Charles University in Prague
Faculty of Mathematics and Physics

MASTER THESIS

Vojtěch Rybář

Finite elements for electromagnetics compatible with de Rham diagram

Department of Numerical Mathematics

Supervisor of the master thesis: prof. Ing. Ivo Doležel, CSc.
Study programme: Mathematics
Specialization: Numerical and Computational Mathematics

Prague, 2011

I would like to express my deepest gratitude to prof. Ing. Ivo Doležel, CSc. for generously supervising this thesis, to RNDr. Pavel Šolín, Ph.D. for inspiration and insights and to RNDr. Tomáš Vejchodský, Ph.D. This thesis would not been finished without his crucial guidance, patience and general support.

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular the fact that the Charles University in Prague has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 paragraph 1 of the Copyright Act.

Prague, April 15, 2011 Vojtěch Rybář

Název práce: Konečné prvky v elektromagnetismu kompatibilní s de Rhamovým diagramem
Autor: Vojtěch Rybář
Katedra (Ústav): Katedra numerické matematiky
Vedoucí diplomové práce: prof. Ing. Ivo Doležel, CSc.

Abstrakt: Předložená práce je věnována konečným prvkům nejnižšího řádu pro řešení časově periodických Maxwelových rovnic ve dvou dimenzích. Úspěšná aproximace těchto rovnic vyžaduje, aby prostory konečných prvků byly kompatibilní s de Rhamovým diagramem. Avšak nejpoužívanější bázové funkce (tzv. Whitneyho funkce) tomuto diagramu nevyhovují. Proto zkonstruujeme kompatibilní báze a studujeme jejich vlastnosti. Protože tato konstrukce není jednoznačná, vyšetřujeme vliv konkrétního výběru báze na podmíněnost příslušných matic konečných prvků. Nakonec využíváme speciální strukturu matic tuhosti, navrhujeme několik iteračních metod a porovnáváme jejich konvergenci.

Klíčová slova: Maxwellovy rovnice, hranové konečné prvky, de Rhamův diagram, báze prostoru konečných prvků

Title: Finite elements for electromagnetics compatible with de Rham diagram
Author: Vojtěch Rybář
Department: Department of Numerical Mathematics
Supervisor: prof. Ing. Ivo Doležel, CSc.

Abstract: The present work is devoted to the lowest-order finite elements for solving time-harmonic Maxwell's equations in two dimensions. Successful approximation of these equations requires the finite element spaces to be compatible with the de Rham diagram. However, the most often used basis functions (the Whitney functions) do not comply with this diagram. Therefore, we construct compatible bases and study their properties. Since the construction is not unique, we investigate the influence of the particular choice on the conditioning of the corresponding finite element matrices. Finally, we utilize the special structure of the stiffness matrices, propose a few iterative schemes, and compare their convergence.

Keywords: Maxwell's equations, edge finite element, de Rham diagram, finite element basis

Contents

Preface

The de Rham diagram is an operator scheme relating the spaces $H^1(\Omega)$, $\boldsymbol{H}(\mathrm{curl}, \Omega)$ and $L^2(\Omega)$ in two spatial dimensions through the gradient and curl operators. The diagram is exact in the sense that the null space of any operator in the diagram exactly coincides with the range of the preceding one. The compatibility of the finite element spaces with the de Rham diagram is essential for their good approximation properties as well as for the analysis (e.g. for interpolation error estimation).

In this thesis we deal with the finite element discretization of the time-harmonic Maxwell's equations. The most widely used lowest-order edge finite elements in two dimensions utilize so-called Whitney functions (named after Hassler Whitney). These functions, however, are not compatible with the de Rham diagram. The main idea of the presented thesis is to construct a basis compatible with this diagram. Such basis consists of two parts: curl-free gradients of scalar functions and nongradient vector fields.

As it turns out, the gradient part of the basis is given by gradients of the standard scalar piecewise linear "hat" vertex functions. The nongradient part consists of a selection of Whitney functions. Such selection corresponds to suitable choice of edges in the finite element triangulation. Our main goal will be to investigate the effect of the choice of the nongradient part on the performance of the finite element method.

Chapter 1

Time-harmonic Maxwell's equations and their discretization

This introductory chapter brings in Maxwell's equations with their time-harmonic variant and their discretization by the finite element method. While the former part is derived from [7], the latter relies on [3] and [11].

1.1 Maxwell's equations

The classical macroscopic electromagnetic field is described in terms of four vector functions of position $\boldsymbol{x} \in \mathbb{R}^3$ and time $t \in \mathbb{R}$:

- electric field strength $\boldsymbol{\mathcal{E}} = \boldsymbol{\mathcal{E}}(\boldsymbol{x}, t)$,
- electric flux density $\boldsymbol{\mathcal{D}} = \boldsymbol{\mathcal{D}}(\boldsymbol{x}, t)$,
- magnetic field strength $\boldsymbol{\mathcal{H}} = \boldsymbol{\mathcal{H}}(\boldsymbol{x}, t)$,
- magnetic flux density $\boldsymbol{\mathcal{B}} = \boldsymbol{\mathcal{B}}(\boldsymbol{x}, t)$.

The electromagnetic field is created by a distribution of sources consisting of static electric charges and a directed flow of the electric charge – the electric current. The distribution of charge is given by the scalar density function ρ, while the currents are described by the vector-valued density function $\boldsymbol{\mathcal{J}}$.

The Maxwell's equations in the differential form then consist of four relations among field variables and sources:

$$\frac{\partial \boldsymbol{\mathcal{B}}}{\partial t} + \nabla \times \boldsymbol{\mathcal{E}} = \mathbf{0}, \qquad \text{(Faraday's law)}$$

$$\nabla \cdot \boldsymbol{\mathcal{D}} = \rho, \qquad \text{(Gauss's law for electricity)}$$

$$\frac{\partial \boldsymbol{\mathcal{D}}}{\partial t} - \nabla \times \boldsymbol{\mathcal{H}} = -\boldsymbol{\mathcal{J}}, \qquad \text{(Ampère's law)}$$

$$\nabla \cdot \boldsymbol{\mathcal{B}} = 0. \qquad \text{(Gauss's law for magnetism)}$$

Provided that charge is conserved, the Gauss' laws could be obtained from the other two. This could be shown by taking the divergences and recalling that $\nabla \cdot (\nabla \times \boldsymbol{F}) = 0$ for any vector field $\boldsymbol{F}$. Indeed, taking the divergence of the Faraday's law we get

$$\nabla \cdot \frac{\partial \boldsymbol{\mathcal{B}}}{\partial t} = 0, \tag{1.1}$$

as applying the same operator on the Ampère's law results in

$$\nabla \cdot \frac{\partial \boldsymbol{\mathcal{D}}}{\partial t} = -\nabla \cdot \boldsymbol{\mathcal{J}}. \tag{1.2}$$

Now, if the charge is conserved and the density $\boldsymbol{\mathcal{D}}$ is sufficiently smooth, then the quantities ρ and $\boldsymbol{\mathcal{J}}$ are in the following relation (using (1.2) and the Gauss's law for electricity):

$$\nabla \cdot \boldsymbol{\mathcal{J}} + \frac{\partial \rho}{\partial t} = 0. \tag{1.3}$$

For sufficiently smooth $\boldsymbol{\mathcal{B}}$ a combination of (1.1) and (1.3) gives

$$\frac{\partial}{\partial t} \nabla \cdot \boldsymbol{\mathcal{B}} = \frac{\partial}{\partial t} (\nabla \cdot \boldsymbol{\mathcal{D}} - \rho) = 0.$$

This implies that if the Gauss's laws hold at one arbitrary time, then they hold for all time.

Nevertheless, numerical schemes must produce approximations that take into account the Gauss's laws in order to be successful in solving discretized problems.

1.2 Time-harmonic Maxwell's equations

The time-dependent system of Maxwell's equations can be reduced to time-harmonic system. This could be done either by the Fourier transformation in time or by assuming that all time-varying quantities are harmonic with frequency $\omega > 0$.

The electromagnetic field is said to be time-harmonic, provided

$$\boldsymbol{\mathcal{E}}(\boldsymbol{x},t) = \mathrm{Re}\left(\exp(-\mathrm{i}\omega t)\hat{\boldsymbol{E}}(\boldsymbol{x})\right), \tag{1.4a}$$

$$\boldsymbol{\mathcal{D}}(\boldsymbol{x},t) = \mathrm{Re}\left(\exp(-\mathrm{i}\omega t)\hat{\boldsymbol{D}}(\boldsymbol{x})\right), \tag{1.4b}$$

$$\boldsymbol{\mathcal{H}}(\boldsymbol{x},t) = \mathrm{Re}\left(\exp(-\mathrm{i}\omega t)\hat{\boldsymbol{H}}(\boldsymbol{x})\right), \tag{1.4c}$$

$$\boldsymbol{\mathcal{B}}(\boldsymbol{x},t) = \mathrm{Re}\left(\exp(-\mathrm{i}\omega t)\hat{\boldsymbol{B}}(\boldsymbol{x})\right), \tag{1.4d}$$

where i stands for the imaginary unit and Re denotes the real part. Thus, $\hat{\boldsymbol{E}}, \hat{\boldsymbol{D}}, \hat{\boldsymbol{H}}$ and $\hat{\boldsymbol{B}}$ are now complex-valued vector functions of the position only.

For consistency, the current density and the charge density are needed to be time-harmonic as well, so we put

$$\boldsymbol{\mathcal{J}}(\boldsymbol{x},t) = \mathrm{Re}\left(\exp(-\mathrm{i}\omega t)\hat{\boldsymbol{J}}(\boldsymbol{x})\right), \tag{1.5a}$$

$$\rho(\boldsymbol{x},t) = \mathrm{Re}\left(\exp(-\mathrm{i}\omega t)\hat{\rho}(\boldsymbol{x})\right). \tag{1.5b}$$

Substituting (1.4) and (1.5) into the Maxwell's equations leads to their time-harmonic variant:

$$-\mathrm{i}\omega\hat{\boldsymbol{B}} + \nabla\times\hat{\boldsymbol{E}} = \boldsymbol{0}, \tag{1.6a}$$

$$\nabla\cdot\hat{\boldsymbol{D}} = \hat{\rho}, \tag{1.6b}$$

$$-\mathrm{i}\omega\hat{\boldsymbol{D}} - \nabla\times\hat{\boldsymbol{H}} = -\hat{\boldsymbol{J}}, \tag{1.6c}$$

$$\nabla\cdot\hat{\boldsymbol{B}} = 0. \tag{1.6d}$$

1.3 Constitutive equations for media

Equations (1.6) must be augmented by certain constitutive laws. The first two constitutive laws couple quantities $\hat{\boldsymbol{E}}$, $\hat{\boldsymbol{D}}$, $\hat{\boldsymbol{H}}$ and $\hat{\boldsymbol{B}}$ with the particular material via the following relations:

$$\hat{\boldsymbol{D}} = \epsilon\hat{\boldsymbol{E}}, \tag{1.7a}$$

$$\hat{\boldsymbol{B}} = \mu\hat{\boldsymbol{H}}. \tag{1.7b}$$

The symbols ϵ and μ denote the electric permitivity and the magnetic permeability, respectively. They characterize the electromagnetic properties of the material. We distinguish the following types:

- in *homogeneous* medium all parameters are independent of the position,
- in *linear* medium the parameters are independent of the electromagnetic field,
- in *isotropic* materials these characteristics do not depend on the direction of the electromagnetic field.

The third constitutive relation takes into account the current that rises by the electromagnetic field itself:

$$\hat{\boldsymbol{J}} = \sigma \hat{\boldsymbol{E}} + \hat{\boldsymbol{J}}_{\boldsymbol{a}}, \tag{1.8}$$

where the non-negative function of position σ is called the conductivity and the vector function $\hat{\boldsymbol{J}}_{\boldsymbol{a}}$ describes the applied current density.

The most common case in practice is the linear, inhomogeneous and isotropic material (e.g. air, copper, etc.). In these cases the parameters ϵ and μ are positive, bounded, scalar functions of position only. For such type of media, the system of time-harmonic Maxwell's equations could be rewritten (using (1.6) together with (1.7) and (1.8)) in the following form:

$$\begin{aligned} -\mathrm{i}\omega\mu\hat{\boldsymbol{H}} + \nabla \times \hat{\boldsymbol{E}} &= 0, \\ \nabla \cdot (\epsilon\hat{\boldsymbol{E}}) &= \frac{1}{\mathrm{i}\omega}\nabla \cdot (\sigma\hat{\boldsymbol{E}} + \hat{\boldsymbol{J}}_{\boldsymbol{a}}), \\ -\mathrm{i}\omega\epsilon\hat{\boldsymbol{E}} + \sigma\hat{\boldsymbol{E}} - \nabla \times \hat{\boldsymbol{H}} &= -\hat{\boldsymbol{J}}_{\boldsymbol{a}}, \\ \nabla \cdot (\mu\hat{\boldsymbol{H}}) &= 0. \end{aligned}$$

Since it is convenient to work with relative parameter values, we define

$$\boldsymbol{E} = \sqrt{\epsilon_0}\hat{\boldsymbol{E}} \quad \text{and} \quad \boldsymbol{H} = \sqrt{\mu_0}\hat{\boldsymbol{H}}, \tag{1.10}$$

where ϵ_0 and μ_0 are the permitivity and permeability of vacuum. Using (1.10) and defining the relative permitivity and permeability by

$$\epsilon_r = \frac{1}{\epsilon_0}\left(\epsilon + \frac{\mathrm{i}\sigma}{\omega}\right) \quad \text{and} \quad \mu_r = \frac{\mu}{\mu_0}, \tag{1.11}$$

we obtain the final version of the first-order system of time-harmonic Maxwell's equations:

$$-\mathrm{i}\kappa\mu_r \boldsymbol{H} + \nabla \times \boldsymbol{E} = 0, \tag{1.12a}$$

$$\nabla \cdot (\epsilon_r \boldsymbol{E}) = -\frac{1}{\kappa^2} \nabla \cdot \boldsymbol{F}, \tag{1.12b}$$

$$-\mathrm{i}\kappa\epsilon_r \boldsymbol{E} - \nabla \times \boldsymbol{H} = -\frac{1}{\mathrm{i}\kappa} \boldsymbol{F}, \tag{1.12c}$$

$$\nabla \cdot (\mu_r \boldsymbol{H}) = 0, \tag{1.12d}$$

where $\boldsymbol{F} = \mathrm{i}\kappa\sqrt{\mu_0}\hat{\boldsymbol{J}}_{\boldsymbol{a}}$ and $\kappa = \omega\sqrt{\epsilon_0\mu_0}$.

It is possible to derive numerical methods for (1.12), but usually the magnetic field $\boldsymbol{H}$ is eliminated by solving (1.12a) for $\boldsymbol{H}$ and substituting into (1.12c) to obtain the second-order time-harmonic Maxwell's system

$$\nabla \times (\mu_r^{-1} \nabla \times \boldsymbol{E}) - \kappa^2 \epsilon_r \boldsymbol{E} = \boldsymbol{F} \tag{1.13}$$

together with (1.12b).

1.4 Interface and boundary conditions

Equations (1.12) or (1.13) do not provide complete description of the electromagnetic field since they are defined in the interior of the computational domain only. They do not hold on boundaries between two different materials where either ϵ_r or μ_r are discontinuous.

Let us consider the case of two media with different electric and magnetic properties separated by an interface S with unit normal $\boldsymbol{\nu}$ pointing from region 2 to region 1. The integral form of Maxwell's equations, which we do not present here, but which can be found for example in [11, Subsection 7.1.2] yields the following conditions for $\nabla \times \boldsymbol{E}$ on the interface S:

$$(\boldsymbol{E}_1 - \boldsymbol{E}_2) \times \boldsymbol{\nu} = 0, \tag{1.14a}$$

$$(\epsilon_{r,1}\boldsymbol{E}_1 - \epsilon_{r,2}\boldsymbol{E}_2) \cdot \boldsymbol{\nu} = \rho_S, \tag{1.14b}$$

where the subscripts denote the limit values of the coefficients and the field variables from the particular side of the interface S. Notice that the quantity ρ_S is known as the interface charge density. Analogously for the magnetic field strength we obtain

$$(\boldsymbol{H}_1 - \boldsymbol{H}_2) \times \boldsymbol{\nu} = \boldsymbol{J}_S, \tag{1.15a}$$

$$(\mu_{r,1}\boldsymbol{H}_1 - \mu_{r,2}\boldsymbol{H}_2) \cdot \boldsymbol{\nu} = 0, \tag{1.15b}$$

where the tangential vector field $\boldsymbol{J}_S$ stands for the interface current density on S.

If the problem takes place in the unbounded domain, then the easiest way of solving it is to restrict the electromagnetic field to a sufficiently large bounded domain Ω by imposing artificial boundary conditions on $S = \partial\Omega$:

$$\boldsymbol{E} \cdot \boldsymbol{\nu} = 0, \tag{1.16a}$$

$$\boldsymbol{H} \cdot \boldsymbol{\nu} = 0. \tag{1.16b}$$

These conditions make the fields $\boldsymbol{E}$ and $\boldsymbol{H}$ tangential to the boundary $\partial\Omega$.

When the material in the outer domain is a *perfect conductor* (with the conductivity being infinite), then we have from (1.8) that if the current density $\boldsymbol{J}$ is to remain bounded then the electric field $\boldsymbol{E}$ has to vanish in the outer domain. This yields to the following boundary condition:

$$\boldsymbol{E} \times \boldsymbol{\nu} = 0. \tag{1.17}$$

If the material on one side of the boundary is not a perfect conductor and if it lets the field to go through only a small distance, then we obtain *imperfect conductor* or *impedance* boundary condition

$$\boldsymbol{\nu} \times \boldsymbol{H} - \lambda(\boldsymbol{\nu} \times \boldsymbol{E}) \times \boldsymbol{\nu} = 0, \tag{1.18}$$

where the impedance λ is a positive function of position on the interface S.

1.5 Weak formulation

Let us assume a bounded simply-connected domain $\Omega \subset \mathbb{R}^3$ with Lipschitz-continuous boundary $\partial\Omega$ that consists of disjoint open parts Γ_P and Γ_I. We are interested in solving (1.13) in Ω, i.e.

$$\nabla \times (\mu_r^{-1} \nabla \times \boldsymbol{E}) - \kappa^2 \epsilon_r \boldsymbol{E} = \boldsymbol{F} \quad \text{for} \quad \forall \boldsymbol{x} \in \Omega, \tag{1.19}$$

with perfect conductor boundary condition on

$$\boldsymbol{E} \times \boldsymbol{\nu} = 0 \quad \text{on} \quad \Gamma_P, \tag{1.20}$$

and impedance boundary condition (1.18) that regarding the normalization (1.10) receives the form

$$\mu_r^{-1}(\nabla \times \boldsymbol{E}) \times \boldsymbol{\nu} - \mathrm{i}k\lambda \boldsymbol{E}_T = \boldsymbol{g} \quad \text{on} \quad \Gamma_I, \tag{1.21}$$

where $E_T = (\boldsymbol{\nu} \times \boldsymbol{E}) \times \boldsymbol{\nu}$.

Testing (1.19) by a sufficiently smooth vector-valued complex functions $\boldsymbol{\Phi} \in \mathbb{C}^3$ and integrating over Ω, we obtain

$$\int_\Omega \left[\nabla \times \left(\mu_r^{-1} \nabla \times \boldsymbol{E}\right) \cdot \boldsymbol{\Phi} - \kappa^2 \epsilon_r \boldsymbol{E} \cdot \boldsymbol{\Phi}\right] \mathrm{d}\boldsymbol{x} = \int_\Omega \boldsymbol{F} \cdot \boldsymbol{\Phi} \, \mathrm{d}\boldsymbol{x},$$

with the scalar product of two complex vectors $\boldsymbol{a} \in \mathbb{C}^n$ and $\boldsymbol{b} \in \mathbb{C}^n$ defined as

$$\boldsymbol{a} \cdot \boldsymbol{b} = \sum_{i=1}^{n} a_i \bar{b}_i.$$

Using Green's theorem together with

$$\begin{aligned} \nabla \cdot (\boldsymbol{a} \times \boldsymbol{b}) &= (\nabla \times \boldsymbol{a}) \cdot \boldsymbol{b} - \boldsymbol{a} \cdot (\nabla \times \boldsymbol{b}), \\ \boldsymbol{a} \cdot (\boldsymbol{b} \times \boldsymbol{c}) &= (\boldsymbol{a} \times \boldsymbol{b}) \cdot \boldsymbol{c}, \end{aligned}$$

we get

$$\begin{aligned} \int_\Omega \left[\left(\mu_r^{-1} \nabla \times \boldsymbol{E}\right) \cdot (\nabla \times \boldsymbol{\Phi}) - \kappa^2 \epsilon_r \boldsymbol{E} \cdot \boldsymbol{\Phi}\right] \mathrm{d}\boldsymbol{x} + \\ + \int_{\partial\Omega} \boldsymbol{\nu} \times \left(\mu_r^{-1} \nabla \times \boldsymbol{E}\right) \cdot \boldsymbol{\Phi}_T \, \mathrm{d}s = \int_\Omega \boldsymbol{F} \cdot \boldsymbol{\Phi} \, \mathrm{d}\boldsymbol{x}, \quad (1.22) \end{aligned}$$

where $\boldsymbol{\Phi}_T = (\boldsymbol{\nu} \times \boldsymbol{\Phi}) \times \boldsymbol{\nu}$ is the tangential projection of vector $\boldsymbol{\Phi}$ to the boundary $\partial\Omega$. Next let us incorporate the boundary conditions (1.20) and (1.21) into (1.22). The perfect boundary condition states that the field $\boldsymbol{E}$ is normal to the boundary Γ_P and (1.20) implies that

$$\boldsymbol{\Phi}_T = 0 \text{ on } \Gamma_P.$$

This choice eliminates the Γ_P-part of the surface integral in (1.22). Applying the impedance boundary condition (1.21) on Γ_I, we obtain

$$\begin{aligned} \int_\Omega \left[\left(\mu_r^{-1} \nabla \times \boldsymbol{E}\right) \cdot (\nabla \times \boldsymbol{\Phi}) - \kappa^2 \epsilon_r \boldsymbol{E} \cdot \boldsymbol{\Phi}\right] \mathrm{d}\boldsymbol{x} - \\ - \int_{\Gamma_I} \mathrm{i}k\lambda \boldsymbol{E}_T \cdot \boldsymbol{\Phi}_T \, \mathrm{d}s = \int_\Omega \boldsymbol{F} \cdot \boldsymbol{\Phi} \, \mathrm{d}\boldsymbol{x} + \int_{\Gamma_I} \boldsymbol{g} \cdot \boldsymbol{\Phi}_T \, \mathrm{d}s. \quad (1.23) \end{aligned}$$

From which we can see that the suitable space for $\boldsymbol{E}$ is

$$\boldsymbol{V} = \{\boldsymbol{E} \in \boldsymbol{H}(\mathrm{curl}, \Omega) \mid \boldsymbol{\nu} \times \boldsymbol{E} = \boldsymbol{0} \text{ on } \Gamma_P\},$$

where the Hilbert space $\boldsymbol{H}(\mathrm{curl}, \Omega)$ contains all $(L^2(\Omega))^3$ functions whose distributional curl lies in $(L^2(\Omega))^3$,

$$\boldsymbol{H}(\mathrm{curl}, \Omega) = \{\boldsymbol{E} \in (L^2(\Omega))^3 \mid \nabla \times \boldsymbol{E} \in (L^2(\Omega))^3\}. \tag{1.24}$$

The space $\boldsymbol{V}$ together with the inner product

$$(\boldsymbol{E}, \boldsymbol{F})_{\boldsymbol{V}} = \int_\Omega (\nabla \times \boldsymbol{E}) \cdot (\nabla \times \boldsymbol{F}) \, \mathrm{d}\boldsymbol{x} + \int_\Omega \boldsymbol{E} \cdot \boldsymbol{F} \, \mathrm{d}\boldsymbol{x} + \int_{\Gamma_I} \boldsymbol{E}_T \cdot \boldsymbol{F}_T \, \mathrm{d}s \tag{1.25}$$

is a Hilbert space. We denote the corresponding norm by $||\boldsymbol{E}||^2_{\boldsymbol{V}} = (\boldsymbol{E}, \boldsymbol{E})_{\boldsymbol{V}}$. Provided $(\boldsymbol{E}, \boldsymbol{F})_\Omega = \int_\Omega \boldsymbol{E} \cdot \boldsymbol{F} \, \mathrm{d}\boldsymbol{x}$, the (1.25) can be rewritten in the following manner:

$$(\boldsymbol{E}, \boldsymbol{F})_{\boldsymbol{V}} = (\nabla \times \boldsymbol{E}, \nabla \times \boldsymbol{F})_\Omega + (\boldsymbol{E}, \boldsymbol{F})_\Omega + (\boldsymbol{E}_T, \boldsymbol{F}_T)_{\Gamma_I}. \tag{1.26}$$

In order to be able to use the standard theory for existence and uniqueness of the weak solution – which will be introduced in the following paragraph – we need to introduce several assumptions on the coefficients and data, see [11]:

1. Ω may be decomposed into n subdomains $\Omega_1, \Omega_2, \ldots, \Omega_n$;
2. $\overline{\Omega} = \bigcup_{i=1}^n \overline{\Omega}_i$; $\Omega_i \cap \Omega_j = \emptyset, \ \forall i \neq j$;
3. Ω_i is connected and has a Lipschitz boundary $\forall i = 1, \ldots, n$;
4. the coefficient μ_r is smooth in each subdomain Ω_i;
5. the coefficient ϵ_r has the following properties:
   - the restriction of ϵ_r to each subdomain Ω_i is a function in $H^3(\Omega_i)$,
   - there exists a constant $C_\epsilon > 0$ such that for each subdomain Ω_i, either $\mathrm{Im}(\epsilon_r) \geq C_\epsilon$ or $\mathrm{Im}(\epsilon_r) = 0, \ \forall i = 1, \ldots, n$;
6. the right-hand side $\boldsymbol{F}$ lies in $(L^2(\Omega))^3$.

Here, we point out that $\mathrm{Im}(\epsilon_r)$ denotes the imaginary part of the complex number ϵ_r.

Finally, under the above assumptions, the weak formulation of (1.19) with (1.20) and (1.21) reads as:

Find the electric field $\boldsymbol{E} \in \boldsymbol{V}$ satisfying

$$a(\boldsymbol{E}, \boldsymbol{\Phi}) = l(\boldsymbol{\Phi}) \quad \text{for all } \boldsymbol{\Phi} \in \boldsymbol{V}, \tag{1.27}$$

where the sesquilinear form $a : \boldsymbol{V} \times \boldsymbol{V} \to \mathbb{C}$ is defined as

$$a(\boldsymbol{E}, \boldsymbol{\Phi}) = (\mu_r^{-1} \nabla \times \boldsymbol{E}, \nabla \times \boldsymbol{\Phi})_\Omega - \kappa^2 (\epsilon_r \boldsymbol{E}, \boldsymbol{\Phi})_\Omega - \mathrm{i}k(\lambda \boldsymbol{E}_T, \boldsymbol{\Psi}_T)_{\Gamma_I}, \tag{1.28}$$

and the linear form $l : \boldsymbol{V} \to \mathbb{C}$ as

$$l(\boldsymbol{\Phi}) = (\boldsymbol{F}, \boldsymbol{\Phi})_\Omega + (\boldsymbol{g}, \boldsymbol{\Phi}_T)_{\Gamma_I}. \tag{1.29}$$

The existence and uniqueness are shown in [7].

1.6 Time harmonic Maxwell's equations in two dimensions

Every 2D problem is equivalent to a 3D problem whose solution does not depend on the last variable, i.e., such that the resulting field has the form

$$\boldsymbol{E} = (E_1(x_1, x_2), E_2(x_1, x_2), 0).$$

With this observation in mind, we test the equation (1.19) with

$$\boldsymbol{\Psi} = (\Psi_1(x_1, x_2), \Psi_2(x_1, x_2), 0).$$

Then we can apply formally the same procedure as in the 3D case to get (1.23). Thus, if we consider the definition

$$\nabla \times \boldsymbol{E} = \frac{\partial \boldsymbol{E}_1}{\partial x_2} - \frac{\partial \boldsymbol{E}_2}{\partial x_1} \tag{1.30}$$

for a 2D vector field $\boldsymbol{E} \in \mathbb{C}^2$, we can define the 2D variant of the $\boldsymbol{H}(\mathrm{curl}, \Omega)$ space:

$$\boldsymbol{H}(\mathrm{curl}, \Omega) = \{\boldsymbol{E} \in (L^2(\Omega))^2 \mid \nabla \times \boldsymbol{E} \in (L^2(\Omega))^2\}. \tag{1.31}$$

The definition of the weak solution of the 2D Maxwell's equations then reads: find $\boldsymbol{E} \in \boldsymbol{V}$ such that

$$a(\boldsymbol{E}, \boldsymbol{\Phi}) = l(\boldsymbol{\Phi}) \quad \text{for all } \boldsymbol{\Phi} \in \boldsymbol{V}, \tag{1.32}$$

with $a : \boldsymbol{V} \times \boldsymbol{V} \to \mathbb{C}$ and $l : \boldsymbol{V} \to \mathbb{C}$ defined in (1.28) and (1.29) for $\boldsymbol{V}$ defined in (1.31). The only difference is that the 2D definition (1.30) of the curl operator has to be considered here.

1.7 Discretization by edge elements

In this section we are moving with our 2D problem to the finite dimensional functional space. For simplicity, we will suppose that Ω is polygonal domain. We cover this domain with a regular mesh $\mathcal{T}_h$ consisting of triangles $\{K\}$. More precisely:

1. $\overline{\Omega} = \bigcup_{K \in \mathcal{T}_h} K$;
2. for each $K \in \mathcal{T}_h$, K is a closed triangle with positive volume;
3. if $K_i \neq K_j$, then $\mathrm{meas}(K_1 \cap K_2) = 0$;
4. every nonempty intersection of $K_i \cap K_j, i \neq j$ can either be a single shared vertex or a whole shared edge.

Let us note that the elements $K \in \mathcal{T}_h$ are traditionally considered as closed sets. However, we will use them in symbols like $H^1(K)$ as a domain of the corresponding Sobolev space with no danger of confusion.

Through the rest of this thesis we will use the following characteristics of the triangulation $\mathcal{T}_h$:

$$\mathcal{V} \ldots \text{the set of all vertices,} \tag{1.33}$$
$$\mathcal{V}_i \ldots \text{the set of all inner vertices} \tag{1.34}$$
$$\mathcal{V}_b \ldots \text{the set of all boundary vertices,} \tag{1.35}$$
$$\mathcal{E} \ldots \text{the set of all edges,} \tag{1.36}$$
$$\mathcal{E}_i \ldots \text{the set of all inner edges,} \tag{1.37}$$
$$\mathcal{E}_b \ldots \text{the set of all boundary edges,} \tag{1.38}$$
$$N_K \ldots \text{number of triangles,} \tag{1.39}$$
$$N_v \ldots \text{number of vertices, i.e. } N_v = |\mathcal{V}|, \tag{1.40}$$
$$N_{iv} \ldots \text{number of inner vertices, i.e. } N_{iv} = |\mathcal{V}_i|, \tag{1.41}$$
$$N_{bv} \ldots \text{number of boundary vertices, i.e. } N_{ib} = |\mathcal{V}_b|, \tag{1.42}$$
$$N_e \ldots \text{number of edges, i.e. } N_e = |\mathcal{E}|, \tag{1.43}$$
$$N_{ie} \ldots \text{number of inner edges, i.e. } N_{ie} = |\mathcal{E}_i|, \tag{1.44}$$
$$N_{be} \ldots \text{number of boundary edges, i.e. } N_{ib} = |\mathcal{E}_b|. \tag{1.45}$$

We remark that the boundary edges lie completely on the boundary $\partial\Omega$, i.e. for any edge we can state that $e \in \mathcal{E}_b$ if and only if $e \subset \partial\Omega$. The inner edge is defined as not being the boundary edge.

In [11, Lemma 7.6 & Remark 7.2] there is presented a useful characterization of the conformity in $\boldsymbol{H}(\mathrm{curl}, \Omega)$. We state it as the following theorem.

Theorem 1.1. *If the vector field $\boldsymbol{E}_h : \Omega \to \mathbb{C}^2$ is a polynomial of degree at most k in each $K \in \mathcal{T}_h$, $\boldsymbol{E}_h \in [P^k(K)]^2$ for all $K \in \mathcal{T}_h$, then $\boldsymbol{E}_h \in \boldsymbol{H}(\mathrm{curl}, \Omega)$ if and only if $\boldsymbol{E}_h \cdot \boldsymbol{\tau}_e$ is continuous along each edge $e \in \mathcal{E}$, where $\boldsymbol{\tau}_e$ is the unit tangential vector to the edge e.*

With the help of Theorem 1.1, we construct the lowest order finite element approximation on $K \in \mathcal{T}_h$ by taking functions with continuous and constant tangential components on the edges of K and by setting the order of polynomial space to one:

$$P = \left\{ \boldsymbol{E}_h \in [P^1(K)]^2 \mid \boldsymbol{E}_h \cdot \boldsymbol{\tau}_{e_i} \text{ is constant for } i = 1, 2, 3 \right\}. \tag{1.46}$$

The last step to complete the standard finite element triplet (K, P, Σ) is the set Σ of linear functionals $L_i : P \to \mathbb{R}$, $i = 1, 2, 3$. We define them as

$$L_i(\boldsymbol{E}_h) = \int_{e_i} \boldsymbol{E}_h \cdot \boldsymbol{\tau}_{e_i} \mathrm{d}\xi, \quad \text{for all } \boldsymbol{E}_h \in P. \tag{1.47}$$

This element was first introduced – although in a different context – by Whitney [15] and it bears his name.

With our polygonal domain Ω covered by a finite element mesh $\mathcal{T}_h$ consisting of triangular Whitney finite elements, the Galerkin subspace $\boldsymbol{V}_h$ of the space $\boldsymbol{V}$ has the form

$$\boldsymbol{V}_h = \{ \boldsymbol{E}_h \in \boldsymbol{V} \mid \boldsymbol{E}_h|_K \in [P^1(K)]^2 \ \forall K \in \mathcal{T}_h \},$$

where we recall that $\mathcal{E}_i$ is the set of all interior edges of the triangulation $\mathcal{T}_h$.

The basis of $\boldsymbol{V}_h$ consists of so-called Whitney functions $\boldsymbol{\psi}_e, e \in \mathcal{E}_i$. The Whitney functions satisfy the following:

$$\boldsymbol{\psi}_e \cdot \boldsymbol{\tau}_e = \frac{1}{|e|} \text{ on } e \in \mathcal{E}_i, \tag{1.48}$$

$$\boldsymbol{\psi}_e \cdot \boldsymbol{\tau}_{e^*} = 0 \text{ on } e^* \in \mathcal{E}_i, e \neq e^*. \tag{1.49}$$

An exhausting description of many aspects of the Whitney elements can be found in [11, Section 7.5].

Finally, with all necessary aspects of the discretization addressed, we can introduce the discrete variational problem:

Find the discrete electric field $\boldsymbol{E}_h \in \boldsymbol{V}_h$ satisfying

$$a(\boldsymbol{E}_h, \boldsymbol{\Phi}_h) = l(\boldsymbol{\Phi}_h) \quad \text{for all } \boldsymbol{\Phi}_h \in \boldsymbol{V}_h. \tag{1.50}$$

As in the continuous case, the existence and uniqueness of such problem is discussed in [7].

Chapter 2

De Rham diagram

In this chapter we will describe mathematical framework for the analysis of the variational formulation and discretization of Maxwell's equations introduced in the previous chapter. To accomplish that, we will rely on [1],[7] and [10].

Those parts that we use in the subsequent numerical experiments will be discussed in more detail.

2.1 De Rham diagram on the continuous level

Considering three dimensions, the three differential operators $\nabla, \nabla\times$ and $\nabla\cdot$ have certain common structure: curls of gradients vanish, curls are divergence-free. This structure persists when we pass to the weak formulation.

In this section we describe the structure of differential operators $\nabla, \nabla\times$ and $\nabla\cdot$ in a rigorous way. To do so, we cannot avoid few definitions, see [1].

Definition 2.1. *The domain in $\mathbb{R}^3$ is contractible if it is simply connected with a connected boundary*

Definition 2.2. *A family of vector spaces (real or complex) $X^0, \ldots, X^d$ and of linear maps A^p from X^{p-1} to X^p, $p = 1, \ldots, d$, forms an exact sequence at the level of X^p if $\mathrm{im}(A^p) = \ker(A^{p+1})$ for $p = 1, \ldots, d-1$, if A^1 is injective in case $p = 0$, and if A^d is surjective in case $p = d$.*

We notice that $\mathrm{im}(A^p)$ and $\ker(A^{p+1})$ stand for the image and kernel of the appropriate operator.

Definition 2.3. *An exact sequence is the one which is exact at all levels.*

It is convenient to discuss sequences with the help of diagrams of the following form:

$$\mathbb{R} \to X^0 \xrightarrow{A^1} X^1 \xrightarrow{A^2} \dots \xrightarrow{A^{d-1}} X^{d-1} \xrightarrow{A^d} X^d \to 0.$$

In such diagrams, arrows are labeled with operators. The image, by any of these operators, of the space left to its arrow, is in the kernel of the next operator on the right.

For contractible domain $\Omega \subset \mathbb{R}^3$ the sequence

$$\mathbb{R} \to C^\infty(\overline{\Omega}) \xrightarrow{\nabla} \boldsymbol{C}^\infty(\overline{\Omega}) \xrightarrow{\nabla\times} \boldsymbol{C}^\infty(\overline{\Omega}) \xrightarrow{\nabla\cdot} C^\infty(\overline{\Omega}) \to 0$$

is exact. Notice that $\boldsymbol{C}^\infty(\overline{\Omega}) = [C^\infty(\overline{\Omega})]^3$. The exactness of this sequence is a direct consequence of Poincaré's lemma from [1]. It states:

Theorem 2.4 (Poincaré's lemma)**.** *Let $\boldsymbol{e}, \boldsymbol{b}$ and q be two vector fields and a function, smooth over a star s-shaped domain Ω, such that $\nabla \times \boldsymbol{e} = 0$ and $\nabla \cdot \boldsymbol{b} = 0$ in Ω. There exists a smooth function ψ and smooth vector fields $\boldsymbol{a}$ and $\boldsymbol{j}$ such that*

$$\boldsymbol{e} = \nabla\psi, \quad \boldsymbol{b} = \nabla \times \boldsymbol{a} \quad \textit{and} \quad q = \nabla \cdot \boldsymbol{j}.$$

For a regular bounded contractible domain, we expect exactness of the following sequence where operators $\nabla, \nabla\times$ and $\nabla\cdot$ are understood in their weak sense:

$$\mathbb{R} \to H^1(\Omega) \xrightarrow{\nabla} \boldsymbol{H}(\mathrm{curl}, \Omega) \xrightarrow{\nabla\times} \boldsymbol{H}(\mathrm{div}, \Omega) \xrightarrow{\nabla\cdot} L^2(\Omega) \to 0. \tag{2.1}$$

This is indeed exact, but the proof rely on some difficult technical results. Therefore, we sketch the proof for level 1 only.

By definition of curl in the classical sense,

$$\ker(\nabla\times, \boldsymbol{C}^\infty(\overline{\Omega})) = \left\{ \boldsymbol{u} \mid \int_\Omega \boldsymbol{u} \cdot (\nabla \times \boldsymbol{\psi}) = 0 \ \forall \boldsymbol{\psi} \in \boldsymbol{C}_0^\infty(\Omega) \right\}.$$

If the domain Ω is contractible, then the Poincaré's lemma implies that $\mathrm{im}(\nabla)$ lies in $C^\infty(\overline{\Omega})$. Consequently,

$$\mathrm{im}(\nabla, C^\infty(\overline{\Omega})) = \boldsymbol{C}^\infty(\overline{\Omega}) \cap \mathrm{im}(\nabla\times, \boldsymbol{C}_0^\infty(\Omega))^\perp.$$

Taking the closures of both sides, we get

$$\overline{\mathrm{im}(\nabla, C^\infty(\overline{\Omega}))} = \ker(\nabla\times, \boldsymbol{H}(\mathrm{curl}, \Omega)).$$

It means that if $\nabla \times \boldsymbol{u} = 0$ in the weak sense, then there is a sequence of functions $\boldsymbol{\phi}_n$, smooth over Ω, such that $\boldsymbol{u} = \lim_{n\to\infty} \nabla\boldsymbol{\phi}_n$.

Suppose now that Ω is bounded. Imposing $\int_\Omega \boldsymbol{\phi}_n = 0$ and using a variant of the Poincaré's inequality [1, Ex. 5.11 & 5.12], we obtain

$$\|\boldsymbol{\phi}_n\| \leq c(\Omega)\|\nabla\boldsymbol{\phi}_n\|,$$

where $c(\Omega)$ depends on Ω only. The sequence $\{\boldsymbol{\phi}_n\}_{i=1}^{\infty}$ is thus a Cauchy sequence.

Let $\boldsymbol{\phi}$ be its limit. Then $\boldsymbol{u} = \nabla\boldsymbol{\phi}$, since ∇ is a closed operator. It results to the exactness at level 1:

$$\ker(\nabla, H^1(\Omega)) = \mathrm{im}(\nabla\times, \boldsymbol{H}(\mathrm{curl}, \Omega)).$$

Similar sketch of the proof for level 2 could be found in [1, Ex. 5.13].

In two spatial dimensions, the 3D exact sequence gives rise to two sequences,

$$\mathbb{R} \to H^1(\Omega) \xrightarrow{\nabla} \boldsymbol{H}(\mathrm{curl}, \Omega) \xrightarrow{\nabla\times} L^2(\Omega) \to 0 \tag{2.2}$$

and

$$\mathbb{R} \to H^1(\Omega) \xrightarrow{\nabla\times} \boldsymbol{H}(\mathrm{div}, \Omega) \xrightarrow{\nabla\cdot} L^2(\Omega) \to 0.$$

Both are obtained by restricting the 3D curl operator, in the first case to vectors $(E_1(x_1, x_2), E_2(x_1, x_2), 0)$, in the second case to $(0, 0, E_3(x_1, x_2))$.

2.2 Discrete de Rham diagram

Apart from curl-conforming finite element space introduced in the previous chapter, there exist also other families of finite element spaces which could be utilized for obtaining numerical solutions of Maxwell's equations. Those are gradient-conforming, divergence-conforming and L^2-conforming finite element spaces.

In this section we will add to diagram (2.1) layer of finite dimensional finite element functional spaces and we will connect it with the original layer via the appropriate finite element interpolation operators:

$$
\begin{array}{ccccccc}
H^1(\Omega) & \xrightarrow{\nabla} & \boldsymbol{H}(\mathrm{curl},\Omega) & \xrightarrow{\nabla\times} & \boldsymbol{H}(\mathrm{div},\Omega) & \xrightarrow{\nabla\cdot} & L^2(\Omega) \\
\bigcup & & \bigcup & & \bigcup & & \\
U & & \boldsymbol{V} & & \boldsymbol{W} & & \\
\downarrow{\scriptstyle \pi_h} & & \downarrow{\scriptstyle \boldsymbol{r}_h} & & \downarrow{\scriptstyle \boldsymbol{w}_h} & & \downarrow{\scriptstyle P_{0,h}} \\
U_h & \xrightarrow{\nabla} & \boldsymbol{V}_h & \xrightarrow{\nabla\times} & \boldsymbol{W}_h & \xrightarrow{\nabla\cdot} & Z_h
\end{array}
\tag{2.3}
$$

where $\boldsymbol{V}, \boldsymbol{V}_h$ and $\boldsymbol{r}_h$ were defined in the previous chapter. The other functional spaces are defined in the following manner:

$$
\begin{aligned}
U &= \{\varphi \in H^1(\Omega)|\ \varphi = 0 \text{ on } \Gamma_p\}, \\
U_h &= \{\varphi_h \in U|\ \varphi_h|_K \in P^1(K)\ \forall K \in \mathcal{T}_h\}, \\
\boldsymbol{W} &= \{\boldsymbol{H} \in \boldsymbol{H}(\mathrm{div},\Omega)|\ \boldsymbol{H}\cdot\boldsymbol{\nu} = 0 \text{ on } \Gamma_p\}, \\
Z_h &= \{q_h \in L^2(\Omega)|\ q_h|_K \in P^0(K)\ \forall K \in \mathcal{T}_h\}.
\end{aligned}
$$

We will not define the divergence conforming finite element space $\boldsymbol{W}_h$ in 3D as well as interpolation operators $\pi_h, \boldsymbol{r}_h, \boldsymbol{w}_h$ and $P_{0,h}$, because we will not use them any further. Our aim only was to show the discrete de Rham diagram in its complete form. Details can be found in [7, Chapter 5].

We mention that inclusions

$$
\nabla U_h \subset \boldsymbol{V}_h \text{ and } \nabla \times \boldsymbol{V}_h \subset \boldsymbol{W}_h
$$

hold and that diagram (2.3) commutes, i.e.

$$
\nabla \pi_h p = \boldsymbol{r}_h \nabla p \quad \text{and} \quad \boldsymbol{w}_h(\nabla \times \boldsymbol{u}) = \nabla \times \boldsymbol{r}_h \boldsymbol{u}.
$$

This commutativity is proved in [7, Theorems 5.49 and 5.40] provided $p \in U$ and $\boldsymbol{u} \in \boldsymbol{V}$ are sufficiently smooth and all interpolants $\pi_h p$ and $\boldsymbol{r}_h \nabla p$ as well as $\boldsymbol{w}_h(\nabla \times \boldsymbol{u})$ and $\nabla \times \boldsymbol{r}_h \boldsymbol{u}$ are well defined.

In 2D the diagram (2.3) transforms to

$$
\begin{array}{ccccc}
H^1(\Omega) & \xrightarrow{\nabla} & \boldsymbol{H}(\mathrm{curl},\Omega) & \xrightarrow{\nabla\times} & L^2(\Omega) \\
\bigcup & & \bigcup & & \\
U & & \boldsymbol{V} & & \\
\downarrow{\scriptstyle \pi_h} & & \downarrow{\scriptstyle \boldsymbol{r}_h} & & \downarrow{\scriptstyle P_{0,h}} \\
U_h & \xrightarrow{\nabla} & \boldsymbol{V}_h & \xrightarrow{\nabla\times} & Z_h
\end{array}
\tag{2.4}
$$

Since diagram (2.4) plays a central role in the theory we develop in Chapter 3 and in examples we present in Chapter 4, we will consider it more precisely. We will proof that it is exact at level 1 and that it commutes. To achieve that, we need to introduce some aspects of $H^1(\Omega)$ conforming finite element spaces and describe their relations to $\boldsymbol{H}(\mathrm{curl}, \Omega)$ conforming edge elements.

Definition 2.5. *The nodal P^1 finite element is a triad $(K, P^1(K), \Sigma)$, where K is a triangle in $\mathbb{R}^2$, $P^1(K)$ is the space of all polynomials of degree one on K and $\Sigma = \{\Sigma_1, \Sigma_2, \Sigma_3\}$ is the set of linear forms $\Sigma_i : P^1(K) \to \mathbb{R}$,*

$$\Sigma_i(p) = p(v_i),\ i = 1, 2, 3,$$

where v_i, $i = 1, 2, 3$ are vertices of triangle K.

For any $p \in H^1(K) \cap C(K)$ we can now define an interpolation $\pi_K p \in P^1(K)$ by requiring that

$$\Sigma_i(p - \pi_k p) = 0,\ \forall i = 1, 2, 3.$$

It is well known (see e.g. [7, Chapter 5]) that the finite element from Definition 2.5 is $H^1(\Omega)$ conforming. That allows us to define the global interpolation $\pi_h : H^1(\Omega) \cap C(K) \to U_h$ by

$$(\pi_h p)|_K = \pi_K p, \quad \forall K \in \mathcal{T}_h. \tag{2.5}$$

The finite dimensional functional space U_h from (2.4) posses a standard basis $\mathcal{B}_U$, which consists of so-called Courant functions that satisfy

$$\varphi_i(B_j) = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j, \end{cases} \tag{2.6}$$

where B_i, B_j, $i, j = 1, \ldots, N_{iv}$ are the interior vertices of $\mathcal{T}_h$.

If we denote

$$\sigma_{ie} = \begin{cases} 1 & \text{if } \boldsymbol{\tau}_e \text{ aims towards } B_i \\ -1 & \text{otherwise,} \end{cases} \tag{2.7}$$

and

$$\omega(B_i) = \{e \in \mathcal{E}_i |\ \exists B_j \in \mathcal{V} \text{ such that } e = \overline{B_i B_j}\},$$

then the definition (2.6) yields

$$\nabla \varphi_i|_e \cdot \boldsymbol{\tau}_e = \begin{cases} \sigma_{ie}/|e| & \text{if } e \in \omega(B_i), \\ 0 & \text{if } e \notin \omega(B_i). \end{cases}$$

This implies together with (1.48) that $\nabla\varphi_i$ could be expressed in a unique way as

$$\nabla\varphi_i = \sum_{e\in\omega(B_i)} \sigma_{ie}\boldsymbol{\psi}_e. \tag{2.8}$$

Having all necessary relations described, we can move to show the properties of diagram (2.3).

Theorem 2.6. *The discrete 2D de Rham diagram is exact at level 1, i.e.*

$$\mathcal{R}_h = \mathrm{im}(\nabla) = \mathrm{span}\{\nabla\varphi_i, B_i \in \mathcal{V}_i\} = \\ = \{\boldsymbol{\psi} \in \boldsymbol{V}_h | \nabla\times\boldsymbol{\psi} = 0\} = \ker(\nabla\times) = \mathcal{K}_h. \tag{2.9}$$

Proof. (a) The inclusion $\mathcal{R}_h \subset \mathcal{K}_h$ follows directly from the well-known fact that $\nabla\times(\nabla\varphi) = 0$ for all $\varphi \in H^1(\Omega)$.
(b) Now we prove the inclusion $\mathcal{K}_h \subset \mathcal{R}_h$. Let $\boldsymbol{\psi} \in \mathcal{K}_h$ be arbitrary but fixed function and $e \in \mathcal{E}$ be any edge of the triangulation $\mathcal{T}_h$.

In $\mathcal{T}_h$, every edge is a line segment connecting two vertices form $\mathcal{V}$, the set of all vertices of $\mathcal{T}_h$. Let e connects vertices B_i and B_j from $\mathcal{V}$, i.e. $e = \overline{B_iB_j}$.

We prescribe a mapping $\alpha : \mathcal{V} \to \mathbb{R},\ \alpha_j = \alpha(B_j)$ by

$$\alpha_j - \alpha_i = \int_{\overline{B_iB_j}} \boldsymbol{\psi}\cdot\boldsymbol{\tau}_{ij}\mathrm{d}s,$$

where $\boldsymbol{\tau}_{ij}$ is a fixed unit tangent vector to $\overline{B_iB_j}$.

Let B_k be such vertex from $\mathcal{V}$ that both line segments $\overline{B_iB_k}$ and $\overline{B_jB_k}$ are in $\mathcal{E}$ (i.e. $\overline{B_iB_j}, \overline{B_jB_k}$ and $\overline{B_kB_i}$ form a triange $\overline{B_iB_jB_k} = K \in \mathcal{T}_h$). With

$$\alpha_k - \alpha_j = \int_{B_kB_j} \boldsymbol{\psi}\cdot\boldsymbol{\tau}_{kj}\mathrm{d}s, \quad \text{and} \quad \tilde{\alpha}_i - \alpha_k = \int_{B_iB_k} \boldsymbol{\psi}\cdot\boldsymbol{\tau}_{ik}\mathrm{d}s,$$

we get

$$\begin{aligned}\tilde{\alpha}_i - \alpha_i &= \tilde{\alpha}_i - \alpha_k + \alpha_k - \alpha_j + \alpha_j - \alpha_i \\ &= \int_{\overline{B_iB_j}} \boldsymbol{\psi}\cdot\boldsymbol{\tau}_{ij}\mathrm{d}s + \int_{\overline{B_kB_j}} \boldsymbol{\psi}\cdot\boldsymbol{\tau}_{kj}\mathrm{d}s + \int_{\overline{B_iB_k}} \boldsymbol{\psi}\cdot\boldsymbol{\tau}_{ik}\mathrm{d}s \\ &= \int_{\partial K} \boldsymbol{\psi}\cdot\boldsymbol{\tau}_{\partial K}\mathrm{d}s = \int_K \nabla\times\psi\ \mathrm{d}\boldsymbol{x} = 0.\end{aligned}$$

Thus α_i is independent of $B_j \in \mathcal{V}$ and the mapping α is well-defined. Now we can define $\phi \in U_h$ by $\phi(B_i) = \alpha_i$ for all $B_i \in \mathcal{V}$. This directly

implies that $\nabla\phi \in \boldsymbol{V}_h$. Since $\phi|_K \in P^1|_K$, then $\nabla\phi|_K \in P^0|_K$ and we can write

$$\nabla\boldsymbol{\phi} \cdot \boldsymbol{\tau}_e |e| = \int_e \nabla\boldsymbol{\phi} \cdot \boldsymbol{\tau}_e \, \mathrm{d}s, \tag{2.10}$$

where $|e|$ is the norm (length) of e. With σ_{je} from (2.7), we get

$$\begin{aligned}\int_e \nabla\boldsymbol{\phi} \cdot \boldsymbol{\tau}_e \, \mathrm{d}s = \sigma_{je} \int_{B_i}^{B_j} \boldsymbol{\phi}' \, \mathrm{d}s = \sigma_{je}(\boldsymbol{\phi}(B_j) - \boldsymbol{\phi}(B_i)) \\ = \alpha_j - \alpha_i = \int_e \boldsymbol{\psi} \cdot \boldsymbol{\tau}_e \mathrm{d}s = \boldsymbol{\psi}|_e \cdot \boldsymbol{\tau}_e. \end{aligned} \tag{2.11}$$

Putting (2.10) and (2.11) together we get $\boldsymbol{\psi} = \nabla\phi$. Since $\boldsymbol{\psi} \in \boldsymbol{V}_h$ was chosen arbitraryly, the inclusion $\mathcal{K}_h \subset \mathcal{R}_h$ is proven. □

Theorem 2.7. *If p is sufficiently smooth that both $\pi_h p$ and $\boldsymbol{r}_h \nabla p$ are well defined, then the equality*

$$\nabla \pi_h p = \boldsymbol{r}_h \nabla p.$$

holds, see diagram (2.4)*.*

Proof. We are going to present the 2D modification of the proof found in [8]. To show the commutativity in diagram (2.3), it suffices to show that all degrees of freedom are equal in all triangles. Let e be an arbitrary edge of an arbitrary triangle $K \in \mathcal{T}_h$. Then

$$\int_e (\nabla \pi_h p - \boldsymbol{r}_h \nabla p) \cdot \boldsymbol{\tau}_e \mathrm{d}s = -\int_e \left(\frac{\partial \pi_h p}{\partial s} - \frac{\partial p}{\partial s} \right) \mathrm{d}s,$$

where we have used (1.47). Now if $e = \overline{B_i B_j}$, then

$$\int_e \left(\frac{\partial \pi_h p}{\partial s} - \frac{\partial p}{\partial s} \right) \mathrm{d}s = (\pi_h p(B_j) - p(B_j)) - (\pi_h p(B_i) - p(B_i)) = 0$$

where we use the vertex interpolation property of p from (2.5).

□

Chapter 3

Finite element basis conforming with de Rham diagram

In Chapter 1 we discretized time harmonic Maxwell's equations by edge elements. We constructed basis $\mathcal{B}_V$ of space $\boldsymbol{V}_h$ containing the Whitney functions $\boldsymbol{\psi}_e$. These functions have the following properties

$$\begin{aligned}\boldsymbol{\psi}_e \cdot \boldsymbol{\tau}_e &= \frac{1}{|e|} \text{ on } e \in \mathcal{E}_i,\\ \boldsymbol{\psi}_e \cdot \boldsymbol{\tau}_{e^*} &= 0 \quad \text{on } e \neq e^*, e^* \in \mathcal{E}_i,\end{aligned}$$

for each edge e from the set of all interior edges $\mathcal{E}_i$, see (1.37). Thus,

$$\mathcal{B}_{\boldsymbol{V}} = \{\psi_e |\ e \in \mathcal{E}_i\}. \tag{3.1}$$

Let us recall the notation (1.33)–(1.45) from Chapter 1, where we established symbols for characterization of $\mathcal{T}_h$, e.g. N_{ie} for the number of all interior edges of the triangulation $\mathcal{T}_h$ and N_{iv} for the number of elements of the set of all interior vertices $\mathcal{V}_i$ of $\mathcal{T}_h$.

From Theorem 2.6 in Chapter 2 it is known that

$$\mathcal{R}_h = \operatorname{span}\{\nabla\varphi_i, B_i \in \mathcal{V}_i\} = \{\boldsymbol{\psi} \in \boldsymbol{V}_h | \nabla \times \boldsymbol{\psi} = 0\} = \mathcal{K}_h$$

is a closed subspace of $\boldsymbol{V}_h$. It is obvious that

$$\mathcal{B}_{\mathcal{K}} = \{\nabla\varphi_i |\ B_i \in \mathcal{V}_i\},$$

is a basis of $\mathcal{K}_h$. From the construction of bases $\mathcal{B}_{\boldsymbol{V}}$ and $\mathcal{B}_{\mathcal{K}}$ it is also obvious, that

$$\dim \mathcal{K}_h = N_{iv}, \quad \text{and} \quad \dim \boldsymbol{V}_h = N_{ie}.$$

From Euler's formula for planar graphs [4, p. 78]:

$$N_K - N_e + N_v = 1$$

where N_K, N_i and N_v are explained in (1.39), (1.43) and (1.40), respectively, we derived the inequality

$$N_{iv} \leq N_{ie}.$$

Let us now divide all interior edges of triangulation $\mathcal{T}_h$ into two subsets $\mathcal{E}_{\text{keep}}$ and $\mathcal{E}_{\text{rem}}$ with the following numbers of elements:

$$|\mathcal{E}_{\text{rem}}| = N_{iv},$$
$$|\mathcal{E}_{\text{keep}}| = N_{ie} - N_{iv}.$$

Now the question is for which interior edges is the set

$$\mathcal{B} = \mathcal{B}_{\mathcal{K}} \cup \{\boldsymbol{\psi}_e |\ e \in \mathcal{E}_{\text{keep}}\}$$

a basis of $\boldsymbol{V}_h$?

Answer is provided in the subsequent theorem which was presented by T. Vejchodský in [13]. To be able to present the theorem clearly, we need to introduce a couple of new symbols: graph $(\mathcal{V}_i \cup \{B^\partial\}, \mathcal{E}^\partial)$ is obtained from $\mathcal{T}_h$ by collapsing the whole boundary to a single vertex B^∂ while preserving the adjacency of inner vertices of $\mathcal{T}_h$. Hence edges from $\mathcal{E}_i$ connecting two inner vertices in $\mathcal{T}_h$ stay unchanged in $\mathcal{E}^\partial$, while edges $e \in \mathcal{E}_i$ such that $e = \overline{B_i B_b},\ B_i \in \mathcal{V}_i,\ B_b \in \mathcal{V}_b$ are transformed to the edge $e = \overline{B_i B^\partial}$ in $\mathcal{E}^\partial$. Similarly, we define also the set of edges $\mathcal{E}^\partial_{\text{rem}}$, which is obtained from $\mathcal{E}_{\text{rem}}$.

Theorem 3.1. *The following statements are equivalent:*

(i) The set $\mathcal{B}$ is a basis of $\boldsymbol{V}_h$.

(ii) The only cycle in $(\mathcal{V},\ \mathcal{E}_b \cup \mathcal{E}_{\text{rem}})$ is $(\mathcal{V}_b, \mathcal{E}_b)$.

(iii) The graph $\mathcal{G}^\partial = (\mathcal{V}_i \cup \{B^\partial\}, \mathcal{E}^\partial_{\text{rem}})$ is a spanning tree in $(\mathcal{V}_i \cup \{B^\partial\}, \mathcal{E}^\partial)$.

Proof. The equivalence of (ii) and (iii) is obvious from the construction of the graph $\mathcal{G}^\partial = (\mathcal{V}_i \cup \{B^\partial\}, \mathcal{E}^\partial_{\text{rem}})$.

Next we want to show that (i) implies (iii). Let $\mathcal{B}$ be a basis in $\boldsymbol{V}_h$ and $\mathcal{G}^\partial$ not a spanning tree. Then $\mathcal{G}^\partial$ has an isolated component $\mathcal{G}_{\text{isol}}$ (a

set of vertices and edges that are not connected with B^∂ by other vertices and edges from $\mathcal{G}^\partial$):

$$\mathcal{G}_{\text{isol}} = (\mathcal{V}_{\text{isol}}, \mathcal{E}_{\text{isol}}) \quad \text{with} \quad \mathcal{V}_{\text{isol}} \subset \mathcal{V}_{\text{i}},\ \mathcal{E}_{\text{isol}} \subset \mathcal{E}_{\text{i}}.$$

Let

$$\varphi = \sum_{B_i \in \mathcal{V}_{\text{isol}}} \varphi_i,$$

then we get

$$\nabla\varphi = \sum_{B_i \in \mathcal{V}_{\text{isol}}} \nabla\varphi_i.$$

Using (2.8) we can continue by

$$\sum_{B_i \in \mathcal{V}_{\text{isol}}} \nabla\varphi_i = \sum_{B_i \in \mathcal{V}_{\text{isol}}} \sum_{e \in \omega(B_i)} \sigma_{ie} \boldsymbol{\psi}_e. \tag{3.2}$$

If we put $\omega^0(\mathcal{V}_{\text{isol}}) = \{e = \overline{B_i B_j} | B_i \in \mathcal{V}_{\text{isol}}, B_j \notin \mathcal{V}_{\text{isol}}\}$ then the sum on the right-hand side of (3.2) can be rewritten to

$$\sum_{B_i \in \mathcal{V}_{\text{isol}}} \sum_{e \in \omega(B_i)} \sigma_{ie} \boldsymbol{\psi}_e = \sum_{e \in \omega^0(\mathcal{V}_{\text{isol}})} \sigma_{ie} \boldsymbol{\psi}_e.$$

Since for edges $e = \overline{B_i B_j}$ with $B_i \in \mathcal{V}_{\text{isol}}$ and $B_j \in \mathcal{V}_{\text{isol}}$ we have

$$\sigma_{ie} = -\sigma_{je},$$

the terms containing these edges vanish from the sum on the right-hand side of (3.2).

This fact implies that the following subset of $\mathcal{B}$:

$$\{\nabla\varphi_i |\ B_i \in \mathcal{V}_{\text{isol}}\} \cup \{\boldsymbol{\psi}_e |\ e \in \omega^0(\mathcal{V}_{\text{isol}})\}$$

is a linearly dependent set, which is in contradiction to $\mathcal{B}$ being a basis.

Finally, let us demonstrate that (iii) implies (i). Let $\mathcal{G}^\partial$ be a spanning tree and $\mathcal{B}$ not a basis in $\boldsymbol{V}_h$. Due to the linear dependence of vectors from $\mathcal{B}$, there exists $\varphi \in U_h$ such that

$$\nabla\varphi = \sum_{B_i \in \mathcal{V}_i} c_i \nabla\varphi_i = \sum_{e \in \mathcal{E}_{\text{keep}}} d_e \boldsymbol{\psi}_e = \boldsymbol{\psi}.$$

In addition there exist $B_k \in \mathcal{V}_i$ such that $c_k \neq 0$ and $e_k \in \mathcal{E}_{\text{keep}}$ such that $d_{e_k} \neq 0$.

Let $e^* \in \mathcal{E}_{\text{rem}},\ e^* = \overline{B_i B_j},\ B_i, B_j \in \mathcal{V}_i$. Then

$$0 = \boldsymbol{\psi}|_{e^*} \cdot \boldsymbol{\tau}_{e^*} = \nabla\varphi|_{e^*} \cdot \boldsymbol{\tau}_{e^*} = \sigma_{ie^*}\frac{c_i - c_j}{|e^*|} = \sigma_{je^*}\frac{c_j - c_i}{|e^*|}, \tag{3.3}$$

since $\boldsymbol{\psi}_e|_{e^*} \cdot \boldsymbol{\tau}_{e^*} = 0$ for all $e \in \mathcal{E}_{\text{keep}}$. From there we get

$$c_i = c_j = c^* \quad \forall B_i, B_j \in \mathcal{V}_i. \tag{3.4}$$

Because $\mathcal{G}^\partial$ is a spanning tree, there is an edge $\tilde{e} \in \mathcal{E}_{\text{rem}}$ such that $\tilde{e} = \overline{B_m B^\partial},\ B_m \in \mathcal{V}_i$. Rewriting (3.3) for edge $\tilde{e}$ we obtain:

$$0 = \boldsymbol{\psi}|_{\tilde{e}} \cdot \boldsymbol{\tau}_{\tilde{e}} = \nabla\varphi|_{\tilde{e}} \cdot \boldsymbol{\tau}_{\tilde{e}} = \sigma_{m\tilde{e}}\frac{c_m}{|\tilde{e}|},$$

which implies that $c_m = 0$. That combined with (3.4) results in

$$c_i = c^* = 0 \quad \forall B_i \in \mathcal{V}_i.$$

Therefore we have a contradiction to an existence of a vertex $B_k \in \mathcal{V}_i$ such that $c_k \neq 0$ that we assumed at the beginning. □

The basis reproducing the discrete kernel of the curl operator constructed according to Theorem 3.1 is by no means unique. In the following chapter we discuss the effect of the choice of the spanning tree on the conditioning of the resulting stiffness matrix.

Chapter 4

Numerical experiments

This chapter is fully devoted to numerical experiments based on the theory developed in previous chapters. It is divided into two parts. The first part investigates the effect of the choice of the spanning tree in the graph on the conditioning of the resulting stiffness matrix, while the second part examines iteration schemes that take advantage of the structure of the stiffness and mass matrices.

4.1 Spanning trees

Theorem 3.1 provides the sufficient condition for the choice of the edge basis functions of the finite element space $\boldsymbol{V}_h$ to be compatible with the de Rham diagram.

Our idea is straightforward. In a model examples we try to investigate all possible spanning trees and we try to find those which minimize and maximize the condition numbers of the corresponding finite element matrices. We hope to find some patterns and general clues which allow an a priori construction of the spanning trees leading to well conditioned finite element matrices.

In search for the spanning tree that would lead to the optimal edge basis functions, we are limited by computational complexity. For that reason we are often unable to evaluate all possible combinations. By adding few extra vertices and edges to a graph, the amount of possible spanning trees grows exponentially.

Moreover, we cannot even use straightforwardly results of the graph theory to obtain the exact number of spanning trees, since our graph would contain multiple edges emerging from the collapse of the whole boundary to one vertex.

In Table 4.1 we illustrate the rate of growth of the number of spanning trees. For selected graphs (see Figure 4.1) we will use a well-known result from the graph theory, the Kirhoff's law [12, p. 138], to find the number of spanning trees in the graph $\mathcal{G}^\partial = (\mathcal{V}_i \cup \{B^\partial\}, \mathcal{E}^\partial_{\rm rem})$ ignoring the multiplicity of edges connecting boundary and inner vertices. These numbers of trees will be compared with the numbers of trees constructed according to Theorem 3.1 in cases, where we have been able to find them all and thus to count them.

graph	N_{iv}	N_{ie}	# trees based on Kirhoff's theorem	# trees
g01	1	6	1	6
g02	7	30	8100	176400
g03	37	132	$\sim 10^{24}$	?
g04	217	702	$\sim 10^{149}$	?

Table 4.1: An illustration of the growth of the number of spanning trees.

In the rest of this section, we will present the actual results, i.e. the condition numbers of matrices obtained by discretization of time-harmonic Maxwell's equations using the bases compatible with the de Rham diagram, see Theorem 3.1. The domain and used meshes are depicted in Figure 4.1 (b)–(d).

In order to present the data clearly, certain additional notation would be useful. We will consider the triangulation $\mathcal{T}_h$ as a graph. Let us denote by

$$E^{\rm all}_{\rm rem} = \{\mathcal{E}_{\rm rem} \subset \mathcal{E}_i |\ \mathcal{G}^\partial = (\mathcal{V}_i \cup \{\mathcal{B}^\partial\}, \mathcal{E}^\partial_{\rm rem}) \text{ is a spanning tree of } \mathcal{T}_h\}$$

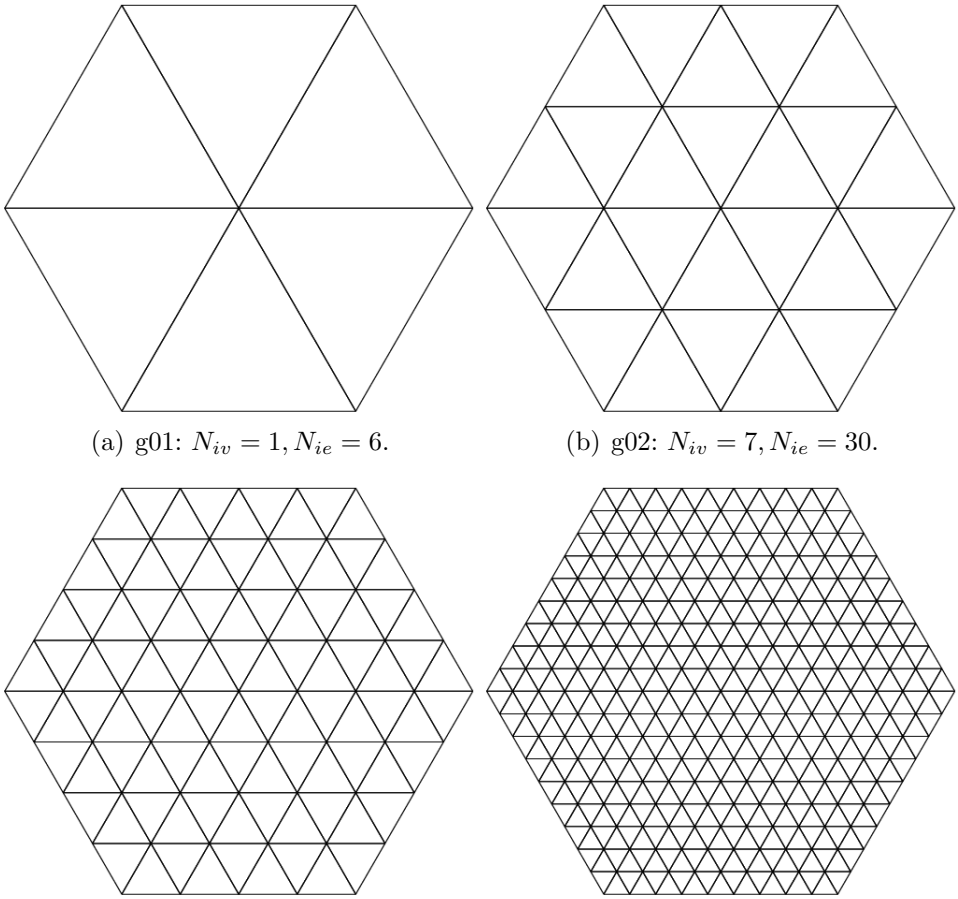
the collection of all sets $\mathcal{E}_{\rm rem}$ of the edges that provide a spanning tree $\mathcal{G}^\partial$. In the examples below we try to test as much as possible situations from $E^{\rm all}_{\rm rem}$. The subset $E_{\rm rem} \subset E^{\rm all}_{\rm rem}$ differs in the particular examples. In Example 1 the set $E_{\rm rem}$ contains all $\mathcal{E}_{\rm rem}$ that lead to spanning trees, i.e. $E_{\rm rem} = E^{\rm all}_{\rm rem}$. Due to enormous number of possibilities, see Figure 4.1, only selected subsets $E_{\rm rem} \subset E^{\rm all}_{\rm rem}$ are considered in Examples 2 and 3.

For each set of edges $\mathcal{E}_{\rm rem}$ we consider its complement $\mathcal{E}_{\rm keep}$ in $\mathcal{E}_{\rm i}$. As shown in Theorem 3.1,

$$\mathcal{B} = \mathcal{B}_{\mathcal{K}} \cup \{\boldsymbol{\psi}_e |\ e \in \mathcal{E}_{\rm keep}\} \tag{4.1}$$

is a basis of $\boldsymbol{V}_h$. By discretizing the variational formulation of the time-harmonic Maxwell's equations (1.23) with $\Gamma_I = \emptyset$ in the space $\boldsymbol{V}_h$ using the basis $\mathcal{B}$, we obtain the system of linear equations

$$(S - \kappa^2 M)\boldsymbol{Y}_h = \boldsymbol{F}_h. \tag{4.2}$$



(a) g01: $N_{iv} = 1, N_{ie} = 6$. (b) g02: $N_{iv} = 7, N_{ie} = 30$.

(c) g03: $N_{iv} = 37, N_{ie} = 132$. (d) g04: $N_{iv} = 217, N_{ie} = 702$.

Figure 4.1: Graphs used for the illustration of the growing rate of spanning trees in Table 4.1.

The matrix $S \in \mathbb{R}^{N_{ie} \times N_{ie}}$ is defined as

$$S_{i,j} = (\mu^{-1} \nabla \times \boldsymbol{\Phi}_i, \nabla \times \boldsymbol{\Phi}_j), \ \boldsymbol{\Phi}_i, \boldsymbol{\Phi}_j \in \mathcal{B}, \ i, j = 1, \ldots, N_{ie}. \tag{4.3}$$

Similarly, the entries of $M \in \mathbb{R}^{N_{ie} \times N_{ie}}$ are given by

$$M_{i,j} = (\boldsymbol{\Phi}_i, \boldsymbol{\Phi}_j), \ \boldsymbol{\Phi}_i, \boldsymbol{\Phi}_j \in \mathcal{B}, \ i, j = 1, \ldots, N_{ie}. \tag{4.4}$$

The vector $\boldsymbol{Y_h} = (Y_1, \ldots, Y_{N_{ie}})^T$ contains the unknown expansions coefficients:

$$\boldsymbol{E}_h = \sum_{i=1}^{N_{ie}} Y_i \boldsymbol{\Phi}_i, \ \boldsymbol{\Phi}_i \in \mathcal{B},$$

and the load vector $\boldsymbol{F}_h = (F_1, \ldots, F_{N_{ie}})^T$ has the following entries:

$$F_i = (\boldsymbol{F}, \boldsymbol{\Phi}_i),\ \boldsymbol{\Phi}_i \in \mathcal{B},\ i = 1, \ldots, N_{ie},$$

where $\boldsymbol{F}$ is from (1.19).

In the subsequent examples we also test the matrix $A_e \in \mathbb{R}^{N_{ie} \times N_{ie}}$, which is obtained by using the Whitney basis functions from $\mathcal{B}_{\boldsymbol{V}}$ (see (3.1)) only. It has the form

$$A_e = S_e - \kappa^2 M_e,$$

where S_e is defined by

$$(S_e)_{ij} = (\mu_r^{-1} \nabla \times \boldsymbol{\psi}_{e_i}, \nabla \times \boldsymbol{\psi}_{e_j}), \quad \boldsymbol{\psi}_{e_i}, \boldsymbol{\psi}_{e_j} \in \mathcal{B}_{\boldsymbol{V}},\ i, j = 1, \ldots, N_{ie},$$

and M_e is defined by

$$(M_e)_{ij} = (\boldsymbol{\psi}_{e_i}, \boldsymbol{\psi}_{e_j}), \quad \boldsymbol{\psi}_{e_i}, \boldsymbol{\psi}_{e_j} \in \mathcal{B}_{\boldsymbol{V}},\ i, j = 1, \ldots, N_{ie}.$$

We will denote by $\mathcal{S}(E_{\text{rem}})$ the set of all matrices S given by (4.3) corresponding to all $\mathcal{E}_{\text{keep}}$ – complements of all $\mathcal{E}_{\text{rem}} \in E_{\text{rem}}$. The sets of similarly obtained matrices M and $S - \kappa^2 M$, will be denoted by $\mathcal{M}(E_{\text{rem}})$ and $\mathcal{A}(E_{\text{rem}})$, respectively.

We will examine the effect of choosing $\mathcal{E}_{\text{rem}} \in E_{\text{rem}}$ on conditioning of matrices from $\mathcal{S}(E_{\text{rem}})$, $\mathcal{M}(E_{\text{rem}})$ and $\mathcal{A}(E_{\text{rem}})$. Since matrices from $\mathcal{S}(E_{\text{rem}})$ are singular (because $\nabla \times (\nabla \varphi) = 0$ for all $\varphi \in H^1(\Omega)$), we will examine only submatrices S_{22} of matrices $S \in \mathcal{S}(E_{\text{rem}})$. The matrix S_{22} is formed from matrix $S \in \mathcal{A}(E_{\text{rem}})$ by taking entries related only to edge basis functions (and not to the gradients of Courant functions), i.e

$$(S_{22})_{ij} = (\mu_r^{-1} \nabla \times \boldsymbol{\psi}_{e_i}, \nabla \times \boldsymbol{\psi}_{e_j}), \quad e_i, e_j \in \mathcal{E}_{\text{keep}},\ i, j = 1, \ldots, N_{ie} - N_{iv}, \tag{4.5}$$

where μ_r was defined in (1.11). For simplicity, we consider $\mu_r = 1$ and $\kappa = 1$ in Examples 1–3 below.

4.1.1 Example 1

In this example we consider a triangulation depicted in 4.1(b). It has 7 interior vertices and 30 interior edges, i.e. $N_{iv} = 7$ and $N_{ie} = 30$. The mesh is very coarse and it would not be most likely used for practical solution of Maxwell's equations. However, we use it as a test example because we are able to enumerate all spanning trees together with resulting condition numbers on available hardware in a reasonable time.

Condition number of the matrix A_e is

$$c(A_e) = 49.3548.$$

In this example we are able to compute all combinations, i.e. $E_{\text{rem}} = E^{\text{all}}_{\text{rem}}$, since $E^{\text{all}}_{\text{rem}}$ consists of 176400 trees only.

Each of calculated condition numbers appears in the results at least twelve times. This is due to the symentry. If we take one arbitrary tree and rotate it by a multiple of 60 degrees or reflect it about the axis connecting antipodal corners of the triangulation, the condition number of the resulting matrix will be every time the same. In order to keep the presentation of the results as simple as possible, we will not mention this fact any more in the presentation of the following results.

Starting with matrix S_{22}, the minimal condition number was obtained for a matrix linked to the tree shown in Figure 4.2. The resulting minimal condition number was

$$\min_{S \in \mathcal{S}(E_{\text{rem}})} c(S_{22}) = 82.29.$$

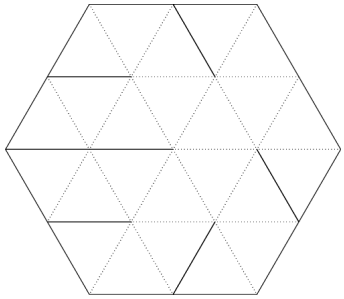


Figure 4.2: The tree leading to the smallest $c(S_{22})$.

The maximal condition number

$$\max_{S \in \mathcal{S}(E_{\text{rem}})} c(S_{22}) = 254.32$$

was attained for two matrices linked to the trees depicted in Figure 4.3.

Moving to matrices $M \in \mathcal{M}(E_{\text{rem}})$, we found the following minimal condition number:

$$\min_{S \in \mathcal{M}(E_{\text{rem}})} c(M) = 628.60.$$

The related tree is presented in Figure 4.4.

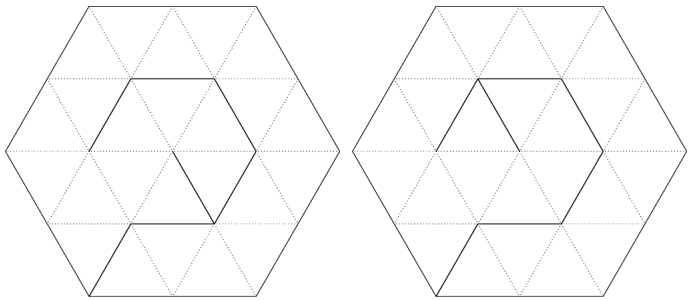


Figure 4.3: The two trees leading to maximal $c(S_{22})$.

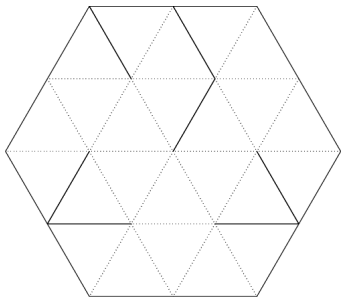


Figure 4.4: The tree connected with minimal $c(M)$.

The tree depicted in Figure 4.5 leads to a matrix with maximal conditional number

$$\max_{S \in \mathcal{M}(E_{\text{rem}})} c(M) = 3778.46$$

among all matrices from $\mathcal{M}(E_{\text{rem}})$.

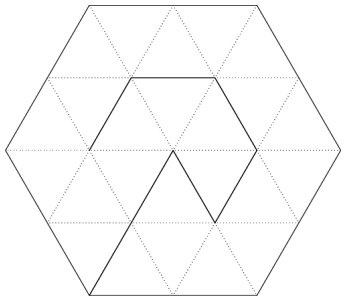


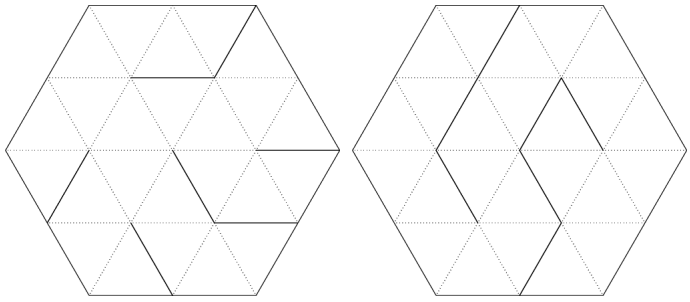
Figure 4.5: The tree corresponding to maximal $c(M)$.

Finally, among matrices from $\mathcal{A}(E_{\text{rem}})$, the minimal and maximal

condition number obtained were

$$\min_{A \in \mathcal{A}(E_{\text{rem}})} c(A) = 103.54,$$
$$\max_{A \in \mathcal{A}(E_{\text{rem}})} c(A) = 334.54.$$

One of the representative trees for both of these extremal values is shown in Figure 4.6.



(a) Min. $c(A)$ tree. (b) Max. $c(A)$ tree.

Figure 4.6: The trees connected with extremal $c(A)$ values.

4.1.2 Example 2

As the second example we have chosen slightly denser triangulation that still allows us to examine the amount of trees comparable to the previous case in a reasonable time. It is the triangulation with 37 interior vertices and 132 interior edges depicted in Figure 4.1(c).

We examined $2 \cdot 10^5$ trees. This is only a small fragment of all possible trees (see Table 4.1), yet the computation took the full capacity of a decent computer workstation for three days.

While constructing the triangulation, we enumerated the edges in ascending manner going from the left to the right and from bellow upwards (see Figure 4.7). The trees were generated by the algorithm described in [9]. This algorithm generates spanning trees of a graph in order of increasing total weight, which is the sum of weights of edges contained in the graph. Since we set all the weights to one, the algorithm generated the first $2 \cdot 10^5$ trees sorted lexicographically according to the numbers of edges.

This time, the matrix A_e had the following condition number:

$$c(A_e) = 244.8501.$$

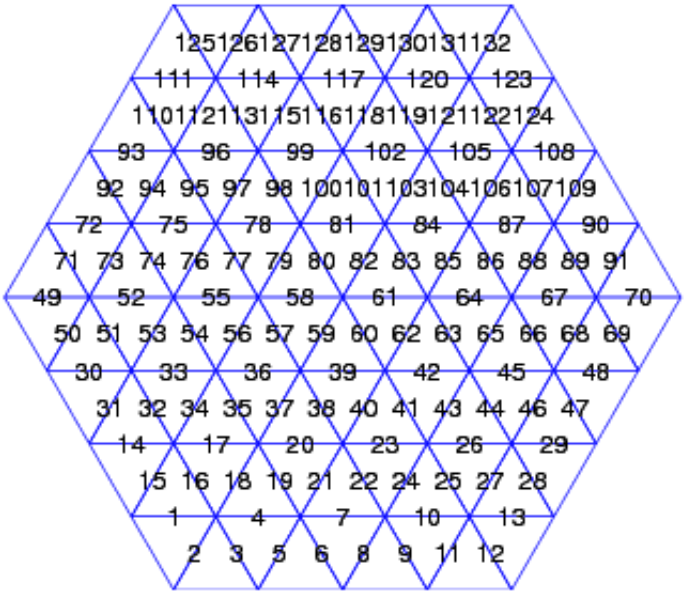


Figure 4.7: Enumeration of edges in triangulation used in Example 2.

As in the previous case, we examine the conditioning of matrices from $\mathcal{S}(\mathcal{E}_{\text{rem}})$. The extremal values

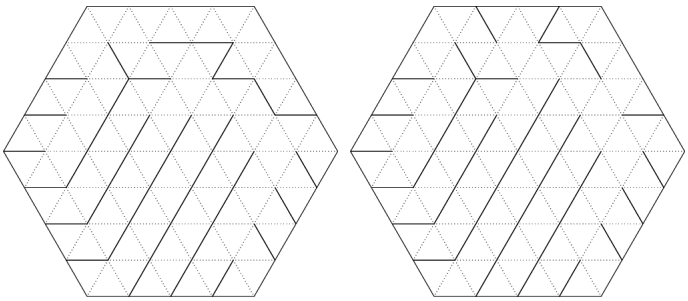
$$\min_{S \in \mathcal{S}(E_{\text{rem}})} c(S_{22}) = 7.7897 \cdot 10^3,$$
$$\max_{S \in \mathcal{S}(E_{\text{rem}})} c(S_{22}) = 1.5706 \cdot 10^4,$$

belong to trees from Figure 4.8.

Proceeding to $\mathcal{M}(E_{\text{rem}})$, we found that trees depicted in Figure 4.9 belong to matrices with extremal values:

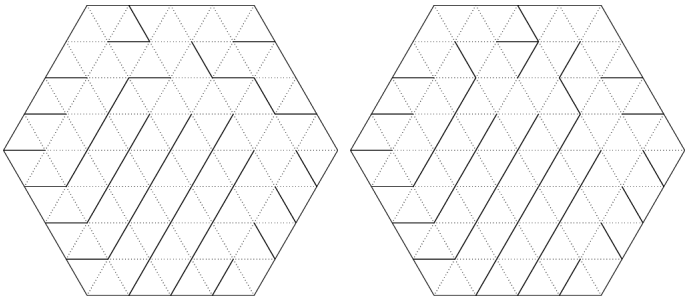
$$\min_{M \in \mathcal{M}(E_{\text{rem}})} c(M) = 1.004 \cdot 10^4,$$
$$\max_{M \in \mathcal{M}(E_{\text{rem}})} c(M) = 1.3957 \cdot 10^4.$$

While analyzing the third set of matrices, $\mathcal{A}(E_{\text{rem}})$, we once again



(a) Min. $c(S_{22})$ tree. (b) Max. $c(S_{22})$ tree.

Figure 4.8: The trees leading to extremal $c(S_{22})$ values in Example 2.



(a) Min. $c(M)$ tree. (b) Max. $c(M)$ tree.

Figure 4.9: The trees bounded with extremal $c(M)$ values in Example 2.

evaluted the extremal values of the condition number:

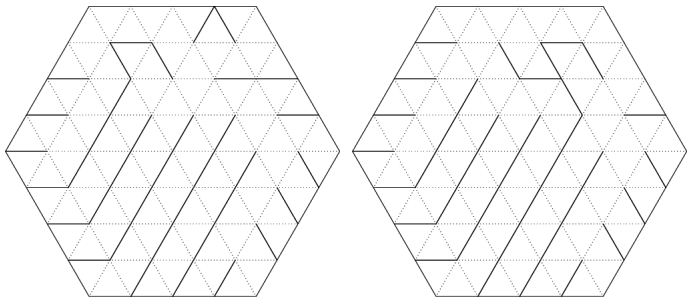
$$\min_{A \in \mathcal{A}(E_{\text{rem}})} c(M) = 1.5464 \cdot 10^5,$$
$$\max_{A \in \mathcal{A}(E_{\text{rem}})} c(M) = 2.2109 \cdot 10^5.$$

Those values belong to the trees depicted in Figure 4.10.

4.1.3 Example 3

Hexagonal triangulation with 217 interior vertices and 702 interior edges served as our final test case. It is depicted in Figure 4.1(d).

Having in mind that due to the computational demands we could not inspect a vast number of trees, we prepared four relatively – to the total amount of trees – small test cases. Each consisted of 10^4 trees, that were prepared by algorithm [9]. In the beginning we assigned to each edge $e \in \mathcal{E}_i$ a weight, then we let the algorithm to generate 10^4 trees in order of increasing total weight, i.e. the sum of assigned weights of all edges in



(a) Min. $c(A)$ tree. (b) Max. $c(A)$ tree.

Figure 4.10: The trees with extremal $c(A)$ values.

a particular tree. The weights of the edges were set in the following for different manners:

1. weight$(e) = 1$ for all $e \in \mathcal{E}_i$,
2. weight$(e) = d$, where d equals to the distance of e from the boundary of the domain,
3. weight$(e) = 1/d$, where d is the same as above,
4. weight(e) was set randomly.

By this setting we expected to obtain in each case different results. For example, in the second case the trees lead to matrices with better conditioning compared to the third case. Such expectation was based on the examination of the previous examples, where it appeared that the trees with shorter branches generates better conditioned matrices.

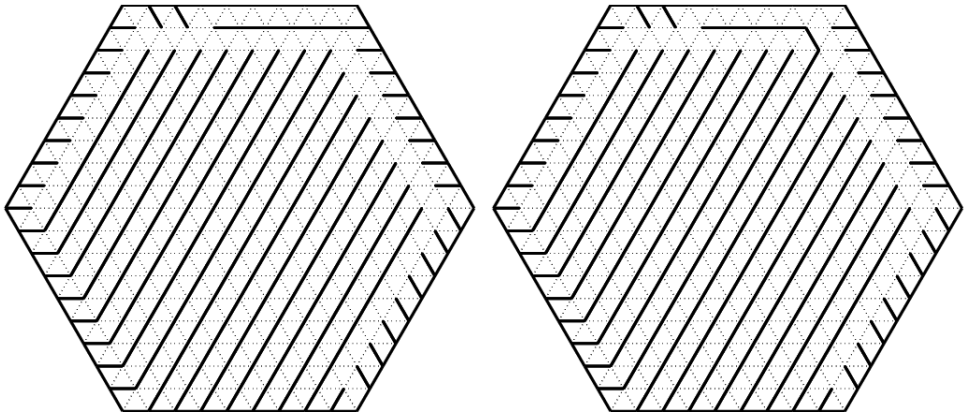
In this example, the results were even more affected by the effect that could have been seen for the results on the previous triangulation: having at our disposal only results of a small portion of all possible trees that do not vary much from each other globally, the resulted condition numbers do not vary much either.

To illustrate the effect stated above, we first introduce the table 4.2 with the results for the first set of trees where weight$(e) = 1$ for all $e \in \mathcal{E}_i$.

set of matrices	$\mathcal{S}(E_{\text{rem}})$	$\mathcal{M}(E_{\text{rem}})$	$\mathcal{A}(E_{\text{rem}})$
min cond.	$8.3769 \cdot 10^3$	$1.0040 \cdot 10^4$	$1.5536 \cdot 10^5$
max cond.	$1.3950 \cdot 10^4$	$1.3081 \cdot 10^4$	$2.1212 \cdot 10^5$

Table 4.2: The extremal resuls for weight(e)=1.

Not only the extremal condition numbers are relatively close, but we can get almost the full dispersion of condition by alternating a single edge. The trees depicted in Figure 4.11 differ in a single edge, yet condition numbers of the related matrices from $\mathcal{A}(E_{\text{rem}})$ are close to both exremal values among all matrices in $\mathcal{A}(E_{\text{rem}})$.



(a) $c(A) = 1.5564 \cdot 10^4$ (b) $c(A) = 2.1004 \cdot 10^4$

Figure 4.11: The largest difference in conditioning between two trees that varies just in a single edge.

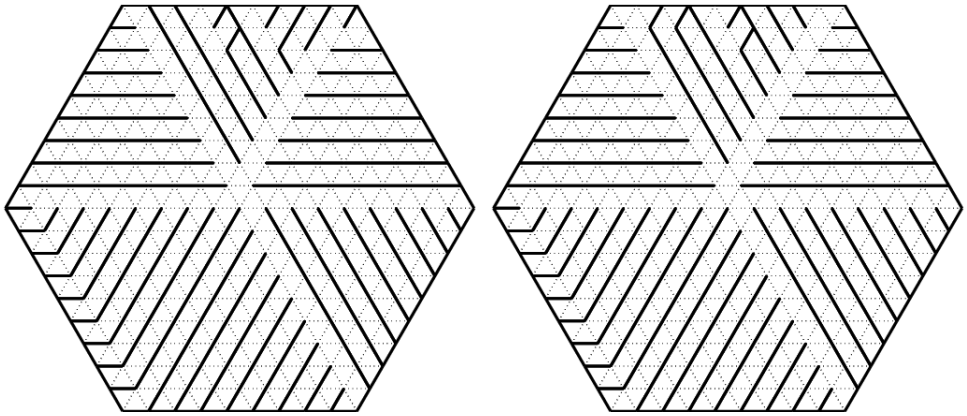
Moving to the set of trees generated with weight$(e) = d$, where d equals to the distance of e from the boundary of triangulation, we found that the results vary even less compared to previous case, as shown in Table 4.3.

set of matrices	$\mathcal{S}(E_{\text{rem}})$	$\mathcal{M}(E_{\text{rem}})$	$\mathcal{A}(E_{\text{rem}})$
min cond.	$3.6591 \cdot 10^3$	$3.1687 \cdot 10^3$	$4.4497 \cdot 10^4$
max cond.	$3.7294 \cdot 10^3$	$3.1691 \cdot 10^3$	$4.5767 \cdot 10^4$

Table 4.3: The extremal results for weight$(e) = d$.

Figure 4.12 shows the two trees leading to matrices with extremal conditioning among all matrices from $\mathcal{A}(E_{\text{rem}})$. These trees provide almost the extremal condition numbers also for matrices from $\mathcal{S}(E_{\text{rem}})$ and $\mathcal{M}(E_{\text{rem}})$. The corresponding condition numbers differ from the extremal ones in the third significant digit at most.

When the weights of edges were set the other way round to the preceding case, i.e. weight$(e) = \frac{1}{d}$, where d equals to the distance of e from the boundary of the domain, we obtained results presented in Table 4.4.



(a) $c(A) = 4.4497 \cdot 10^4$ (b) $c(A) = 4.5767 \cdot 10^4$

Figure 4.12: Trees with extremal $c(A)$ values.

set of matrices	$\mathcal{S}(E_{\text{rem}})$	$\mathcal{M}(E_{\text{rem}})$	$\mathcal{A}(E_{\text{rem}})$
min cond.	$9.1761 \cdot 10^3$	$1.1653 \cdot 10^4$	$1.4650 \cdot 10^5$
max cond.	$1.4952 \cdot 10^4$	$1.14556 \cdot 10^4$	$2.0593 \cdot 10^5$

Table 4.4: The extremal results for weight$(e) = 1/d$.

Despite the broader differences in the results compared to the previous set, the trees corresponding to extremal results of $\mathcal{A}(E_{\text{rem}})$ are depicted in Figure 4.13. The trees corresponding to the extremal condition numbers of matrices from $\mathcal{S}(E_{\text{rem}})$ and $\mathcal{M}(E_{\text{rem}})$ would look very similarly.

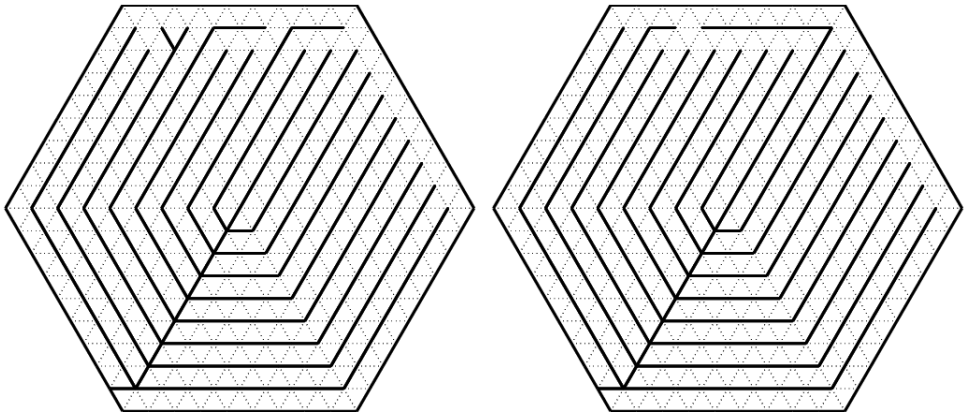
The most diverse results were obtained when the weights of edges were set randomly, see Table 4.5.

set of matrices	$\mathcal{S}(E_{\text{rem}})$	$\mathcal{M}(E_{\text{rem}})$	$\mathcal{A}(E_{\text{rem}})$
min cond.	$1.4952 \cdot 10^4$	$2.1341 \cdot 10^4$	$6.7073 \cdot 10^4$
max cond.	$3.4119 \cdot 10^4$	$4.0948 \cdot 10^4$	$2.1304 \cdot 10^5$

Table 4.5: The extremal results for weight(e) set randomly.

Figure 4.14 shows two trees leading to minimal and maximal values of c(A). In this figure we emphasize using bold edges, that in the left panel the most of the edges form one large branch, while in the right panel they are divided into two branches.

Finally, Table 4.6 summarizes the extremal values among all four cases.



(a) $c(A) = 1.4650 \cdot 10^5$ (b) $2.0593 \cdot 10^5$

Figure 4.13: Trees with extremal $c(A)$ values.

set of matrices	$\mathcal{S}(E_{\text{rem}})$	$\mathcal{M}(E_{\text{rem}})$	$\mathcal{A}(E_{\text{rem}})$
min cond.	$3.6591 \cdot 10^3$	$3.1687 \cdot 10^3$	$4.4497 \cdot 10^4$
max cond.	$3.4119 \cdot 10^4$	$4.0948 \cdot 10^4$	$2.1304 \cdot 10^4$

Table 4.6: Summarized results of Example 3.

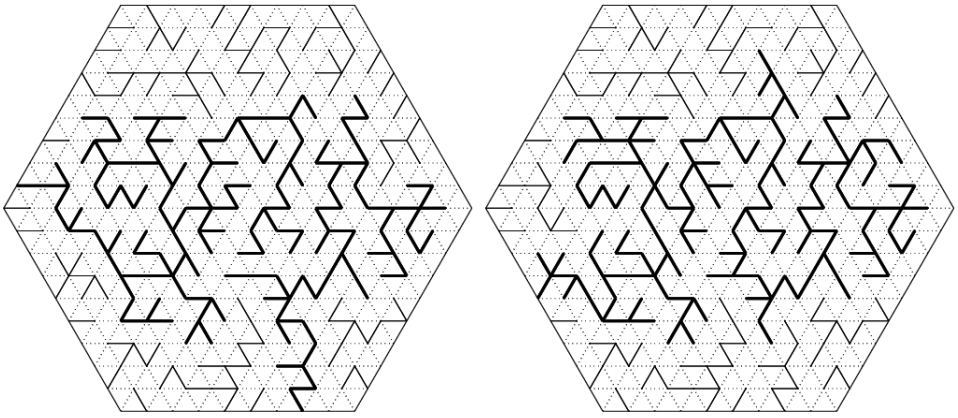
4.1.4 Analysis of the results and conclusions

This subsection tries to interpret the results presented so far in this chapter. The analysis will mostly rely on the statistical tool for measuring dependence in observed data – the correlation coefficient. It is a number between -1 and 1. If there is no relationship between the data, the correlation coefficient is close to zero. The stronger is the direct linear dependence among the two sets of data, the closer the correlation coefficient is to 1. On the other hand, the stronger is the negative linear dependence among the two sets of data, the closer the correlation coefficient is to -1. Further details can be found e.g. in [14, page 461].

First, we will present the rate of dependence, measured by the correlation coefficient, among conditioning of the matrices from the sets $\mathcal{S}(E_{\text{rem}}), \mathcal{M}(E_{\text{rem}})$ and $\mathcal{A}(E_{\text{rem}})$.

Let us denote by $\boldsymbol{C}_S, \boldsymbol{C}_M$ and $\boldsymbol{C}_A$ the vectors, which have in the i-th row, $i = 1, ..., |E_{\text{rem}}|$, the condition number related to the same $\mathcal{E}_{\text{rem}} \in E_{\text{rem}}$.

Then in Example 1 the correlation coefficients among matrices from



(a) $c(A) = 6.7073 \cdot 10^4$ (b) $c(A) = 2.1304 \cdot 10^5$

Figure 4.14: Trees with extremal $c(A)$ values.

sets $\mathcal{S}(E_{\text{rem}}), \mathcal{M}(E_{\text{rem}})$ and $\mathcal{A}(E_{\text{rem}})$ were:

$$\begin{aligned} \text{corr}(\boldsymbol{C}_S, \boldsymbol{C}_M) &= 0.7597, \\ \text{corr}(\boldsymbol{C}_S, \boldsymbol{C}_A) &= 0.9970, \\ \text{corr}(\boldsymbol{C}_M, \boldsymbol{C}_A) &= 0.7321. \end{aligned}$$

In Example 2, we obtained these results:

$$\begin{aligned} \text{corr}(\boldsymbol{C}_S, \boldsymbol{C}_M) &= 0.6148, \\ \text{corr}(\boldsymbol{C}_S, \boldsymbol{C}_A) &= 0.3189, \\ \text{corr}(\boldsymbol{C}_M, \boldsymbol{C}_A) &= 0.5491. \end{aligned}$$

Finally, in Example 3 the dependence measured by the correlation coefficient was

$$\begin{aligned} \text{corr}(\boldsymbol{C}_S, \boldsymbol{C}_M) &= 0.9688, \\ \text{corr}(\boldsymbol{C}_S, \boldsymbol{C}_A) &= 0.6215, \\ \text{corr}(\boldsymbol{C}_M, \boldsymbol{C}_A) &= 0.4898. \end{aligned}$$

From these results one cannot make any conclusions of dependence among sets $\mathcal{S}(E_{\text{rem}}), \mathcal{M}(E_{\text{rem}})$ and $\mathcal{A}(E_{\text{rem}})$.

One can also pose the following question: Can we deduce from geometrical qualities of the selected tree $\mathcal{G}^{\partial} = (\mathcal{V}_i \cup \{\mathcal{B}^{\partial}\}, \mathcal{E}^{\partial}_{\text{rem}})$, whether the related matrices S, M and A are better- or worse-conditioned than the most of matrices from $\mathcal{S}(E_{\text{rem}}), \mathcal{M}(E_{\text{rem}})$ and $\mathcal{A}(E_{\text{rem}})$?

From the figures so far presented in this chapter, it seems that the trees with shorter branches lead to better-conditioned matrices. Even despite the existence of counterexamples, e.g. trees in Figure 4.8.

However, it has proved extremely hard to formulate this hypothesis in a rigorous way. As an example we provide the following criterion:
Criterion 1. Consider all pairs of boundary edges. For every pair, find the shortest path from one to another, stepping only on edge midpoints (no edges can be skipped). It is forbidden to step on edges which belong to a spanning tree. The length of the path is the number of interior edges one steps on. Thus for every pair of edges one has a natural number. The maximum of these numbers over all pairs is the criterion.

Its value is lower for trees with shorter branches, nevertheless the dependence to the conditioning is not too strong, for example the correlation coefficient between the criterion and the conditioning of matrices from $\mathcal{A}(E_{\text{rem}})$ is 0.3065 in Example 1 and even worse in Examples 2 and 3.

We have tried many other more or less similar criteria, but always with the same inconclusive results.

4.2 Iteration schemes

For now, let us consider that we have picked some particular $\mathcal{E}_{\text{rem}} \in E_{\text{rem}}$. This $\mathcal{E}_{\text{rem}}$ can be chosen based on suggestions from the previous section. Using the complement $\mathcal{E}_{\text{keep}} \subset \mathcal{E}_i$ of $\mathcal{E}_{\text{rem}}$ to construct the basis (4.1), the final linear system obtained from discretization of the variational formulation of the time-harmonic Maxwell's equations (1.23) with $\Gamma_I = \emptyset$ in the space $\boldsymbol{V}_h$ has the following form:

$$(S - \kappa^2 M)\boldsymbol{Y}_h = \boldsymbol{F}_h, \tag{4.6}$$

where S was defined in (4.3) and M in (4.4).

If we rearrange the elements of the basis $\mathcal{B}$ in the way that the gradients of scalar hat functions are at the first N_{iv} positions, then according to (4.5) the matrix S has the following structure:

$$S = \begin{pmatrix} 0 & 0 \\ 0 & S_{22} \end{pmatrix}, \tag{4.7}$$

where $S_{22} \in \mathbb{R}^{(N_{ie}-N_{iv})\times(N_{ie}-N_{iv})}$ is symmetric positive definite. Since it will be used later, we rewrite the remaining matrices and vectors in (4.6)

in the same 2-by-2 block structure:

$$M = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix}, \quad \boldsymbol{Y}_h = \begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \end{pmatrix}, \quad \boldsymbol{F}_h = \begin{pmatrix} \boldsymbol{F}_1 \\ \boldsymbol{F}_2 \end{pmatrix}, \tag{4.8}$$

where $M_{11} \in \mathbb{R}^{N_{iv} \times N_{iv}}$, $M_{12}, M_{21}^T \in \mathbb{R}^{N_{iv} \times N_{ie}}$, $M_{22} \in \mathbb{R}^{(N_{ie}-N_{iv}) \times (N_{ie}-N_{iv})}$, $\boldsymbol{Y}_1, \boldsymbol{F}_1 \in \mathbb{R}^{N_{iv}}$ and $\boldsymbol{Y}_2, \boldsymbol{F}_2 \in \mathbb{R}^{N_{ie}}$.

The matrix M is symmetric positive definite as well. However, the system matrix $S - \kappa^2 M$ is in general symmetric but indefinite. That makes it difficult to solve by iterative solvers.

The goal of this section is to examine iterative methods that takes advantage of the structure of the matrix S.

We present the methods first. If not mentioned otherwise, all subproblems are solved by a direct sparse solver. Since we use the system Matlab for numerical experiments, we use its "backslash" command for solving the subproblems.

Method A. Start with an initial guess $\boldsymbol{Y}_h^{(0)}$ and compute $\boldsymbol{Y}_h^{(k+1)}$, $k = 0, 1, 2, \ldots$ by

$$-\kappa^2 M \boldsymbol{Y}_h^{(x+1)} = \boldsymbol{F}_h - S \boldsymbol{Y}_h^{(x)}.$$

This method does not exploit the special 2×2 block structure of matrix S, nevertheless in each iteration a system with symmetric positive matrix M has to be solved.

Method B. With (4.7) and (4.8) we can reformulate (4.6) as

$$-\kappa^2 M_{11} \boldsymbol{Y}_1 - \kappa^2 M_{12} \boldsymbol{Y}_2 = \boldsymbol{F}_1, \tag{4.9}$$

$$S_{22} \boldsymbol{Y}_2 - \kappa^2 M_{21} \boldsymbol{Y}_1 - \kappa^2 M_{22} \boldsymbol{Y}_2 = \boldsymbol{F}_2. \tag{4.10}$$

Starting with an initial guess $\boldsymbol{Y}_2^{(0)}$, we compute $\boldsymbol{Y}_2^{(k+1)}$ from (4.9) and $\boldsymbol{Y}_1^{(k+1)}$ from (4.10), $k = 0, 1, 2, \ldots$, as follows:

$$-\kappa^2 M_{11} \boldsymbol{Y}_1^{(k+1)} = \boldsymbol{F}_1 + \kappa^2 M_{12} \boldsymbol{Y}_2^{(k)}, \tag{4.11}$$

$$(S_{22} - \kappa^2 M_{22}) \boldsymbol{Y}_2^{(k+1)} = \boldsymbol{F}_2 + \kappa^2 M_{21} \boldsymbol{Y}_1^{(k+1)}. \tag{4.12}$$

Although this method takes into account the block structure of S, the matrix $S_{22} - \kappa^2 M_{22}$ from (4.12) is generally indefinite.

Method C. This method is close to method B, only the matrix needed to get $\boldsymbol{Y}_2^{(k+1)}$ is in this case symmetric positive definite. From initial guess $\boldsymbol{Y}_2^{(0)}$ we get $\boldsymbol{Y}_1^{(k+1)}$ and $\boldsymbol{Y}_2^{(k+1)}$, $k = 0, 1, 2, \ldots$, solving

$$-\kappa^2 M_{11} \boldsymbol{Y}_1^{(k+1)} = \boldsymbol{F}_1 + \kappa^2 M_{12} \boldsymbol{Y}_2^{(k)}, \tag{4.13}$$

$$S_{22} \boldsymbol{Y}_2^{(k+1)} = \boldsymbol{F}_2 + \kappa^2 M_{21} \boldsymbol{Y}_1^{(k+1)} + \kappa^2 M_{22} \boldsymbol{Y}_2^{(k)}. \tag{4.14}$$

Method D. By another minor modification of (4.12) we can get system of symmetric positive definite matrices once again. Let us start with $\boldsymbol{Y}_2^{(0)}$, then we get $\boldsymbol{Y}_1^{(k+1)}$ and $\boldsymbol{Y}_2^{(k+1)}$, $k = 0, 1, 2, \ldots$, from

$$-\kappa^2 M_{11}\boldsymbol{Y}_1^{(k+1)} = \boldsymbol{F}_1 + \kappa^2 M_{12}\boldsymbol{Y}_2^{(k)},$$
$$-\kappa^2 M_{22}\boldsymbol{Y}_2^{(k+1)} = \boldsymbol{F}_2 - S_{22}\boldsymbol{Y}_2^{(k)} + \kappa^2 M_{21}\boldsymbol{Y}_1^{(k+1)}.$$

Method E. This is a modification of method C that keeps the first block of 2×2 block structure of S_{22} nulled: while $\boldsymbol{Y}_1$ is obtained from (4.13) using a sparse direct solver as in metod C, $\boldsymbol{Y}_2$ is obtained from (4.14) using conjugated gradients method.

Having the algorithms described, we can move to numerical experiments. We chose three hexagonal meshes with 702, 5112 and 10302 inner edges. The first one is the same as in Example 1, the other two were obtained from the first one by its uniform refinements.

Each time we tested the methods described above for fixed matrices S and M from (4.6) and variable κ^2. For comparison, we tried out Matlab build-in iterative solvers `bicq`, `bicgstab`, `cg`, `cgs`, `gmres`, `lsqr` and `qmr` on solving of the same underlying problem discretized by edge elements only.

4.2.1 Test 1

In this test the matrices S and M are from $\mathbb{R}^{702\times 702}$. Edge basis functions related to the tree depicted in Figure 4.12(a) were used for their construction. We considered the method as convergent if its relative residual drops below 10^{-6} in at most 702 iterations. The upper limit for the number of iterations of conjugate gradients in method E was set to 217 (which in this test equals to N_{iv}).

The convergence of the tested method is described in Table 4.7, while the convergence of the traditional methods in Table 4.8.

		method				
κ^2	$\mathrm{cond}(S-\kappa^2 M)$	A	B	C	D	E
10^{-5}	1.63×10^7	–	3	3	–	–
10^{-4}	1.63×10^6	–	4	4	–	–
10^{-3}	1.62×10^5	–	9	5	–	–
10^{-2}	1.57×10^4	–	–	11	–	–
10^{-1}	1.19×10^4	–	–	–	–	–
10^{0}	4.45×10^4	–	–	–	–	–
10^{1}	4.56×10^4	–	–	–	–	–
10^{2}	3.58×10^3	19	–	–	–	–
10^{3}	3.20×10^3	6	–	–	–	–
10^{4}	3.17×10^3	4	–	–	–	–
10^{5}	3.17×10^3	3	–	–	–	–

Table 4.7: Conditioning and the number of iterations needed for convergence for the tested methods in Test 1.

κ^2	cond$(S-\kappa^2 M)$	`bicg`	`bicgstab`	`cg`	`cgs`	`gmres`	`lsqr`	`qmr`
10^{-5}	2.39×10^6	51	42	51	41	50	181	50
10^{-4}	2.39×10^5	51	42	51	41	50	181	50
10^{-3}	2.39×10^4	51	42	51	41	50	181	50
10^{-2}	2.39×10^3	52	42	52	42	51	182	51
10^{-1}	6.99×10^2	56	44	–	45	55	186	55
10^0	3.47×10^3	84	125	–	–	84	226	84
10^1	2.29×10^2	192	339	–	–	182	274	185
10^2	3.81×10^0	12	8	–	7	12	24	12
10^3	2.09×10^0	8	5	–	5	8	13	8
10^4	2.00×10^0	8	5	–	4	8	12	8
10^5	1.99×10^0	8	5	–	4	8	12	8

Table 4.8: Conditioning and the number of iterations needed for convergence for the traditional methods in Test 1.

4.2.2 Test 2

This time, the matrices S and M have dimensions 5112×5112. Moving to larger matrix implied using its sparse form and instead of direct computing the conventional 2-norm condition number we present only the lower 1-norm estimate implemented in Matlab function `condest`. The number of conjugated gradients steps in method E was left to 217 as raising it had no implications on convergence at all.

The rate of convergence in Test 2 can be found in Tables 4.9 and 4.10.

κ^2	cond$(S-\kappa^2 M)$	method				
		A	B	C	D	E
10^{-5}	1.86×10^8	–	7	4	–	–
10^{-4}	1.79×10^7	–	–	6	–	–
10^{-3}	1.64×10^6	–	–	11	–	–
10^{-2}	5.56×10^6	–	–	–	–	–
10^{-1}	3.21×10^6	–	–	–	–	–
10^{0}	1.47×10^6	–	–	–	–	–
10^{1}	5.39×10^6	–	–	–	–	–
10^{2}	2.69×10^5	23	–	–	–	–
10^{3}	2.28×10^5	7	–	–	–	–
10^{4}	2.25×10^5	4	–	–	–	–
10^{5}	2.05×10^5	7	–	–	–	–

Table 4.9: Conditioning and the number of iterations needed for convergence for the tested methods in Test 2.

4.2.3 Test 3

In this case all results were obtained in the same way as in the previous one, only the matrices S and M were enlarged to the size 10302×10302.

The results of Test 3 can be found in Tables 4.11 and 4.12.

4.2.4 Conclusions

Apart from Methods A and B, where the former does not take into account the block structure of matrices S and M and in the latter occurs indefinite matrix, only Method C converged and only in the case of small κ^2. In this case the rate of convergence was faster compared to the traditional widely used methods.

Methods D and E were found not to converge in any of the test cases. Method A works well for large values of κ^2 and it provides comparable results to the traditional methods. However, this method is not advantageous, because it does not take into account the block structure of matrices S and M.

Method B converges for very small values of κ^2 only. Moreover, its drawback is that a system with an indefinite matrix have to be solved in each iteration, see (4.12).

The most successful seems to be Method C. However, it converges for small values of κ^2 only. On the other hand, if it converges then the number of needed iterations is substantially smaller then the number of

κ^2	cond$(S-\kappa^2 M)$	bicg	bicgstab	cg	cgs	gmres	lsqr	qmr
10^{-5}	1.32×10^7	137	115	137	108	135	1249	135
10^{-4}	1.33×10^6	137	109	137	108	135	1249	135
10^{-3}	1.28×10^5	138	123	138	111	136	1251	136
10^{-2}	2.12×10^4	147	129	–	126	144	1269	144
10^{-1}	2.06×10^4	185	375	–	–	184	1353	184
10^{0}	1.36×10^4	475	1165	–	–	465	2152	466
10^{1}	1.10×10^4	2071	2861	–	–	1536	1953	1948
10^{2}	4.79×10^0	12	7	–	7	12	24	12
10^{3}	2.47×10^0	8	5	–	4	8	12	8
10^{4}	2.35×10^0	8	4	–	4	8	12	8
10^{5}	2.33×10^0	8	4	–	4	8	12	8

Table 4.10: Conditioning and the number of iterations needed for convergence for the traditional methods in Test 2.

iterations of all the traditional methods.

κ^2	$\text{cond}(S-\kappa^2 M)$	method				
		A	B	C	D	E
10^{-5}	3.71×10^8	–	13	5	–	–
10^{-4}	3.42×10^7	–	–	7	–	–
10^{-3}	1.41×10^7	–	–	17	–	–
10^{-2}	1.16×10^7	–	–	–	–	–
10^{-1}	5.80×10^6	–	–	–	–	–
10^{0}	3.06×10^6	–	–	–	–	–
10^{1}	1.08×10^7	–	–	–	–	–
10^{2}	4.88×10^5	23	–	–	–	–
10^{3}	4.40×10^5	7	–	–	–	–
10^{4}	4.62×10^5	4	–	–	–	–
10^{5}	4.20×10^5	7	–	–	–	–

Table 4.11: Conditioning and the number of iterations needed for convergence for the new methods in Test 3.

κ^2	$\text{cond}(S-\kappa^2 M)$	`bicg`	`bicgstab`	`cg`	`cgs`	`gmres`	`lsqr`	`qmr`
10^{-5}	1.35×10^7	195	148	195	159	191	2463	191
10^{-4}	1.37×10^6	195	149	195	147	192	2464	192
10^{-3}	1.50×10^5	199	151	199	163	195	2470	195
10^{-2}	7.23×10^4	208	202	–	179	203	2494	203
10^{-1}	1.72×10^4	327	1357	–	–	308	3121	310
10^{0}	1.96×10^4	893	1884	–	–	855	3787	880
10^{1}	1.06×10^4	4238	10617	–	–	3427	4776	4084
10^{2}	4.79×10^0	11	7	–	6	11	23	11
10^{3}	2.47×10^0	8	5	–	4	8	12	8
10^{4}	2.35×10^0	8	4	–	4	8	12	8
10^{5}	2.33×10^0	8	4	–	4	8	12	8

Table 4.12: Conditioning and the number of iterations needed for convergence for the traditional methods in Test 3.

Summary

On the way to accomplish the goal that we set ourselves in the preface, we focused on creating clear and self-contained mathematical document.

We started in Chapter 1 with introducing three dimensional Maxwell's equations in their differential form. We transformed these equations to the time-harmonic form and covered important topics such as media characteristics and interface and boundary conditions. This helped us to introduce the weak formulation of time-harmonic Maxwell's equations. After a short note on its transformation to two dimensions, we discretized the 2D time-harmonic Maxwell's equations using the Whitney basis.

The second chapter was fully devoted to the de Rham diagram. We started with the necessary definitions and presented the diagram in a form relating the spaces of continuous functions. From there we went to the version of the digram that relates the functional spaces used in weak formulation of time-harmonic Maxwell's equations. When we added the layer of finite dimensional finite element spaces to the diagram, we finally obtained its complete form. In the 2D version of the complete diagram we proved the exactness at one certain level and the commutativity.

In Chapter 3 we explained how to construct bases of finite dimensional finite element subspaces of $\boldsymbol{H}(\mathrm{curl}, \Omega)$ compatible with the de Rham diagram. It was obvious from the construction that this basis is definitely not unique.

Finally, in the last chapter, we examined the practical aspects of the newly constructed basis. In its first part we tested on three examples which of the newly constructed bases have the best properties, i.e. leads to finite element matrices with the best conditioning. The second part examined several iteration schemes that take advantage of the structure of stiffness and mass matrices and compared their results to the traditionally used methods.

Several authors addressed similar problems in related context, see e.g. [2],[5] or [6]. While the undoubted importance in the theory, the practical aspects of the suggested finite element method of the lowest-order for the

time-harmonic Maxwell's Equations compatible with de Rham diagram are slightly more problematic. For example, the conditioning of stiffness and mass matrices is worse compared to the standard Whitney basis.

As early as during the introductory familiarization with the subject of the thesis, when we tested the mesh with three interior edges only, we proposed a hypothesis. It states: trees with shorter and plainer branches lead to better conditioned matrices than tress with less but longer and richer branches.

In the experiments described in Examples 1–3 in the fourth chapter we collected enough data to confirm the hypothesis. As described in the second part subsection 4.1.4, it has proved extremely hard to formulate the hypothesis in a rigorous way.

As a secondary result, we emphasize one iteration method from the set of tested iteration schemes. This iteration method takes advantage of the special 2-by-2 block structure of the stiffness and mass matrices from the finite element discretization where the newly created basis was used. The method converges for small values of κ^2 only. In this case, the number of needed iterations is notably smaller then the number of iterations needed by the traditional methods.

Bibliography

[1] Bossavit A.: Computational Electromagnetism, Academic Press, London, 1997.

[2] Cendes Z.J., Mandes J.B.: Tree-cotree decomposition for first-order complete tangential vector finite elements, Internat. J. Numer Methods Eng. 40 (1997), 1667–1685.

[3] Davidson D.B.: Computational Electromagnetics for RF and Microwave Engineering, Second Edition, Cambridge University Press, Cambridge, 2010.

[4] Harris J.M., Hirst J.L., Mossinghoff, M.J.: Combinatorics and graph theory, Springer, New York, 2008.

[5] Igarashi I.: On the Property of the Curl-Curl Matrix in Finite Element Analysis With Edge Elements, IEEE Transactions on Magnetics 37 (2001), 3129–3132.

[6] Kaveh A., Ghaderi I.: Conditioning of structural stiffness matrices, Computers & Structures 63 (1997), 719–725.

[7] Monk P.: Finite Element Methods for Maxwell's Equations, Clarendon Press, Oxford, 2002.

[8] Nédelec J.C.: Mixed finite elements in $\mathbb{R}^3$. Numer. Math. 35 (1980), 315–41.

[9] Sörensen K., Janssens G. K.: An algorithm to generate all spanning trees of a graph in order of increasing cost, Pesqui. Oper. 25 (2005), 219–229.

[10] Stein, E., Borst, R. and Hughes, T.J.R.: Encyclopedia of computational mechanics: Fundamentals, Vol. 1, John Wiley & Sons, Chichester, 2004.

[11] Šolín P.: Partial Differential Equations and the Finite Element Method, Wiley-Interscience, New Jersey, 2004.

[12] Tutte W. T.: Graph Theory, Cambridge University Press, Cambridge, 2001.

[13] Vejchodský T.: On Edge-Elements Reproducing the Discrete Kernel of the Curl Operator, paper presented at ESCO 2008, Jetřichovice, Czech Republic, June 2008.

[14] Verma A.K., Ajit, S., Karanki D.R.: Reliability and Safety Engineering, Springer, 2010.

[15] Whitney H.: Geometric Integration Theory, Princeton University Press, Princeton, 1957.