylov subspace methods," in *8th SIAM Conference on Applied Linear Algebra*, 2003.

[20] Y. Hida, X. S. Li, and D. H. Bailey, "Library for double-double and quad-double arithmetic," *NERSC Division, Lawrence Berkeley National Laboratory*, 2007.

[21] W. Mulder, "A multigrid solver for 3D electromagnetic diffusion," *Geophysical Prospecting*, vol. 54, no. 5, pp. 633–649, 2006.

[22] T. Kolev and P. Vassilevski, "Some experience with a H1-based auxiliary space AMG for H(curl) problems," *Lawrence Livermore Nat. Lab., Livermore, CA, Rep. UCRL-TR-221841*, 2006.

[23] I. V. Oseledets, "Lower bounds for separable approximations of the Hilbert kernel," *Sbornik: Mathematics*, vol. 198, no. 3, p. 425, 2007.