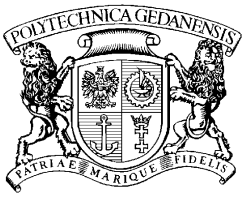
Classical and hybrid techniques for the analysis of microwave waveguides and resonators

PhD Thesis

by

Jacek Mielewski

Faculty of Electronics, Telecommunications and Informatics
of the
Technical University of Gdańsk



Supervisor: Professor Michał Mrozowski

Gdańsk 1999

To my Wife and my Parents

Contents

Symbol conventions

General symbols

$\mathbf{L}$	—	operator
u	—	function
j	—	imaginary unity
$\underline{\underline{A}}$	—	matrix
$\underline{\underline{I}}$	—	unit matrix
$\underline{\underline{a}}$	—	tensor
$\underline{a}$	—	vector of parameters
$\vec{A}$	—	vector function
$A_{(\cdot)}$	—	$(\cdot)$ component of a vector function
$\hat{n}$	—	unit vector normal to a surface
$\hat{(\cdot)}$	—	unit vector in $(\cdot)$ direction
$(\cdot)^*$	—	complex conjugation
$(\cdot)^{-1}$	—	inverse of an operator
$(\cdot)^T$	—	transpose of an operator
$(\cdot)^H$	—	Hermitian transpose of an operator
$\Big(u, v\Big)$	—	standard inner product $(= \int_\Omega uv^* \, d\Omega)$
$(\cdot) \times (\cdot)$	—	vector product of two vectors
$(\cdot) \cdot (\cdot)$	—	scalar product of two vectors

Physical quantities

$\vec{D}$	—	electric flux density
$\vec{B}$	—	magnetic flux density
$\vec{E}$	—	electric field intensity
$\vec{H}$	—	magnetic field intensity
$\underline{\underline{\epsilon}}$	—	permittivity tensor of the medium
$\underline{\underline{\mu}}$	—	permeability tensor of the medium
ϵ_0	—	permittivity of the vacuum
μ_0	—	permeability of the vacuum
v_c	—	speed of light in the vacuum
ω	—	angular frequency
f	—	frequency
β	—	propagation constant
$k_{(\cdot)}$	—	wavenumber in $(\cdot)$ direction

Operators

$\nabla \times (\cdot)$	—	rotation operator
$\nabla \cdot (\cdot)$	—	divergence operator
$\nabla(\cdot)$	—	gradient operator

Chapter 1

Introduction

1.1 Motivation and background

A dynamic progress in the field of high-frequency technology influences the methodology of construction of microwave, millimeter-wave, and optical-wave devices. In order to improve the parameters of devices, engineers explore structures of new shapes, filled with atypical materials. An example of such a tendency is the evolution of the shape of nonreciprocal microwave ferrite phase shifter. The original structure, proposed in [116] and shown in Fig. 1.1(a), consisted of two vertical ferrite slabs inserted within a rectangular waveguide. This construction evolved [49, 78] into a complex structure involving a waveguide with nonuniform cross-section filled with a ferrite toroid (Fig. 1.1(b)). Another application of nonuniform structures are integrated optics circuits. An example of application of new materials of complex anisotropic properties are low-loss resonators [40, 54, 60, 61] or filters (crystals, e.g. sapphire) or various nonreciprocal or surface-wave structures (ferrites, chiroferrites) [49, 60, 78].

One of the most fundamental problems associated with the design of passive components such as filters, couplers or transitions and in the application of resonant methods for characterization of the materials used in the microwave and millimeter-wave bands is an eigenvalue analysis. Solution of an eigenproblem gives the information about the resonant spectrum of a resonator, the length of a wave in a waveguide and the field patterns for the corresponding modes. Due to the large constructional complexity of the devices and complex properties of the materials used in the microwave and higher frequency ranges analytical methods of analysis cannot be usually applied. In consequence, numerical methods of analysis of electromagnetic fields have became one of the most intensively explored research topics for many research groups in the world. This field of research is being intensively developed, which results in a growing number of publications.

The numerical analysis of resonators and waveguides consists of three steps: analytical formulation of an operator problem, projection of the problem into a finite dimensional space, resulting in a matrix problem, and, finally, numerical solution of the matrix problem. Efficiency of the whole solution procedure requires the selection of a proper approach for each of the steps.

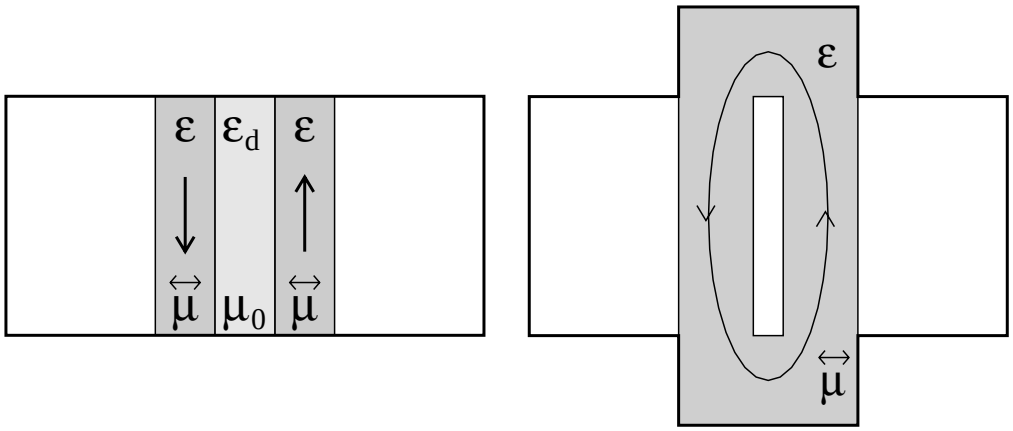


Figure 1.1: Cross-sections of a simple rectangular waveguide phase shifter structure proposed in [116] (a) and the structure based on grooved waveguide proposed in [49, 78] (b).

Various analytical formulations for waveguide and resonator structures can be derived from Maxwell's equations, resulting in canonical or noncanonical (nonstandard) eigenvalue problems. Among the canonical eigenproblems we can distinguish standard problems, of the form

$$\mathbf{A}u = \lambda u \tag{1.1}$$

and generalized problems, of the form

$$\mathbf{A}u = \lambda \mathbf{B}u \tag{1.2}$$

where $\mathbf{A}$, $\mathbf{B}$ are linear operators, u is an eigenfunction and λ is the corresponding eigenvalue. An example of nonstandard eigenproblem is a quadratic eigenproblem, i.e.

$$\lambda^2 \mathbf{A}u + \lambda \mathbf{B}u + \mathbf{C}u = 0 \tag{1.3}$$

It can be shown [98] that this particular problem can easily be transformed into a generalized one with a number of unknowns doubled. The following operator formula

$$\mathbf{M}(\lambda)u = 0 \tag{1.4}$$

describes the most general form of the noncanonical eigenproblem. Since all the above formulations originate from Maxwell's equations, they are mathematically fully equivalent. However, from the efficiency point of view they are not. Numerical solution of eigenproblem (1.1) or (1.2) by the standard method of moment (see Sec. 3.1.2) or Rayleigh-Ritz method (see Sec. 3.1.1) lead to standard or generalized matrix eigenvalue problems, while the solution of (1.4) usually leads to the linear system which can be solved by searching zeroes of the operator matrix determinant [79]. Since the latter technique is, in general,

much less efficient than the former one, canonical eigenproblems are of greater interest. Another aspect affecting the efficiency is the number of unknowns (field components) involved in the formulation, which determines the size of the operator matrix of the eigenproblem that has to be solved. In consequence, canonical eigenproblems may be preferred over the nonstandard quadratic ones which are equivalent to the solution of a double sized generalized eigenproblem. Another possibility, which can result in considerable time and memory savings, arises in the solution of homogeneous 3D problems using 2D approach whenever it is possible (see Sec. 2.3.2).

1.1.1 Classical methods of analysis

As already pointed out, numerical solution of an eigenvalue problem requires the conversion of its analytical form into the matrix one. This process relies on representing of the infinite dimensional operator eigenproblem in the finite dimensional space and is referred to as *projection*. There are two main projection methods, the method of moments and the Rayleigh-Ritz method. The main difference between various projection techniques based on these methods is the set of basis functions involved. Among the most frequently used functions are simple Maxwellian and polynomial ones [48, 59, 80, 81]. Depending on the physical problem and basis functions, the projection methods result in the eigenproblems with matrix (matrices) of specific properties. The most considerable features of the matrices, which decide on the efficiency of the numerical methods used to solve the problem, include small size, symmetry and sparsity. Therefore, the selection of a proper conversion method becomes very important for efficiency. A main drawback of the methods involving the simple entire domain Maxwellian basis functions is that they can only be used to the analysis of structures of highly regular shape and hence other, more versatile, conversion methods become indispensable. The simplest and the most popular of them are the finite element method (FEM) [2, 3, 7, 19, 21, 23, 24, 30–34, 38, 44, 48, 55, 56, 66–68, 89–92, 107, 114, 117, 120, 121] and the finite difference frequency domain method (FDFD) [1, 9, 11–14, 18, 25, 37, 40, 46, 52, 53, 64, 77, 82, 84, 94, 101–106, 112, 113, 119, 122–125, 127], using the polynomial basis functions which are nonzero only locally. Main advantages of both conversion methods are that they can easily cope with irregularly shaped external and internal boundaries and they result in sparse matrices. One of the most important differences between these methods is that the FEM results in a generalized problem, while the FDFD in a standard one with highly structured matrix. This fact can considerably influence the efficiency of the solution of the matrix eigenproblem.

Since the memory and time requirements of various numerical solution methods strongly depend on the matrix size, symmetry and sparsity, the selection of a suitable numerical algorithm for a given matrix problem should take into consideration these matrix topological and spectral properties. Moreover, the features of the eigensolver, such as fast convergence and ability to select particular eigenvalues within the entire spectrum, which are also important in terms of memory and time savings, should also be considered. Traditional numerical methods, such as QR or QZ, are very popular in solving problems with

small and dense matrices, arising in the Rayleigh-Ritz method and method of moments involving the entire domain basis functions. However, when FEM or FDFD methods are used, producing very large and sparse matrices, QR and QZ algorithms require very large computational effort. It stems from the fact, that these algorithms compute all eigenvalues and operate on all matrix elements. In this case the application of iterative methods, such as the *power method*, *subspace iteration* or *Arnoldi/Lanczos methods* [39, 98], which can benefit from the sparse character of the matrices and compute only selected eigenvalues, can substantially improve efficiency of computations. Basic versions of iterative algorithms compute the eigenvalues of the largest magnitude, which are usually out of interest as far as electromagnetic eigenproblems are concerned. In order to obtain convergence to the required eigenvalues and accelerate the algorithms various *spectral transformations* are employed. Typical examples include shift and shift-invert techniques or polynomial filtering. Another important aspect of efficient numerical calculations by means of the iterative algorithms is the possibility of utilization of capabilities offered by modern superscalar and parallel computer systems. Recently, new versatile iterative algorithms have been released [8,65], which use *reverse communication* procedure for performing matrix-vector products. In consequence, many system specific properties of the modern computers can be used to improve efficiency. One such algorithm is the *implicitly restarted Arnoldi method* [108].

Spurious solutions. An important aspect of numerical analysis of electromagnetic fields are nonphysical spurious solutions, occurring as a consequence of chosen formulation and projection method used. They can arise when vector basis functions that satisfy Maxwell's curl equations but do not satisfy Maxwell's divergence equations [24,31,69] are used in the projection. In consequence, nonphysical solutions that do not fulfill the divergence equations can be obtained and they are called *spurious*. The best way of avoiding spurious solutions is to choose a formulation which is not able to generate them. However, it is not always possible (e.g. in the case of 3D inhomogeneous problems) and then some methods of their elimination are required. If this is the case, two general approaches have been reported. The first one is based on the modification of the analytical formulation, while the second one relies on the application of proper basis and trial functions in the projection method. An example of the former approach is *penalty method*, in which so called *penalty term* is added to the operator equation provided that the spurious solutions are shifted into the spectrum region which is out of interest. This technique is often employed in conjunction with the FEM [48,55,56,90,91] and the FDFD [1,13,84,101–103,125] methods. An example of application of the second technique are *edge elements* [7, 48, 117, 121], which are a kind of the FEM. In this method solution a solution vector is represented as a series of vector polynomial functions obeying locally Gauss law, thereby ensuring that the solution vector fulfills the divergence equations.

1.1.2 Hybrid methods of analysis

In spite of the great versatility of FEM and FDFD methods there are many structures which are difficult to analyze, e.g. structures filled with anisotropic materials such as ferrites or chiroferrites. In general, classical analysis of anisotropic structures requires formulations where additional field components are used. It leads to larger matrix problems and much longer computational times. As an alternative, hybrid methods, such as the coupled mode method [5, 43], can be used. They are based on expansion of the fields into a series of entire domain basis functions. In the case of the coupled mode method, basis functions are computed as the solutions for the simplified original problem (e.g. isotropic). The difference between the parameters of the original structure and the basis one is treated as a perturbation. Another, new class of hybrid algorithms are eigenfunction expansion algorithms [85–88,95], which, compared to the classical methods, can dramatically increase efficiency of computation of waveguide dispersion characteristics without deteriorating the accuracy. These algorithms are based on expansion of the fields into a series of eigenfunctions computed for certain frequency or propagation constant points. The advantage of such choice of basis functions is that they fulfill all inner and outer boundary conditions. In consequence, the number of the functions taken into expansion can be substantially reduced leading to the problems of very small size.

1.2 State of the knowledge in the methods improving efficiency of numerical solution of electromagnetic eigenvalue problems

1.2.1 Analysis of waveguides and resonators in Cartesian coordinates

The most versatile methods of electromagnetic field analysis include the FEM and FDFD methods (see Sec. 1.1.1). The selection of publications devoted to the FEM [2, 3, 7, 19, 21, 23, 24, 30–34, 38, 44, 48, 55, 56, 66–68, 89–92, 107, 114, 117, 120, 121] and the FDFD [1, 9, 11–14, 18, 25, 37, 40, 46, 52, 53, 64, 77, 82, 84, 94, 101–106, 112, 113, 119, 122–125, 127] numerical analysis methods is very wide. Depending on the analyzed structure type and properties of the filling media, various analytical formulations and projection methods are used, resulting in specific matrix problems. Next, appropriate numerical solvers have to be used depending on characteristic properties of the matrices.

First electromagnetic applications of the FEM [3,107] and the FDFD [11,12,25] methods concerned analysis of waveguides using scalar formulations in terms of longitudinal components of electric E_z or magnetic H_z field. These simple formulations could only be used to the analysis of homogeneous structures. Resulting symmetric eigenproblems were solved by some less efficient iterative methods, such as *successive overrelaxation* [11,12,25], requiring good starting estimation of the eigenvalue of interest and the corresponding eigenvector.

In order to analyze inhomogeneous waveguides a vectorial formulation in terms of E_z and H_z fields were used. It was used in conjunction with the FEM [2, 19, 21] and the FDFD [18, 46, 106] methods, resulting in symmetric eigenproblem. In the most cases, numerical solution methods were analogous to the ones used for the scalar case, but some FDFD codes [18, 106] reported application of *bisection* method for computing of eigenvalues and *inverse iteration* for determination of the corresponding eigenvectors. This approach was still far from efficient computations.

Since the formulation using only longitudinal field components could not be used for analysis of general anisotropic structures, other formulations based on full vector electric $\vec{E}$ or/and magnetic $\vec{H}$ fields were incorporated. In context of the FEM [7, 55, 56, 89–91, 117, 120, 121] and the FDFD [1, 9, 13, 84, 101–103, 122, 125] methods they resulted (except [1]) in symmetric eigenproblems. These problems were solved using standard bisection [7,122] and QR [1] methods. However, other methods such as subspace iteration [9,89–91,101–103,117,125] or Lanczos method [120] began to be used also. The subspace iteration was the first projection method[1] used for solution of electromagnetic waveguide and resonator eigenvalue problems. In contrast to previously used methods, the projection methods could compute a few eigenvalues/eigenvectors of interest at one run, making the computations much more efficient. The algorithm of Lanczos was another efficient projection method, intended for symmetric eigenproblems.

Concurrently to full vector formulations, the ones in terms of transverse electric $\vec{E}_t$ and/or magnetic $\vec{H}_t$ fields were used for homogeneous structures such as waveguides and some resonators. These formulations resulted in the eigenproblem that was not symmetric. This fact was the main drawback of the transverse formulations, because nonsymmetric eigenproblem is much harder to solve, compared to a symmetric one, and initially there was no efficient nonsymmetric eigensolvers developed. First and even many later implementations of the FEM [31, 32, 44] and FDFD [14, 104, 105, 123, 124] using transverse formulations, involved standard shifted power [123, 124], inverse iteration [44] or QR/QZ [14, 31, 32, 104, 105] algorithms. Incorporation of the subspace iteration into the FEM [30, 33, 34, 66–68] and FDFD [1, 102, 103] codes substantially improved efficiency of the analysis. Another more efficient iterative algorithm used in conjunction with the FDFD method [37] was the Arnoldi method — nonsymmetric version of the Lanczos method.

1.2.2 Analysis of cylindrically symmetric cavities

Many researchers were particularly interested in the analysis of cylindrically symmetric cavities. All formulations used in FEM [24, 38, 114] and FDFD [40, 64, 82, 94, 113] approaches took into account rotational symmetry of the structure. Assuming $e^{-jm\phi}$ field dependence (where m is the azimuthal mode index) 3D problems were reduced into the 2D case (in the r-z plane). In [24,114], full vector $\vec{H}$ formulations were used in the FEM,

[1] Projection used in this context is related to the projection of a finite dimensional space onto another one of smaller dimension. This kind of projection is described in Sec. A.1.4 in Appendix A.

resulting in generalized eigenproblems with symmetric matrices. Analogous formulation used in the FDFD [94] gave standard eigenproblem with nonsymmetric matrix. A full vector formulation [82] was also used for the analysis of the TE and TM modes (of azimuthal index $m = 0$), resulting in generalized eigenproblem with symmetric matrices. These particular modes can also be analyzed using the formulations for azimuthal H_ϕ or E_ϕ component only and some authors [38, 40, 64, 113] involved these more appropriate scalar formulations. In contrast to full vector formulations [24, 94, 114], for the analysis of hybrid modes (with $m \neq 0$), the formulations in terms of the radial r- and axial z-components were used in [40, 64, 113], resulting in standard eigenproblems with nonsymmetric matrices. Numerical solution of the eigenproblems resulting in the most of the discussed above approaches [24, 38, 40, 82, 113] were performed by means of the subspace iteration method and the QR method was reported only once [94].

1.2.3 Modern iterative eigenvalue solvers

Current literature shows a great interest in algorithms solving large and sparse matrix eigenproblems in the more efficient way than the QR/QZ methods. Among them are various variants of the subspace iteration or the Arnoldi/Lanczos methods. In most cases, the authors develop their own routines and algorithms, but some of them adapt already existing codes. In the case of proprietary codes, it is very difficult to compare the efficiency with other codes. The situation is different when it comes to existing library procedures such as the ARPACK implementation of the implicitly restarted Arnoldi method. Recent studies[2] [65] have revealed that the efficiency of this algorithm depends on the problem which is to be solved. In the majority of considered cases the efficiency is higher than that offered by other numerical methods. However, there are no informations concerning efficiency of this algorithm with respect to electromagnetic eigenproblems resulting from particular methods of conversion.

1.2.4 Spectral transformations

Spectral transformations can be used to improve performance of iterative eigensolvers. They are also required to obtain convergence to the eigenvalues of interest. The simplest approach employed in many cases is shifting. A disadvantage of such approach is relatively slow convergence (see App. A for discussion). Another technique reported by some authors [24, 30, 33, 34, 44, 66–68, 102, 117] is shift-invert, which is also used for accelerating convergence of numerical solvers. The main disadvantage is that the shift-invert technique requires the inversion of the matrix or its decomposition and subsequent solutions of a linear system. Acceleration methods which does not require matrix conversion are based on application of polynomial preconditioners such as Chebyshev or digital finite impulse response (FIR) filters. Incorporation of the Chebyshev preconditioning in the solution of electromagnetic eigenproblems has been recently reported by some authors using the

[2] Available via anonymous ftp from `ftp://info.mcs.anl.gov/pub/tech_reports/reports/P547.ps.Z`

subspace iteration method [37, 40, 101, 102, 113]. To the best our knowledge, FIR filters have not been used for this purpose yet.

1.2.5 Hybrid methods

Application of the hybrid methods such as coupled mode method is widely described in the literature [5, 43]. This method allows one to analyze structures filled with complex materials at the cost not much higher than the cost of computing basis functions. Recently elaborated methods, based on expansion of fields into a series of eigenfunctions, constitute a new direction of research. This approach can be viewed as a generalization of the concept of an electromagnetic basis optimization [51,57,58,81]. Among the eigenfunction expansion methods one may distinguish a method which uses eigenfunctions evaluated at the cutoff [95] and a method using the *asymptotic waveform evaluation* technique [85], which is based on expanding the fields into the Taylor series. Another kind of algorithms [86–88], in which fields are expanded into a series of eigenfunctions calculated for arbitrary frequencies or propagation constants, seems to be particularly attractive. These algorithms generate matrix eigenproblem of small size in a very simple way. The time needed to solve such problem can be neglected compared to the time necessary for calculation of basis functions. In consequence, based on the solutions in a few points, dispersion characteristics of a waveguide can be evaluated very fast. Application of these algorithms was so far very limited due to their innovative character.

1.3 Scope and goal of the thesis

The goal of this thesis is to develop the methods, which substantially accelerate numerical solution of various electromagnetic eigenvalue problems. This is realized by:

- choice of a proper analytical formulation,
- choice of an adequate conversion method resulting in eigenproblem that can be efficiently solved,
- application of modern iterative algorithms able to solve large sparse matrix eigenvalue problems at low cost (in terms of the memory and time),
- application of various preconditioners (shift-invert, Chebyshev polynomials, FIR digital filters) for accelerating the convergence of numerical eigensolvers,
- application of hybrid methods to the analysis of structures filled with complex materials and fast determination of dispersion characteristics of waveguides.

This thesis makes the claim that, the most efficient calculation of the modes in various classes of electromagnetic problems can be realized by the application of the Krylov space methods based on iterative computation. This technique makes efficient use of memory on parallel systems, and can be easily enhanced by the application of spectral

transformations. Another claim of this thesis is that the hybrid conversion methods can be very fast and accurate tool for the analysis of anisotropic structures and determination of dispersion characteristics of inhomogeneously loaded waveguides.

The original contributions of this thesis are the following:

- implementation of the Arnoldi method to the solution of dense matrix eigenvalue problems resulting from classical conversion methods such as Galerkin method, FEM and FDFD and performance comparison with other numerical solvers such as QR and subspace iteration,
- analysis of the performance of Chebyshev preconditioning implemented in the Arnoldi method applied to the solution of the eigenproblem arising in the FDFD analysis of microwave resonators,
- application of the finite impulse response (FIR) digital filters as preconditioners in electromagnetic eigenproblems where the eigenvalues from the middle of the spectrum are of interest,
- application of the coupled mode method to the efficient analysis of a waveguide partially filled with a ferrite,
- application of the novel eigenfunction expansion technique to fast determination of dispersion characteristics of dielectric and ferrite guides.

1.4 Chapter outline

In order to obtain a framework for the development of effective solution methods of microwave waveguide and resonator eigenproblems, classification of possible formulations in terms of the number of the involved field components and potential possibility of generation of spurious solutions is presented in Chapter 2. Next, various classical and hybrid methods of conversion of the operator problems into the matrix one are described in Chapter 3. Chapter 4 summarizes the most important methods of the eigenproblem solution, such as QR, QZ, subspace iteration or Arnoldi/Lanczos methods. Some typical examples of the analysis, involving various classical (Galerkin method, FEM, FDFD) and hybrid (coupled mode, eigenfunction expansion) conversion methods and various direct (QR) and iterative (subspace iteration, Arnoldi method) numerical solvers, showing the efficiency of the particular methods of analysis are discussed in Chapter 5. The tests include the application of various preconditioners (shift-invert, Chebyshev polynomials, digital FIR filters), applied in order to accelerate the convergence of the numerical methods. The results presenting the most effective numerical methods for the analysis of the microwave waveguides and resonators are summarized at the conclusions stage in Chapter 6. Due to lack of a comprehensive comparison of modern eigensolvers in the electromagnetic literature a short review of the numerical methods is presented in the Appendix A. Additionally, an experimental verification of the analysis results of a complex nonreciprocal ferrite phase shifter structure is described in Appendix B.

Chapter 2

Classification of electromagnetic eigenvalue problems

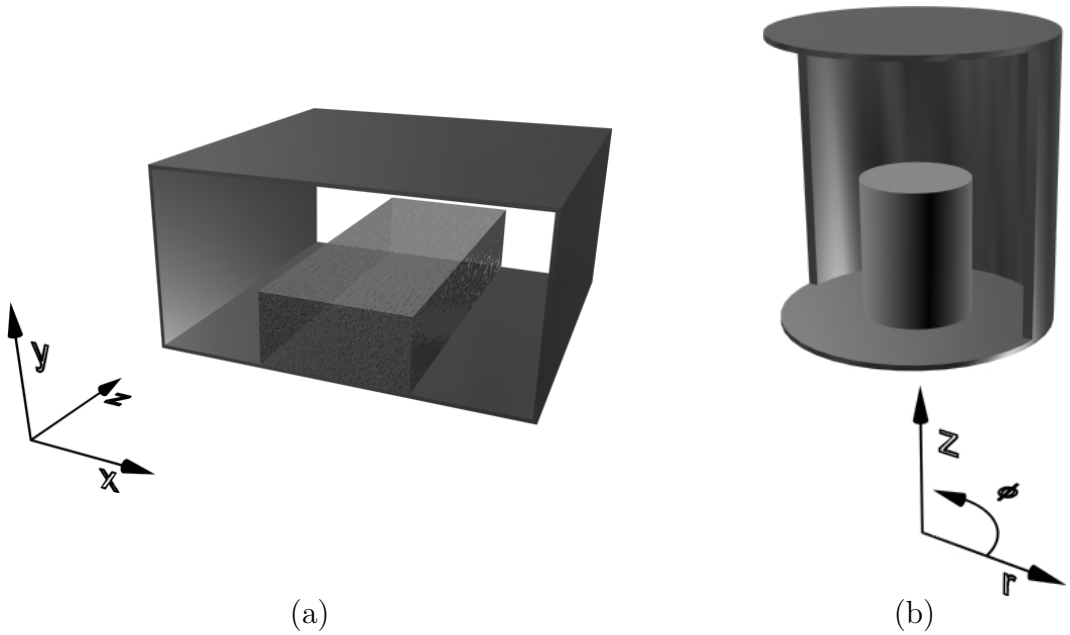
2.1 Definitions and assumptions

Electromagnetic problems can be divided into two main classes: *open* problems and *closed* problems. The former category concerns the analysis of electromagnetic waves excited and propagated in a free space, while the latter involves the problems related to analysis of electromagnetic waves in enclosed structures, very often bounded by perfect electric conductor (PEC) screens and perfect magnetic conductor (PMC) screens. The predominant approach to the solution of open problems is via an integral equation. For closed structures either differential or integral formulations are used. The differential formulation is more versatile when it comes to inhomogeneous media. Therefore in the analysis of waveguides and resonators we concentrate on differential formulations.

Depending on presence or absence of electric and/or magnetic sources in an analyzed region, we can distinguish two types of problems: *deterministic* and *eigenvalue* ones. The former problems rely on computation of electromagnetic fields being a response to some excitation, while the latter group is related to problems of electromagnetic field determination in a source-free regions. In spite of the fact that any physical electromagnetic field has deterministic nature, the synthesis and design of microwave components (e.g. transitions, filters, couplers, phase shifters, resonators, etc.) can be realized by solving the eigenvalue problem. We confine our succeeding discussion to this type of problems.

Formulations for waveguide and resonator structures can also be classified according to the type of functions describing electromagnetic field. Two main classes of the formulations are those based on field components and the ones involving appropriately chosen scalar and/or vector potential functions. These functions are constructed so that they can describe all fields in a particular coordinate system. Both classes of formulations mentioned above are fully equivalent and only the former one, i.e. involving fields rather than potentials, will be further discussed. It should be noted, however, that in some particular cases (pointed out in the text), potential formulations can be advantageous and lead to simpler problems.

A majority of resonator and waveguide structures used in practice conform well to the Cartesian or cylindrical coordinates. Accordingly, we limit the details of the discussion to these two coordinate systems. Moreover, we assume that all considered structures are bounded by PEC or PMC walls. The structures where the media parameters are not functions of position are called *homogeneous*. There are many practical structures that are homogeneous only in one particular direction (see Fig. 2.1). The directions perpendicular to this direction are called *transverse*. We will consider only the waveguide structures that are homogeneous in the propagation direction (called also *longitudinal*). It is moreover assumed that the wave in the guide propagates always along the z-axis. In the steady state, variation of the fields in any homogeneous direction can be described using propagation constant β or azimuthal mode index m. This fact is used to reduce the order of many formulations, as shown later on in this chapter.



(a) (b)

Figure 2.1: Examples of homogeneous structures: (a) in the z-direction in the Cartesian coordinate system (a waveguide) and (b) in the ϕ-direction in the cylindrical coordinate system (a rotationally symmetric resonator).

The formulations will be called *symmetric* or *self adjoint* when the operators that are involved in them are symmetric with respect to the standard inner product[1] defined in a Hilbert space. For any operator $\mathbf{A}$, the symmetry implies that

$$\Big(Au, v\Big) = \Big(u, Av\Big) \tag{2.1}$$

Some vectorial resonator and waveguide formulations can generate numerical solutions which do not fulfill all Maxwell's equations (in particular the divergence ones). These

[1]We define the standard inner product as $\Big(u, v\Big) = \int_\Omega uv^* \, d\Omega$

solutions are called *spurious* or *nonphysical*, while the formulations which are able to generate them we will call *potentially spurious*. The formulations which are not potentially spurious will be called *spurious-free*.

We are interested in the simplest and the most effective approaches so we do not take into account formulations leading to nonstandard eigenproblems. Only the simplest resonator and waveguide formulations, i.e. standard and generalized, will be presented below preceded by an introduction covering basic electromagnetic concepts.

2.2 Electromagnetic background

Two assumptions are essential for eigenvalue problems. The first one is the absence of electromagnetic sources, while the second one is the steady state of electromagnetic fields. This latter feature implies the time harmonic $e^{j\omega t}$ dependence of the fields, where ω is angular frequency. The following Maxwell's equations describe the stationary fields in a source-free homogeneous region

$$\nabla \times \vec{E} = -j\omega\vec{B} \tag{2.2}$$
$$\nabla \times \vec{H} = j\omega\vec{D} \tag{2.3}$$
$$\nabla \cdot \vec{D} = 0 \tag{2.4}$$
$$\nabla \cdot \vec{B} = 0 \tag{2.5}$$

where $\vec{E}$ and $\vec{H}$ denote respectively the electric and magnetic field intensities, while $\vec{D}$ and $\vec{B}$ are electric and magnetic flux densities. Equations (2.2) and (2.3) are usually called *curl equations* while (2.4) and (2.5) are *divergence equations*.

The relations between flux densities and field intensities are referred to as *material equations* (2.6) and (2.7)

$$\vec{D} = \underline{\underline{\epsilon}} \cdot \vec{E} \tag{2.6}$$
$$\vec{B} = \underline{\underline{\mu}} \cdot \vec{H} \tag{2.7}$$

where $\underline{\underline{\epsilon}}$ and $\underline{\underline{\mu}}$ denote tensors of electric permittivity and magnetic permeability. In the most general form, the tensors are expressed by

$$\underline{\underline{\epsilon}} = \begin{bmatrix} \epsilon_{11} & \epsilon_{12} & \epsilon_{13} \\ \epsilon_{21} & \epsilon_{22} & \epsilon_{23} \\ \epsilon_{31} & \epsilon_{32} & \epsilon_{33} \end{bmatrix} \quad , \quad \underline{\underline{\mu}} = \begin{bmatrix} \mu_{11} & \mu_{12} & \mu_{13} \\ \mu_{21} & \mu_{22} & \mu_{23} \\ \mu_{31} & \mu_{32} & \mu_{33} \end{bmatrix} \tag{2.8}$$

where subscripts $\{1, 2, 3\}$ correspond to the coordinates in a particular coordinate system (e.g. $\{x, y, z\}$ in the Cartesian one or $\{r, \phi, z\}$ in the cylindrical one). The tensors describing lossless materials are Hermitian [79], i.e.

$$\underline{\underline{\epsilon}} = \underline{\underline{\epsilon}}^{H} \quad , \quad \underline{\underline{\mu}} = \underline{\underline{\mu}}^{H} \tag{2.9}$$

On the boundary between different media, the following equations for continuity of the fields can be derived [6] from equations (2.2–2.5)

$$\hat{n} \times (\vec{E}_2 - \vec{E}_1) = 0 \tag{2.10}$$
$$\hat{n} \times (\vec{H}_2 - \vec{H}_1) = 0 \tag{2.11}$$
$$\hat{n} \cdot (\vec{D}_2 - \vec{D}_1) = 0 \tag{2.12}$$
$$\hat{n} \cdot (\vec{B}_2 - \vec{B}_1) = 0 \tag{2.13}$$

where $\hat{n}$ is a unit vector normal to the boundary. Equations (2.10) and (2.11) express continuity of the tangential components of $\vec{E}$ and $\vec{H}$, while (2.12) and (2.13) express continuity of the normal components of $\vec{D}$ and $\vec{B}$.

In the vicinity of PEC screen, the boundary conditions take the form [79]

$$\left.\begin{array}{r} \nabla \cdot (\underline{\underline{\epsilon}} \cdot \vec{E}) \\ \hat{n} \times \underline{\underline{\epsilon}}^{-1} \cdot \nabla \times \vec{H} \\ \hat{n} \cdot (\underline{\underline{\mu}} \cdot \vec{H}) \end{array}\right\} = 0 \tag{2.14}$$

$$\hat{n} \times \vec{E} \quad \text{or}$$

while the conditions for PMC boundary are

$$\left.\begin{array}{r} \nabla \cdot (\underline{\underline{\mu}} \cdot \vec{H}) \\ \hat{n} \times \underline{\underline{\mu}}^{-1} \cdot \nabla \times \vec{E} \\ \hat{n} \cdot (\underline{\underline{\epsilon}} \cdot \vec{E}) \end{array}\right\} = 0 \tag{2.15}$$

$$\hat{n} \times \vec{H} \quad \text{or}$$

2.3 Resonator problems

The problem of resonator analysis consists in calculating a set of frequencies ω and corresponding electromagnetic fields, satisfying Maxwell's equations and boundary conditions for a given resonator structure. Depending on the resonator shape and media properties various three-dimensional (3D) formulations derived from Maxwell's equations are possible. In the case of homogeneous structure, a 3D problem can be transformed into an equivalent two-dimensional (2D) or even one-dimensional (scalar) problem. All these classes of formulations are discussed in the following sections.

2.3.1 Full vector formulations

The most straightforward full vector formulation can be derived directly from (2.2) and (2.3) by respectively multiplying them with j and $-j$. It results in the following standard eigenproblem with ω being the eigenvalue

$$\begin{bmatrix} 0 & -j\nabla\times \\ j\nabla\times & 0 \end{bmatrix} \begin{bmatrix} \underline{\underline{\epsilon}}^{-1} & 0 \\ 0 & \underline{\underline{\mu}}^{-1} \end{bmatrix} \begin{bmatrix} \vec{D} \\ \vec{B} \end{bmatrix} = \omega \begin{bmatrix} \vec{D} \\ \vec{B} \end{bmatrix} \tag{2.16}$$

Using the methodology described in [79], it can easily be shown that this formulation is not symmetric, because the operator on the left side is not symmetric (with respect to the standard inner product defined by (2.1)).

In order to obtain a symmetric formulation, two solutions are possible. The first one is to express (2.16) in terms of $\vec{E}$ and $\vec{H}$ fields, what results in the following generalized eigenproblem

$$\begin{bmatrix} 0 & -j\nabla\times \\ j\nabla\times & 0 \end{bmatrix} \begin{bmatrix} \vec{E} \\ \vec{H} \end{bmatrix} = \omega \begin{bmatrix} \underline{\underline{\epsilon}} & 0 \\ 0 & \underline{\underline{\mu}} \end{bmatrix} \begin{bmatrix} \vec{E} \\ \vec{H} \end{bmatrix} \tag{2.17}$$

It can easily be shown, that this problem is symmetric for lossless structures bounded by any combination of PEC and PMC screens, because both operators, on the left and on the right side, are symmetric.

The second solution is to multiply both sides of (2.16) with matrix $\mathbf{M}$, defined as

$$\mathbf{M} = \begin{bmatrix} 0 & -\hat{z}\times \\ \hat{z}\times & 0 \end{bmatrix} \tag{2.18}$$

This leads to another generalized eigenproblem in the form of

$$\begin{bmatrix} -j\hat{z}\times\nabla\times & 0 \\ 0 & -j\hat{z}\times\nabla\times \end{bmatrix} \begin{bmatrix} \underline{\underline{\epsilon}}^{-1} & 0 \\ 0 & \underline{\underline{\mu}}^{-1} \end{bmatrix} \begin{bmatrix} \vec{D} \\ \vec{B} \end{bmatrix} = \omega \begin{bmatrix} 0 & -\hat{z}\times \\ \hat{z}\times & 0 \end{bmatrix} \begin{bmatrix} \vec{D} \\ \vec{B} \end{bmatrix} \tag{2.19}$$

It was shown in [79], that under the same conditions as above eigenproblem (2.19) is symmetric.

The main disadvantage of formulations (2.16), (2.17), and (2.19) is that they result in the problems with complex operators involving six field components and therefore their numerical solution may be very costly. In order to alleviate these drawbacks we can eliminate $\vec{H}$ or $\vec{E}$ from Maxwell's equations getting the formulations for $\vec{D}$ or $\vec{B}$ fluxes only

$$\nabla\times\underline{\underline{\mu}}^{-1}\cdot\nabla\times\underline{\underline{\epsilon}}^{-1}\cdot\vec{D} = \omega^2\vec{D} \tag{2.20}$$

$$\nabla\times\underline{\underline{\epsilon}}^{-1}\cdot\nabla\times\underline{\underline{\mu}}^{-1}\cdot\vec{B} = \omega^2\vec{B} \tag{2.21}$$

Formulations (2.20) and (2.21) are called *vector wave equations* or *curl-curl equations* and involve only three field components. They can result in real eigenproblems (for real $\underline{\underline{\epsilon}}$ and $\underline{\underline{\mu}}$). However, they are not symmetric if written in this form.

Symmetry can be obtained after rewriting (2.20) and (2.21) in terms of $\vec{E}$ and $\vec{H}$ fields, as generalized eigenproblems

$$\nabla\times\underline{\underline{\mu}}^{-1}\cdot\nabla\times\vec{E} = \omega^2\underline{\underline{\epsilon}}\cdot\vec{E} \tag{2.22}$$

$$\nabla\times\underline{\underline{\epsilon}}^{-1}\cdot\nabla\times\vec{H} = \omega^2\underline{\underline{\mu}}\cdot\vec{H} \tag{2.23}$$

These problems become symmetric for lossless structures [48].

All full vector formulations presented in this section are very general and applicable to either anisotropic or inhomogeneous structures. However, they are potentially spurious, because their derivation does not incorporate the divergence equations. In order to alleviate this problem the penalty method [1, 13, 48, 55, 56, 84, 90, 91, 101–103, 125] can be applied.

2.3.2 Homogeneous resonators

If the structure is isotropic and homogeneous then vector wave equations, derived in the previous section, can be decoupled, leading to scalar wave equation of the Helmholtz type [6]

$$\nabla^2\Phi = -k^2\Phi \tag{2.24}$$

where $k^2 = \omega^2\epsilon\mu$ and Φ is some scalar function. In the Cartesian coordinates, Φ can be associated with any component of $\vec{E}$, $\vec{H}$, $\vec{D}$ or $\vec{B}$, while in the cylindrical coordinates, such a formulation is possible only for z-components of these fields. It should be noted, that in the spherical coordinate system, scalar formulation (2.24) is possible only for appropriately chosen scalar potential function Φ (see [6] for the detailed discussion).

Formulations of type (2.24) are very desirable because they can be solved at relatively low cost (only one component is involved and the eigenproblem is real and symmetric for lossless media). In practice, the structures homogeneous only in the particular direction are much more common. In this case, the corresponding eigenproblems can also be substantially simplified. Consider a few the most practical cases.

2.3.2.1 z-direction homogeneity

For the resonator homogeneous in the z-direction, variation of electromagnetic fields along the z-axis is given by the $e^{-j\beta z}$ factor. Propagation constant β along the z-axis is expressed by

$$\beta = \frac{p\pi}{c} \tag{2.25}$$

where c is the length of the resonator along the z-axis and $p = 0, 1, 2, \ldots$ is the mode index in the z-direction. Therefore, for any particular value of β the analysis of the 3D resonator can be simplified to the form of a 2D problem, provided that any waveguide ω-formulation described in Sec. 2.4 can be used.

2.3.2.2 x- or y-direction homogeneity (Cartesian coordinates)

In the case of homogeneity in the x- or y-direction in the Cartesian coordinate system, the formulations for the z-direction homogeneity can be used after a simple transformation of coordinates.

An alternative approach is to derive analogous formulations to the ones described in Sec. 2.4 for the x or y being longitudinal direction.

2.3.2.3 Cylindrical symmetry

In the structure with cylindrical symmetry, i.e. homogeneous in the azimuthal ϕ-direction, the field dependence is described by the term $e^{-jm\phi}$, where $m \in \{0, \pm1, \pm2, \ldots\}$ is the azimuthal mode index. Any field $\vec{A}$ can be decomposed into the transverse $\vec{A}_t$ and the azimuthal A_ϕ parts as follows

$$\vec{A} = \vec{A}_t + \hat{\phi}A_\phi \qquad , \qquad \vec{A}_t = \hat{r}A_r + \hat{z}A_z \tag{2.26}$$

The gradient operator $\nabla(\cdot)$ can also be split so that

$$\nabla = \nabla_t - j\frac{m}{r}\hat{\phi} \qquad , \qquad \nabla_t = \hat{r}\frac{\partial}{\partial r} + \hat{z}\frac{\partial}{\partial z} \tag{2.27}$$

Similarly, the divergence operator $\nabla \cdot (\cdot)$ splits up in the following way

$$\nabla \cdot \vec{A} = \nabla_t \cdot \vec{A}_t - j\frac{m}{r}A_\phi \qquad , \qquad \nabla_t \cdot \vec{A}_t = \frac{1}{r}\frac{\partial}{\partial r}(rA_r) + \frac{\partial}{\partial z}A_z \tag{2.28}$$

The rotation operator $\nabla \times (\cdot)$ can also be decomposed as follows

$$\nabla \times \vec{A} = \nabla_t \times \vec{A} - j\frac{m}{r}\hat{\phi} \times \vec{A} \qquad , \qquad \nabla_t \times \vec{A} = \nabla \times \vec{A}\Big|_{\frac{\partial}{\partial\phi}\vec{A}=0} = \nabla \times \vec{A}\Big|_{m=0} \tag{2.29}$$

Writing it in the matrix format we get

$$\begin{aligned} \nabla \times \vec{A} &= \begin{bmatrix} -j\frac{m}{r}\hat{\phi}\times & \nabla_t \times \hat{\phi} \\ \hat{\phi}\nabla_t \times & 0 \end{bmatrix} \begin{bmatrix} \vec{A}_t \\ A_\phi \end{bmatrix} \\ &= \begin{bmatrix} -j\frac{m}{r}\hat{\phi} \times (\cdot) & -\hat{\phi} \times \frac{1}{r}\nabla_t(r(\cdot)) \\ -r\nabla_t \cdot \frac{1}{r}\hat{\phi} \times (\cdot) & 0 \end{bmatrix} \begin{bmatrix} \vec{A}_t \\ A_\phi \end{bmatrix} \end{aligned} \tag{2.30}$$

Assuming that material tensors $\underline{\underline{\epsilon}}$ and $\underline{\underline{\mu}}$ have the form

$$\underline{\underline{\epsilon}} = \begin{bmatrix} \underline{\underline{\epsilon}}_{tt} & 0 \\ 0 & \epsilon_{\phi\phi} \end{bmatrix} \qquad , \qquad \underline{\underline{\mu}} = \begin{bmatrix} \underline{\underline{\mu}}_{tt} & 0 \\ 0 & \mu_{\phi\phi} \end{bmatrix} \tag{2.31}$$

one can formulate the eigenproblem in terms of the transverse fields only. The derivation is based on the decompositions (2.26), (2.30) and (2.31) applied to eigenproblem (2.20), which results in

$$\begin{aligned} &\begin{bmatrix} -j\frac{m}{r}\hat{\phi}\times & \nabla_t \times \hat{\phi} \\ \hat{\phi}\nabla_t \times & 0 \end{bmatrix} \begin{bmatrix} \underline{\underline{\mu}}_{tt}^{-1} & 0 \\ 0 & \mu_{\phi\phi}^{-1} \end{bmatrix} \begin{bmatrix} -j\frac{m}{r}\hat{\phi}\times & \nabla_t \times \hat{\phi} \\ \hat{\phi}\nabla_t \times & 0 \end{bmatrix} \begin{bmatrix} \underline{\underline{\epsilon}}_{tt}^{-1} & 0 \\ 0 & \epsilon_{\phi\phi}^{-1} \end{bmatrix} \begin{bmatrix} \vec{D}_t \\ D_\phi \end{bmatrix} \\ &\qquad = \omega^2 \begin{bmatrix} \vec{D}_t \\ D_\phi \end{bmatrix} \end{aligned} \tag{2.32}$$

The equation describing transverse part is then

$$\begin{aligned} -\frac{m^2}{r^2}\hat{\phi} \times \underline{\underline{\mu}}_{tt}^{-1} \cdot \hat{\phi} \times \underline{\underline{\epsilon}}_{tt}^{-1} \cdot \vec{D}_t + \nabla_t \times \mu_{\phi\phi}^{-1}\nabla_t \times \underline{\underline{\epsilon}}_{tt}^{-1} \cdot \vec{D}_t & \\ -j\frac{m}{r}\hat{\phi} \times \underline{\underline{\mu}}_{tt}^{-1} \cdot \nabla_t \times \hat{\phi}\epsilon_{\phi\phi}^{-1}D_\phi &= \omega^2\vec{D}_t \end{aligned} \tag{2.33}$$

Using (2.28), divergence equation (2.4) can take the form

$$D_\phi = -j\frac{r}{m}\nabla_t \cdot \vec{D}_t \tag{2.34}$$

Substituting (2.34) into (2.33) we eliminate D_ϕ and get the final formulation for transverse electric flux density

$$\begin{aligned} \Bigg[-\frac{m^2}{r^2}\hat{\phi} \times \underline{\underline{\mu}}_{tt}^{-1} \cdot \hat{\phi} \times \underline{\underline{\epsilon}}_{tt}^{-1} \cdot (\cdot) + \nabla_t \times \mu_{\phi\phi}^{-1}\nabla_t \times \underline{\underline{\epsilon}}_{tt}^{-1} \cdot (\cdot) & \\ -\frac{1}{r}\hat{\phi} \times \underline{\underline{\mu}}_{tt}^{-1} \cdot \nabla_t \times \hat{\phi}\epsilon_{\phi\phi}^{-1}r\nabla_t \cdot (\cdot)\Bigg]\vec{D}_t &= \omega^2\vec{D}_t \end{aligned} \tag{2.35}$$

Starting from eigenproblem (2.21) and using divergence equation (2.5) similar eigenproblem for transverse magnetic flux density $\vec{B}_t$ can be derived

$$\left[-\frac{m^2}{r^2}\hat{\phi}\times\underline{\underline{\epsilon}}_{tt}^{-1}\cdot\hat{\phi}\times\underline{\underline{\mu}}_{tt}^{-1}\cdot(\cdot)+\nabla_t\times\epsilon_{\phi\phi}^{-1}\nabla_t\times\underline{\underline{\mu}}_{tt}^{-1}\cdot(\cdot)\right.$$
$$\left.-\frac{1}{r}\hat{\phi}\times\underline{\underline{\epsilon}}_{tt}^{-1}\cdot\nabla_t\times\hat{\phi}\mu_{\phi\phi}^{-1}r\nabla_t\cdot(\cdot)\right]\vec{B}_t = \omega^2\vec{B}_t \tag{2.36}$$

For azimuthally invariant modes ($m = 0$) eigenproblems (2.35) and (2.36) can be substantially simplified. However, in this special case, alternative scalar formulations are also possible. They can be derived starting from Maxwell's equations (2.2) and (2.3), which can be split using (2.30) into the following equations

$$\nabla_t\times\hat{\phi}\epsilon_{\phi\phi}^{-1}D_\phi = -j\omega\vec{B}_t \tag{2.37}$$
$$\hat{\phi}\nabla_t\times\underline{\underline{\epsilon}}_{tt}^{-1}\cdot\vec{D}_t = -j\omega B_\phi \tag{2.38}$$

and

$$\nabla_t\times\hat{\phi}\mu_{\phi\phi}^{-1}B_\phi = j\omega\vec{D}_t \tag{2.39}$$
$$\hat{\phi}\nabla_t\times\underline{\underline{\mu}}_{tt}^{-1}\cdot\vec{B}_t = j\omega D_\phi \tag{2.40}$$

Isolating $\vec{B}_t$ from (2.37) and substituting it into (2.40) we get the following standard scalar eigenproblem for D_ϕ (TE modes)

$$\hat{\phi}\nabla_t\times\underline{\underline{\mu}}_{tt}^{-1}\cdot\nabla_t\times\hat{\phi}\epsilon_{\phi\phi}^{-1}D_\phi = \omega^2 D_\phi \tag{2.41}$$

Analogously, separating $\vec{D}_t$ from (2.39) and replacing it in (2.38) we get a standard scalar eigenproblem for B_ϕ (TM modes)

$$\hat{\phi}\nabla_t\times\underline{\underline{\epsilon}}_{tt}^{-1}\cdot\nabla_t\times\hat{\phi}\mu_{\phi\phi}^{-1}B_\phi = \omega^2 B_\phi \tag{2.42}$$

It is worth noting, that formulations (2.35) and (2.36) are spurious-free because the divergence equation was used in their derivation, while formulations (2.41) and (2.42) are spurious-free because they are scalar.

2.4 Waveguide problems

Analysis of waveguides consists in computing a set of pairs $\{\omega, \beta\}$ and corresponding electromagnetic fields, which satisfy Maxwell's equations and boundary conditions for an analyzed structure. Two classes of formulations can be distinguished, *β-formulations* and *ω-formulations* [79]. In the former case, the eigenvalue is a function of propagation constant β and angular frequency ω is a parameter. In the latter case, a function of ω is the eigenvalue, while β is the parameter.

As has been assumed, the waveguide is uniform in the longitudinal direction z and thus the directions perpendicular to z are transverse directions. The wave propagating

in the $+z$ direction is governed by the factor $e^{-j\beta z}$. Any field $\vec{A}$ can now be decomposed into the transverse $\vec{A}_t$ and the longitudinal A_z parts, so that

$$\vec{A} = \vec{A}_t + \hat{z}A_z \tag{2.43}$$

where

$$\vec{A}_t = \hat{x}A_x + \hat{y}A_y \qquad \text{or} \qquad \vec{A}_t = \hat{r}A_r + \hat{\phi}A_\phi \tag{2.44}$$

in the Cartesian or cylindrical coordinates respectively. The nabla operator ∇ can be split so that

$$\nabla = \nabla_t - j\beta\hat{z} \tag{2.45}$$

where

$$\nabla_t = \hat{x}\frac{\partial}{\partial x} + \hat{y}\frac{\partial}{\partial y} \qquad \text{or} \qquad \nabla_t = \hat{r}\frac{\partial}{\partial r} + \hat{\phi}\frac{1}{r}\frac{\partial}{\partial \phi} \tag{2.46}$$

in the case of Cartesian or cylindrical coordinates respectively. The divergence operator $\nabla \cdot (\cdot)$ is then

$$\nabla \cdot \vec{A} = \nabla_t \cdot \vec{A}_t - j\beta A_z \tag{2.47}$$

where

$$\nabla_t \cdot \vec{A}_t = \frac{\partial}{\partial x}A_x + \frac{\partial}{\partial y}A_y \qquad \text{or} \qquad \nabla_t \cdot \vec{A}_t = \frac{1}{r}\frac{\partial}{\partial r}(rA_r) + \frac{1}{r}\frac{\partial}{\partial \phi}A_\phi \tag{2.48}$$

in the Cartesian or cylindrical coordinate system. Similarly, the rotation operator $\nabla \times (\cdot)$ can be written as

$$\nabla \times \vec{A} = \nabla_t \times \vec{A} - j\beta\hat{z} \times \vec{A} \qquad , \qquad \nabla_t \times \vec{A} = \nabla \times \vec{A}\Big|_{\frac{\partial}{\partial z}\vec{A}=0} = \nabla \times \vec{A}\Big|_{\beta=0} \tag{2.49}$$

and writing it in the matrix format we get

$$\begin{aligned} \nabla \times \vec{A} &= \begin{bmatrix} -j\beta\hat{z}\times & \nabla_t \times \hat{z} \\ \hat{z}\nabla_t \times & 0 \end{bmatrix} \begin{bmatrix} \vec{A}_t \\ A_z \end{bmatrix} \\ &= \begin{bmatrix} -j\beta\hat{z}\times & -\hat{z} \times \nabla_t \\ -\nabla_t \cdot \hat{z}\times & 0 \end{bmatrix} \begin{bmatrix} \vec{A}_t \\ A_z \end{bmatrix} \end{aligned} \tag{2.50}$$

Material tensors $\underline{\underline{\epsilon}}$ and $\underline{\underline{\mu}}$ are decomposed as follows

$$\underline{\underline{\epsilon}} = \begin{bmatrix} \underline{\underline{\epsilon}}_{tt} & \underline{\epsilon}_{tz} \\ \underline{\epsilon}_{zt} & \epsilon_{zz} \end{bmatrix} \qquad , \qquad \underline{\underline{\mu}} = \begin{bmatrix} \underline{\underline{\mu}}_{tt} & \underline{\mu}_{tz} \\ \underline{\mu}_{zt} & \mu_{zz} \end{bmatrix} \tag{2.51}$$

To simplify further notations, the following symbols are additionally defined

$$\underline{\underline{\kappa}} \triangleq \underline{\underline{\epsilon}}^{-1} = \begin{bmatrix} \underline{\underline{\kappa}}_{tt} & \underline{\kappa}_{tz} \\ \underline{\kappa}_{zt} & \kappa_{zz} \end{bmatrix} \qquad , \qquad \underline{\underline{\nu}} \triangleq \underline{\underline{\mu}}^{-1} = \begin{bmatrix} \underline{\underline{\nu}}_{tt} & \underline{\nu}_{tz} \\ \underline{\nu}_{zt} & \nu_{zz} \end{bmatrix} \tag{2.52}$$

2.4.1 Full vector formulations

Applying decomposition (2.50) to Maxwell's curl equations one obtains

$$\nabla_t \times \vec{E} - j\beta\hat{z} \times \vec{E} = -j\omega\underline{\underline{\mu}} \cdot \vec{H} \tag{2.53}$$

$$\nabla_t \times \vec{H} - j\beta\hat{z} \times \vec{H} = j\omega\underline{\underline{\epsilon}} \cdot \vec{E} \tag{2.54}$$

Multiplying (2.53) and (2.54) by $-j$ and j and rearranging the terms, the following generalized β-eigenvalue problem can be derived

$$\begin{bmatrix} \omega\underline{\underline{\epsilon}} & j\nabla_t\times \\ -j\nabla_t\times & \omega\underline{\underline{\mu}} \end{bmatrix} \begin{bmatrix} \vec{E} \\ \vec{H} \end{bmatrix} = \beta \begin{bmatrix} 0 & -\hat{z}\times \\ \hat{z}\times & 0 \end{bmatrix} \begin{bmatrix} \vec{E} \\ \vec{H} \end{bmatrix} \tag{2.55}$$

It was shown in [79], that this problem is symmetric for lossless structures. It can be transformed into a nonsymmetric standard eigenproblem if both sides of (2.55) are multiplied by the operator $\mathbf{M}$ defined with (2.18).

A corresponding ω-formulations can be written after the operator decomposition (2.50) applied to (2.16), (2.17), and (2.19). We obtain then

$$\begin{bmatrix} 0 & -j\nabla_t \times (\cdot) - \beta\hat{z} \times (\cdot) \\ j\nabla_t \times (\cdot) + \beta\hat{z} \times (\cdot) & 0 \end{bmatrix} \begin{bmatrix} \underline{\underline{\epsilon}}^{-1} & 0 \\ 0 & \underline{\underline{\mu}}^{-1} \end{bmatrix} \begin{bmatrix} \vec{D} \\ \vec{B} \end{bmatrix} = \omega \begin{bmatrix} \vec{D} \\ \vec{B} \end{bmatrix} \tag{2.56}$$

$$\begin{bmatrix} 0 & -j\nabla_t \times (\cdot) - \beta\hat{z} \times (\cdot) \\ j\nabla_t \times (\cdot) + \beta\hat{z} \times (\cdot) & 0 \end{bmatrix} \begin{bmatrix} \vec{E} \\ \vec{H} \end{bmatrix} = \omega \begin{bmatrix} \underline{\underline{\epsilon}} & 0 \\ 0 & \underline{\underline{\mu}} \end{bmatrix} \begin{bmatrix} \vec{E} \\ \vec{H} \end{bmatrix} \tag{2.57}$$

and

$$\begin{bmatrix} -j\hat{z} \times \nabla_t \times (\cdot) - \beta\hat{z} \times \hat{z} \times (\cdot) & 0 \\ 0 & -j\hat{z} \times \nabla_t \times (\cdot) - \beta\hat{z} \times \hat{z} \times (\cdot) \end{bmatrix} \cdot \begin{bmatrix} \underline{\underline{\epsilon}}^{-1} & 0 \\ 0 & \underline{\underline{\mu}}^{-1} \end{bmatrix} \begin{bmatrix} \vec{D} \\ \vec{B} \end{bmatrix} = \omega \begin{bmatrix} 0 & -\hat{z}\times \\ \hat{z}\times & 0 \end{bmatrix} \begin{bmatrix} \vec{D} \\ \vec{B} \end{bmatrix} \tag{2.58}$$

For lossless structures and real β formulations (2.57) and (2.58) are both symmetric [79].

All formulations given above involve six field components and result in complex eigenproblems. The reduced size formulations, involving only one full vector field can be derived using (2.20) and (2.21). Splitting the fields and operators into the transverse and longitudinal parts (with equations (2.43), (2.50) and (2.52)) we get the following nonsymmetric standard eigenvalue problems with ω^2 being an eigenvalue. For $\vec{D}$ field we have

$$\begin{bmatrix} -j\beta\hat{z}\times & -\hat{z} \times \nabla_t \\ -\nabla_t \cdot \hat{z}\times & 0 \end{bmatrix} \begin{bmatrix} \underline{\underline{\nu}}_{tt} & \underline{\nu}_{tz} \\ \underline{\nu}_{zt} & \nu_{zz} \end{bmatrix} \begin{bmatrix} -j\beta\hat{z}\times & -\hat{z} \times \nabla_t \\ -\nabla_t \cdot \hat{z}\times & 0 \end{bmatrix} \begin{bmatrix} \underline{\underline{\kappa}}_{tt} & \underline{\kappa}_{tz} \\ \underline{\kappa}_{zt} & \kappa_{zz} \end{bmatrix} \begin{bmatrix} \vec{D}_t \\ D_z \end{bmatrix} = \omega^2 \begin{bmatrix} \vec{D}_t \\ D_z \end{bmatrix} \tag{2.59}$$

and for $\vec{B}$ field

$$\left[\begin{array}{cc} -j\beta\hat{z}\times & -\hat{z}\times\nabla_t \\ -\nabla_t\cdot\hat{z}\times & 0 \end{array}\right]\left[\begin{array}{cc} \underline{\underline{\kappa}}_{tt} & \underline{\kappa}_{tz} \\ \underline{\kappa}_{zt} & \kappa_{zz} \end{array}\right]\left[\begin{array}{cc} -j\beta\hat{z}\times & -\hat{z}\times\nabla_t \\ -\nabla_t\cdot\hat{z}\times & 0 \end{array}\right]\left[\begin{array}{cc} \underline{\underline{\nu}}_{tt} & \underline{\nu}_{tz} \\ \underline{\nu}_{zt} & \nu_{zz} \end{array}\right]\left[\begin{array}{c} \vec{B}_t \\ B_z \end{array}\right] = \omega^2\left[\begin{array}{c} \vec{B}_t \\ B_z \end{array}\right] \quad (2.60)$$

Analogous symmetric formulations (for lossless media and real β) can be derived from (2.22) and (2.23).

The main disadvantage of all full vector waveguide formulations discussed in this section is that they result in eigenproblems with complex operators involving three or six fields components, provided that their numerical solution can be expensive. Moreover, analogously to the full vector resonator formulations discussed in Sec. 2.3.1, the waveguide ones are also potentially spurious.

2.4.2 Transverse component formulations

Applying decomposition (2.43) to equations (2.53) and (2.54) and extracting the transverse and longitudinal parts one can eliminate z-component of the fields. In consequence, the following β-formulation for transverse fields $\vec{E}_t$ and $\vec{H}_t$ is obtained in the form of a complex generalized eigenproblem

$$\left[\begin{array}{cc} \mathbf{A}_{\mathbf{T}ee} & \mathbf{A}_{\mathbf{T}eh} \\ \mathbf{A}_{\mathbf{T}he} & \mathbf{A}_{\mathbf{T}hh} \end{array}\right]\left[\begin{array}{c} \vec{E}_t \\ \vec{H}_t \end{array}\right] = \beta\left[\begin{array}{cc} 0 & -\hat{z}\times \\ \hat{z}\times & 0 \end{array}\right]\left[\begin{array}{c} \vec{E}_t \\ \vec{H}_t \end{array}\right] \quad (2.61)$$

where

$$\begin{aligned} \mathbf{A}_{\mathbf{T}ee} &= \omega\underline{\underline{\epsilon}}_{tt}\cdot(\cdot) - \frac{1}{\omega}\nabla_t\times\mu_{zz}^{-1}\nabla_t\times(\cdot) - \omega\underline{\epsilon}_{tz}\cdot\underline{\epsilon}_{zt}\epsilon_{zz}^{-1}(\cdot) \\ \mathbf{A}_{\mathbf{T}eh} &= -j\underline{\epsilon}_{tz}\cdot\epsilon_{zz}^{-1}\nabla_t\times(\cdot) - j\nabla_t\times\mu_{zz}^{-1}\hat{z}\cdot\underline{\mu}_{zt}\cdot(\cdot) \\ \mathbf{A}_{\mathbf{T}he} &= j\underline{\mu}_{tz}\cdot\mu_{zz}^{-1}\nabla_t\times(\cdot) + j\nabla_t\times\epsilon_{zz}^{-1}\hat{z}\cdot\underline{\epsilon}_{zt}\cdot(\cdot) \\ \mathbf{A}_{\mathbf{T}hh} &= \omega\underline{\underline{\mu}}_{tt}\cdot(\cdot) - \frac{1}{\omega}\nabla_t\times\epsilon_{zz}^{-1}\nabla_t\times(\cdot) - \omega\underline{\mu}_{tz}\cdot\underline{\mu}_{zt}\mu_{zz}^{-1}(\cdot) \end{aligned} \quad (2.62)$$

The details of the derivation can be found in [79]. Eigenproblem (2.61) can easily be transformed to the standard one when multiplied by the operator $\mathbf{M}$.

A corresponding ω-formulation, for $\vec{E}_t$ and $\vec{H}_t$ fields, can also be derived [79]. However, it results in a quadratic generalized eigenvalue problem.

Further simplifications of (2.61) are possible for strictly bidirectional guides [79]. For this type of structures $\underline{\epsilon}_{tz} = \underline{\epsilon}_{zt} = 0$ and $\underline{\mu}_{tz} = \underline{\mu}_{zt} = 0$ and therefore

$$\underline{\underline{\epsilon}} = \left[\begin{array}{cc} \underline{\underline{\epsilon}}_{tt} & 0 \\ 0 & \epsilon_{zz} \end{array}\right] \quad , \quad \underline{\underline{\mu}} = \left[\begin{array}{cc} \underline{\underline{\mu}}_{tt} & 0 \\ 0 & \mu_{zz} \end{array}\right] \quad (2.63)$$

Moreover

$$\underline{\underline{\kappa}} = \left[\begin{array}{cc} \underline{\underline{\kappa}}_{tt} & 0 \\ 0 & \kappa_{zz} \end{array}\right] \quad , \quad \underline{\underline{\nu}} = \left[\begin{array}{cc} \underline{\underline{\nu}}_{tt} & 0 \\ 0 & \nu_{zz} \end{array}\right] \quad (2.64)$$

where

$$\begin{aligned} \underline{\underline{\kappa}}_{tt} &= \underline{\underline{\epsilon}}_{tt}^{-1} \quad , \quad \kappa_{zz} = \epsilon_{zz}^{-1} \\ \underline{\underline{\nu}}_{tt} &= \underline{\underline{\mu}}_{tt}^{-1} \quad , \quad \nu_{zz} = \mu_{zz}^{-1} \end{aligned} \tag{2.65}$$

In consequence, formulation (2.61) for transverse fields $\vec{E}_t$ and $\vec{H}_t$ obtains a simpler form

$$\begin{bmatrix} \mathbf{A}_{\mathbf{T}ee} & 0 \\ 0 & \mathbf{A}_{\mathbf{T}hh} \end{bmatrix} \begin{bmatrix} \vec{E}_t \\ \vec{H}_t \end{bmatrix} = \beta \begin{bmatrix} 0 & -\hat{z}\times \\ \hat{z}\times & 0 \end{bmatrix} \begin{bmatrix} \vec{E}_t \\ \vec{H}_t \end{bmatrix} \tag{2.66}$$

where

$$\begin{aligned} \mathbf{A}_{\mathbf{T}ee} &= \omega \underline{\underline{\epsilon}}_{tt} \cdot (\cdot) - \frac{1}{\omega} \nabla_t \times \nu_{zz}^{-1} \nabla_t \times (\cdot) \\ \mathbf{A}_{\mathbf{T}hh} &= \omega \underline{\underline{\mu}}_{tt} \cdot (\cdot) - \frac{1}{\omega} \nabla_t \times \kappa_{zz}^{-1} \nabla_t \times (\cdot) \end{aligned} \tag{2.67}$$

The symmetry of (2.66) was proven in [79] for lossless case.

The corresponding ω-formulation for transverse components is derived in [79]

$$\begin{bmatrix} \mathbf{B}_{\mathbf{T}dd} & 0 \\ 0 & \mathbf{B}_{\mathbf{T}bb} \end{bmatrix} \begin{bmatrix} \vec{D}_t \\ \vec{B}_t \end{bmatrix} = \omega \begin{bmatrix} 0 & -\hat{z}\times \\ \hat{z}\times & 0 \end{bmatrix} \begin{bmatrix} \vec{D}_t \\ \vec{B}_t \end{bmatrix} \tag{2.68}$$

where

$$\begin{aligned} \mathbf{B}_{\mathbf{T}dd} &= \beta \underline{\underline{\kappa}}_{tt} \cdot (\cdot) - \frac{1}{\beta} \nabla_t \kappa_{zz} \nabla_t \cdot (\cdot) \\ \mathbf{B}_{\mathbf{T}bb} &= \beta \underline{\underline{\nu}}_{tt} \cdot (\cdot) - \frac{1}{\beta} \nabla_t \nu_{zz} \nabla_t \cdot (\cdot) \end{aligned} \tag{2.69}$$

It can be shown [79] that (2.68) is symmetric for lossless media and real β.

It should be noted, that eigenproblems (2.66) and (2.68) become real for real tensors $\underline{\underline{\epsilon}}$ and $\underline{\underline{\mu}}$ and they can be transformed to nonsymmetric standard problems if premultiplied by $\overline{\mathbf{M}}$.

In the case of strictly bidirectional guides, electric $\vec{E}_t$ and magnetic $\vec{H}_t$ transverse fields can be decoupled. One of the possibilities is to eliminate $\vec{E}_t$ or $\vec{H}_t$ field from one of the equations in (2.66) using another one. This results in the following equivalent differential equations

$$\hat{z} \times \mathbf{A}_{\mathbf{T}hh} \hat{z} \times \mathbf{A}_{\mathbf{T}ee} \vec{E}_t = -\beta^2 \vec{E}_t \tag{2.70}$$

$$\hat{z} \times \mathbf{A}_{\mathbf{T}ee} \hat{z} \times \mathbf{A}_{\mathbf{T}hh} \vec{H}_t = -\beta^2 \vec{H}_t \tag{2.71}$$

An analogous ω-formulation is based on (2.68)

$$\hat{z} \times \mathbf{B}_{\mathbf{T}bb} \hat{z} \times \mathbf{B}_{\mathbf{T}dd} \vec{D}_t = -\omega^2 \vec{D}_t \tag{2.72}$$

$$\hat{z} \times \mathbf{B}_{\mathbf{T}dd} \hat{z} \times \mathbf{B}_{\mathbf{T}bb} \vec{B}_t = -\omega^2 \vec{B}_t \tag{2.73}$$

One of the main disadvantages of formulations (2.70–2.73) is that they involve differential operators of fourth order, which can cause difficulties in some methods of conversion into

the numerical problem (due to requirement for continuity of the differentiated functions) [93]. Another disadvantage is that these formulations are possibly potentially spurious because their derivation has not involved the divergence equations.

Another way of eliminating transverse fields is to start from equations (2.59) or (2.60) for bidirectional guide. The equation for $\vec{D}$ field has the form

$$\begin{bmatrix} -j\beta\hat{z}\times & -\hat{z}\times\nabla_t \\ -\nabla_t\cdot\hat{z}\times & 0 \end{bmatrix} \begin{bmatrix} \underline{\underline{\mu}}_{tt}^{-1} & 0 \\ 0 & \mu_{zz}^{-1} \end{bmatrix} \begin{bmatrix} -j\beta\hat{z}\times & -\hat{z}\times\nabla_t \\ -\nabla_t\cdot\hat{z}\times & 0 \end{bmatrix} \begin{bmatrix} \underline{\underline{\epsilon}}_{tt}^{-1} & 0 \\ 0 & \epsilon_{zz}^{-1} \end{bmatrix} \begin{bmatrix} \vec{D}_t \\ D_z \end{bmatrix} = \omega^2 \begin{bmatrix} \vec{D}_t \\ D_z \end{bmatrix} \quad (2.74)$$

It is easy to find out that the transverse part of the above equation involves the term depending on the longitudinal field component. This term can be eliminated by applying divergence equation (2.4) in the decomposed form

$$D_z = \frac{1}{j\beta}\nabla_t\cdot\vec{D}_t \quad (2.75)$$

derived using (2.47). This approach leads to the following generalized ω-formulations for $\vec{D}_t$ field (see [79] or [86] for the details of the derivation)

$$\left[\beta^2\underline{\underline{\mu}}_{tt}^{-1}\cdot\hat{z}\times\underline{\underline{\epsilon}}_{tt}^{-1}\cdot(\cdot)+\nabla_t\mu_{zz}^{-1}\hat{z}\nabla_t\times\underline{\underline{\epsilon}}_{tt}^{-1}\cdot(\cdot)+\underline{\underline{\mu}}_{tt}^{-1}\cdot\nabla_t\times\hat{z}\epsilon_{zz}^{-1}\nabla_t\cdot(\cdot)\right]\vec{D}_t = \omega^2\hat{z}\times\vec{D}_t \quad (2.76)$$

An analogous approach can be used for $\vec{B}_t$ field, leading to

$$\left[\beta^2\underline{\underline{\epsilon}}_{tt}^{-1}\cdot\hat{z}\times\underline{\underline{\mu}}_{tt}^{-1}\cdot(\cdot)+\nabla_t\epsilon_{zz}^{-1}\hat{z}\nabla_t\times\underline{\underline{\mu}}_{tt}^{-1}\cdot(\cdot)+\underline{\underline{\epsilon}}_{tt}^{-1}\cdot\nabla_t\times\hat{z}\mu_{zz}^{-1}\nabla_t\cdot(\cdot)\right]\vec{B}_t = \omega^2\hat{z}\times\vec{B}_t \quad (2.77)$$

Premultiplying the above equations respectively by $-\underline{\underline{\mu}}_{tt}\cdot(\cdot)$ and $-\underline{\underline{\epsilon}}_{tt}\cdot(\cdot)$ and rearranging the terms we can obtain equivalent β-formulations for $\vec{E}_t$ and $\vec{H}_t$ fields

$$\left[\omega^2\underline{\underline{\mu}}_{tt}\cdot\hat{z}\times\underline{\underline{\epsilon}}_{tt}\cdot(\cdot)-\underline{\underline{\mu}}_{tt}\cdot\nabla_t\mu_{zz}^{-1}\hat{z}\nabla_t\times(\cdot)-\nabla_t\times\hat{z}\epsilon_{zz}^{-1}\nabla_t\cdot\underline{\underline{\epsilon}}_{tt}\cdot(\cdot)\right]\vec{E}_t = \beta^2\hat{z}\times\vec{E}_t \quad (2.78)$$

$$\left[\omega^2\underline{\underline{\epsilon}}_{tt}\cdot\hat{z}\times\underline{\underline{\mu}}_{tt}\cdot(\cdot)-\underline{\underline{\epsilon}}_{tt}\cdot\nabla_t\epsilon_{zz}^{-1}\hat{z}\nabla_t\times(\cdot)-\nabla_t\times\hat{z}\mu_{zz}^{-1}\nabla_t\cdot\underline{\underline{\mu}}_{tt}\cdot(\cdot)\right]\vec{H}_t = \beta^2\hat{z}\times\vec{H}_t \quad (2.79)$$

The last four eigenproblems are not symmetric, but it can be shown [79, 86], that transposition of eigenproblem (2.76) results in eigenproblem of form (2.77) and transposition of (2.78) gives (2.79).

Generalized eigenproblems (2.76–2.79) can be also transformed into the standard ones by multiplying them with $-\hat{z}\times(\cdot)$.

2.4.3 Longitudinal component formulations

For isotropic structures a formulation in terms of E_z and H_z fields is also possible. The derivation can be found in [118]. It results in a pair of equations in the form of nonstandard eigenvalue problems

$$\begin{aligned}\nabla_t^2 E_z + k_t^2 E_z + \epsilon^{-1}\nabla_t\epsilon\cdot\nabla_t E_z - \frac{\omega^2}{k_t^2}\nabla_t(\epsilon\mu)\cdot\left(\nabla_t E_z - \frac{\beta}{\omega\epsilon}\hat{z}\times\nabla_t H_z\right) &= 0\\ \nabla_t^2 H_z + k_t^2 H_z + \mu^{-1}\nabla_t\mu\cdot\nabla_t H_z - \frac{\omega^2}{k_t^2}\nabla_t(\epsilon\mu)\cdot\left(\nabla_t H_z + \frac{\beta}{\omega\mu}\hat{z}\times\nabla_t E_z\right) &= 0\end{aligned} \quad (2.80)$$

where $k_t^2 = \omega^2\epsilon\mu - \beta^2$.

Another formulation can be found in [48], resulting in the following pair of equations, equivalent to (2.80)

$$\begin{aligned}\beta\omega\nabla_t\cdot\frac{1}{k_t^2}\hat{z}\times\nabla_t H_z - \omega^2\nabla_t\cdot\frac{\epsilon}{k_t^2}\nabla_t E_z - \omega^2\epsilon E_z &= 0\\ -\beta\omega\nabla_t\cdot\frac{1}{k_t^2}\hat{z}\times\nabla_t E_z - \omega^2\nabla_t\cdot\frac{\mu}{k_t^2}\nabla_t H_z - \omega^2\mu H_z &= 0\end{aligned} \quad (2.81)$$

This set of equations can be rearranged in order to get generalized ω-eigenproblem

$$\begin{bmatrix} -\nabla_t\cdot\frac{\epsilon}{\mu\epsilon-\delta^2}\nabla_t & \delta\nabla_t\cdot\frac{1}{\mu\epsilon-\delta^2}\hat{z}\times\nabla_t \\ -\delta\nabla_t\cdot\frac{1}{\mu\epsilon-\delta^2}\hat{z}\times\nabla_t & -\nabla_t\cdot\frac{\mu}{\mu\epsilon-\delta^2}\nabla_t \end{bmatrix}\begin{bmatrix} E_z \\ H_z \end{bmatrix} = \omega^2\begin{bmatrix} \epsilon & 0 \\ 0 & \mu \end{bmatrix}\begin{bmatrix} E_z \\ H_z \end{bmatrix} \quad (2.82)$$

where $\delta = \beta/\omega$. This problem is symmetric in case of lossless media and real β (and ω) [48].

The main drawback of formulation (2.82) is that it is based on wave equations only and its derivation does not involve the divergence equations. In consequence, this formulation is potentially spurious. Moreover, it is not defined for $\delta = \sqrt{\mu\epsilon}$, thus this case should be avoided.

It should be noted, that in the cutoff $\delta = 0$ and formulation (2.82) splits off into two scalar equations (2.87) and (2.88).

2.4.4 Scalar formulations

At critical points ($\beta = 0$ or $\omega = 0$), Maxwell's equations (2.2–2.5) can be simplified. In the case of bidirectional guides, scalar eigenproblems can be formulated in terms of TE- and TM-type modes (with respect to the z-axis).

At cutoff ($\beta = 0$), Maxwell's equations (2.2) and (2.3) can be split into two pairs of equations

$$\nabla_t\times\hat{z}E_z = -j\omega\vec{B}_t \quad (2.83)$$

$$\hat{z}\nabla_t\times\underline{\underline{\epsilon}}_{tt}^{-1}\cdot\vec{D}_t = -j\omega\mu_{zz}H_z \quad (2.84)$$

and

$$\nabla_t\times\hat{z}H_z = j\omega\vec{D}_t \quad (2.85)$$

$$\hat{z}\nabla_t\times\underline{\underline{\mu}}_{tt}^{-1}\cdot\vec{B}_t = j\omega\epsilon_{zz}E_z \quad (2.86)$$

Isolating $\vec{D}_t$ from (2.85) and substituting it into (2.84) we get the following scalar eigenproblem for H_z (TE modes)

$$\hat{z}\nabla_t \times \underline{\underline{\epsilon}}_{tt}^{-1} \cdot \nabla_t \times \hat{z}H_z = \omega^2 \mu_{zz} H_z \tag{2.87}$$

Analogously, separating $\vec{B}_t$ from (2.83) and replacing it in (2.86) we get a scalar eigenproblem for E_z (TM modes)

$$\hat{z}\nabla_t \times \underline{\underline{\mu}}_{tt}^{-1} \cdot \nabla_t \times \hat{z}E_z = \omega^2 \epsilon_{zz} E_z \tag{2.88}$$

In the case of static solution ($\omega = 0$), the transverse parts of Maxwell equations (2.2) and (2.3) fulfill the following equations

$$-j\beta\hat{z} \times \vec{E}_t - \hat{z} \times \nabla_t E_z = 0 \tag{2.89}$$
$$-j\beta\hat{z} \times \vec{H}_t - \hat{z} \times \nabla_t H_z = 0 \tag{2.90}$$

Divergence equations (2.4) and (2.5) can be written in the following form

$$\nabla_t \cdot \underline{\underline{\epsilon}}_{tt} \cdot \vec{E}_t - j\beta\epsilon_{zz} E_z = 0 \tag{2.91}$$
$$\nabla_t \cdot \underline{\underline{\mu}}_{tt} \cdot \vec{H}_t - j\beta\mu_{zz} H_z = 0 \tag{2.92}$$

Computing $\vec{E}_t$ from (2.89) and substituting the result into (2.91) one can get the following scalar eigenproblem for E_z (TE modes)

$$\nabla_t \cdot \underline{\underline{\epsilon}}_{tt} \cdot \nabla_t E_z = \beta^2 \epsilon_{zz} E_z \tag{2.93}$$

while taking $\vec{H}_t$ from (2.90) and replacing it in (2.92) we obtain a scalar eigenproblem for H_z (TM modes)

$$\nabla_t \cdot \underline{\underline{\mu}}_{tt} \cdot \nabla_t H_z = \beta^2 \mu_{zz} H_z \tag{2.94}$$

Eigenproblems (2.87–2.88) and (2.93–2.94) are symmetric for lossless media.

2.4.5 Homogeneous waveguides

Consider a homogeneous strictly bidirectional waveguide filled with the material of special type of anisotropy (e.g. gyrotropic medium magnetized in z-direction), defined by the following tensors

$$\underline{\underline{\epsilon}} = \begin{bmatrix} \epsilon & j\epsilon_a & 0 \\ -j\epsilon_a & \epsilon & 0 \\ 0 & 0 & \epsilon_{zz} \end{bmatrix} \quad , \quad \underline{\underline{\mu}} = \begin{bmatrix} \mu & j\mu_a & 0 \\ -j\mu_a & \mu & 0 \\ 0 & 0 & \mu_{zz} \end{bmatrix} \tag{2.95}$$

Starting from wave equations (2.20) and (2.21) in the decomposed form and using the divergence equations, the formulation in terms of longitudinal E_z and H_z components

can be derived [41]. It results in a pair of equations in the form of nonstandard complex eigenvalue problems

$$\begin{aligned}\nabla_t^2 E_z - \beta^2 \frac{\epsilon_{zz}}{\epsilon} E_z + \omega^2 \epsilon_{zz} \mu_{\text{eff}} E_z - j\omega\beta\mu_{zz}\left(\frac{\epsilon_a}{\epsilon} + \frac{\mu_a}{\mu}\right) H_z &= 0 \\ \nabla_t^2 H_z - \beta^2 \frac{\mu_{zz}}{\mu} H_z + \omega^2 \mu_{zz} \epsilon_{\text{eff}} H_z + j\omega\beta\epsilon_{zz}\left(\frac{\epsilon_a}{\epsilon} + \frac{\mu_a}{\mu}\right) E_z &= 0 \end{aligned} \tag{2.96}$$

where

$$\epsilon_{\text{eff}} = \epsilon - \frac{\epsilon_a^2}{\epsilon} \quad , \quad \mu_{\text{eff}} = \mu - \frac{\mu_a^2}{\mu} \tag{2.97}$$

Dividing equations (2.96) by ω^2 and rearranging the terms we get the following generalized complex ω-eigenproblem for longitudinal field components

$$\begin{bmatrix} \mu_{\text{eff}}(\cdot) - \delta^2 \frac{1}{\epsilon}(\cdot) & -j\delta\left(\frac{\epsilon_a}{\epsilon} + \frac{\mu_a}{\mu}\right)(\cdot) \\ j\delta\left(\frac{\epsilon_a}{\epsilon} + \frac{\mu_a}{\mu}\right)(\cdot) & \epsilon_{\text{eff}}(\cdot) - \delta^2 \frac{1}{\mu}(\cdot) \end{bmatrix} \begin{bmatrix} D_z \\ B_z \end{bmatrix} = -\frac{1}{\omega^2} \begin{bmatrix} \kappa_{zz} & 0 \\ 0 & \nu_{zz} \end{bmatrix} \nabla_t^2 \begin{bmatrix} D_z \\ B_z \end{bmatrix} \tag{2.98}$$

where $\delta = \beta/\omega$. A similar generalized complex β-eigenproblem for longitudinal field components can be obtained by dividing equations (2.96) with β^2

$$\begin{bmatrix} \frac{1}{\delta^2}\mu_{\text{eff}}(\cdot) - \frac{1}{\epsilon}(\cdot) & -j\frac{1}{\delta}\left(\frac{\epsilon_a}{\epsilon} + \frac{\mu_a}{\mu}\right)(\cdot) \\ j\frac{1}{\delta}\left(\frac{\epsilon_a}{\epsilon} + \frac{\mu_a}{\mu}\right)(\cdot) & \frac{1}{\delta^2}\epsilon_{\text{eff}}(\cdot) - \frac{1}{\mu}(\cdot) \end{bmatrix} \begin{bmatrix} D_z \\ B_z \end{bmatrix} = -\frac{1}{\beta^2} \begin{bmatrix} \kappa_{zz} & 0 \\ 0 & \nu_{zz} \end{bmatrix} \nabla_t^2 \begin{bmatrix} D_z \\ B_z \end{bmatrix} \tag{2.99}$$

Due to incorporation of the divergence equations in the derivation process of formulations (2.96), (2.98) and (2.99) they do not generate spurious solutions. Moreover, it can be easily shown that for lossless media an real β (and ω) (2.98) and (2.99) are symmetric.

It is worth noting that in the case of isotropic materials $\epsilon_{zz} = \epsilon$, $\mu_{zz} = \mu$ and $\epsilon_a = \mu_a = 0$. Therefore equations (2.96) split into the two independent scalar equations of type (2.24).

2.4.5.1 x- or y-direction homogeneity (Cartesian coordinates)

Consider a waveguide homogeneous in the y- and z-directions or x- and z-directions and call them transverse directions. If lossless bidirectional media with tensors $\underline{\underline{\epsilon}}$ and $\underline{\underline{\mu}}$ of form analogous to (2.63) are taken into account, then a simplified formulation for the components in the distinguished direction can be derived.

For the waveguide homogeneous in y- and z-directions the tensors should have the form

$$\underline{\underline{\epsilon}} = \begin{bmatrix} \epsilon_{xx} & 0 & 0 \\ 0 & \epsilon & j\epsilon_a \\ 0 & -j\epsilon_a & \epsilon \end{bmatrix} \quad , \quad \underline{\underline{\mu}} = \begin{bmatrix} \mu_{xx} & 0 & 0 \\ 0 & \mu & j\mu_a \\ 0 & -j\mu_a & \mu \end{bmatrix} \tag{2.100}$$

and for the waveguide homogeneous in the x- and z-directions the following

$$\underline{\underline{\epsilon}} = \begin{bmatrix} \epsilon & 0 & -j\epsilon_a \\ 0 & \epsilon_{yy} & 0 \\ j\epsilon_a & 0 & \epsilon \end{bmatrix} \quad , \quad \underline{\underline{\mu}} = \begin{bmatrix} \mu & 0 & -j\mu_a \\ 0 & \mu_{yy} & 0 \\ j\mu_a & 0 & \mu \end{bmatrix} \tag{2.101}$$

The examples of the materials possessing such properties are gyrotropic media respectively magnetized in the x- and y-direction.

First, consider the case of the structure homogeneous in y- and z-directions. Decompose the fields and the nabla operator ∇ in the following way

$$\vec{A} = \hat{x}A_x + \vec{A}_t \qquad , \qquad \vec{A}_t = \hat{y}A_y + \hat{z}A_z \tag{2.102}$$

$$\nabla = \hat{x}\frac{\partial}{\partial x} + \nabla_t \qquad , \qquad \nabla_t = \hat{y}\frac{\partial}{\partial y} + \hat{z}\frac{\partial}{\partial z} \tag{2.103}$$

Starting from the decomposed form of wave equations (2.20) and (2.21) and using the divergence equations the following eigenproblem for D_x and B_x can be derived

$$\begin{bmatrix} \frac{\partial}{\partial x}\epsilon^{-1}\frac{\partial}{\partial x}(\cdot) + \omega^2\mu_{\text{eff}}(\cdot) & \omega\left(\frac{\mu_a}{\mu}\frac{\partial}{\partial x} + \frac{\partial}{\partial x}\frac{\epsilon_a}{\epsilon}\right)(\cdot) \\ -\omega\left(\frac{\epsilon_a}{\epsilon}\frac{\partial}{\partial x} + \frac{\partial}{\partial x}\frac{\mu_a}{\mu}\right)(\cdot) & \frac{\partial}{\partial x}\mu^{-1}\frac{\partial}{\partial x}(\cdot) + \omega^2\epsilon_{\text{eff}}(\cdot) \end{bmatrix} \begin{bmatrix} D_x \\ B_x \end{bmatrix} = k_t^2 \begin{bmatrix} \kappa_{xx} & 0 \\ 0 & \nu_{xx} \end{bmatrix} \begin{bmatrix} D_x \\ B_x \end{bmatrix} \tag{2.104}$$

where $k_t^2 = -\nabla_t^2$ and for a rectangular waveguide

$$k_t^2 = k_y^2 + \beta^2 \tag{2.105}$$

The details of the derivation along with the proof of symmetry (in the lossless case) can be found in [79].

For the structure homogeneous in x- and z-directions and the material tensors defined by (2.100) an analogous approach can be found for D_y and B_y fields

$$\begin{bmatrix} \frac{\partial}{\partial y}\epsilon^{-1}\frac{\partial}{\partial y}(\cdot) + \omega^2\mu_{\text{eff}}(\cdot) & \omega\left(\frac{\mu_a}{\mu}\frac{\partial}{\partial y} + \frac{\partial}{\partial y}\frac{\epsilon_a}{\epsilon}\right)(\cdot) \\ -\omega\left(\frac{\epsilon_a}{\epsilon}\frac{\partial}{\partial y} + \frac{\partial}{\partial y}\frac{\mu_a}{\mu}\right)(\cdot) & \frac{\partial}{\partial y}\mu^{-1}\frac{\partial}{\partial y}(\cdot) + \omega^2\epsilon_{\text{eff}}(\cdot) \end{bmatrix} \begin{bmatrix} D_y \\ B_y \end{bmatrix} = k_t^2 \begin{bmatrix} \kappa_{yy} & 0 \\ 0 & \nu_{yy} \end{bmatrix} \begin{bmatrix} D_y \\ B_y \end{bmatrix} \tag{2.106}$$

where $k_t^2 = -\nabla_t^2$ and for a rectangular waveguide

$$k_t^2 = k_x^2 + \beta^2 \tag{2.107}$$

Symmetry of this formulation can be deduced by analogy to (2.104).

It should be noted that formulations (2.96), (2.98) and (2.99) can easily be obtained from (2.104) or (2.106) if we assume that the structure is homogeneous in all directions.

Formulations (2.104) and (2.106) do not generate spurious solutions because their derivation involves the divergence equations.

2.5 Choice of formulation

Numerical analysis of resonator and waveguide structures relies on a conversion (projection) method applied to an analytical formulation and numerical solution of the resulting matrix eigenproblem. Efficiency of any numerical eigensolver substantially depends on the properties of the matrix operators. Since these properties are implied by the type

of the conversion method used and the analytical formulation involved, a choice of the formulation can indirectly influence the efficiency of the overall analysis.

The most important factors which depend on analytical formulation and influence the efficiency are: the number of field components involved, the operator symmetry and the ability of the formulation to generate spurious solutions. The first factor is crucial for the numerical efficiency due to the fact that for almost all conversion methods discussed in the next chapter, the size of the resulting matrix is directly proportional to the number of components in the formulation. The symmetry of the formulation is also important, because many conversion (projection) methods can preserve this feature and result in a symmetric matrix eigenproblem, which in many cases can be solved faster than the nonsymmetric one. The last factor influences efficiency because coping with spurious solutions in formulations that are potentially spurious can require additional effort. Simple identification of spurious solutions generated by the corresponding matrix eigenproblem requires additional computations, while the elimination of the solutions using the method based on the modifications of the analytical formulation (such as penalty method [1, 13, 48, 55, 56, 84, 90, 91, 101–103, 125]) can strongly deteriorate the efficiency of iterative eigensolvers due to worsening of the convergence rate. Moreover, the modification of the original operator (as in the penalty method) can also influence the quality of the solution by increasing its numerical error. In contrast to both methods discussed above, the elimination of spurious solutions via selection of proper basis and testing functions in the projection may not require an additional computational effort. This is the case when entire subdomain expansion of the functions (see Chapter 3) is used in the projection (e.g. edge elements used in FEM [7, 48, 117, 121]). However, when the entire domain expansion method (see Chapter 3) is used, the determination of the proper functions defined over the entire domain is, in general, difficult and expensive. It should be noted that, incorporation of a spurious-free formulation completely removes the problem of spurious solutions. Based on these observations let us consider a few particular structure types and suggest the appropriate formulations.

The most unsuitable situation is in the case of the most general inhomogeneous resonator structures, where all corresponding six-component (2.16, 2.17, 2.19) and three-component (2.20–2.23) formulations are able to generate spurious eigenvalues. Since the six- and three-component formulations are fully equivalent the latter ones are preferred, because they lead to smaller matrix problems. If the resonator has cylindrical symmetry and is filled with an isotropic or strictly bidirectional medium the resonator problem can be reduced to a waveguide-type problem or to a scalar one. As a result, appropriate two- or one-component spurious-free formulations can be used.

When alternative (equivalent) dual formulations are possible (e.g. (2.20) and (2.21) or (2.22) and (2.23)) the choice between them can be governed by the continuity of particular field components. This is due to the fact that the functions representing fields or fluxes in the conversion methods such as the Rayleigh-Ritz method, Galerkin method, FEM or FDFD have to be continuous in order to apply differential operators to. In consequence, in the analysis of the structures filled with, for example, non-magnetic materials the

formulations for magnetic field or flux components may be preferred over the formulations for the electric components.

In the analysis of waveguides, all full vector formulations can be regarded as not competitive to transverse component formulations because they involve more components and/or can generate spurious solutions. For the most general media, the spurious-free formulation (2.61), involving four transverse components can be used. However, for strictly bidirectional media, more efficient two-component formulations (2.76–2.79) are preferable. Moreover, scalar formulations (2.87) and (2.88) or scalar formulations (2.93) and (2.94) can be yet more efficient for cutoff and static cases, respectively. For the structures filled with isotropic medium, longitudinal component formulation (2.82), alternative to (2.76–2.79), is not competitive because it is not spurious-free.

In the case of homogeneous and strictly bidirectional waveguides (but not isotropic), reduced spurious-free two-component formulations (2.98–2.99), (2.104) and (2.106) can also be used. For isotropic structures these formulations are obviously reducible to the scalar ones.

Chapter 3

Conversion to matrix eigenvalue problems

Numerical techniques of solving partial differential equations described in the previous chapter are mostly based on conversion of the operator equations into matrix eigenvalue problems. All the methods of conversion rely on projection of an infinite dimensional analytical problem into a finite dimensional space. The differences between the projection methods concern two aspects: the projection technique and choice of the projection space i.e. basis functions.

The methods of conversion can be divided into two general classes: *classical* and *hybrid* methods. The former methods involve simple Maxwellian or polynomial basis functions [59] which are usually evaluated analytically. The methods of this kind are capable of analyzing efficiently structures of relatively simple geometries or filled with isotropic materials. In the hybrid methods the solution is obtained in two steps. The first step involves solution of a simpler eigenproblem, usually by means of any classical method. The resulting eigenvectors are then used in the second step as composite Maxwellian basis functions [59] in order to approximate the field of the problem at hand. The hybrid methods are specially intended for analysis of the structures with complex shape and media parameters.

The most popular classical and hybrid conversion methods will be described in the following sections.

3.1 Classical methods

A general form of the operator eigenproblems obtained in the previous chapter is

$$\mathbf{A}v = \lambda \mathbf{B}v \tag{3.1}$$

The most popular methods for conversion operator problems of form (3.1) into the matrix eigenvalue problems are the Rayleigh-Ritz method and the method of moments. Basis functions used in both methods can be defined over the entire domain (*entire domain expansion*) or over particular parts of the domain (*entire subdomain expansion*). Since

simple entire domain expansion functions cannot well approximate fields on an irregular boundary the classical conversion methods using the entire domain expansion are restricted to the structures of simple shape (conforming to the coordinates). In order to analyze structures with irregular boundaries, the methods using the entire subdomain expansion, where basis functions are locally defined over the subdomains can be used. An example of such a versatile entire subdomain expansion method is the finite element method. Another universal technique of this kind is the finite difference frequency domain method, which is based on the discretization of the differential operators. All these classical methods are briefly described in the succeeding sections.

3.1.1 Rayleigh-Ritz method

Rayleigh-Ritz (RR) method is closely related to *variational techniques* [81, 100], which are based on minimizing variation of a *functional*, which takes up different form depending on the symmetry of the problem. The variation of functional should vanish in the *stationary point*. Finding this point is equivalent to finding the solution of the problem. RR method relies on the application of *Rayleigh-Ritz procedure* (compare section A.1.4) to the stationarity criterion. RR procedure performs a projection of an operator onto a finite dimensional space. This results in a matrix eigenvalue problem.

3.1.1.1 Symmetric eigenproblem

Let us consider the case where operators involved in problem (3.1) are symmetric.

The functional appropriate for the symmetric problem has the following form

$$\mathbf{F} = \Big(\mathbf{A}v, v\Big) - \lambda\Big(\mathbf{B}v, v\Big) \tag{3.2}$$

The first variation of $\mathbf{F}$ about the exact solution v, caused by a small perturbation dv, can be found as

$$d\mathbf{F} = \Big(\mathbf{A}v, dv\Big) - \lambda\Big(\mathbf{B}v, dv\Big) + \Big(\mathbf{A}\,dv, v\Big) - \lambda\Big(\mathbf{B}\,dv, v\Big) \tag{3.3}$$

Equating (3.3) to zero leads to the stationarity criterion, which, using the symmetry properties of eigenproblem (3.1), can be transformed to the following form [79]

$$d\mathbf{F} = 2\Big(\mathbf{A}v - \lambda\mathbf{B}v, dv\Big) = 0 \tag{3.4}$$

Since the stationarity criterion (3.4) is known, the RR procedure can be applied. The solution vector v is expanded into a series of admissible (satisfying boundary conditions) *basis functions* (or *trial functions*) $\{u_i\}_{i=1,\ldots,n}$

$$v = \sum_{i=1}^{n} \alpha_i u_i \tag{3.5}$$

Therefore, the perturbation dv can be written as

$$dv = \sum_{j=1}^{n} d\alpha_j \, u_j \tag{3.6}$$

Substituting (3.5) and (3.6) into (3.4) and using the symmetry properties one can get

$$\sum_{j=1}^{n} \left[d\alpha_j \sum_{i=1}^{n} \alpha_i \Big(\mathbf{A}u_i - \lambda \mathbf{B}u_i, u_j \Big) \right] = 0 \tag{3.7}$$

This equation is satisfied if and only if

$$\forall_{j=1,\ldots,n} \sum_{i=1}^{n} \alpha_i \Big(\mathbf{A}u_i - \lambda \mathbf{B}u_i, u_j \Big) = 0 \tag{3.8}$$

This results in the following symmetric generalized matrix eigenproblem

$$\underline{\underline{A}} \, \underline{v} = \lambda \underline{\underline{B}} \, \underline{v} \tag{3.9}$$

where $\underline{v} = [\alpha_1, \ldots, \alpha_n]^T$ and elements of matrices $\underline{\underline{A}}$ and $\underline{\underline{B}}$ are defined as

$$A_{ji} = \Big(\mathbf{A}u_i, u_j \Big) \qquad , \qquad B_{ji} = \Big(\mathbf{B}u_i, u_j \Big) \tag{3.10}$$

The matrices $\underline{\underline{A}}$ and $\underline{\underline{B}}$ are dense when the entire domain basis functions $\{u_i\}_{i=1,\ldots,n}$ are used.

It should be noted, that for **B**-orthonormal[1] set of the basis functions matrix $\underline{\underline{B}} = \underline{\underline{I}}$ and eigenproblem (3.9) becomes a standard one.

Having solved eigenproblem (3.9) for the vector of expansion coefficients $\underline{v}$, the approximate solution v can be found from (3.5).

3.1.1.2 Nonsymmetric eigenproblem

Classical variational approach to nonsymmetric eigenproblems requires definition of the problem transposed to (3.1)

$$^t\mathbf{A}\tilde{v} = \lambda^{*t}\mathbf{B}\tilde{v} \tag{3.11}$$

where $^t\mathbf{A}$ and $^t\mathbf{B}$ are the adjoint operators $\mathbf{A}$ and $\mathbf{B}$, respectively.

A suitable variational functional for problem (3.1) can be found as

$$\mathbf{F} = \Big(\mathbf{A}v, \tilde{v} \Big) - \lambda \Big(\mathbf{B}v, \tilde{v} \Big) \tag{3.12}$$

The first variation of $\mathbf{F}$, caused by small perturbations dv and $d\tilde{v}$, can be found as

$$d\mathbf{F} = \Big(\mathbf{A}v, d\tilde{v} \Big) - \lambda \Big(\mathbf{B}v, d\tilde{v} \Big) + \Big(\mathbf{A}\, d\tilde{v}, v \Big) - \lambda \Big(\mathbf{B}\, d\tilde{v}, v \Big) \tag{3.13}$$

[1] **B**-orthonormality of functions $\{u_i\}_{i=1,\ldots,n}$ implies that they are normalized and orthogonal with respect to the **B**-dot product, i.e. $\forall_{i=j} \Big(\mathbf{B}u_i, u_j \Big) = 1$ and $\forall_{i \neq j} \Big(\mathbf{B}u_i, u_j \Big) = 0$

It can be transformed into the form of a stationarity criterion

$$d\mathbf{F} = \Big(\mathbf{A}v - \lambda\mathbf{B}v, d\tilde{v}\Big) + \Big(dv, {}^t\mathbf{A}\tilde{v} - \lambda^{*t}\mathbf{B}\tilde{v}\Big) = 0 \tag{3.14}$$

Application of the RR procedure involves the expansion of both v and $\tilde{v}$. Functions v and dv are expanded using (3.5) and (3.6), while $\tilde{v}$ and $d\tilde{v}$ as

$$\tilde{v} = \sum_{i=1}^{n} \tilde{\alpha}_i \tilde{u}_i \tag{3.15}$$

$$d\tilde{v} = \sum_{j=1}^{n} d\tilde{\alpha}_j \, \tilde{u}_j \tag{3.16}$$

Substituting (3.5), (3.6), (3.15) and (3.16) into (3.14) one gets

$$\sum_{j=1}^{n} \left[d\tilde{\alpha}_j \sum_{i=1}^{n} \alpha_i \Big(\mathbf{A}u_i - \lambda\mathbf{B}u_i, \tilde{u}_j\Big) + d\alpha_j \sum_{i=1}^{n} \tilde{\alpha}_i \Big(u_j, {}^t\mathbf{A}\tilde{u}_i - \lambda^{*t}\mathbf{B}\tilde{u}_i\Big) \right] = 0 \tag{3.17}$$

This is valid if the following equations are simultaneously satisfied

$$\forall_{j=1,\ldots,n} \quad \textstyle\sum_{i=1}^{n} \alpha_i \Big(\mathbf{A}u_i - \lambda\mathbf{B}u_i, \tilde{u}_j\Big) = 0 \tag{3.18}$$

$$\forall_{j=1,\ldots,n} \quad \textstyle\sum_{i=1}^{n} \tilde{\alpha}_i \Big(u_j, {}^t\mathbf{A}\tilde{u}_i - \lambda^{*t}\mathbf{B}\tilde{u}_i\Big) = 0 \tag{3.19}$$

These result in a pair of generalized dense matrix eigenproblems. The eigenproblem arising from (3.18) is

$$\underline{\underline{A}}\,\underline{v} = \lambda \underline{\underline{B}}\,\underline{v} \tag{3.20}$$

where matrices $\underline{\underline{A}}$ and $\underline{\underline{B}}$ are, in general, nonsymmetric and their elements are defined as

$$A_{ji} = \Big(\mathbf{A}u_i, \tilde{u}_j\Big) \qquad , \qquad B_{ji} = \Big(\mathbf{B}u_i, \tilde{u}_j\Big) \tag{3.21}$$

The second equivalent nonsymmetric eigenproblem, arising from (3.19), is

$$\underline{\underline{C}}\,\underline{\tilde{v}} = \lambda^* \underline{\underline{D}}\,\underline{\tilde{v}} \tag{3.22}$$

where $\underline{\tilde{v}} = [\tilde{\alpha}_1, \ldots, \tilde{\alpha}_n]^T$ and elements of matrices $\underline{\underline{C}}$ and $\underline{\underline{D}}$ are defined as

$$C_{ji} = \Big(u_j, {}^t\mathbf{A}\tilde{u}_i\Big) \qquad , \qquad D_{ji} = \Big(u_j, {}^t\mathbf{B}\tilde{u}_i\Big) \tag{3.23}$$

As in the symmetric case, generalized eigenproblems (3.20) and (3.22) become standard ones if functions $\{u_i\}_{i=1,\ldots,n}$ and $\{\tilde{u}_j\}_{j=1,\ldots,n}$ are biorthonormal[2] with respect to the **B**-dot product.

The approximate solution v can be computed using (3.5), once the solution $\underline{v}$ of eigenproblem (3.20) is found.

The approach described above can sometimes be difficult to implement due to the need for determination of expansion functions $\tilde{u}_i$ for the adjoint problem. The adjoint fields are not always easily computable because they should satisfy boundary conditions for the transposed problem, which are generally different from the conditions for the original problem [33,79]. This inconvenience can be avoided when using the *local potential method*.

[2]Biorthonormality of the two sets of functions $\{u_i\}_{i=1,\ldots,n}$ and $\{v_j\}_{j=1,\ldots,n}$ implies that they are normalized and orthogonal, i.e. $\forall_{i=j}\Big(u_i, \tilde{u}_j\Big) = 1$ and $\forall_{i\neq j}\Big(u_i, \tilde{u}_j\Big) = 0$

Local potential method. This technique is also called *generalized entropy* [100]. It relies on constructing a variational functional for nonsymmetric problem using the methodology of symmetric problems. In order to find the stationary state, it is temporarily assumed that the part of the functional related to the non-self-adjoint operator remains unchanged under the perturbation.

Consider the case when operator $\mathbf{B}$ in (3.1) is symmetric, while $\mathbf{A}$ is nonsymmetric. Decompose $\mathbf{A}$ into the symmetric $\mathbf{A_s}$ and nonsymmetric $\mathbf{A_n}$ parts

$$\mathbf{A} = \mathbf{A_s} + \mathbf{A_n} \tag{3.24}$$

Assuming that the component of the functional related to the non-self-adjoint part of the operator is kept constant under the perturbation, the functional for the self-adjoint problem can be written as [33]

$$\mathbf{F} = \Big(\mathbf{A_s}v, v\Big) - \lambda\Big(\mathbf{B}v, v\Big) + 2\Big(n, v\Big) \tag{3.25}$$

where

$$n = \mathbf{A_n}v|_{v=v_0} = \text{const} \tag{3.26}$$

Functional $\mathbf{F}$ is called the *local potential* and the corresponding variational expression can be found as [33]

$$d\mathbf{F} = 2\Big(\mathbf{A_s}v + n - \lambda\mathbf{B}v, dv\Big) = 0 \tag{3.27}$$

One can see, that releasing now constraint (3.26) we obtain exactly (3.4).

Application of the RR procedure results in the eigenproblem analogous to (3.9) with the difference that matrix $\underline{\underline{A}}$ is nonsymmetric.

3.1.2 Method of moments

The method of moments [42, 80] is similar to the RR procedure in the sense of expanding an unknown function into a series of *basis functions*. Taking the inner product of an eigenproblem with a set of *testing functions* or *weighting functions* results in a set of equations, which can be written in the form of a matrix eigenproblem.

Consider eigenproblem (3.1) and expand an unknown function v into a series of basis functions $\{u_i\}_{i=1,\dots,n}$

$$v = \sum_{i=1}^{n} \alpha_i u_i \tag{3.28}$$

where $\{u_i\}_{i=1,\dots,n}$ form a complete set in the domain of the problem. Consider also a set of testing functions $\{w_j\}_{j=1,\dots,n}$ spanning the range of $\mathbf{A}$.

Taking the inner products of eigenproblem (3.1) with the testing functions we get the following set of equations

$$\forall_{j=1,\dots,n} \sum_{i=1}^{n} \alpha_i \Big(\mathbf{A}u_i, u_j\Big) = \lambda \sum_{i=1}^{n} \alpha_i \Big(\mathbf{B}u_i, u_j\Big) \tag{3.29}$$

This can be written in the form of generalized eigenproblem

$$\underline{\underline{A}}\,\underline{v} = \lambda \underline{\underline{B}}\,\underline{v} \tag{3.30}$$

where matrices elements are defined as

$$A_{ji} = \Big(\mathbf{A}u_i, w_j\Big) \qquad , \qquad B_{ji} = \Big(\mathbf{B}u_i, w_j\Big) \tag{3.31}$$

and $\underline{v} = [\alpha_1, \ldots, \alpha_n]^T$. The matrices $\underline{\underline{A}}$ and $\underline{\underline{B}}$ are dense when the basis and testing functions defined over the entire domain are involved. Eigenproblem (3.30) is, in general, nonsymmetric. However, it can become symmetric for self-adjoint operators and adequately chosen basis and testing functions. Moreover, this generalized eigenproblem can become a standard one, if the basis and testing functions are biorthonormal with respect to the **B**-dot product.

Different types of method of moments can be distinguished, depending on the testing functions selected. These of main interest are: the *Galerkin method*, *least squares method* and *point matching method*.

3.1.2.1 Galerkin method

In the Galerkin method (GM) testing functions are selected so that they are equal to the basis functions, i.e.

$$\forall_{i=1,\ldots,n} \quad w_i = u_i \tag{3.32}$$

In consequence, the elements of matrices $\underline{\underline{A}}$ and $\underline{\underline{B}}$ from eigenproblem (3.30) are defined as

$$A_{ji} = \Big(\mathbf{A}u_i, u_j\Big) \qquad , \qquad B_{ji} = \Big(\mathbf{B}u_i, u_j\Big) \tag{3.33}$$

Therefore, the matrix eigenproblem becomes symmetric if the operators **A** and **B** are symmetric as well. This generalized problem can become a standard one, if functions $\{u_i\}_{i=1,\ldots,n}$ are **B**-orthonormal.

It should be noted, that the eigenproblems resulting from the Galerkin method are identical to the eigenproblems resulting from the standard Rayleigh-Ritz procedure in the symmetric case (Sec. 3.1.1.1) and from the local potential method in the nonsymmetric case (Sec. 3.1.1.2). In this sense, the GM and RR methods are fully equivalent.

3.1.2.2 Least squares method

In the least squares method (LS) testing functions are selected so that

$$\forall_{i=1,\ldots,n} \quad w_i = \mathbf{A}u_i \tag{3.34}$$

It implies that the elements of matrices $\underline{\underline{A}}$ and $\underline{\underline{B}}$ are defined as

$$A_{ji} = \Big(\mathbf{A}u_i, \mathbf{A}u_j\Big) \qquad , \qquad B_{ji} = \Big(\mathbf{B}u_i, \mathbf{A}u_j\Big) \tag{3.35}$$

In consequence, $\underline{\underline{A}}$ is symmetric and positive definite, while $\underline{\underline{B}}$ is nonsymmetric.

3.1.2.3 Point matching method

In the point matching method, called also the *collocation technique*, testing functions are selected so that

$$\forall_{i=1,\ldots,n} \quad w_i = \delta(r - r_i) \tag{3.36}$$

where $\delta(r-r_i)$ is the Dirac delta function at point i, located in the domain at coordinates r_i. In this case, the dot products defining elements of matrices $\underline{\underline{A}}$ and $\underline{\underline{B}}$ correspond to the values of basis functions at specific sampling points $\{r_j\}_{j=1,\ldots,n}$, i.e.

$$A_{ji} = \mathbf{A}u_i(r_j) \qquad , \qquad B_{ji} = \mathbf{B}u_i(r_j) \tag{3.37}$$

A main disadvantage of this method is that the condition number of the resulting matrix strongly depends on the choice of sampling points. Therefore, other choices of testing functions are preferred.

3.1.3 Finite element method

The finite element method (FEM) [2,3,7,19,21,23,24,30–34,38,44,48,55,56,66–68,89–92, 107,114,117,120,121] is an entire subdomain expansion method exploiting the concepts of the Rayleigh-Ritz or the Galerkin methods for specific type of locally defined basis (and testing) functions.

The FEM is based on a discretization of the domain, say Ω. The domain is divided into m disjoint small subdomains Ω_i (called *finite elements*), in a way that their union is the entire domain Ω, i.e.

$$\forall_{i\neq j} \quad \Omega_i \cap \Omega_j = \varnothing \qquad \text{and} \qquad \bigcup_{i=1}^{m} \Omega_i = \Omega \tag{3.38}$$

In the most popular finite element approach, (called *nodal finite elements*) the problem is formulated in terms of the unknown function v at *nodal points*, located on the boundaries or inside the elements. In the simplest case, shown in Fig. 3.1, nodes are located at the vertices of the elements.

The solution v of the problem is constructed as a superposition of functions v^i, locally defined over each element i

$$v = \bigcup_{i=1}^{m} v^i \tag{3.39}$$

where each function v^i is defined only inside the element i.

The local solution v^i is approximated by a linear combination of *interpolation functions* (or *shape functions*) N_j^i

$$v^i = \sum_{j=1}^{n^i} \alpha_j^i N_j^i \tag{3.40}$$

where n^i is the number of nodes within the element i.

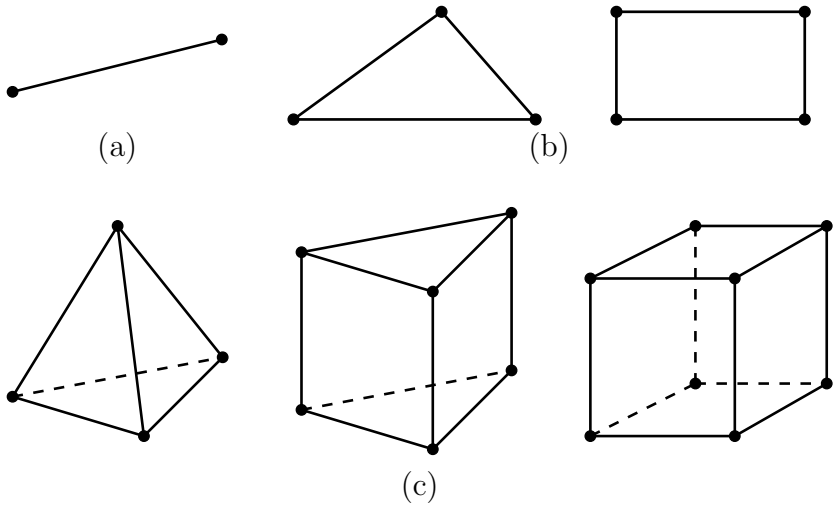


Figure 3.1: Basic finite elements: one-dimensional (a), two-dimensional (b), and three-dimensional (c).

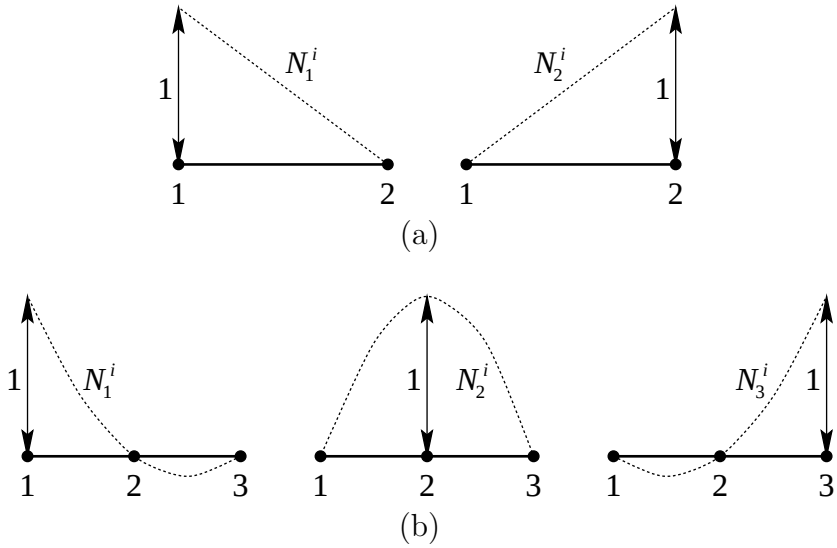


Figure 3.2: One-dimensional interpolation functions: (a) linear, (b) quadratic.

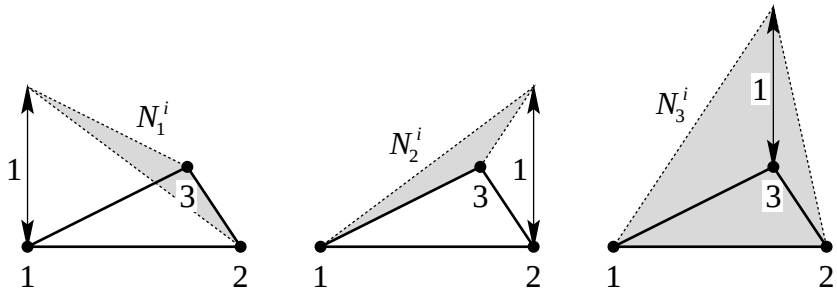


Figure 3.3: Two-dimensional linear interpolation functions for triangular element.

The interpolation functions $\{N_j^i\}_{j=1,\ldots,n^i}$ are constructed so that they vanish outside the element i, and within this element they are polynomials satisfying the following property

$$N_j^i(r_k^i) = \delta_{jk} = \begin{cases} 1, & \text{for} \quad j = k \\ 0, & \text{for} \quad j \neq k \end{cases} \tag{3.41}$$

where r_k^i denotes coordinates of k-th node in the element i. Examples of one- and two-dimensional interpolation functions are shown in Figs. 3.2 and 3.3.

Due to property (3.41), the solution v (3.39) at the k-th nodal point of the element i (of coordinates r_k^i) can be evaluated as the following union

$$v(r_k^i) = \bigcup_{i=1}^{m} \sum_{j=1}^{n^i} \alpha_j^i N_j^i(r_k^i) = \alpha_k^i \tag{3.42}$$

It means, that the expansion coefficient corresponding to the interpolation function associated to the node k of element i is precisely the value of the function v at this point.

In order to apply Galerkin or RR method, the function v should be written in the form of global functions expansion. Substituting (3.40) into (3.39) and using some global numbering scheme [33] we get

$$v = \bigcup_{i=1}^{m} \sum_{j=1}^{n^i} \alpha_j^i N_j^i = \sum_{k=1}^{n} \alpha_k u_k \tag{3.43}$$

where n is the total number of nodes within the domain Ω, α_k is the expansion coefficient corresponding to the k-th node (in the global numbering scheme) and u_k is the basis function, which is the union of all interpolation functions associated with the k-th node. Examples of such one- and two-dimensional basis functions are shown in Fig. 3.4.

Application of either the GM or RR approach to an operator problem of form (3.1) and using basis functions described by (3.43) leads to a matrix eigenvalue problem in the form of (3.9) (or (3.30)). Since the computation of matrix elements involves calculation of inner products with the functions being only locally nonzero, the matrices $\underline{\underline{A}}$ and $\underline{\underline{B}}$ are sparse.

It should be noted, that the calculation of each inner product is equivalent to a sum of inner products in the form of $\left(\mathbf{A}N_j^i, N_l^k\right)$ and $\left(\mathbf{B}N_j^i, N_l^k\right)$. Therefore, the incorporation of low order basis functions for formulations with higher order differential operators may result in distribution functions. In order to avoid problems with computing integrals, the order of used basis functions cannot be lower than the order of differential operators involved.

Finite element method can also be formulated in terms of a vector basis defined over the edges of elements rather than the scalar basis defined over the nodal points. Such formulation is called *edge elements* [7, 48, 117, 121]. It has the advantage that it a-priori eliminates spurious solutions caused by the violation of divergence equation.

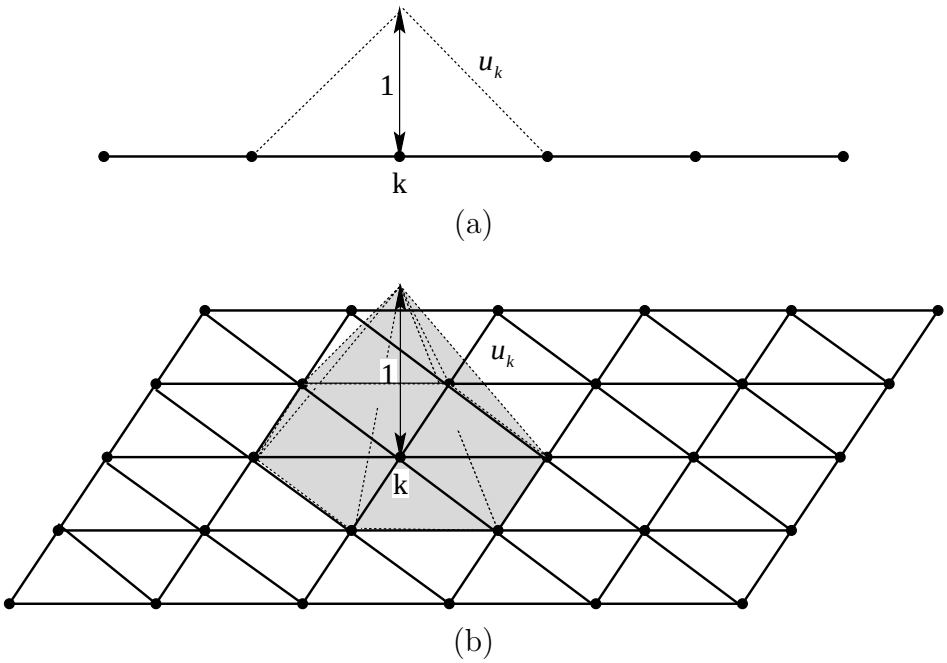


Figure 3.4: Linear basis functions, corresponding to node k, for (a) one-dimensional elements, and (b) two-dimensional triangular elements.

3.1.4 Finite difference frequency domain method

The finite difference frequency domain (FDFD) method is based on discretization of the domain and differential operators. The operators are approximated by a set of algebraic equations relating the value of an unknown field at a discrete point in the domain to the values at some neighboring points. Many different versions of FDFD method were described in the literature [1, 9, 11–14, 18, 25, 37, 40, 46, 52, 53, 64, 77, 82, 84, 94, 101–106, 112, 113, 119, 122–125, 127]. They differ mainly in the way of approximating the fields and their derivatives, and in the form of operator being discretized.

3.1.4.1 Finite difference approximation

In the FDFD, the domain is discretized using a set of grid points, defining a *grid space*. This is similar to the discretization used in the FEM method, in a sense that the grid points can be nodal points defining elements in the FEM. One of the factors influencing the process of approximation is the type of grid used. Typical *nonuniform* (*graded*) grids are presented in Fig. 3.5. The main advantage of such grids is simplicity of implementation of finite differences and conservation of symmetry property of the operators after the conversion to the matrix problem.

In general, the domain can be discretized using elements of various shapes and sizes forming *unconstrained (irregular)* or *nested* meshes, shown in Fig. 3.6. Such grids can better conform to complex geometries at the cost of more difficult implementation and/or loss of matrix symmetry in symmetric problems. Another aspect of symmetry will be

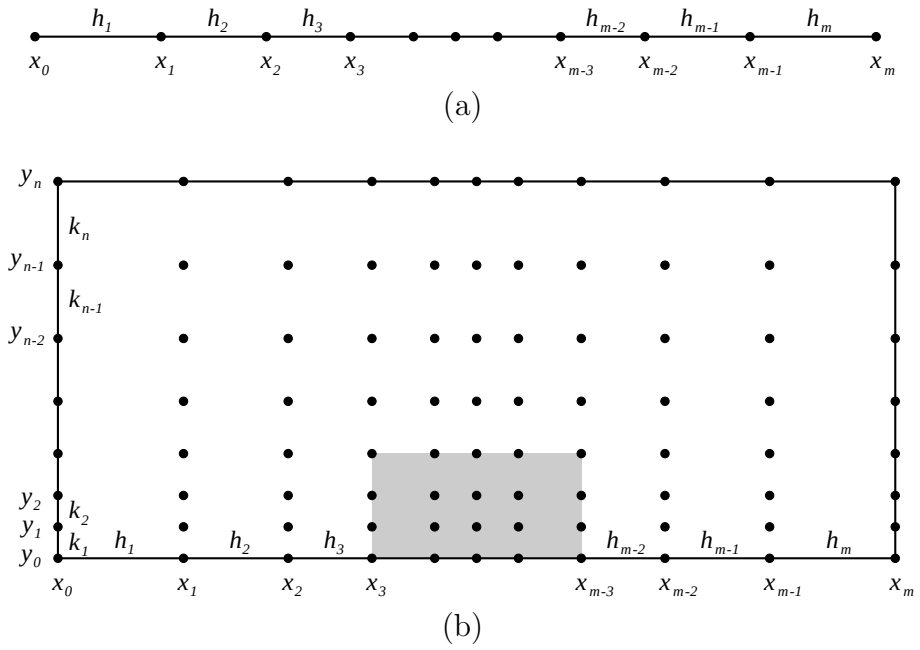


Figure 3.5: Typical (a) one-dimensional, and (b) two-dimensional nonuniform (graded) grids in the Cartesian coordinates.

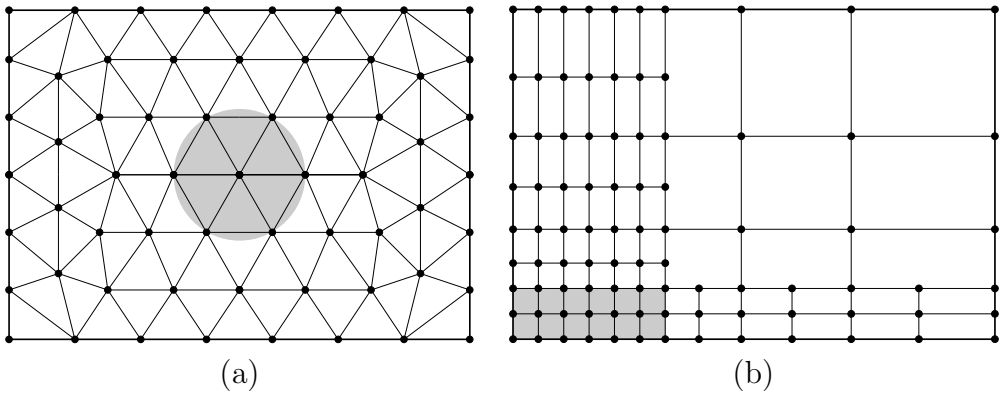


Figure 3.6: Examples of two-dimensional (a) unconstrained (irregular) and (b) nested grid.

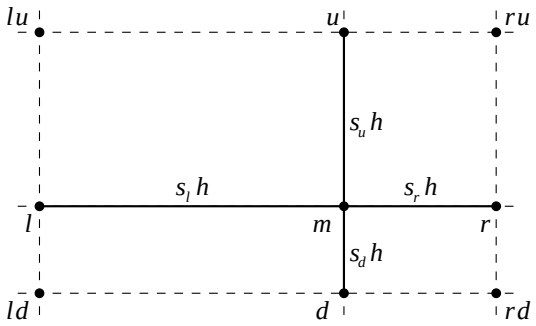


Figure 3.7: Two-dimensional nonuniform rectangular grid in the vicinity of point m.

discussed in Sec. 3.1.4.2.

Consider the simplest two-dimensional case with a nonuniform rectangular grid in the vicinity of an arbitrary point m. This situation is shown in Fig. 3.7. At point m, finite difference approximations of derivatives of an unknown function v can be derived using the values of v at the neighboring points. Expanding v in the points r, u, l, and d into Taylor series about the point m we get [77]

$$v_r = v_m + s_r h \left(\frac{\partial v}{\partial x}\right)_m + \frac{s_r^2 h^2}{2}\left(\frac{\partial^2 v}{\partial x^2}\right)_m + \mathcal{O}(h^3) \tag{3.44}$$

$$v_u = v_m + s_u h \left(\frac{\partial v}{\partial y}\right)_m + \frac{s_u^2 h^2}{2}\left(\frac{\partial^2 v}{\partial y^2}\right)_m + \mathcal{O}(h^3) \tag{3.45}$$

$$v_l = v_m - s_l h \left(\frac{\partial v}{\partial x}\right)_m + \frac{s_l^2 h^2}{2}\left(\frac{\partial^2 v}{\partial x^2}\right)_m + \mathcal{O}(h^3) \tag{3.46}$$

$$v_d = v_m - s_d h \left(\frac{\partial v}{\partial y}\right)_m + \frac{s_d^2 h^2}{2}\left(\frac{\partial^2 v}{\partial y^2}\right)_m + \mathcal{O}(h^3) \tag{3.47}$$

Eliminating $\partial^2 v/\partial x^2$ and $\partial^2 v/\partial y^2$ between (3.44), (3.46) and (3.45), (3.47) we first obtain partial derivatives of v in the form of

$$\left(\frac{\partial v}{\partial x}\right)_m = \frac{1}{h}\frac{v_r - v_l}{s_r + s_l} + \mathcal{O}(h) \tag{3.48}$$

$$\left(\frac{\partial v}{\partial y}\right)_m = \frac{1}{h}\frac{v_u - v_d}{s_u + s_d} + \mathcal{O}(h) \tag{3.49}$$

and, analogously, eliminating $\partial v/\partial x$ and $\partial v/\partial y$ we get the following second partial derivatives of v

$$\left(\frac{\partial^2 v}{\partial x^2}\right)_m = \frac{2}{h^2}\left(\frac{v_r}{s_r(s_r + s_l)} - \frac{v_m}{s_r s_l} + \frac{v_l}{s_l(s_r + s_l)}\right) + \mathcal{O}(h) \tag{3.50}$$

$$\left(\frac{\partial^2 v}{\partial y^2}\right)_m = \frac{2}{h^2}\left(\frac{v_u}{s_u(s_u + s_d)} - \frac{v_m}{s_u s_d} + \frac{v_d}{s_d(s_u + s_d)}\right) + \mathcal{O}(h) \tag{3.51}$$

Mixed derivatives can be derived in the same manner.

Considerable simplifications arise for uniform grid, i.e. $s_r = s_l = s_x$ and $s_u = s_d = s_y$. In this case, first and second derivatives can be found as

$$\left(\frac{\partial v}{\partial x}\right)_m = \frac{v_r - v_l}{2 s_x h} + \mathcal{O}(h^2) \tag{3.52}$$

$$\left(\frac{\partial v}{\partial y}\right)_m = \frac{v_u - v_d}{2 s_y h} + \mathcal{O}(h^2) \tag{3.53}$$

and

$$\left(\frac{\partial^2 v}{\partial x^2}\right)_m = \frac{v_r - 2v_m + v_l}{s_x^2 h^2} + \mathcal{O}(h^2) \tag{3.54}$$

$$\left(\frac{\partial^2 v}{\partial y^2}\right)_m = \frac{v_u - 2v_m + v_d}{s_y^2 h^2} + \mathcal{O}(h^2) \tag{3.55}$$

Note, that the error of approximations (3.52), (3.53) (3.54), and (3.55) is reduced by one order. Such type of approximations is called *second order central differences* and, for regular grid, it offers the lower order of error compared to other types of approximations such as *forward* or *backward* differences [99,119]. Higher order approximations, i.e. fourth, can also be used [99,119]. They result in higher accuracy at the cost of much more complex implementation of boundary conditions.

In electromagnetic problems involving vector fields, finite differences can be implemented in two ways. One approach is to use a single collocated mesh for electric and magnetic fields, while the second approach uses a dual grid system. Below, we briefly discuss both techniques.

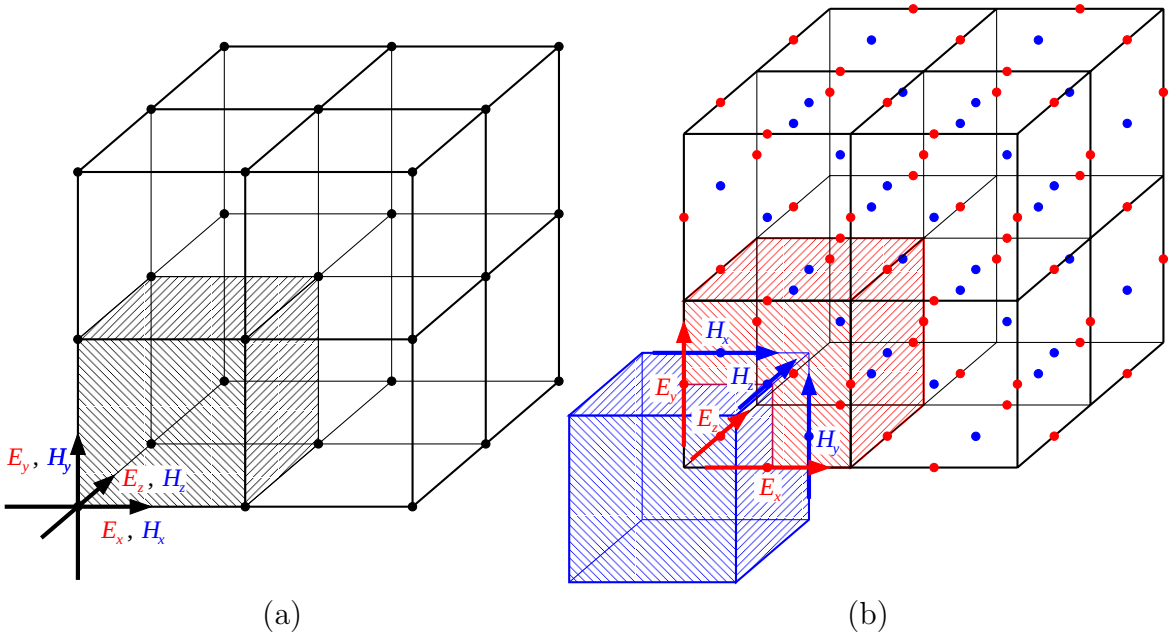


Figure 3.8: Examples of three-dimensional (a) single and (b) dual grid.

Single grid approach. In this approach, a single collocated mesh, shown in Fig. 3.8(a), is used for all field components. All electric and magnetic fields of interest are defined in the same points of the domain. This approach is sometimes called *condensed node* [1].

Main disadvantages of the single grid approach is that application of unconstrained grids may lead to the loss of matrix symmetry and that it may cause difficulties in dealing with discontinuities.

Dual grid approach. This approach is based on dual Yee's mesh [127]. It uses dual grid system, shown in Fig. 3.8(b), defined so that electric and magnetic cells are shifted half a cell one with respect to the other. In consequence, the electric and magnetic fields are defined along the edges of electric and magnetic cells at the points where the cell edges intersect with dual magnetic or electric cell walls, respectively.

The advantage of this approach is that even unconstrained meshes used may lead to the symmetric operator matrices. Additionally, discontinuities are easier to deal with.

3.1.4.2 Formulation of operator problem

Depending on the final form of operator equation, two main approaches can be distinguished in finite difference techniques. The first one is based on the Rayleigh-Ritz method and variational formulation of the operator equation, while the second one relies on direct discretization of the operator problem. Both formulations can lead to standard eigenproblems of form (3.1) with highly diagonally structured matrix. However, the higher the order of approximations used, the greater the number of diagonals.

RR/Galerkin approach. This approach [77, 119] is based on the RR method applied for an adequate variational functional, which results in the matrix eigenproblem of form (3.9) or (3.30). The elements of matrices $\underline{\underline{A}}$ and $\underline{\underline{B}}$ are then expressed by inner products (3.10) or (3.33), involving differential operators. In the case of rectangular grid, field derivatives can be approximated by central differences. It should be noted that this approach is analogous to the Galerkin method applied for operator problem with piecewise linear basis and testing functions.

Application of the RR/Galerkin procedure shifts the FDFD method very close to the FEM method. In the case when the same rectangular grid is used for both methods, the only difference relies on the type of the approximation of the field and corresponding derivatives used [69]. Application of the RR/Galerkin variational approach has been reported in [18, 106] in context of the formulation for longitudinal E_z and H_z components and in [84] in context of a full vector $\vec{H}$ formulation. In all the reported cases a single grid was involved.

Very important advantage of this approach is that it preserves the symmetry of the operator equation, even for nested grids and curved boundaries [119].

Direct discretization. This approach [77, 119] relies on direct discretization of operators in the eigenproblem, so that the derivatives are approximated by finite differences. In this approach a standard matrix eigenvalue problem is usually obtained.

In order to preserve the symmetry of the operators, boundary conditions should be adequately implemented, e.g. the Neumann boundary condition using the concept of fictious grid points [77]. In order to preserve the symmetry for the case of curved boundaries between different media, a concept of effective permittivity/permeability [13, 16, 50, 64, 122–124] should be incorporated. It allows one to use constrained meshes even in the case of boundaries which do not conform to the grid.

The direct discretization is the most frequently used method. Various implementations of this method used single and dual grids. The single grid approach was implemented in various scalar [11, 12, 25, 40, 113], longitudinal [46], transverse [1, 14, 37, 104, 105], and full vector formulations [1]. The dual grid approach was used in various scalar [64], transverse [64, 123, 124], and full vector formulations [9, 13, 101–103, 122, 125].

It should be noted that for nested meshes, such as shown in Fig. 3.6(b), direct discretization method always leads to a nonsymmetric problem. Another important observation is that application of the direct discretization method for grids defined in cylindrical coordinates leads to nonsymmetric eigenproblems [40, 64, 113, 119].

3.2 Hybrid methods

Hybrid methods are based on the Rayleigh-Ritz procedure or method of moments, in which trial functions are the solutions for simplified structures or the solutions for specific (ω, β) points. These trial solutions, used in the expansion of the unknown field, can be evaluated analytically or by means of any standard method. This concept can be viewed as a generalization of the concept of an electromagnetic basis described in [51, 57, 58, 81], where entire domain basis functions were optimized in order to reduce the number of the functions required for certain accuracy of approximation of the field in an analyzed structure.

It should be noted, that application of specific entire domain basis functions in the expansion can also lead to considerable simplification of relations describing final operator matrix elements.

3.2.1 Coupled mode method

The coupled mode method [5, 43] is a kind of entire domain expansion method, which is based on RR procedure applied to an adequate functional (*perturbation formula*). The fields are expanded into a series of *basis fields*, which are usually computed for a *basis structure* of the same geometry, filled with isotropic materials. It leads to elimination of differential operators from the final eigenproblem.

This method is very useful when the medium within the structure has complex parameters (i.e. is anisotropic, gyrotropic, lossy, etc.) and the analysis involving any classical approach leads to computationally intensive problems.

Due to perturbational character of the coupled mode method, it can be efficiently used only for the structures of parameters that are relatively weakly perturbed in relation to the basis structure.

3.2.1.1 Resonator problems

In the case of resonator analysis, formulation (2.22) is usually used as a starting point. Assuming that the structure is lossless, this formulation is symmetric and adequate functional can be found using (3.2) in the form of

$$\left(\nabla \times \underline{\underline{\mu}}^{-1} \cdot \nabla \times \vec{E}, \vec{E}\right) = \omega^2 \left(\underline{\underline{\epsilon}} \cdot \vec{E}, \vec{E}\right) \tag{3.56}$$

Using vector identities and properties of $\vec{E}$ field on PEC or PMC screens (3.56) can be written as [43]

$$\int_V \underline{\underline{\mu}}^{-1} \cdot (\nabla \times \vec{E}^*) \cdot (\nabla \times \vec{E})\, dv = \omega^2 \int_V \vec{E}^* \cdot \underline{\underline{\epsilon}} \cdot \vec{E}\, dv \tag{3.57}$$

The application of RR procedure involves the expansion

$$\vec{E} = \sum_{i=1}^{n} \alpha_i \vec{e}_i \tag{3.58}$$

where basis functions $\vec{e}_i$ obey the wave equation (2.22), i.e.

$$\nabla \times \underline{\underline{\mu}}_i^{-1} \cdot \nabla \times \vec{e}_i = \omega_i^2 \underline{\underline{\epsilon}}_i \cdot \vec{e}_i \tag{3.59}$$

where $\underline{\underline{\epsilon}}_i$ and $\underline{\underline{\mu}}_i$ are assumed to be those for lossless media (in the basis guide).

The resulting generalized eigenproblem is of form (3.9) with elements defined by

$$\begin{aligned} A_{ij} &= \int_V \underline{\underline{\mu}}_j^{-1} \cdot (\nabla \times \vec{e}_i^*) \cdot (\nabla \times \vec{e}_j)\, dv = \omega_j^2 \int_V \vec{e}_i^* \cdot \underline{\underline{\epsilon}}_j \cdot \vec{e}_j\, dv \\ B_{ij} &= \int_V \vec{e}_i^* \cdot \underline{\underline{\epsilon}} \cdot \vec{e}_j\, dv \end{aligned} \tag{3.60}$$

It can be seen, that if we consider basis functions computed for the same basis structure ($\underline{\underline{\epsilon}}_i = \underline{\underline{\epsilon}}_j = \underline{\underline{\epsilon}}_b$), they are orthogonal with respect to **B**-inner product (weighted with $\underline{\underline{\epsilon}}_b$), i.e.

$$\left(\vec{e}_j, \vec{e}_i\right)_{\mathbf{B}} = \int_V \vec{e}_i^* \cdot \underline{\underline{\epsilon}}_b \cdot \vec{e}_j\, dv \tag{3.61}$$

and matrix $\underline{\underline{B}}$ becomes diagonal. Such a problem is easy transformable to a standard one.

The analogous dual formulation, for $\vec{H}$ field, can be derived, starting from (2.23).

3.2.1.2 Waveguide problems

In the case of waveguide analysis, the following perturbation formula can be derived from symmetric (for lossless media and real ω) formulation (2.61)

$$\begin{aligned} &\left(\omega \underline{\underline{\epsilon}} \cdot \vec{E} + j\nabla_t \times \vec{H}, \vec{E}\right) + \left(\omega \underline{\underline{\mu}} \cdot \vec{H} - j\nabla_t \times \vec{E}, \vec{H}\right) \\ &\qquad = \beta \left[\left(-\hat{z} \times \vec{H}, \vec{E}\right) + \left(\hat{z} \times \vec{E}, \vec{H}\right)\right] \end{aligned} \tag{3.62}$$

Using the definition of inner product (3.62) can be written in the following form [43]

$$\begin{aligned} &\int_S \left[\vec{E}^* \cdot (\omega \underline{\underline{\epsilon}} \cdot \vec{E} + j\nabla_t \times \vec{H}) + \vec{H}^* \cdot (\omega \underline{\underline{\mu}} \cdot \vec{H} - j\nabla_t \times \vec{E})\right] ds \\ &\qquad = \beta \int_S \hat{z} \cdot (\vec{E} \times \vec{H}^* + \vec{E}^* \times \vec{H})\, ds \end{aligned} \tag{3.63}$$

In the RR procedure the following expansion is used

$$\left[\begin{array}{c} \vec{E} \\ \vec{H} \end{array} \right] = \sum_{i=1}^{n} \alpha_i \left[\begin{array}{c} \vec{e} \\ \vec{h} \end{array} \right]_i = \sum_{i=1}^{n} \alpha_i \left[\begin{array}{c} \vec{e}_i \\ \vec{h}_i \end{array} \right] \tag{3.64}$$

where basis functions $\vec{e}_i$ and $\vec{h}_i$ fulfill adequate equations derived from (2.61)

$$\omega \underline{\underline{\epsilon}}_i \cdot \vec{e}_i + j\nabla_t \times \vec{h}_i = -\beta_i \hat{z} \times \vec{h}_i \tag{3.65}$$

$$\omega \underline{\underline{\mu}}_i \cdot \vec{h}_i - j\nabla_t \times \vec{e}_i = \beta_i \hat{z} \times \vec{e}_i \tag{3.66}$$

where $\underline{\underline{\epsilon}}_i$ and $\underline{\underline{\mu}}_i$ characterize lossless media.

The resulting generalized eigenproblem is of form (3.9) with elements defined by [5,43]

$$\begin{aligned} A_{ij} &= \beta_j B_{ij} + \omega \int_S [\vec{e}_i^* \cdot (\underline{\underline{\epsilon}} - \underline{\underline{\epsilon}}_j) \cdot \vec{e}_j + \vec{h}_i^* \cdot (\underline{\underline{\mu}} - \underline{\underline{\mu}}_j) \cdot \vec{h}_j] \, ds \\ B_{ij} &= \int_S \hat{z} \cdot (\vec{e}_i^* \times \vec{h}_j + \vec{e}_j \times \vec{h}_i^*) \, ds \end{aligned} \tag{3.67}$$

This problem can be reduced into a standard one, if we take the set of basis functions $[\vec{e}_i, \vec{h}_i]^T$ computed for the same lossless basis structure (and the same real ω). Since eigenproblem (2.61) is symmetric for lossless media, the basis functions are orthogonal with respect to the **B**-inner product, defined by

$$\begin{aligned} \left(\left[\begin{array}{c} \vec{e}_j \\ \vec{h}_j \end{array} \right], \left[\begin{array}{c} \vec{e}_i \\ \vec{h}_i \end{array} \right] \right)_{\mathbf{B}} &= \left(-\hat{z} \times \vec{h}_j, \vec{e}_i \right) + \left(\hat{z} \times \vec{e}_j, \vec{h}_i \right) \\ &= \int_S \hat{z} \cdot (\vec{e}_i^* \times \vec{h}_j + \vec{e}_j \times \vec{h}_i^*) \, ds \end{aligned} \tag{3.68}$$

Therefore matrix $\underline{\underline{B}}$ becomes diagonal.

3.2.2 Eigenfunction expansion methods

This family of methods is especially useful for fast evaluation of dispersion characteristics of waveguide structures. Eigenfunction expansion (EE) methods are based on application of the method of moments to any waveguide eigenproblem with basis and testing functions chosen in a special way. The entire domain basis functions are the solutions of the eigenproblem for particular frequencies or propagation constants, while the testing functions are corresponding solutions of the transposed problem. In contrast to standard method of moments involving usual harmonic basis functions, expansion functions involved in EE methods fulfill all interior boundary conditions and may be regarded as a perfect electrodynamic basis [51, 57, 58, 81]. As a result, the number of trial functions, which should be taken to obtain a required accuracy, is relatively small. It leads to very small matrix problem for each desired frequency or propagation constant point. It should be noted, that basis functions can also be computed at critical points ($\omega = 0$ or $\beta = 0$), where the solution is less expensive. Therefore the whole EE method can be yet more efficient.

In fact, three specific techniques incorporating eigenfunction expansion have been proposed recently. The first one [95], uses basis (and testing) functions which are evaluated for the cutoff case. The second approach [85], incorporates a technique called the *asymptotic waveform evaluation*, which uses the Taylor series or Padé approximation to interpolate particular modes over a frequency band around a selected frequency point. The most general EE approach, which can employ the eigenfunctions computed at arbitrary points on the dispersion diagram has been developed with author's participation and was reported in texts [86–88]. Derivation of the algorithms obtained by means of this latter method is shortly summarized below.

3.2.2.1 Formulation of matrix eigenproblem

Consider the problem of analysis of a lossless and (strictly) bidirectional waveguide with permittivity and permeability tensors given by (2.63). The wave propagation in such a guide may be described by any of the equations (2.76–2.79).

Express problems (2.76) and (2.78) in the following operator form

$$\mathbf{L}u + \omega^2\mathbf{G}u - \beta^2\mathbf{S}u = 0 \tag{3.69}$$

where u denotes $\vec{E}_t$ or $\vec{D}_t$, $\mathbf{L}$ represents differential part of the operators and $\mathbf{S}$, $\mathbf{G}$ are operators that either involve the media parameters (e.g. $\mathbf{S}$ for $\vec{D}_t$ or $\mathbf{G}$ for $\vec{E}_t$) or are simple $\hat{z}\times$ operators.

Basis functions are the solutions of the above problem at n discrete points so that we know triads $\{\omega_i^2, \beta_i^2, u_i(\omega_i, \beta_i)\}_{i=1,\ldots,n}$, which satisfy equation

$$\mathbf{L}u_i = -\omega_i^2\mathbf{G}u_i + \beta_i^2\mathbf{S}u_i \tag{3.70}$$

within domains determined by the boundary conditions.

Since equations (2.77) and (2.79) are respectively transpositions of (2.76) and (2.78), corresponding solutions of the first two of the above equations (i.e. $\vec{H}_t$ or $\vec{B}_t$) are taken as testing functions w. In fact, these problems need not to be solved explicitly because the i-th testing function ($\vec{h}_{ti}$ or $\vec{b}_{ti}$) can be evaluated with Maxwell equations applied for the i-th basis function ($\vec{e}_{ti}$ or $\vec{d}_{ti}$).

The application of the method of moments with the expansion into a series of basis functions

$$u(\omega, \beta) = \sum_{i=1}^{n} \alpha_i(\omega, \beta) u_i \tag{3.71}$$

and corresponding testing functions w_i leads to the following matrix equation

$$\underline{\underline{G}}\,(\omega^2\underline{\underline{I}} - \underline{\underline{\Omega}}^2)\underline{a} = \underline{\underline{S}}\,(\beta^2 - \underline{\underline{Z}}^2)\underline{a} \tag{3.72}$$

where $\underline{\underline{\Omega}} = \mathrm{diag}[\omega_i^2]$, $\underline{\underline{Z}}^2 = \mathrm{diag}[\beta_i^2]$, $\underline{a} = [\alpha_1, \alpha_2, \ldots, \alpha_n]^T$ and the elements of matrices $\underline{\underline{G}}$ and $\underline{\underline{S}}$ are given by

$$G_{ki} = \Big(\mathbf{G}u_i, v_k\Big) = \int_S \hat{z} \cdot (\vec{d}_{ti} \times \vec{b}_{tk})\, ds \tag{3.73}$$

$$S_{ki} = \Big(\mathbf{S}u_i, v_k\Big) = \int_S \hat{z} \cdot (\vec{e}_{ti} \times \vec{h}_{tk})\, ds \tag{3.74}$$

where S denotes the cross-section of the guide. It should be noted, that this algorithm requires the calculation of integrals involving only the z-components of electromagnetic momenta (3.73) and Poynting vectors (3.74). Equation (3.72) can easily be transformed in order to obtain a generalized matrix eigenproblem with either ω^2 or β^2 being an eigenvalue.

In general, the points for calculating the basis, i.e. ω_i and the corresponding β_i can be arbitrary selected from the points of dispersion diagram. Consider a few particular choices. Suppose all expansion functions are calculated for the same β_0. Points B, F, G in Fig. 5.18 on page 94 are an example of the points where such modal fields are calculated for $\beta_0 = 0$. Using the biorthogonality property of the basis and testing functions it is easy to show that matrix $\underline{\underline{G}}$ becomes diagonal. Moreover, $\underline{\underline{Z}}^{\,2} = \beta_0^2 \underline{\underline{I}}$ and generalized problem (3.72) can easily be transformed to the the following standard β-eigenproblem

$$\underline{\underline{A}}\,\underline{a} = \frac{1}{\beta^2 - \beta_0^2}\underline{a} \tag{3.75}$$

where the elements of matrix $\underline{\underline{A}}$ are given by an extremely simple formula

$$A_{ki} = \frac{1}{\omega^2 - \omega_k^2} s_{ki} \tag{3.76}$$

with

$$s_{ki} = \frac{\int_S \hat{z} \cdot (\vec{e}_{ti} \times \vec{h}_{tk})\, ds}{\int_S \hat{z} \cdot (\vec{d}_{tk} \times \vec{b}_{tk})\, ds} \tag{3.77}$$

Alternatively, one may transform (3.72) to the standard ω-eigenproblem

$$\underline{\underline{B}}\,\underline{a} = \omega^2 \underline{a} \tag{3.78}$$

with

$$B_{ki} = (\beta^2 - \beta_0^2) s_{ki} + \omega_k^2 \delta_{ki} \tag{3.79}$$

where δ_{ik} is the Kronecker symbol.

Another pair of algorithms can be obtained, if eigensolutions of (3.70), used as basis functions, are evaluated for $\omega = \omega_0$. Points C, H, I in Fig. 5.18 from page 94 are an example of the points where such modal fields are calculated for $\omega_0 = 2\pi 10\text{Grad/s}$. Such a choice of basis functions leads to the diagonalization of matrix $\underline{\underline{S}}$ and yields two algorithms given by the standard matrix eigenproblems

$$\underline{\underline{C}}\,\underline{a} = \frac{1}{\omega^2 - \omega_0^2}\underline{a} \tag{3.80}$$

$$\underline{\underline{D}}\,\underline{a} = \beta^2 \underline{a} \tag{3.81}$$

with the elements of matrices $\underline{\underline{C}}$ and $\underline{\underline{D}}$ given by

$$C_{ki} = \frac{1}{\beta^2 - \beta_k^2} g_{ki} \tag{3.82}$$

$$D_{ik} = (\omega^2 - \omega_k^2) g_{ki} + \beta_k^2 \delta_{ki} \tag{3.83}$$

Table 3.1: Classification of eigenfunction expansion algorithms.

Algorithm	Unknown	Basis	Eigenvalue	Eqn.
β-GS	$\beta(\omega)$	$\{u_i, \omega_i, \beta_i\}$	β^2	(3.72)
ω-GS	$\omega(\beta)$	$\{u_i, \omega_i, \beta_i\}$	ω^2	(3.72)
β-S	$\beta(\omega)$	$\{u_i, \omega_i, \beta_0\}$	$(\beta^2 - \beta_0^2)^{-1}$	(3.75)
ω-S	$\omega(\beta)$	$\{u_i, \omega_i, \beta_0\}$	ω^2	(3.78)
ω-G	$\omega(\beta)$	$\{u_i, \omega_0, \beta_i\}$	$(\omega^2 - \omega_0^2)^{-1}$	(3.80)
β-G	$\beta(\omega)$	$\{u_i, \omega_0, \beta_i\}$	β^2	(3.81)

with g_{ki} defined as

$$g_{ki} = \frac{\int_S \hat{z} \cdot (\vec{d}_{ti} \times \vec{b}_{tk})\, ds}{\int_S \hat{z} \cdot (\vec{e}_{tk} \times \vec{h}_{tk})\, ds} \tag{3.84}$$

Equations (3.72), (3.75), (3.78), (3.80), and (3.81) describe six algorithms which differ one from another by the choice of basis functions and the selection of an unknown and a parameter. For convenience all algorithms are summarized in Table 3.1. The following convention is used to designate the algorithms. The first letter denotes the type of dispersion characteristics generated by the algorithm while one or two letters following the dash indicate the quantity (S for Poynting vectors and G for electromagnetic momenta) required to evaluate the matrix elements.

Note, that the procedures described above can be adopted to other formulations than (2.76) or (2.78), forming new families of algorithms differing one from another by the choice of the basis functions.

3.2.2.2 Speedup of eigenfunction expansion algorithms

To show the advantages of using a hybrid approach let us discuss the speedup that can be expected in the eigenfunction expansion algorithm.

In the most general case (algorithms β-GS and ω-GS) the calculation of the matrices elements involve computation of the coupling between electromagnetic field of basis modes, in the form of momenta or Poynting vectors. The evaluation of the matrices can be further simplified by a special choice of the basis and testing functions and the application of orthogonality relations. Moreover, it can be shown [86] that in the practical implementation of the algorithms the momenta or/and Poynting vector matrices are evaluated only once before the ω- or β-sweep. Therefore, this cost can be neglected and the total computational effort of the algorithms is determined by the cost of the evaluation of the basis functions and solution of a small dense matrix eigenvalue problem in as many points as required.

The basis can be evaluated with an arbitrary numerical or analytical technique. Since it is calculated at a few frequency or propagation constant points, at the most, even a

time consuming method may be used to this end. If the computational effort for solving the small dense eigenfunction expansion problem is lower than the computation workload in a standard algorithm, the new approach gives a speedup of

$$S_{EE} = \frac{nt_s}{mt_{sb} + nt_{ee}} \tag{3.85}$$

where n is a number of computation points, t_s is the time of a single solution in a standard approach, m is a number of the points where the basis fields are evaluated, t_{sb} is the solution time at the single basis point, and t_{ee} is the solution time of the small dense problem at one point. If the basis is determined using the standard method ($t_{sb} = t_s$) and $t_s \gg t_{ee}$ expression (3.85) tends to

$$S_{EE} \approx \frac{n}{m} \tag{3.86}$$

It is evident that the time savings may be significant, especially when the number of points n is large.

Chapter 4

Solution of matrix eigenvalue problems

In general, two kinds of matrix eigenvalue problems arise from the techniques described in the previous chapter: standard eigenproblems of the form

$$\underline{\underline{A}}\,\underline{v} = \lambda \underline{v} \tag{4.1}$$

and generalized eigenproblems of the form

$$\underline{\underline{A}}\,\underline{v} = \lambda \underline{\underline{B}}\,\underline{v} \tag{4.2}$$

A number of numerical techniques can be applied to each given particular problem (4.1) or (4.2). A detailed description of several the most frequently used solution methods can be found in Appendix A. Here we only discuss the most important features of these methods. We can distinct two general classes of algorithms, i.e. the ones based on *matrix transformations* and the *iterative* ones. The algorithms representing the former class include: *QR*, *bisection* and *Jacobi methods* that are intended for standard eigenproblems, and *QZ* method that is a generalization of the QR method for generalized eigenproblems. A common idea of these algorithms is that they apply various matrix transformations to operator matrices in order to obtain canonical forms such as Schur or diagonal ones [39,98]. Once they are found, the solution of the eigenproblem can easily be determined. A severe drawback of these methods is that performing matrix transformations requires that the matrix must be stored explicitly in a dense format, even if it has a sparse character. It should also be noted that for the solution of nonsymmetric problems only QR (QZ) method can be applied and this method always computes all eigenvalues, even though only a few ones are usually required for a particular purpose.

The class of iterative algorithms is represented by *subspace iteration*, *Arnoldi*, *Lanczos* and *nonsymmetric Lanczos* methods that are intended for standard as well as for generalized eigenproblems. These algorithms are based on a projection of the operator matrix of size N into a subspace of a small size l ($l \ll N$). This subspace is constructed in such a way that it contains only a few eigenvalues. Computation of a basis for this subspace is performed iteratively and involves products of the matrix operator and some

vectors. In contrast to the methods based on matrix transformations the operator matrix can be stored in a sparse format and the advantages of sparse matrix computations leading to considerable time and memory savings can be fully utilized. Since the basic versions of iterative methods compute the eigenvalues of the largest magnitude (which are usually out of interest) various *spectral transformations* should be performed in order to obtain convergence to the required eigenvalues (see Sec. A.1.2 in App. A). The advantage of applying the spectral transformations lies in the fact that they can also be used for accelerating the algorithms by increasing their convergence rate. This process is called *preconditioning* and typical examples are shift and shift-invert techniques or polynomial filtering using e.g. Chebyshev or digital finite impulse response (FIR) filters (polynomial *preconditioners*).

Solution of generalized eigenproblems of form (4.2) is much more complex. General methodology of the solution depends on the size and symmetry of the problem. Small eigenproblems can be solved with QR-like methods. In the nonsymmetric case ($\underline{\underline{A}}$ or $\underline{\underline{B}}$ nonsymmetric), QZ method is usually used, which can be viewed as a generalization of the QR method that performs implicit QR iteration on $\underline{\underline{A}}\,\underline{\underline{B}}^{-1}$ matrix. However, for the symmetric case ($\underline{\underline{A}}$, $\underline{\underline{B}}$ symmetric and $\underline{\underline{A}}$ or $\underline{\underline{B}}$ definite) another approach is usually adopted. The generalized eigenproblem is transformed into a standard one with using Cholesky decomposition and then the QR algorithm is directly applied (see Appendix A for details).

Similarly to the standard eigenproblems, the large generalized ones are usually solved by means of iterative methods, such as the subspace iteration or the Arnoldi/Lanczos methods. In order to apply these methods the eigenproblems are initially transformed into standard ones. Depending on the symmetry, different factorizations are used for that purpose: LU, for nonsymmetric case, and Cholesky, for symmetric case. The cost of the factorizations can be relatively low when the matrices are structured and sparse.

4.1 Numerical implementations of matrix eigensolvers

Some of the numerical methods described above are available in the form of public domain software. In particular, these are implementations of QR, subspace iteration and Arnoldi methods.

4.1.1 QR method

One of the first public domain libraries containing implementations of QR and QZ methods was EISPACK (1976). In 1992 it was superseded by LAPACK (*Linear Algebra PACKage*) library. LAPACK routines are written so that as much as possible of the computation is performed by calls to highly efficient BLAS library (*Basic Linear Algebra Subprograms*). Fortran codes of all these libraries are accessible via Internet from `http://www.netlib.org` or by anonymous ftp from `ftp://ftp.netlib.org` in directories `eispack`, `lapack` and `blas`, respectively. However, many high-performance computer

manufacturers, such as SUN, SGI, IBM et al., offer specially coded and optimized for a given machine versions of LAPACK and BLAS libraries. Usually, they give much better performance than their model Fortran implementations from Netlib.

The implementations of QR and QZ methods from LAPACK library are also optimized for performance for various kinds of dense matrix eigenproblems: standard or generalized, symmetric or nonsymmetric, real or complex, full, band or tridiagonal. These algorithms are easy to use and straightforward in parameter selection. The specific LAPACK routine, which is intended for solving nonsymmetric standard real and full dense eigenproblems, is called `DGEEV`.

4.1.2 Subspace iteration and Arnoldi method

Many implementations of subspace iteration and Arnoldi methods have been written so far. However, only four codes, listed in Table 4.1, are available in public domain and have a sufficient quality level. An extensive comparison of these codes was performed by Lehoucq and Scott in [65][1]. This comparison showed that the performance of all considered codes depends on the eigenproblem solved and on user-defined parameters of computational routines. In many cases, the best results could be obtained with ARPACK that implements the implicitly restarted Arnoldi method (see App. A).

Table 4.1: Public domain library-quality implementations of subspace iteration (SI) and Arnoldi method (ERA — explicitly restarted and IRA — implicitly restarted).

Code	Method	Year	Availability
`LOPSI`	SI	1981	`ftp://ftp.netlib.org/toms/570` or `http://www.netlib.org/toms/570`
`SRRIT`	SI	1993	`ftp://ftp.netlib.org/toms/776` or `http://www.netlib.org/toms/776`
`ARNCHEB`	ERA	1993	`ftp://ftp.cerfacs.fr` in directory `pub/algo/software/Qualcomp/Arncheb`
`ARPACK`	IRA	1995	`ftp://ftp.netlib.org/scalapack/arpack96.tgz` or `http://www.netlib.org/scalapack/arpack96.tgz`

The ARPACK package is the collection of particularly versatile routines, incorporating *reverse communication* scheme, which allows one to perform certain operations outside the subroutine. The ARPACK routines require the program to provide product of the operator matrix and a vector indicated by the procedure. The matrix is not passed to the routine, what leaves to the user the decisions on the manner in which the matrix is stored and the matrix-vector product is realized. It allows the application of fast and efficiently optimized mathematical subroutines intended for a particular format of the matrix. Great

[1] Available via anonymous ftp from `ftp://info.mcs.anl.gov/pub/tech_reports/reports/P547.ps.Z`

versatility of the ARPACK routines relies also on the possibility of choosing an end of the spectrum where the eigenvalues are to be found.

The implementation of the real version of the implicitly restarted Arnoldi algorithm (see App. A) from the ARPACK is called `DNAUPD`. This routine requires many formal parameters to be selected. Their detailed description can be found in extensive documentation of the ARPACK. Those of the parameters which are responsible for definition of the problem, such as problem size N, problem type (standard or generalized) `BMAT`, number of eigenvalues to compute k (denoted by `NEV` in the code) or the spectrum end of interest `WHICH`, are obvious. However, the parameters responsible for efficiency and accuracy are not straightforward in selection. The most important for efficiency is Krylov subspace[2] size l (denoted by `NCV` in the code), which should be chosen in dependence on N, k (`NEV`) and the properties of the matrix resulting from the particular projection method. In order to properly select l (`NCV`) some tests are usually required. The most important for accuracy of computed eigenvalues is `TOL` parameter, which is related to Ritz estimates r_j (see Sec. A.4).

[2]see Appendix A for the definition of Krylov subspace

Chapter 5

Examples of analysis

In the preceding chapters we developed out a set of tools that are required for analysis of microwave waveguide and resonator structures. Our aim was to select approaches that could lead to the most effective analysis methods.

In Chapter 2, we found analytical formulations possessing features important for efficiency of the numerical solution, i.e. small number of components involved, suppression of generation of spurious solutions (spurious-free property), symmetry of the operators. We found out that from this point of view the most effective approach to the analysis of the most general waveguides was offered by spurious-free formulations involving four transverse components. If the waveguides are strictly bidirectional spurious-free formulations involving two transverse components can be used instead. Application of scalar formulations is always the most effective approach due to small size of the resulting eigenproblem and the fact that they do not generate spurious solutions. However, scalar waveguide formulations are only possible in the case of fully homogeneous and isotropic structures and in the solution of static and cutoff problems. The most efficient analysis of resonators with rotational symmetry relies on using appropriate spurious-free waveguide formulations. Unfortunately, analysis of general inhomogeneous resonators requires application of full vector six- or three-component formulations, which are potentially spurious.

In Chapter 3, we discussed various classical and hybrid conversion methods of the analytical eigenproblem to a matrix one from the point of view of their versatility and topological properties of the resulting problems, such as the type, size and symmetry. We found out that classical conversion methods such as the Rayleigh-Ritz method or the method of moments involving basis (and testing) functions defined over the entire domain can only be applied to the analysis of the structures of relatively simple geometries. These methods result in problems with medium size and dense matrices. The finite element method, which can be viewed as a kind of the Rayleigh-Ritz or Galerkin methods involving locally defined basis functions, is much more versatile and produces eigenproblems with large and sparse matrices. Another classical method, the finite difference frequency domain, is also well suited for analysis of structures with complex geometries. The matrices resulting from this method are very large and highly diagonally structured (i.e. sparse). It was also found out that hybrid methods, such as the coupled

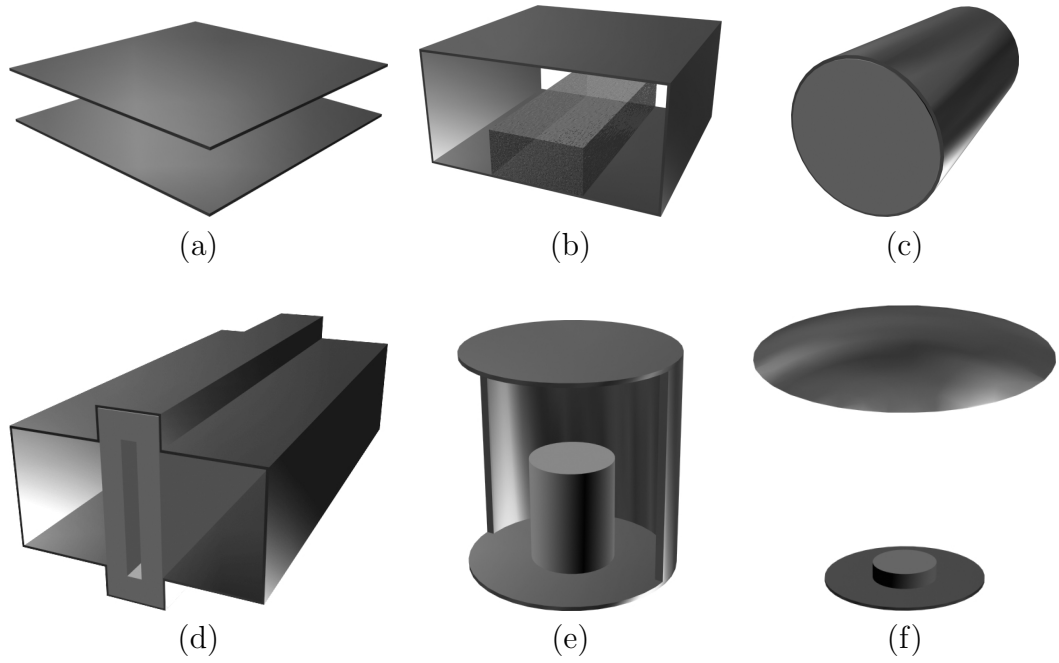


Figure 5.1: The structures analyzed during the tests: (a) parallel plate guide, (b) image line, (c) circular waveguide, (d) grooved waveguide with toroidal inlay, (e) rotationally symmetric resonator, (f) rotationally symmetric open resonator.

mode method and eigenfunction expansion methods, produce problems with small and dense matrices because they involve basis (and testing) functions that are "optimized" in the sense of obeying internal and external boundary conditions. It seemed that the coupled mode method could be well suited for analysis of structures containing media of complex properties, while eigenfunction expansion methods could be advantageous in computation of dispersion characteristics of waveguides.

In Chapter 4 (and Appendix A), we presented the potential of various numerical methods for solving matrix eigenproblems. We were specially interested in the methods available in the form of public domain software. We found that among the methods based on matrix transformations, only the QR method could be applied to symmetric as well as nonsymmetric problems and had robust and stable implementations. Due to the fact that the QR method computes all eigenvalues and keep entire matrix in the memory, solution of large problems is very expensive. Another kind of methods, the iterative ones, much better conforms to the problems with large and sparse matrices. The most mature codes were available for implementations of the Arnoldi and the subspace iteration methods. These methods compute only selected eigenvalues and can easily incorporate spectral transformation techniques for accelerating the convergence.

To illustrate the application of different methods discussed in Chapters 2–4, compare performance, accuracy and efficiency we analyzed several waveguide and resonator structures shown in Fig. 5.1 using appropriate formulations for each given structure, various conversion methods and appropriate matrix eigenproblem solution methods. The results

of the tests, shown in the following sections, were grouped according to the classical and hybrid conversion methods used. The examples of the classical methods include the Galerkin method, finite element method and finite difference frequency domain method, while the examples of hybrid methods include the coupled mode method and eigenfunction expansion method.

There were several purposes of the selected tests. The first of them was to show efficiency of the Galerkin method in the analysis of the structures of simple geometry but filled with the medium of complex properties, such as chiroferrite. In order to show high efficiency of the iterative methods based on projection onto the Krylov space, such as the Arnoldi method, a set of tests concerning the analysis of an image line was performed. The tests included three classical conversion methods such as the Galerkin method, finite element method and finite difference frequency domain method. The aim of two next tests concerning the analysis of rotationally symmetric closed and open resonators was to show additional speedups that can be obtained in consequence of using various spectral transformations with iterative solvers. To show high efficiency of hybrid conversion methods in the analysis of anisotropic structures and fast determination of dispersion characteristics of waveguides the tests of the coupled mode method and an eigenfunction expansion method were respectively made. The former method was used to the analysis of a grooved waveguide filled with toroidal magnetized ferrite, while the latter one was applied to the analysis of an image line and a circular waveguide loaded with anisotropic magnetic medium.

5.1 Galerkin method

To show that the Galerkin method can be very effective in the analysis of the structures of a simple shape, containing material of complex properties, it is applied to the analysis of a homogeneous parallel plate chiroferrite waveguide.

5.1.1 Parallel plate chiroferrite waveguide

Problem and structure. In this example we will show the application of the Galerkin method to the investigation of the properties of the parallel plate waveguide shown in Fig. 5.2, containing a composite medium having both chiral and gyrotropy properties. The chiroferrite medium is weakly magnetized in the z-direction thus the permeability tensor $\underline{\underline{\mu}}$ is given by (2.95), where $\mu = \mu_{zz} = \mu_0$ and $\mu_a = \mu_0 \gamma M_s / f$ (gyromagnetic constant $\gamma = 2.8\,[\mathrm{MHz/Oe}]$, saturation magnetization $M_s = 1000\,\mathrm{Gauss}$). The remaining parameters of the structure are: $d = 10\,\mathrm{mm}$, $\epsilon = \epsilon_0$, $\xi_c = 0.001\mho$. We concentrate on the determination of the dispersion characteristic for the fundamental mode.

We can distinguish two groups of modes appearing in the investigated structure, namely the even and the odd [70] ones. They can be computed independently, by selecting adequate basis functions into expansion. The proper selection of the basis functions is discussed in the paragraph devoted to the conversion to a matrix problem.

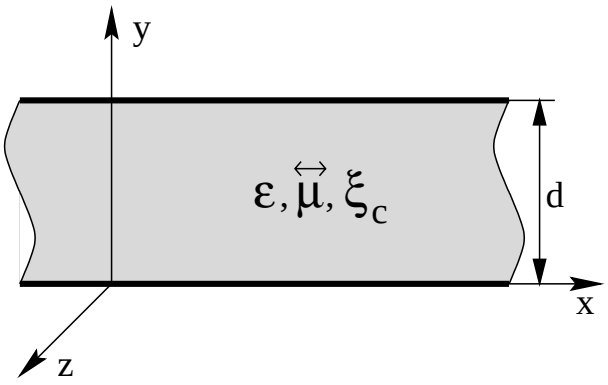


Figure 5.2: Dimensions of the parallel plate waveguide structure.

Formulation. In Chapter 2, only anisotropic media were considered. In the case of chiral media, it is necessary to develop a separate formulation, because the relations between flux densities and field intensities involve additional couplings between the electric and magnetic components.

For chiral media, material equations (2.6) and (2.7) take the form

$$\begin{aligned} \vec{D} &= \underline{\underline{\epsilon}} \cdot \vec{E} + j\xi_c \vec{B} \\ &= (\underline{\underline{\epsilon}} + \xi_c^2 \underline{\underline{\mu}}) \cdot \vec{E} + j\xi_c \underline{\underline{\mu}} \cdot \vec{H} \qquad (5.1) \\ \vec{B} &= \underline{\underline{\mu}} \cdot \vec{H} - j\xi_c \underline{\underline{\mu}} \cdot \vec{E} \\ &= \underline{\underline{\mu}} \cdot (\vec{H} - j\xi_c \vec{E}) \qquad (5.2) \end{aligned}$$

where ξ_c is the chirality admittance. Therefore, steady state Maxwell's equations (2.2) and (2.3) can be expanded to the form

$$\begin{aligned} \nabla \times \vec{E} &= -j\omega \vec{B} \\ &= -j\omega \underline{\underline{\mu}} \cdot \vec{H} + \omega \xi_c \underline{\underline{\mu}} \cdot \vec{E} \qquad (5.3) \\ \nabla \times \vec{H} &= j\omega \vec{D} \\ &= j\omega (\underline{\underline{\epsilon}} + \xi_c^2 \underline{\underline{\mu}}) \cdot \vec{E} + \omega \xi_c \underline{\underline{\mu}} \cdot \vec{H} \qquad (5.4) \end{aligned}$$

To derive the eigenproblem we use the formulation for transverse electric and magnetic fields, analogous to (2.66). In the derivation we assume that permittivity and permeability tensors are described by (2.63). Applying the decomposition (2.50) to equations (5.3) and (5.4) and isolating the transverse and longitudinal parts one can eliminate the z-component of the fields. In consequence, the following β-formulation for transverse $\vec{E}_t$ and $\vec{H}_t$ fields is obtained in the form of a complex generalized eigenproblem

$$\begin{bmatrix} \mathbf{A}_{\mathbf{T}ee} & \mathbf{A}_{\mathbf{T}eh} \\ \mathbf{A}_{\mathbf{T}he} & \mathbf{A}_{\mathbf{T}hh} \end{bmatrix} \begin{bmatrix} \vec{E}_t \\ \vec{H}_t \end{bmatrix} = \beta \begin{bmatrix} 0 & -\hat{z}\times \\ \hat{z}\times & 0 \end{bmatrix} \begin{bmatrix} \vec{E}_t \\ \vec{H}_t \end{bmatrix} \qquad (5.5)$$

where

$$\begin{aligned}
\mathbf{A_{Tee}} &= \omega(\underline{\underline{\epsilon}}_{tt} + \xi_c^2 \underline{\underline{\mu}}_{tt}) \cdot (\cdot) - \frac{1}{\omega}\nabla_t \times \frac{\epsilon_{zz} + \xi_c^2 \mu_{zz}}{\mu_{zz}\epsilon_{zz}} \nabla_t \times (\cdot) \\
\mathbf{A_{Teh}} &= j\omega\xi_c \underline{\underline{\mu}}_{tt} \cdot (\cdot) + j\frac{1}{\omega}\nabla_t \times \frac{\xi_c}{\epsilon_{zz}} \nabla_t \times (\cdot) \\
\mathbf{A_{The}} &= -j\omega\xi_c \underline{\underline{\mu}}_{tt} \cdot (\cdot) - j\frac{1}{\omega}\nabla_t \times \frac{\xi_c}{\epsilon_{zz}} \nabla_t \times (\cdot) \\
\mathbf{A_{Thh}} &= \omega \underline{\underline{\mu}}_{tt} \cdot (\cdot) - \frac{1}{\omega}\nabla_t \times \frac{1}{\epsilon_{zz}} \nabla_t \times (\cdot)
\end{aligned} \tag{5.6}$$

Taking into account the homogeneity of the structure and the form of the material tensors, problem (5.5) can be transformed into the following standard eigenproblem

$$\begin{bmatrix} \mathbf{A_{ee}} & \mathbf{A_{eh}} \\ \mathbf{A_{he}} & \mathbf{A_{hh}} \end{bmatrix} \begin{bmatrix} \vec{E}_t \\ \vec{H}_t \end{bmatrix} = \beta \begin{bmatrix} \vec{E}_t \\ \vec{H}_t \end{bmatrix} \tag{5.7}$$

where

$$\begin{aligned}
\mathbf{A_{ee}} &= j\xi_c \hat{z} \times \left[\omega \underline{\underline{\mu}}_{tt} \cdot (\cdot) + \frac{1}{\omega\epsilon}\nabla_t \times \nabla_t \times (\cdot) \right] \\
\mathbf{A_{eh}} &= -\hat{z} \times \left[\omega \underline{\underline{\mu}}_{tt} \cdot (\cdot) - \frac{1}{\omega\epsilon}\nabla_t \times \nabla_t \times (\cdot) \right] \\
\mathbf{A_{he}} &= \hat{z} \times \left[\omega(\epsilon + \xi_c^2 \underline{\underline{\mu}}_{tt}) \cdot (\cdot) - \frac{\epsilon + \xi_c^2 \mu}{\omega\mu\epsilon}\nabla_t \times \nabla_t \times (\cdot) \right] \\
\mathbf{A_{hh}} &= j\xi_c \hat{z} \times \left[\omega \underline{\underline{\mu}}_{tt} \cdot (\cdot) + \frac{1}{\omega\epsilon}\nabla_t \times \nabla_t \times (\cdot) \right]
\end{aligned} \tag{5.8}$$

Conversion to matrix problem. We now apply the Galerkin method using independent expansion of $\vec{E}_t$ and $\vec{H}_t$ fields

$$\begin{aligned}
\vec{E}_t &= \sum_{i=1}^{N_e} a_i \vec{e}_{ti} \\
\vec{H}_t &= \sum_{i=1}^{N_h} b_i \vec{h}_{ti}
\end{aligned} \tag{5.9}$$

where the basis functions $\vec{e}_{ti}$ and $\vec{h}_{ti}$ are selected to be the modes existing in non-chiral ($\xi_c = 0$) and non-gyrotropic ($\underline{\underline{\mu}} = \mu$) structure, obeying the following equations derived from (5.7) and (5.8)

$$\begin{aligned}
-\hat{z} \times \left(\omega\mu(\cdot) - \frac{1}{\omega\epsilon}\nabla_t \times \nabla_t \times (\cdot) \right) \vec{h}_{ti} &= \beta_i \vec{e}_{ti} \\
\hat{z} \times \left(\omega\epsilon(\cdot) - \frac{1}{\omega\mu}\nabla_t \times \nabla_t \times (\cdot) \right) \vec{e}_{ti} &= \beta_i \vec{h}_{ti}
\end{aligned} \tag{5.10}$$

The solutions of (5.10) are TEM, TE and TM modes which, when normalized, are given by

$$\vec{e}_t^{\,\mathrm{TEM}} = \begin{cases} e_x = 0 \\ e_y = \sqrt{\frac{1}{d}} \end{cases} \quad , \quad \vec{h}_t^{\,\mathrm{TEM}} = \begin{cases} h_x = -\sqrt{\frac{1}{d}} \\ h_y = 0 \end{cases} \tag{5.11}$$

$$\vec{e}_{ti}^{\,\mathrm{TE}} = \begin{cases} e_x = \sqrt{\frac{2}{d}}\sin(k_{yi}y) \\ e_y = 0 \end{cases} \quad , \quad \vec{h}_{ti}^{\,\mathrm{TE}} = \begin{cases} h_x = 0 \\ h_y = \sqrt{\frac{2}{d}}\sin(k_{yi}y) \end{cases} \tag{5.12}$$

$$\vec{e}_{ti}^{\,\mathrm{TM}} = \begin{cases} e_x = 0 \\ e_y = \sqrt{\frac{2}{d}}\cos(k_{yi}y) \end{cases} \quad , \quad \vec{h}_{ti}^{\,\mathrm{TM}} = \begin{cases} h_x = -\sqrt{\frac{2}{d}}\cos(k_{yi}y) \\ h_y = 0 \end{cases} \tag{5.13}$$

where $k_{yi} = i\pi/d$. Note that $\hat{z} \times \vec{e}_{ti} = \vec{h}_{ti}$ and the basis functions are normalized so that

$$\int_S \hat{z} \cdot (\vec{e}_{ti} \times \vec{h}^*_{ti})\, ds = \int_S \vec{e}_{ti} \cdot \vec{e}^{\,*}_{ti}\, ds = \int_S \vec{h}^*_{ti} \cdot \vec{h}_{ti}\, ds = 1 \tag{5.14}$$

forming a complete set of orthonormal functions on the transverse plane S of the basis guide.

It can be shown that for the wave propagating in the z-direction the basis functions also fulfill the following conditions

$$\begin{aligned} \nabla_t \times \nabla_t \times \vec{e}_{ti} &= \left\{ \begin{array}{c} 1^{\mathrm{TE}} \\ 0^{\mathrm{TM}} \end{array} \right\} k_{yi}^2 \vec{e}_{ti} \\ \nabla_t \times \nabla_t \times \vec{h}_{ti} &= \left\{ \begin{array}{c} 0^{\mathrm{TE}} \\ 1^{\mathrm{TM}} \end{array} \right\} k_{yi}^2 \vec{h}_{ti} \end{aligned} \tag{5.15}$$

The mode of interest (the dominant one) is an even mode. The even modes are computed when only the TEM, even TM_{2k} and odd TE_{2k-1} basis functions are taken into expansion. The odd modes are computed when only the odd TM_{2k-1} and even TE_{2k} basis functions are considered [70]. Since $\{\vec{e}_{ti}\}_{i=1,\ldots,N_e} \cup \{\vec{h}_{ti}\}_{i=1,\ldots,N_h}$ form an orthonormal set of functions, application of the Galerkin method leads to the following standard eigenvalue problem of size $N = N_e + N_h$

$$\underline{\underline{A}}\,\underline{h} = \lambda \underline{h} \tag{5.16}$$

where $\lambda = \beta$, $\underline{h} = [a_1, \ldots, a_{N_e}, b_1, \ldots, b_{N_h}]^T$ and elements of $\underline{\underline{A}}$ are defined with appropriate inner products which using (5.10) and (5.15) can be expressed as

$$\begin{aligned} A_{j,i} &= \left(\mathbf{A_{ee}} e_{ti}, e_{tj}\right) = \int_S \vec{e}^{\,*}_{tj} \cdot \mathbf{A_{ee}} \vec{e}_{ti}\, ds \\ &= -\omega\mu_a\xi_c\delta_{ji} + j\omega\mu\xi_c \left[1 + \left\{ \begin{array}{c} 1^{\mathrm{TE}} \\ 0^{\mathrm{TM}} \end{array} \right\} \left(\frac{k_{yi}}{k}\right)^2\right] \int_S \vec{e}^{\,*}_{tj} \cdot \vec{h}_{ti}\, ds \\ A_{j,N_e+i} &= \left(\mathbf{A_{eh}} h_{ti}, e_{tj}\right) = \int_S \vec{e}^{\,*}_{tj} \cdot \mathbf{A_{eh}} \vec{h}_{ti}\, ds \\ &= \beta_i\delta_{ji} - j\omega\mu_a \int_S \vec{e}^{\,*}_{tj} \cdot \vec{h}_{ti}\, ds \end{aligned}$$

$$
\begin{aligned}
A_{N_e+j,i} &= \left(\mathbf{A_{he}} e_{ti}, h_{tj}\right) = \int_S \vec{h}^*_{tj} \cdot \mathbf{A_{he}} \vec{e}_{ti}\, ds \\
&= \left(1 + \mu \frac{\xi_c^2}{\epsilon}\right) \beta_i \delta_{ji} + j\omega\mu_a \xi_c^2 \int_S \vec{h}^*_{tj} \cdot \vec{e}_{ti}\, ds \\
A_{N_e+j,N_e+i} &= \left(\mathbf{A_{hh}} h_{ti}, h_{tj}\right) = \int_S \vec{h}^*_{tj} \cdot \mathbf{A_{hh}} \vec{h}_{ti}\, ds \\
&= -\omega\mu_a \xi_c \delta_{ji} - j\omega\mu\xi_c \left[1 + \left\{ \begin{array}{c} 0^{\mathrm{TE}} \\ 1^{\mathrm{TM}} \end{array} \right\} \left(\frac{k_{yi}}{k}\right)^2\right] \int_S \vec{h}^*_{tj} \cdot \vec{e}_{ti}\, ds \qquad (5.17)
\end{aligned}
$$

where δ_{ij} is the Kronecker symbol and $k = \omega\sqrt{\mu\epsilon}$. The matrix $\underline{\underline{A}}$ is dense, nonsymmetric and complex.

It should be noted that exactly the same eigenproblem can be obtained using a coupled mode formalism [70].

Solution of matrix eigenproblem. For the solution of eigenproblem (5.16) we use EISPACK numerical implementation of the QR method (see Sec. 4.1.1), intended for complex and nonsymmetric matrices. The computations were performed on a PC computer.

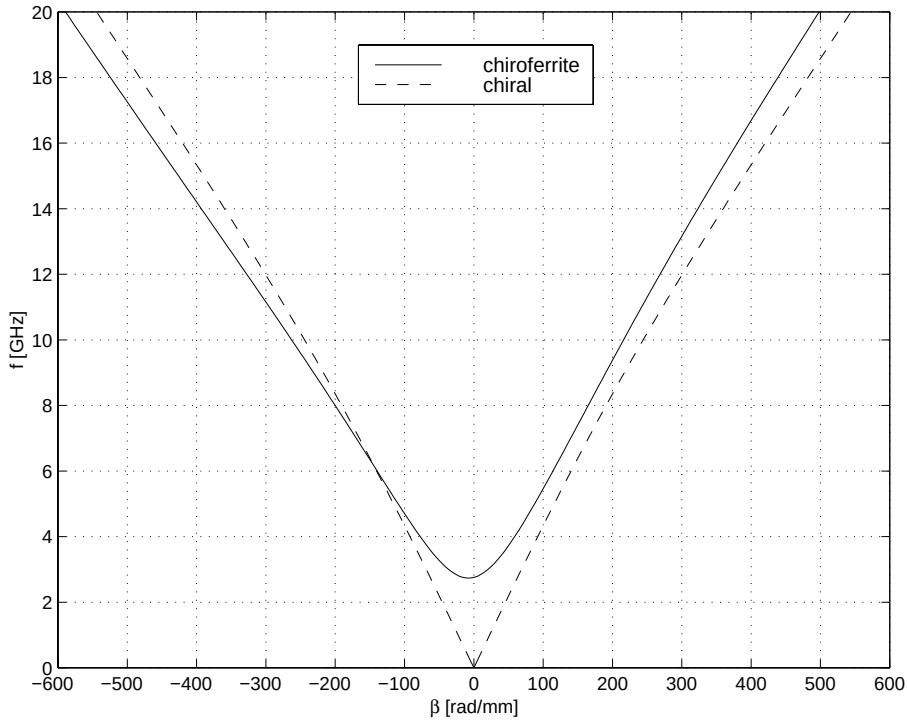


Figure 5.3: Dispersion characteristics of the fundamental mode in the parallel plate waveguide filled with chiroferrite and chiral medium.

Results. The dispersion characteristic of the dominant even mode calculated for $N = 40$ ($N_e = N_h = 20$) is shown in Fig. 5.3. The results for chiral structure ($M_s = 0 \Rightarrow \mu_a = 0$) are also presented for comparison. The nonreciprocal character of the dominant mode

can be observed, which is in contrast to the gyrotropic or chiral structures having equal propagation constants for two opposite directions of propagation. It can be also seen that when the gyrotropy properties are neglected the fundamental mode appears without cutoff frequency.

In the test described above we have shown that the Galerkin method can be very powerful in the analysis of structures containing materials of complex properties. The elements of the resulting matrix can easily be determined if the structure has a simple geometry and the basis (and testing) functions can be determined analytically. If high accuracy is not required the number of basis functions taken into expansion can be low. Then the size of the matrix is small the QR method can be used for determination of the eigenvalues and the corresponding eigenvectors. However, when high accuracies are demanded, large matrices are obtained and very much time and memory consuming QR method can be superseded by an iterative method such as Arnoldi. A comparison of performance of these two eigensolvers is made in the next test concerning the analysis of an image line.

5.1.2 Inhomogeneous rectangular waveguide loaded with a dielectric slab

Problem and structure. For a given frequency $f = 12\,\mathrm{GHz}$, we look for propagation constants β of four dominant even modes in the structure of dimensions: $a = 15.8\,\mathrm{mm}$, $b = 7.9\,\mathrm{mm}$, $w = 6.9\,\mathrm{mm}$, $h = 3.2\,\mathrm{mm}$, and relative permittivity of the slab $\epsilon_r = 9$. Since this structure is symmetric in x-direction we can independently compute even (for PEC symmetry plane) or odd modes (for PMC symmetry plane). This requires only even or odd basis functions to be taken into expansion. Initially, we investigate a complex problem of parameter selection of an Arnoldi method iterative solver in context of computation efficiency. A performance of a QR and the Arnoldi solvers is also compared. Moreover, we test the advantages of on-line matrix-vector computations using the iterative Arnoldi solver.

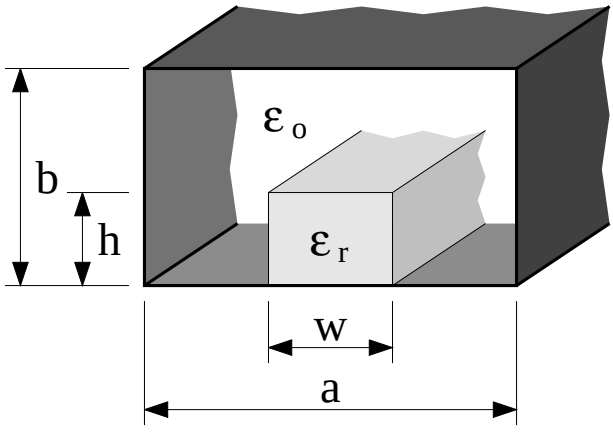


Figure 5.4: Dimensions of the image line structure.

Formulation. Analysis is based on formulation (2.79) for transverse magnetic fields, which do not generate spurious solutions. Since the waveguide is loaded with non-magnetic isotropic materials then $\underline{\underline{\mu}}_{tt} = \mu_{zz} = \mu_0$ and $\underline{\underline{\epsilon}}_{tt} = \epsilon_{zz} = \epsilon(x,y)$. Using vector identities (2.79) can be transformed into the following form

$$\left[\nabla_t^2(\cdot) + \omega^2\mu_0\epsilon(\cdot) + \frac{1}{\epsilon}\nabla_t\epsilon \times \nabla_t \times (\cdot)\right]\vec{H}_t = \beta^2\vec{H}_t \tag{5.18}$$

Decomposing fields and operators in (5.18) we get

$$\begin{bmatrix} \mathbf{A_{xx}} & \mathbf{A_{xy}} \\ \mathbf{A_{yx}} & \mathbf{A_{yy}} \end{bmatrix}\begin{bmatrix} H_x \\ H_y \end{bmatrix} = \beta^2\begin{bmatrix} H_x \\ H_y \end{bmatrix} \tag{5.19}$$

where

$$\begin{aligned}
\mathbf{A_{xx}} &= \nabla_t^2(\cdot) + \omega^2\mu_0\epsilon(\cdot) - \frac{1}{\epsilon}\frac{\partial}{\partial y}\epsilon\frac{\partial}{\partial y}(\cdot) \\
\mathbf{A_{xy}} &= \frac{1}{\epsilon}\frac{\partial}{\partial y}\epsilon\frac{\partial}{\partial x}(\cdot) \\
\mathbf{A_{yx}} &= \frac{1}{\epsilon}\frac{\partial}{\partial x}\epsilon\frac{\partial}{\partial y}(\cdot) \\
\mathbf{A_{yy}} &= \nabla_t^2(\cdot) + \omega^2\mu_0\epsilon(\cdot) - \frac{1}{\epsilon}\frac{\partial}{\partial x}\epsilon\frac{\partial}{\partial x}(\cdot)
\end{aligned} \tag{5.20}$$

Conversion to matrix problem. In the Galerkin method involved in the previous example, vector fields are expanded using vector basis (and testing) functions (see (5.9)). In order to improve accuracy of the solution, each component (H_x and H_y) of the vector field $\vec{H}_t$ can be expanded into series separately, as follows

$$\begin{aligned}
H_x &= \sum_{i=1}^{N_x} a_i h_{xi} \\
H_y &= \sum_{i=1}^{N_y} b_i h_{yi}
\end{aligned} \tag{5.21}$$

where a_i, b_i are the expansion coefficients and

$$\begin{aligned}
h_{xi} &= \frac{2}{\sqrt{ab}}\sin\frac{m_x\pi x}{a}\cos\frac{m_y\pi y}{b} \quad ; \quad m_x = 1,2,\dots\ ,\ m_y = 0,1,\dots \\
h_{yi} &= \frac{2}{\sqrt{ab}}\cos\frac{n_x\pi x}{a}\sin\frac{n_y\pi y}{b} \quad ; \quad n_x = 0,1,\dots\ ,\ n_y = 1,2,\dots
\end{aligned} \tag{5.22}$$

The even modes can be found by taking only even basis functions (m_x, n_x even) into the expansion. Since $\{h_{xi}\}_{i=1,\dots,N_x} \cup \{h_{yi}\}_{i=1,\dots,N_y}$ are chosen so that they form an orthonormal set of functions on transverse plane S of the basis guide, application of the Galerkin method leads to the standard eigenvalue problem

$$\underline{\underline{A}}\,\underline{h} = \lambda\underline{h} \tag{5.23}$$

where $\lambda = \beta^2$, $\underline{h} = [a_1, \ldots, a_{N_x}, b_1, \ldots, b_{N_y}]^T$ and elements of $\underline{\underline{A}}$ are defined with

$$\begin{aligned} A_{j,i} &= \left(\mathbf{A_{xx}} h_{xi}, h_{xj}\right) = \int_S \vec{h}^*_{xj} \cdot \mathbf{A_{xx}} \vec{h}_{xi}\, ds \\ A_{j,N_x+i} &= \left(\mathbf{A_{xy}} h_{yi}, h_{xj}\right) = \int_S \vec{h}^*_{xj} \cdot \mathbf{A_{xy}} \vec{h}_{yi}\, ds \\ A_{N_x+j,i} &= \left(\mathbf{A_{yx}} h_{xi}, h_{yj}\right) = \int_S \vec{h}^*_{yj} \cdot \mathbf{A_{yx}} \vec{h}_{xi}\, ds \\ A_{N_x+j,N_x+i} &= \left(\mathbf{A_{yy}} h_{yi}, h_{yj}\right) = \int_S \vec{h}^*_{yj} \cdot \mathbf{A_{yy}} \vec{h}_{yi}\, ds \end{aligned} \tag{5.24}$$

The resulting matrix $\underline{\underline{A}}$ is dense and nonsymmetric. Its size $N = N_x + N_y$ is usually moderate.

Solution of matrix eigenproblem. For the solution of dense matrix problems the QR method is usually used. To show how the application of the iterative methods can improve efficiency of the solution we test two computer implementations of the QR and the Arnoldi methods described in Chapter 4. For the solution of eigenproblem (5.23) we use the routines `DGEEV` and `DNAUPD` from LAPACK and ARPACK packages. No spectral transformations are applied to the matrix. Matrix-vector computations, required by `DNAUPD`, are performed using `DGEMV` routine from BLAS. The eigenvalues of interest ($\lambda = \beta^2$) are those of the largest real part [26, 29] (this information is passed to `DNAUPD` routine).

The computations are performed on a super-scalar SGI Power Challenge XL computer with four R8000 processors, each equipped with 1 MB of fast cache memory.

Results. During the tests we computed four dominant even modes whose normalized propagation constants β, calculated for $N = 1800$ ($N_x = N_y = 900$), are shown in Table 5.1.

Table 5.1: Four dominant normalized propagation constants β/k_0 (k_0 is the wavenumber for a plane wave in the free space) of even modes of image line shown in Fig. 5.4, computed using Galerkin method. Parameters of the structure are: $a = 15.8\,\text{mm}$, $b = 7.9\,\text{mm}$, $w = 6.9\,\text{mm}$, $h = 3.2\,\text{mm}$, $\epsilon_r = 9$, and $f = 12\,\text{GHz}$.

Eigenvalue no.	GM ($N = 1800$, even basis funct.)
1	$1.0274 + j0.0000$
2	$0.3647 + j1.2937$
3	$0.3647 - j1.2937$
4	$0.0000 + j1.6822$

We realized that for a given problem size N the computation time of `DNAUPD` routine was highly dependent on the size of the Krylov subspace l and `TOL` parameter (see

Sec. 4.1.2). A number of tests [26] were made, where l and `TOL` were tuned in order to efficiently apply the routine. It was found that for this type of eigenproblem the time-optimum subspace size l should fulfill the following condition

$$4 < \frac{l-k}{\sqrt[4]{N}} < 12 \tag{5.25}$$

The value of l in the middle of the range was therefore

$$l_0 = k + 8\sqrt[4]{N} \tag{5.26}$$

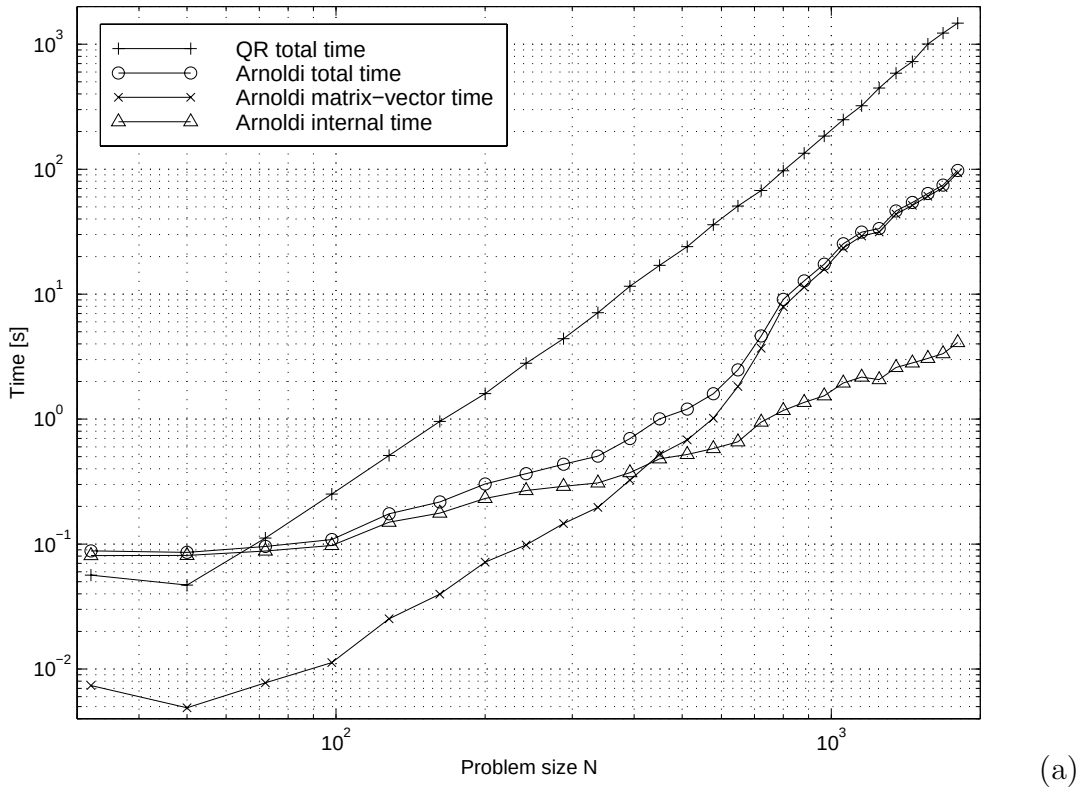
In our case ($k = 4$) the time-optimum subspace size l_0 was in the range $25 < l_0 < 56$ for $50 < N < 1800$.

Numerical tests [26] also revealed that the selection of a too small value of `TOL` resulted only in a growth of computation time. For example, setting `TOL` $= 10^{-6}$ led to the accuracy of the 1-st eigenvalue equal to 10^{-14} (which was very close to the maximum available accuracy) and was 10^{-10} for the next three eigenvalues. However, calculation of all eigenvalues with the accuracy of 10^{-14} required setting `TOL` $= 10^{-11}$, which enlarged total computation time by 50%, while setting `TOL` $= 10^{-16}$ doubled the time offering no improvement in accuracy.

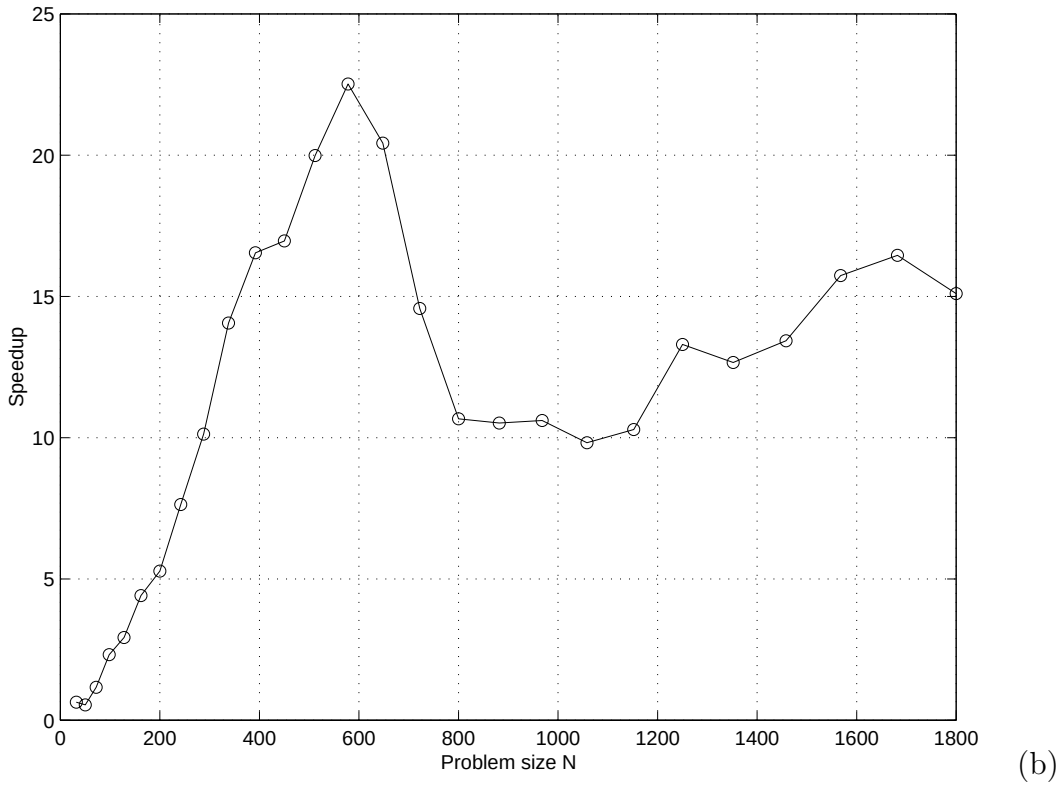
In the remaining tests performed by the group with author's participation [26–29] we used $l = l_0$ and `TOL` $= 10^{-6}$. The QR and the Arnoldi methods were ran on one to four R8000 processors. The computation times for the one processor case are shown in Fig. 5.5(a). These results were computed using `complib.sgimath` optimized library, which included all LAPACK and BLAS routines (`DGEEV` and `DGEMV` among others). It can be seen, that the Arnoldi method offers solution faster than the QR method, for $N > 70$. Moreover, for $N > 450$ the time spent in matrix-vector operations dominates the total Arnoldi computation time. Therefore, for large values of N, efficiency of the Arnoldi method is governed mainly by the efficiency of the matrix-vector product. In Fig. 5.5(b) we can see a speedup of the Arnoldi method over the QR. Large speedup (≈ 23) for $N = 600$ is caused by very efficient utilization of cache memory by `DGEMV` routine from `complib.sgimath` library, while deterioration of the speedup for larger values of N is caused by *out-of-cache effect* [26–29]. When the amount of required memory approaches size of the cache much slower conventional memory is used. Since memory usage in the Arnoldi method is smaller than in the QR algorithm the cache can be effectively used for larger values of N. However, for $N > 600$ both algorithms definitely works outside the cache. This effect can be used to optimize efficiency of the Arnoldi solver on multiprocessor computers.

Other interesting aspects of parallel computations are comprehensively described in [26–29]. One of the tests verified that, in contrast to `DGEEV`, `DGEMV` can be very effectively parallelized. It means that efficiency of the QR method is almost independent on the number of processors used, while the Arnoldi method proves to be very well scalable.

Accurate computations require the solution of large problems. They demand large amount of memory ($\mathcal{O}(N^2)$) if matrix $\underline{\underline{A}}$ is explicitly stored. This amount cannot be



(a)



(b)

Figure 5.5: Computation times for QR and Arnoldi methods ran on one R8000 processor (a). The corresponding speedup of the Arnoldi method over the QR (b).

reduced if QR method is used. However, when an iterative method, such as Arnoldi, is applied a matrix-vector product in inner products (5.24) can be computed *on-line*, i.e. anytime when matrix $\underline{\underline{A}}$ is required it can be computed using lookup tables containing transcendental functions (integrals of sine and cosine functions) instead of being recalled directly from memory. Since the size of the lookup tables is about $\sqrt{N/2} \times \sqrt{N/2}$ and six of them are required for full description of the problem then the number of double elements stored is therefore reduced by the factor of $2N/3$. However, this leads to much longer time spent in matrix-vector operations [26].

This example showed that for efficient application of the implicitly restarted Arnoldi method its parameters should be selected with care. Once they are properly chosen, the Arnoldi method offers significant speedup over the QR method. Additional speedup can be gained when the computer architecture aspects, such as the size of cache memory, are taken into account. It was also shown that in the case when computer memory is critical, the matrix can be computed on-line but the computational cost is much higher.

The conclusions of the discussion are also applicable to other waveguide structures analyzed by means of the Galerkin method and all types of computed modes (even, odd or both).

5.2 Finite element method

To compare computational aspects of the solution of eigenproblems resulting from various methods of conversion we apply the finite element method (FEM) to the analysis of the image line structure analyzed previously by means of the Galerkin method (in Sec. 5.1.2). The FEM is a technique generating generalized eigenproblem with sparse matrices. Thus, comparison of performance of two sparse matrix eigensolvers, i.e. the subspace iteration and the Arnoldi method, can also be made. Moreover, great efficiency of spectral transformation in the case of a generalized eigenproblem solution will be demonstrated.

5.2.1 Inhomogeneous rectangular waveguide loaded with a dielectric slab

Problem and structure. In this test we compute propagation constants of four dominant even modes of the image line shown in Fig. 5.4, using a FEM code developed at UCL London [33]. Symmetry of the structure is taken into account and only a half of it is analyzed, resulting in possibility of computation of even and odd modes separately. In order to compare efficiency of a subspace iteration solver versus an Arnoldi one we replaced the original UCL eigensolver based on the subspace iteration method with the one involving the implicitly restarted Arnoldi method.

Formulation. As the starting point, the spurious-free formulation (2.79) for transverse magnetic fields H_t is used. For non-magnetic materials ($\underline{\underline{\mu}}_{tt} = \mu_{zz} = \mu_0$) premultiplication

of (2.79) by $\hat{z} \times \underline{\underline{\epsilon}}_{tt}^{-1} \cdot (\cdot)$ leads to the following nonsymmetric generalized eigenproblem

$$\left[-\omega^2 \mu_0(\cdot) - \hat{z} \times \nabla_t \kappa_{zz} \hat{z} \nabla_t \times (\cdot) - \hat{z} \times \underline{\underline{\kappa}}_{tt} \cdot \nabla_t \times \hat{z} \nabla_t \cdot (\cdot)\right] \vec{H}_t = \beta^2 \hat{z} \times \underline{\underline{\kappa}}_{tt} \cdot \hat{z} \times \vec{H}_t \quad (5.27)$$

where $\underline{\underline{\kappa}}_{tt}$ and κ_{zz} are given by (2.65).

Conversion to matrix problem. In the UCL implementation of the FEM method mixed order triangular elements are used, i.e. elements of the first order are used in homogeneous regions of the structure, while the second order elements are used on the boundaries between different media. Rayleigh-Ritz method with local potential and equivalent Galerkin method (see Sec. 3.1) applied to (5.27) result in the generalized matrix eigenproblem of the form [33]

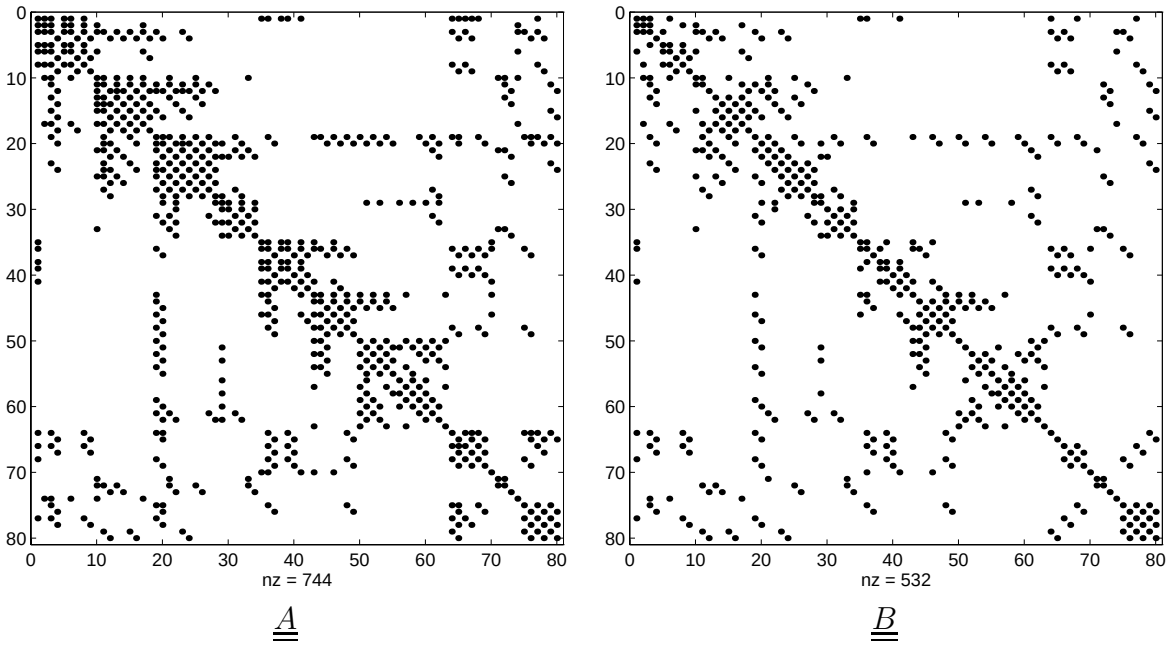
$$\underline{\underline{A}}\,\underline{h} = \lambda \underline{\underline{B}}\,\underline{h} \quad (5.28)$$

where $\lambda = -\beta^2$ and

$$\underline{\underline{A}} = \begin{bmatrix} \underline{\underline{A}}_{xx} & \underline{\underline{A}}_{xy} \\ \underline{\underline{A}}_{yx} & \underline{\underline{A}}_{yy} \end{bmatrix} \quad , \quad \underline{\underline{B}} = \begin{bmatrix} \underline{\underline{B}}_{xx} & \underline{\underline{B}}_{xy} \\ \underline{\underline{B}}_{yx} & \underline{\underline{B}}_{yy} \end{bmatrix} \quad , \quad \underline{h} = \begin{bmatrix} \underline{h}_x \\ \underline{h}_y \end{bmatrix} \quad (5.29)$$

Symbols $\underline{h}_x$, $\underline{h}_y$ denote vectors of expansion coefficients for x and y magnetic field components, respectively. The definitions of elements of all $\underline{\underline{A}}$'s and $\underline{\underline{B}}$'s submatrices and details of the derivation can be found in [33].

Matrices $\underline{\underline{A}}$ and $\underline{\underline{B}}$, shown in Fig. 5.6, are sparse and, in general, nonsymmetric and complex. However, for lossless media $\underline{\underline{B}}$ becomes symmetric and $\underline{\underline{A}}$ is also symmetric, when the structure is additionally homogeneous.



$\underline{\underline{A}}$ $\underline{\underline{B}}$

Figure 5.6: Sparsity patterns for matrices $\underline{\underline{A}}$ and $\underline{\underline{B}}$ resulting from FEM.

Solution of matrix eigenproblem. Since the matrices are large and sparse we test two sparse matrix iterative solvers, i.e. the subspace iteration solver [33], originally implemented in the UCL code, and the alternative implicitly restarted Arnoldi method solver, `DNAUPD` from ARPACK.

We decided to apply a very efficient shift-invert preconditioner, which can be used with a generalized problem without any additional cost (see Sec. A.1.2.2 for discussion). The shift-invert strategy can be implemented as in Alg. 17 from page 136. It relies on implicit solution of standard eigenproblem

$$(\underline{\underline{A}} - \sigma\underline{\underline{B}})^{-1}\underline{\underline{B}}\,\underline{h} = \frac{1}{\lambda - \sigma}\underline{h} \tag{5.30}$$

where σ is the shift. In the modified eigenproblem (5.30), the matrix operator is the product of the inverse of matrix $(\underline{\underline{A}} - \sigma\underline{\underline{B}})$ and matrix $\underline{\underline{B}}$. The inversion is not computed directly. Instead, a sparse LU decomposition of the matrix is performed once before the iteration and when the product $\underline{y} = (\underline{\underline{A}} - \sigma\underline{\underline{B}})^{-1}\underline{\underline{B}}\,\underline{x}$ is required, a linear system of equations $(\underline{\underline{A}} - \sigma\underline{\underline{B}})\underline{y} = \underline{\underline{B}}\,\underline{x}$ is solved.

The convergence rate in the shift-invert mode strongly depends on the shift σ (see Sec. A.1.2.2 for discussion). In the analysis of waveguides it is convenient to choose the shift so that $\sigma > k_0^2\epsilon_{max}$ where k_0 is the wavenumber for a plain wave in vacuum and ϵ_{max} is the maximal permittivity of the materials within the structure. In this case, the dominant modes correspond to the eigenvalues of (5.30) possessing the largest magnitude.

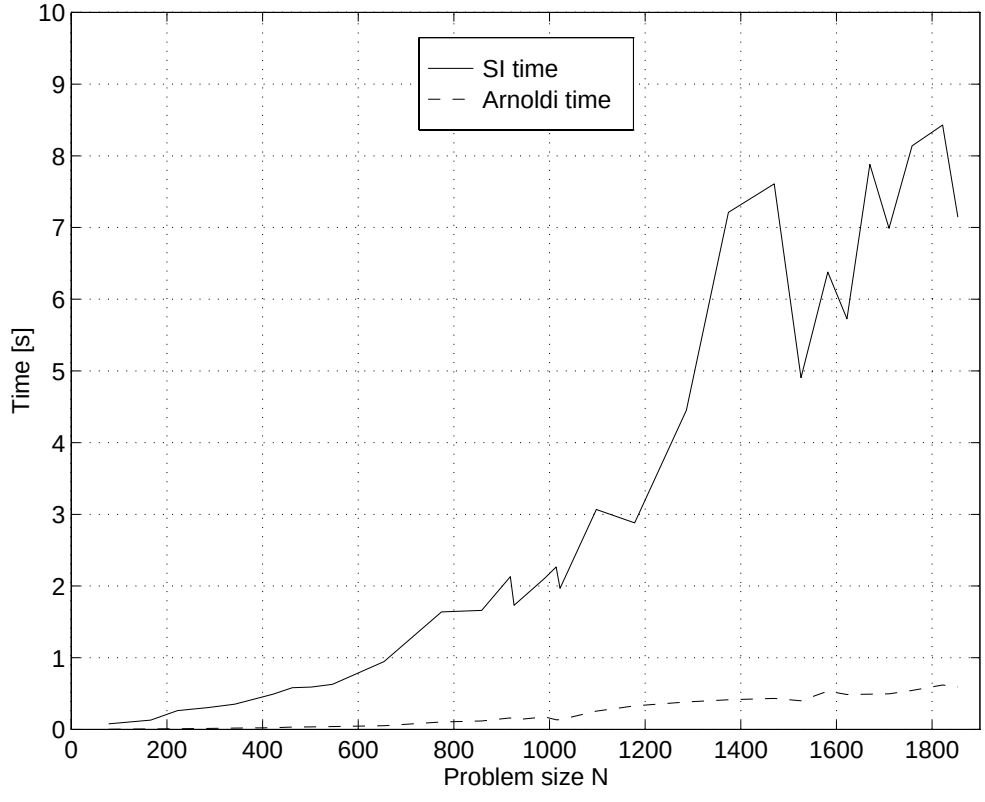
Our computations are performed on the same super-scalar SGI Power Challenge XL computer as mentioned in Sec. 5.1, using one R8000 processor.

Results. During the tests we computed four dominant even modes whose normalized propagation constants β, calculated for $N = 2046$, are shown in Table 5.2.

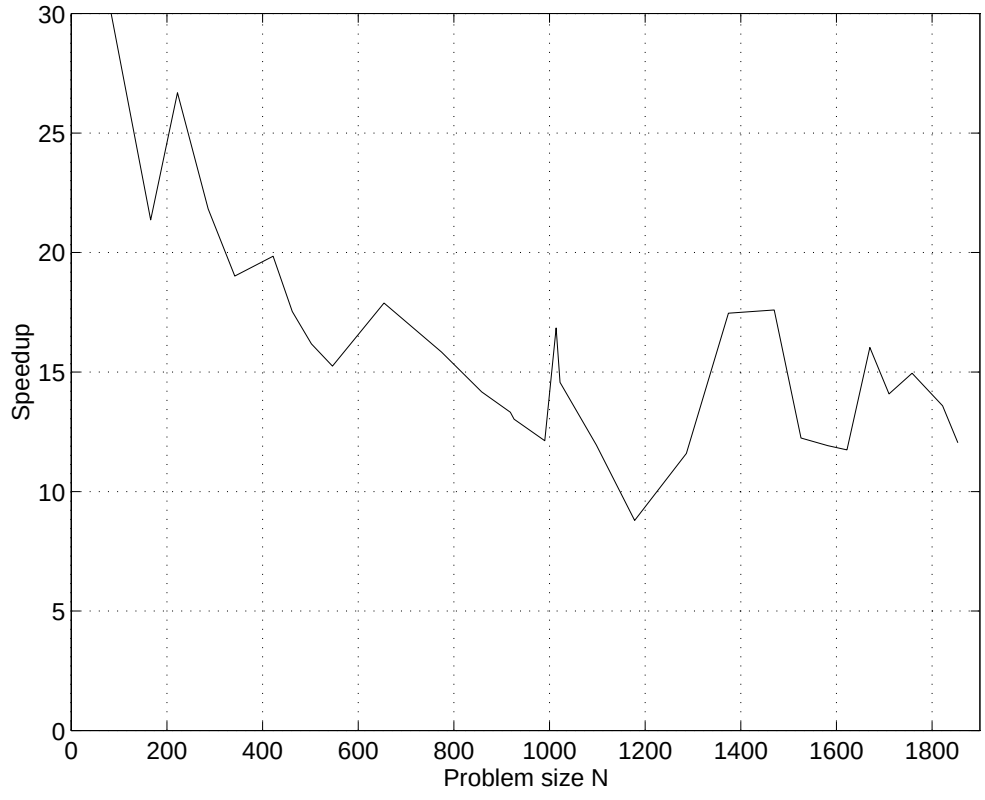
Table 5.2: Four dominant normalized propagation constants β/k_0 (k_0 is the wavenumber for a plane wave in the free space) of even modes of image line shown in Fig. 5.4, computed using FEM method. Parameters of the structure are: $a = 15.8\,\text{mm}$, $b = 7.9\,\text{mm}$, $w = 6.9\,\text{mm}$, $h = 3.2\,\text{mm}$, $\epsilon_r = 9$, and $f = 12\,\text{GHz}$.

Eigenvalue no.	FEM ($N = 2046$, half the struct.)
1	$1.0192 + j0.0000$
2	$0.3264 + j1.2862$
3	$0.3264 - j1.2862$
4	$0.0000 + j1.6839$

Parameters of the Arnoldi and subspace iteration solvers are selected so that the same shift σ is applied, the size of the subspace is fixed at $l = 8$ in each of the solvers and the relative accuracy of computed eigenvalues is 10^{-6}. Due to convergence problems of the subspace iteration we have to set σ as high as $12k_0^2$ in the case of even modes.



(a)



(b)

Figure 5.7: Calculation times of four dominant even modes, computed by means of subspace iteration and Arnoldi methods (a) and speedup of the Arnoldi method over the subspace iteration (b).

Calculation times for both solvers are presented in Fig. 5.7(a). The Arnoldi algorithm is much faster than the subspace iteration one. Speedup of the Arnoldi method over subspace iteration, presented in Fig. 5.7(b), is greater than 12 for almost all problem sizes. Very similar observations can be made for calculation of odd modes. In this case, the shift was set to $\sigma = 16k_0^2$ and corresponding speedup was approximately 8 [73, 74].

Better performance of the implicitly restarted Arnoldi method comes from two factors. One reason is a faster convergence rate, which results in fewer iterations before the convergence criteria are satisfied, i.e. 9 iterations versus ≈ 45. The second reason is that the Arnoldi iterations involve fewer products of the operator matrix and a vector demanding very costly solutions of the linear system $(\underline{\underline{A}} - \sigma\underline{\underline{B}})\underline{y} = \underline{\underline{B}}\,\underline{x}$ (they dominate the solution time of both iterative methods). The Arnoldi method requires the system to be solved ≈ 30 times, while the 45 iterations in the subspace iteration method is associated with 360 solutions. Comparison of the number of solution steps gives the figure of 12 which is in agreement with the data in Fig. 5.7(b).

It should be noted, that the results shown in Fig. 5.7(a) and (b) do not take into account the time spent initially in the LU decomposition of the matrix. This time for $N > 400$ becomes greater than the Arnoldi time and for $N > 1800$ it is almost 10 times larger (e.g. for $N = 1856$ the LU time is approximately 5s, compared to 0.6s of the Arnoldi time). Additional factor influencing total computational time of the eigensolvers is partial loss of the sparsity of the operator matrix that is stored in the form of the decomposition matrices $\underline{\underline{L}}$ and $\underline{\underline{U}}$ (see Sec. A.5.2). Therefore, additional $\mathcal{O}(N)$ of memory is also required.

The results for the case of computation of four dominant odd modes are similar. The detailed description can be found in [73, 74].

It was shown that in the solution of generalized eigenproblems generated in the analysis of waveguides by means of the FEM method the Arnoldi method is much faster than the subspace iteration method. Compared to a standard eigenproblem, iterative solution of a generalized one requires additional effort (the time and memory) of computing the matrix decomposition and subsequent solution of a linear system. However, due to application of the shift-invert preconditioner a convergence rate of the solver is improved. In consequence, the total time used by the Arnoldi solver (including the LU decomposition time) in the case of sparse matrix FEM eigenproblem is much smaller than the time needed by the solver applied to the dense matrix problem of the same size, obtained by means of the Galerkin method. A comparison of both solution times for $N \approx 1800$ is presented in Table 5.4 on page 79.

5.3 Finite difference frequency domain method

Another technique which results in eigenproblems with sparse matrices is the finite difference frequency domain method (FDFD). However, in contrast to the FEM the eigenproblems produced by the FDFD are standard rather than generalized ones and the matrices are additionally highly diagonally structured (see Fig. 5.9). In Sec. 5.3.1, we apply the

FDFD method to the analysis of the image line structure previously analyzed by means of the Galerkin (Sec. 5.1.2) and FEM (Sec. 5.2.1) methods. This offers the possibility to compare computational aspects of the solution of the eigenproblems resulting from various methods of conversion.

5.3.1 Inhomogeneous rectangular waveguide loaded with a dielectric slab

Problem and structure. We are interested in propagation constants β of eight dominant modes of the image line shown in Fig. 5.4. Symmetry of the structure is not taken into account, because all (even and odd) modes are of interest. To compare the performance of the Arnoldi solver operating on highly sparse matrices with the performance of the QR method and with the Arnoldi solver applied to the matrices resulting from the other conversion methods, the time-optimum parameters of the Arnoldi solver are first found.

Formulation. In our analysis, we involve the spurious-free formulation (2.78) for transverse electric fields. It is premultiplied by $-\hat{z}\times(\cdot)$ before the conversion to matrix problem, resulting in the standard eigenproblem of form

$$\left[-\omega^2\hat{z}\times\underline{\underline{\mu}}_{tt}\cdot\hat{z}\times\underline{\underline{\epsilon}}_{tt}\cdot(\cdot)+\hat{z}\times\underline{\underline{\mu}}_{tt}\cdot\nabla_t\mu_{zz}^{-1}\hat{z}\nabla_t\times(\cdot)+\nabla_t\hat{z}\epsilon_{zz}^{-1}\nabla_t\cdot\underline{\underline{\epsilon}}_{tt}\cdot(\cdot)\right]\vec{E}_t = \beta^2\vec{E}_t \quad (5.31)$$

Conversion to matrix problem. Computational domain is discretized using constrained $N_x \times N_y$ Yee's grid shown in Fig. 5.8, where N_x and N_y respectively denote the number of grid points in the x- and y-direction. The grid is nonuniform (graded) and chosen so that the edges of electric cells conform to the boundaries of the slab.

Direct discretization method is applied to equation (5.31). Each of the differential operators is estimated using the second order (3.52–3.53) or the first order (3.48–3.49) approximations of derivatives, accordingly to the homogeneity of the grid in the discretized regions. Effective permittivity concept [13, 122–124] is used for estimation of the permittivity in the points on the dielectric boundary. Since the structure is filled with isotropic medium the operators corresponding to the effective permittivity and permeability are diagonal.

The developed code performs a discretization of equation (5.31), which leads to the standard eigenproblem in the form of [72]

$$\underline{\underline{A}}\,\underline{e} = \lambda\underline{e} \quad (5.32)$$

where $\lambda = \beta^2$ and

$$\underline{\underline{A}} = \begin{bmatrix} \underline{\underline{A}}_{xx} & \underline{\underline{A}}_{xy} \\ \underline{\underline{A}}_{yx} & \underline{\underline{A}}_{yy} \end{bmatrix} \quad , \quad \underline{e} = \begin{bmatrix} \underline{e}_x \\ \underline{e}_y \end{bmatrix} \quad (5.33)$$

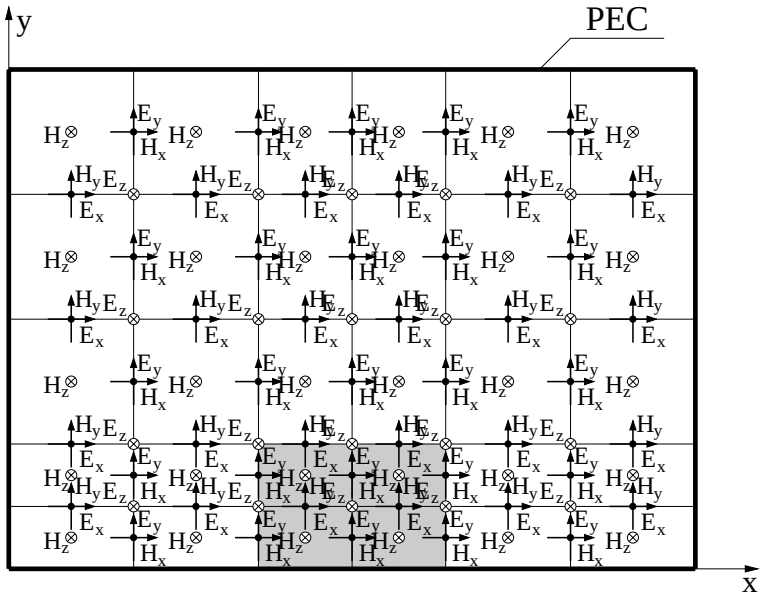
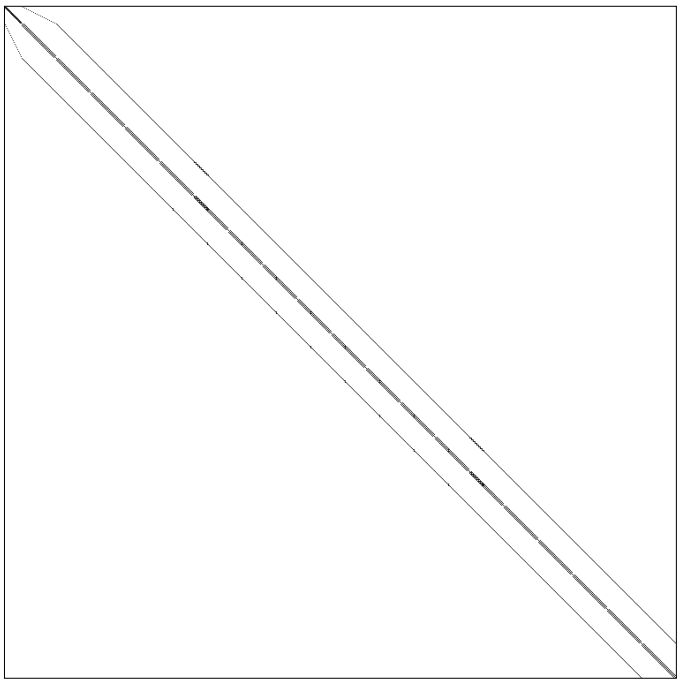


Figure 5.8: An example of nonuniform (graded) 6×4 Yee's grid applied to the cross-section of an image line.



NX = 20, NY = 20, N = 760, NZ = 3744

Figure 5.9: Sparsity pattern for matrix $\underline{\underline{A}}$ resulting from FDFD.

Symbols $\underline{e}_x$, $\underline{e}_y$ denote vectors of E_x and E_y field intensities in the discretization points. The resulting matrix $\underline{\underline{A}}$, shown in Fig. 5.9, is highly diagonally structured, sparse and nonsymmetric. Its size $N \approx 2N_xN_y$ is usually very large.

Solution of matrix eigenproblem. Problem (5.32) is solved using `DNAUPD` implementation of the Arnoldi method. Eight largest real part eigenvalues ($k = 8$) are investigated. No spectral transformations are applied to the matrix. Efficient realization of matrix-vector multiplication for sparse matrices is realized by `AMUX` subroutine from SPARSKIT package[1].

Computations are performed on the same super-scalar SGI Power Challenge XL computer as mentioned in Sections 5.1.2 and 5.2.1, on one R8000 processor.

Results. Normalized propagation constants β of eight dominant modes computed during the tests, calculated for $N = 1740$ ($N_x = N_y = 30$), are shown in Table 5.3.

Table 5.3: Eight dominant normalized propagation constants β/k_0 (k_0 is the wavenumber for a plane wave in the free space) of even modes of image line shown in Fig. 5.4, computed using the FDFD method. Parameters of the structure are: $a = 15.8\,\text{mm}$, $b = 7.9\,\text{mm}$, $w = 6.9\,\text{mm}$, $h = 3.2\,\text{mm}$, $\epsilon_r = 9$, and $f = 12\,\text{GHz}$.

Eigenvalue no.	FDFD ($N = 1740$, whole struct.)
1	$1.9762 + 0.0000$
2	$1.0264 + 0.0000$
3	$0.5583 + 0.0000$
4	$0.0000 + 1.2665$
5	$0.3710 + 1.2985$
6	$0.3710 - 1.2985$
7	$0.0000 + 1.6789$
8	$0.0000 + 1.8735$

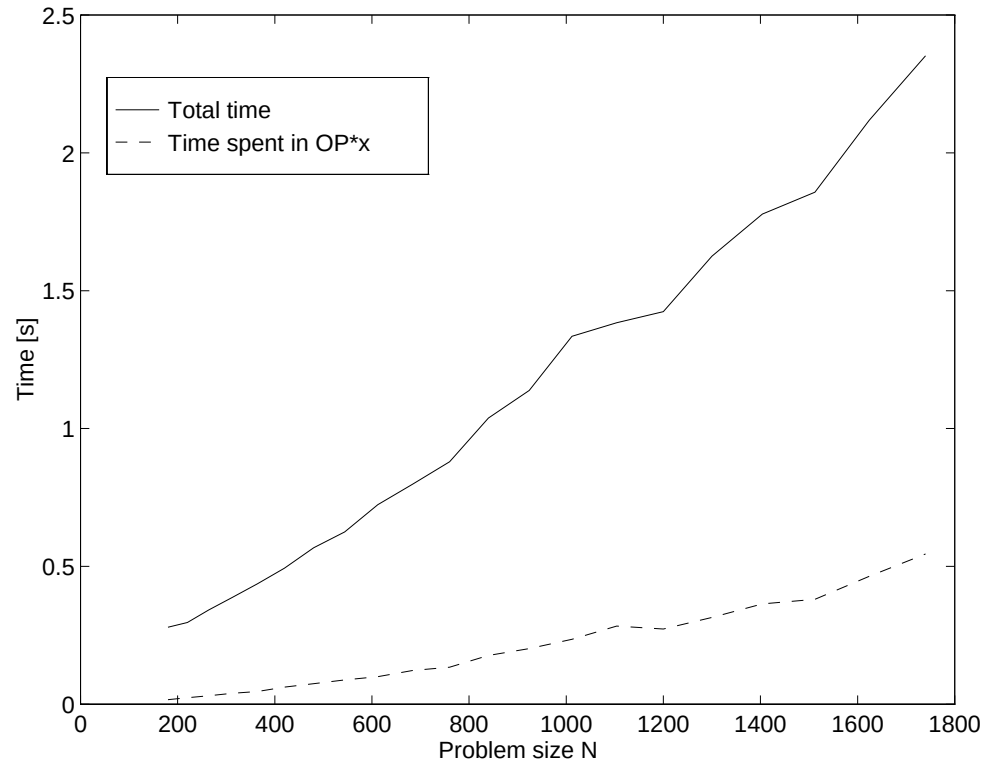
For the number of grid lines N_x and N_y varying from 10 to 30, the time-optimum size l of the Krylov subspace was experimentally selected to be [72]

$$l = k + 2\sqrt[4]{N} \tag{5.34}$$

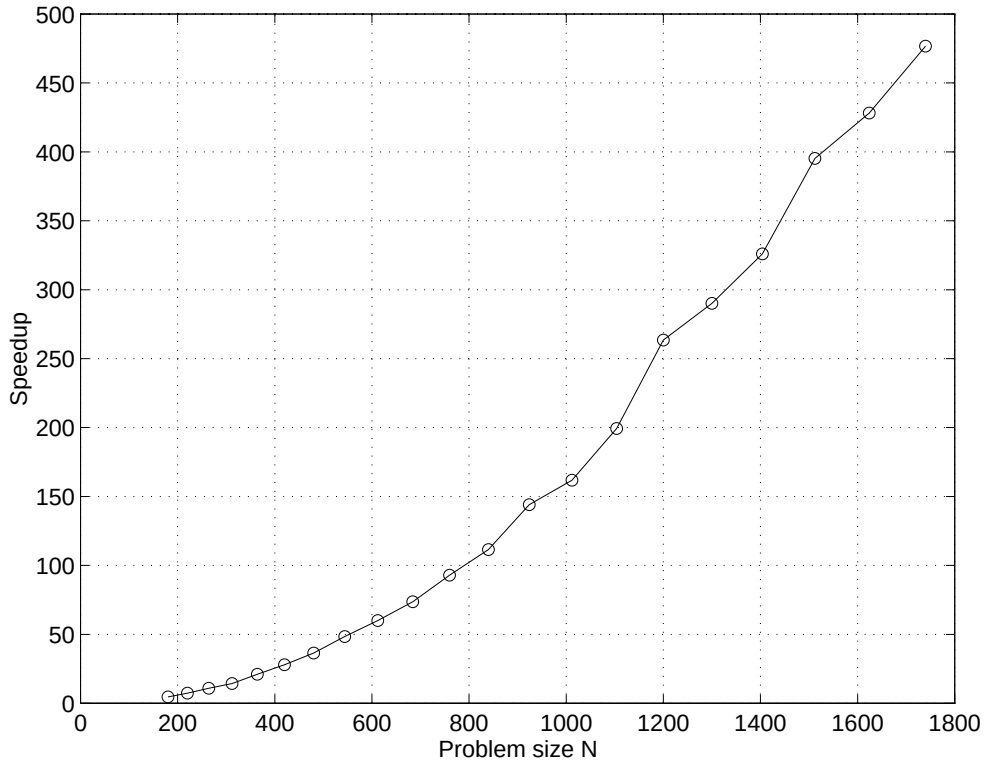
where N is the problem size and $k = 8$ is the number of eigenvalues to be calculated.

Computation times of the `DNAUPD` subroutine for l defined by (5.34) and `TOL` $= 10^{-6}$ are presented in Fig. 5.10(a). Comparison of the computation times of the Arnoldi procedure versus the QR method is shown in Fig. 5.10(b). We can see dramatic performance improvement when `DNAUPD` is used. It stems from the fact that for sparse matrices the dependence of calculation time on problem size N is nearly linear, in contrast to $\mathcal{O}(N^3)$

[1] Available via anonymous ftp from `ftp://ftp.cs.umn.edu/dept/sparse/SPARSKIT2.tar.gz` or as a link from Yousef Saad home page `http://www.cs.umn.edu/~saad`.



(a)



(b)

Figure 5.10: Computation times of Arnoldi `DNAUPD` subroutine (a) and speedup over QR `DGEEV` eigensolver (b).

dependence observed for the QR method where the matrices have to be converted into a dense format.

5.3.2 Comparison of classical conversion methods

Various approaches discussed in previous sections are compared in Table 5.4. In contrast to the Galerkin and FEM methods, the time spent in the product of a very sparse matrix resulting from the FDFD and a vector is small compared to the total computational time of the FDFD. In effect, the strategy incorporating the FDFD and Arnoldi methods is the fastest from all strategies summarized in Table 5.4. The approach solving the generalized eigenproblem generated by the FEM is 2 times more expensive (due to the large computation time of the LU decomposition). In fact, this speedup is even greater, because the time for the FDFD/Arnoldi method corresponds to the computation of 8 eigenvalues, while for the FEM/Arnoldi method only to the 4 ones. Additionally, the application of the LU decomposition requires $\mathcal{O}(N)$ more memory for storing the matrix in the decomposed form. Other strategies are much slower, e.g. the Galerkin method with the Arnoldi solver is for the same size of the problem is 40 times slower, while this figure is as high as 500 or 600 when the QR eigensolver is used.

Table 5.4: Approximate solution times of eigenproblems of size $N \approx 1800$ arising in the analysis of an image line by means of the Galerkin method (GM), FEM and FDFD.

Conversion method	GM	FDFD	GM	FEM	FEM	FDFD
Matrix	dense	sparse	dense	sparse	sparse	sparse
Eigensolver	QR	QR	Arnoldi	LU+ SI	LU+Arnoldi	Arnoldi
Spectral transformation	none	none	none	shift-invert	shift- invert	none
Eigenvalues computed	N	N	4	4	4	8
Time [s]	1500	1250	100	5 + 8	5 + 0.6	2.5
Acceleration	**600**	**500**	**40**	**5**	**2**	**1**

It was shown in this test that *computational complexity* of the Arnoldi algorithm is $\mathcal{O}(N)$ for sparse matrices, in contrast to $\mathcal{O}(N^3)$ dependence reported for the QR method. The amount of memory required by the Arnoldi algorithm (*memory usage*) is $\mathcal{O}(N)$ for sparse matrices and $\mathcal{O}(N^2)$ for dense ones, while in order to apply the QR method $\mathcal{O}(N^3)$ of the memory is needed. A comparison of time and memory requirements of all tested solvers is shown in Table 5.5 along with the computational complexity and memory usage for other eigensolvers discussed in App. A.

It can be seen that, in general, the memory used by the methods based on matrix transformations to store the matrix is $\mathcal{O}(N^2)$. This is due to the fact that in order to perform the transformations the matrices must be stored explicitly in a dense format. In

Table 5.5: Comparison of computational complexity and memory usage of various numerical algorithms computing eigenvalues and corresponding eigenvectors in dependence on matrix format. Symbols "−" denote, that the case is not applicable.

Numerical methods		Computational complexity			Memory usage					
					Algorithm			Matrix		
		dense	on-line	sparse	dense	on-line	sparse	dense	on-line	sparse
Matrix transformation	QR (QZ)	$\mathcal{O}(N^3)$	–	–	$\mathcal{O}(N^2)$	–	–	$\mathcal{O}(N^2)$	–	–
	bisection	$\mathcal{O}(N^2)$	–	–	$\mathcal{O}(N^2)$	–	–	$\mathcal{O}(N)$	–	–
	Jacobi	$\mathcal{O}(N^3)$	–	–	$\mathcal{O}(N^2)$	–	–	$\mathcal{O}(N^2)$	–	–
Iterative	subspace iteration Arnoldi/Lanczos nonsymmetric Lanczos	$\mathcal{O}(N^2)$		$\mathcal{O}(N)$	$\mathcal{O}(N)$			$\mathcal{O}(N^2)$	$\mathcal{O}(N)$	

the bisection method, the requirements for the memory are lower, because the method is only applicable to tridiagonal matrices. Computational complexity of this method is also smaller, because this algorithm is able to compute only selected eigenvalues. A common constraint of bisection and Jacobi algorithms is that they can be applied only to symmetric problems. It should be noted, that when accurate engineering calculations are required bisection and Jacobi algorithms are less effective than the QR method (see App. A for discussion).

In the iterative algorithms the information about the matrix is acquired in the form of matrix-vector product. In this case the matrix has not to be stored in a dense format, but in an appropriate sparse format or even it may not to be explicitly stored at all if on-line matrix-vector computations are performed. In this last case, the elements of the matrix can be computed using, for instance, lookup tables (see Sec. 5.1.2). If the matrix is sparse or on-line matrix computations are performed the memory required for the storage of the matrix elements (or lookup tables) is only $\mathcal{O}(N)$. Since the iterative algorithms compute only selected eigenvalues (and corresponding eigenvectors), the extra memory used by the algorithm is also $\mathcal{O}(N)$. Computational complexity of the iterative algorithms is $\mathcal{O}(N)$ if sparse matrix problems are solved, and especially, if the matrix is highly diagonally structured (as for the FDFD method). For dense matrices and on-line matrix computations this complexity is $\mathcal{O}(N^2)$. Among iterative algorithms included in Table 5.5, the nonsymmetric Lanczos algorithm requires the least memory. However, compared to the subspace iteration and Arnoldi methods, it is not so stable (see App. A for details).

We can also see that the application of the QR method always requires the largest

computer resources, while the smallest requirements for the time and the memory can be expected when an iterative eigensolver is applied to sparse matrices. If the iterative solver is used the matrix can also be computed on-line. Application of this strategy can be very advantageous when the memory is critical and eigenproblems with dense matrices (e.g. resulting from the Galerkin method) are solved. In this case the requirements for the memory are much smaller, i.e. $\mathcal{O}(N)$ rather than $\mathcal{O}(N^2)$.

The next two tests, discussed in Sections 5.3.3 and 5.3.4, concern acceleration of the iterative eigensolvers incorporated in the analysis of rotationally symmetric resonators by means of the FDFD method. The acceleration can be obtained when advanced spectral transformations involving Chebyshev polynomials and finite impulse response (FIR) digital filters are applied, which both improve the convergence of the solver to the required eigenvalues.

5.3.3 Rotationally symmetric dielectric resonator

In this test we show how the application of the preconditioning technique based on Chebyshev polynomials can speed up the solution of the eigenproblem arising in the analysis of inhomogeneous resonators. An advantage of this technique over the shift-invert one is that it does not require computation of the matrix inverse or decomposition (see Sec. A.1.2). In this test we also compare performance of the preconditioned Arnoldi solver with the performance of a subspace iteration solver, reported in the literature. Moreover, the accuracy of the developed FDFD code incorporating dual Yee's grid is compared with the accuracy of a FDFD code described in the literature [113] that used the single grid approach.

Problem and structure. We are interested in accuracy of computing of the smallest resonant frequencies f of the rotationally symmetric resonator shown in Fig. 5.11(a), characterized by the following parameters: $D = 0.68\,\text{in} = 172.72\,\text{mm}$, $H = 0.3\,\text{in} = 76.2\,\text{mm}$, $b = 1.02\,\text{in} = 259.08\,\text{mm}$, $l = 0.6\,\text{in} = 152.4\,\text{mm}$, $\epsilon_1 = \epsilon_2 = 1$, $\underline{\underline{\epsilon}} = 35.74$. Since this structure is symmetric in the z-direction we can analyze only one half and compute independently even (with PEC symmetry plane) or odd (with PMC symmetry plane) modes.

Formulation. Depending on the modes of interest (hybrid, TE or TM), the analysis is based on three different spurious-free formulations. If hybrid modes ($m > 0$) are to be found we apply formulation (2.35) for transverse electric flux, while if TE or TM modes ($m = 0$) are of interest scalar formulations (2.41) for D_ϕ and (2.42) for B_ϕ are used, respectively. Moreover, we use the nomenclature in which even hybrid and odd hybrid modes are distinguished and denoted with HE or EH symbols.

Conversion to matrix problem. One half of the $\phi = \text{const}$ plane of the resonator is discretized using regular $N_r \times N_z$ Yee's grid, shown in Fig.5.11(b), where N_r and N_z

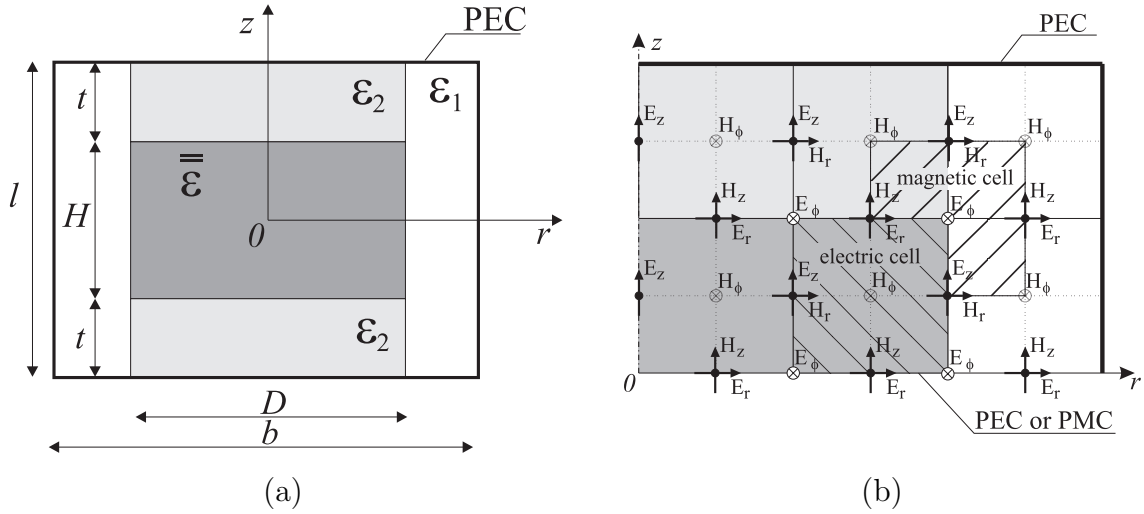


Figure 5.11: Dimensions of the rotationally symmetric dielectric resonator structure (a) and an example of uniform 3×2 grid applied to a half of the ϕ = const plane of the resonator (b).

denote the number of grid points in the r- and z-direction, respectively. The size of the grid is chosen so that all dielectric boundaries coincide with the edges of electric cells.

Differential operators in each of the eigenproblems (2.35), (2.41) or (2.42) are directly discretized, using the second order approximations. The concept of effective permittivity [64] is applied on dielectric boundaries.

This leads to the eigenproblems in the form

$$\underline{\underline{A}}\,\underline{v} = \lambda \underline{v} \tag{5.35}$$

It should be stressed that matrix eigenproblem (5.35) is a standard one and $\lambda = \omega^2 = 4\pi^2 f^2$. For eigenproblem (2.35)

$$\underline{\underline{A}} = \begin{bmatrix} \underline{\underline{L}}^{D}_{rr} & \underline{\underline{L}}^{D}_{rz} \\ \underline{\underline{L}}^{D}_{zr} & \underline{\underline{L}}^{D}_{zz} \end{bmatrix} \quad , \quad \underline{v} = \begin{bmatrix} \underline{d}_r \\ \underline{d}_z \end{bmatrix} \tag{5.36}$$

while

$$\underline{\underline{A}} = \underline{\underline{L}}^{D}_{\phi\phi} \quad , \quad \underline{v} = \underline{d}_\phi \tag{5.37}$$

$$\underline{\underline{A}} = \underline{\underline{L}}^{B}_{\phi\phi} \quad , \quad \underline{v} = \underline{b}_\phi \tag{5.38}$$

for eigenproblems (2.41) and (2.42), respectively. Symbols $\underline{d}_*$ and $\underline{b}_*$ denote vectors of D_* and B_* flux densities in the discretization points. Matrix $\underline{\underline{A}}$ is sparse and nonsymmetric. Its size $N \approx 2N_rN_z$ in the case (5.36) and $N \approx N_rN_z$ in the case (5.37) or (5.38).

Solution of matrix eigenproblem. Resulting eigenvalue problems are solved using `DNAUPD` implementation of the Arnoldi method. Two approaches are used in our tests. The first one is to use no preconditioning and investigate the eigenvalues of the smallest

magnitude (or equivalently of the smallest real part). In this case, each Arnoldi iteration requires calculation of a matrix-vector product $l-k$ times at the most (except for the first one which needs l products), where k is the number of eigenvalues to find (`NEV`) and l is the size of the Krylov subspace (`NCV`). Internally, each iteration of the Arnoldi algorithm requires the solution of a $l \times l$ eigenproblem by means of the QR method. External calculation of the matrix-vector product and internal solution of the $l \times l$ eigenproblem dominate the total time of the Arnoldi algorithm.

The second approach is to apply Chebyshev preconditioner (see Sec. A.1.2.3) and compute the eigenvalues of the largest magnitude (after the transformation the smallest real part eigenvalues become the largest magnitude ones). Chebyshev polynomial of order q is applied to an operator matrix via a recurrence formula (see Alg. 6 in App. A). Therefore, each time when the product of the operator matrix and the vector is required, q matrix-vector products are calculated. Since the eigenvalues of the operator matrix are known, corresponding eigenvalues of $\underline{\underline{A}}$ are computed using the Rayleigh quotient (compare Alg. 5 in App. A).

Results. Using appropriate PEC or PMC conditions in the central section of the structure, we compute resonant frequencies f for each type of modes independently. In consequence, an equidistant 51×30 grid is applied to one half of the structure. The size of the resulting eigenproblem is $N \approx 3000$ for hybrid modes, while $N \approx 1500$ for $\mathrm{TE}_0 m$ and $\mathrm{TM}_0 n$ modes. Calculated resonant frequencies are presented in Table 5.6 along with the results for the FDFD method based on a single grid [113] and the mode matching method [128]. The FDFD with Yee's grid appears to be more accurate (except for TE_{01} and TM_{01} modes) than the single grid formulation and the maximum difference relative to the mode matching method is 0.22%.

Table 5.6: Comparison of resonant frequencies [GHz] for the resonator loaded with a dielectric rod, shown in Fig. 5.11(a). Parameters of the structure are: $D = 0.68\,\mathrm{in}$, $H = 0.3\,\mathrm{in}$, $b = 1.02\,\mathrm{in}$, $l = 0.6\,\mathrm{in}$, $\epsilon_1 = 1$, $\epsilon_2 = 1$, $\underline{\underline{\epsilon}} = 35.74$.

Mode	TE_{01}	EH_{11}	HE_{11}	TM_{01}	HE_{21}	EH_{21}
Present FDFD	3.433	4.229	4.318	4.541	5.000	5.323
FD-SIC [113]	3.429	4.205	4.310	4.542	4.992	5.311
Mode matching [128]	3.428	4.224	4.326	4.551	5.00	5.33
Difference rel. to [128]	−0.15%	−0.12%	0.18%	0.22%	0.00%	0.13%

An analogous comparison of resonant frequencies for two anisotropic dielectric resonator structures can be found in [76] and [75]. The conclusions concerning the accuracy of our approach are similar and the maximum difference relative to the mode matching method does not exceed 0.25%.

To test the efficiency of the Arnoldi solver and compare it with the SI we compute two dominant ($k = 2$) TE_0 odd modes for the structure described above, with accuracy to

the third decimal. The results are presented in Table 5.7. For unaccelerated case we can observe that as the size of subspace l increases, the number of iterations decreases and so does the number of matrix-vector products. Nevertheless, for larger values of l the total computation time increases due to more costly internal $l \times l$ eigenproblem solutions. On the other hand, when l decreases, the number of required matrix-vector products rapidly grows up and the total time also increases.

Table 5.7: Computational cost of analysis of the isotropic dielectric resonator in unaccelerated case and calculation times for the optimal order q of acceleration polynomial for the problem of size $N \approx 1500$.

l	4	5	10	15	20	25	30	35	40	45	50
iterations	5343	1483	155	84	52	37	28	30	20	20	17
mat-vec	10688	4443	1228	1083	933	850	782	987	760	860	816
time [s]	106.6	43.2	12.6	12.7	12.8	13.3	13.4	17.9	15.3	18.7	19.1
acc. time [s]	3.9	3.2	2.9	3.1	3.1	3.1	3.5	2.9	3.3	3.8	4.3
(q_{opt})	(60)	(20)	(10)	(20)	(20)	(10)	(30)	(20)	(20)	(20)	(20)

These trade-offs can easily be mitigated when the Arnoldi method is combined with the Chebyshev acceleration technique. Implementation of this technique requires definition of c and e parameters in accordance with equations (A.27). For comparison purposes, these parameters are selected in the manner described in [113]. Application of the preconditioner causes that the number of iterations taken is significantly reduced. For example, for $10 \leq p \leq 20$ the decrease is by the factor of ≈ 15 for $q = 10$ and ≈ 30 for $q = 20$. This causes that the total calculation time is dominated by the time spent in matrix-vector operations. The number of computed products is shown in Fig. 5.12 for a few different values of Chebyshev polynomial order q. We see that there are many cases where we need only 700 products to get convergence, especially for $10 \leq p \leq 35$ and $q = 10$. This result is 30% better than the one reported in [113], where the subspace iteration solver needed 1000 products. For large values of q the Arnoldi algorithm makes only 1 iteration, resulting in pq matrix-vector products. Thus, if l in not sufficiently small the calculations are inefficient.

As can be seen in Table 5.7, the optimal total calculation time for the unaccelerated case was 12.6 s for $l = 10$ on our SGI Power Challenge computer used. The shortest times are observed for l and q offering the smallest number of matrix-vector products i.e. 2.9 s for $q = 10$ and l the same as above, which gives the speedup factor > 4. All the points located in Fig. 5.12 below a thick dotted line indicate the total calculation times smaller than 4 s. In all these points the speedups are greater than 4. Much more impressive speedups are offered for small memory calculations ($l = 4$) causing that for $50 \leq q \leq 90$ the computation times are at most 40% longer than the optimal one. General observations for other mode types (TM, HE, EH) are similar.

It was shown in this test that the performance of the Arnoldi solver with Chebyshev

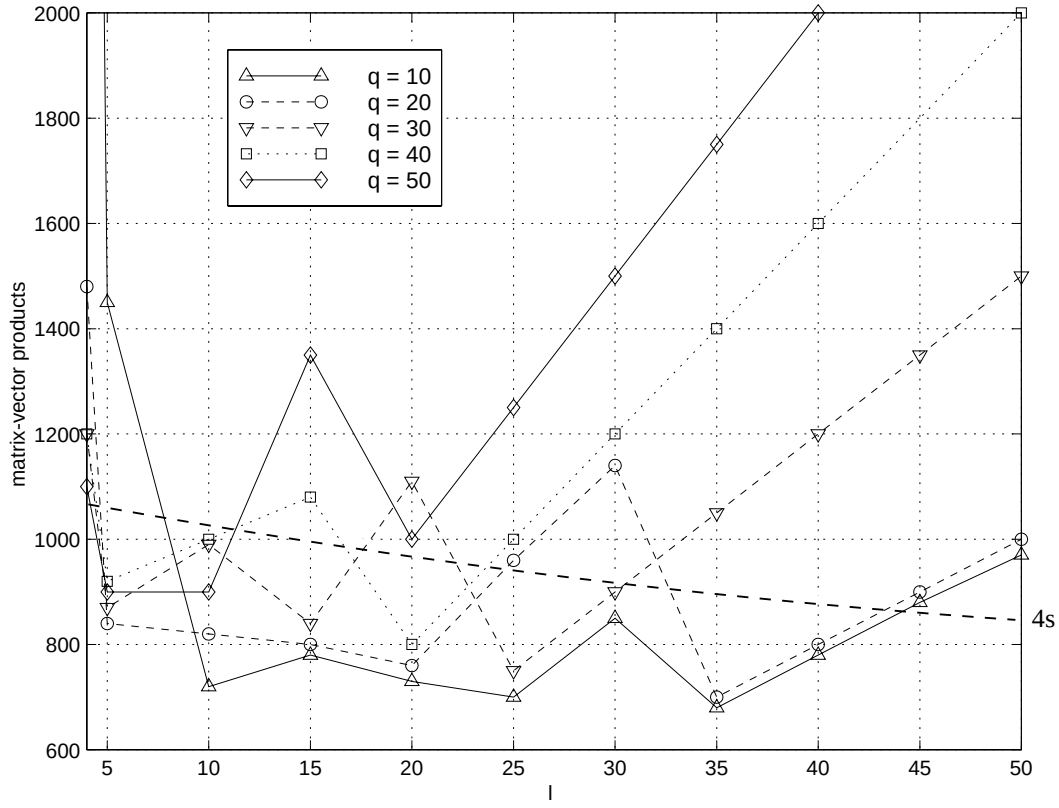


Figure 5.12: The number of matrix-vector products taken in the analysis of the isotropic dielectric resonator for different values of l and q.

preconditioning strongly depends on the choice of the subspace size l and the polynomial order q. Once they are properly chosen, the implicitly restarted Arnoldi method can offer considerable memory and time savings and can be more efficient than the subspace iteration in the FDFD analysis of rotationally symmetric resonator structures. It was also found that the FDFD resonator formulation based on Yee's dual grid is more accurate in implementing boundary conditions than the one based on single grid.

5.3.4 Rotationally symmetric open resonator

In this test we show application of another powerful preconditioning technique, based on finite impulse response (FIR) digital filters, to the solution of the eigenproblem arising in the analysis of high order modes of a rotationally symmetric inhomogeneously filled open resonator. An alternative approach that is capable of computing the resonant frequencies of the high order modes is the shift-invert. However, it requires computation of the matrix inverse or decomposition, what is not possible in practice if the matrix has a very large size. The FIR filter preconditioning overcomes this difficulty because analogously to the Chebyshev preconditioning it does not require computation of any matrix inversion

or decomposition. In this test we also investigate the accuracy of a correction for the computed resonant frequencies, due to the numerical dispersion.

Problem and structure. We investigate resonant frequency of quasi $\mathrm{TEM}_{0,0,20}$ mode (in a Gaussian beam notation) of a typical inhomogeneously loaded open resonator, shown in Fig. 5.13. The bottom and the top covers of this rotationally symmetric resonator are respectively plane and spherical metallic mirrors. Since the analytical solutions exist for the structure in Fig. 5.13, the accuracy of the computed resonant frequencies can easily be verified.

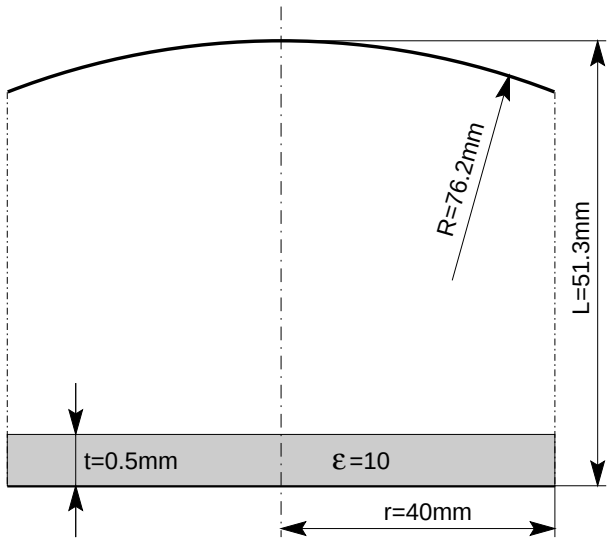


Figure 5.13: Dimensions of the open resonator structure loaded with a dielectric disc.

Formulation. The investigated quasi $\mathrm{TEM}_{0,0,20}$ mode corresponds to the mode of azimuthal mode index $m = 1$. We use the same formulation for hybrid modes as in the previous example, i.e. (2.35) for transverse electric flux $\vec{D}_t$.

Conversion to matrix problem. We use the discretization scheme analogous to the one described in Sec. 5.3.3. The scheme involves constrained dual Yee's grid. We also apply the effective permittivity concept on dielectric boundaries, which, in this case, do not coincide with the edges of electric cells. Moreover, a conformal technique is used to model curved mirrors [16, 50, 64].

This approach leads again to the standard eigenproblems of form (5.35), where $\underline{\underline{A}}$ and $\underline{v}$ are described with (5.36).

Solution of matrix eigenproblem. Since the size of the resonator is large compared to the length of the wave in vacuum λ, large matrix eigenvalue problem has to be solved and numerical dispersion may introduce large errors into computer simulations. Moreover, the mode of interest is relatively high, which means that one has to find eigenvalues located

far from both spectrum ends. To overcome these difficulties polynomial preconditioning, involving a FIR bandpass digital filter, combined with the Arnoldi method (`DNAUPD`) was used. The polynomial preconditioning serves two goals. Firstly, it allows one to select modes having resonance in a prescribed frequency range and, secondly, it accelerates the convergence of the solver (see Sec. A.1.2).

To explain the polynomial character of the FIR filter involved we show the details of the filter construction. Since design procedures for FIR filters are developed in the normalized ω-domain, a transformation of the problem from the frequency domain f should precede the design. It can be realized by the following transformation

$$\omega = \arccos\left(\frac{f-c}{e}\right) \tag{5.39}$$

where c is the center frequency of the matrix spectrum and e is the radius of the matrix spectrum. Therefore, the filter is, in fact, designed for a center frequency $\omega_0 = \arccos(\frac{f_0-c}{e})$. Frequency response of the linear phase FIR filter of order $2q$ can be written in the form

$$H(\omega) = \sum_{n=-q}^{q} h(n)e^{-j\omega n} = \sum_{n=0}^{q} b_n \cos(n\omega) \tag{5.40}$$

where b_n are coefficients related to the impulse response $h(n)$ of the filter. Thus, using (5.39) we get

$$H(f) = \sum_{n=0}^{k} b_n \cos\left[n \arccos\left(\frac{f-c}{e}\right)\right] = \sum_{n=0}^{k} b_n T_n\left(\frac{f-c}{e}\right) \tag{5.41}$$

where T_n is the Chebyshev polynomial of degree n.

In our case, the linear phase 80-th order FIR filter ($q = 40$) with a flat monotonically decreasing pass band and an equiripple stop band, shown in Fig. 5.14, was applied [126]. For a desired mode, the center frequency f_0 of the filter is calculated from an approximate formula (discussed below). The Arnoldi algorithm finds several eigenvalues around the center frequency f_0. The desired one is identified by examination of the corresponding fields.

It should be noted that the approach with polynomial filtering avoids matrix inversion/decomposition, which are both very ill conditioned tasks. Moreover, in the iterative method, such as the Arnoldi method, polynomial preconditioning of the matrix can be realized with a recursive scheme, involving matrix-vector products only.

Results. Numerical dispersion can be reduced by increasing the order of the FDFD approximation or decreasing the discretization step. However, both approaches lead to large complication of the code and/or increase of computation time. Instead, we can compute a correction, based on the results for a homogeneous (empty) structure. In our case, comparison of the calculated frequency with the one computed from analytical approach gives the numerical dispersion error Δf of order -0.183% [126].

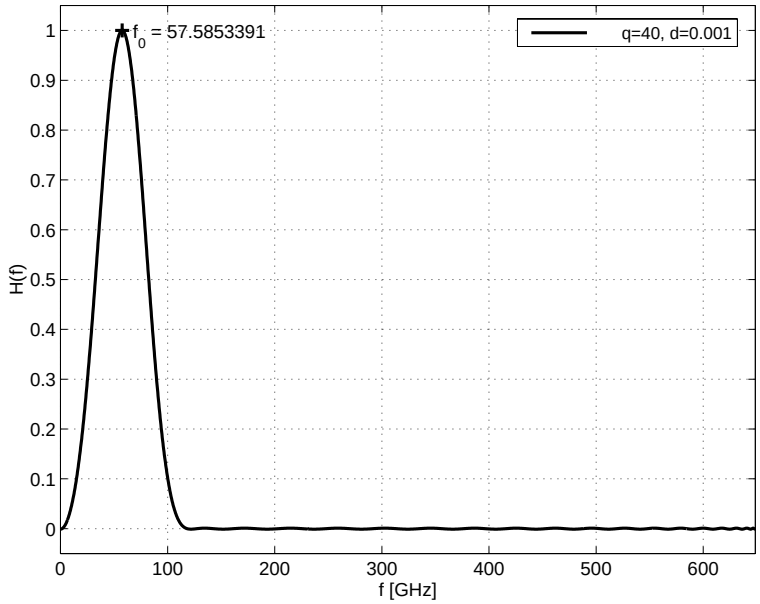


Figure 5.14: Frequency response of a 80-th order FIR filter, designed for the center frequency $f_0 = 57.5853391\,\mathrm{GHz}$ and the ripple in a stop band $d = 0.001$.

The grid size, which is used for computation of the resonant frequency f_{FD} of quasi $\mathrm{TEM}_{0,0,20}$ mode for the structure shown in Fig. 5.13, is chosen to be $\Delta r = \Delta z = 0.158\,\mathrm{mm}$ ($\approx \lambda/30$ for the resonant frequency). The matrix problem to be solved involves 160.000 unknowns! As a preconditioner we use the FIR filter shown in Fig. 5.14, designed for the center frequency $f_0 = f_{\mathrm{theo}} + \Delta f$, where $f_{\mathrm{theo}} = 57.6909135\,\mathrm{GHz}$ is computed from an analytical formula for a single dielectric layer case. Since the majority of the volume is air the dispersion error is assumed to be the same as for the homogeneous case. Accordingly, the computed frequency $f_{\mathrm{FD}} = 57.5778935\,\mathrm{GHz}$ is corrected by +0.183%. The corrected resonant frequency $f_{\mathrm{corr}} = 57.6836978\,\mathrm{GHz}$ is only 0.013% off the value given by the analytical approach. This validates the use of the dispersion correction also for the multilayer case. Field plots of the computed $\mathrm{TEM}_{0,0,20}$ mode are shown in Fig. 5.15.

Application of this digital FIR filter preconditioning technique allows us to compute the eigenvalues located in the middle of the spectrum at a reasonable time, 8 hours. One can consider this time as high, but it should be born in mind that the size of the matrix resulting from the FDFD is as high as 160.000 and application of another preconditioning technique, e.g. shift-invert, for this purpose would lead to much higher computational times and memory requirements and could be not possible in practice.

It was shown in this test that rotationally symmetric inhomogeneous open resonators can be accurately analyzed using the FDFD method. Numerical dispersion of high quasi TEM_{00q} modes can be effectively corrected by considering the numerical dispersion in a homogeneous resonator. It was also verified that resonant frequencies of high order modes of a rotationally symmetric open resonators can be efficiently computed using the FDFD method and the Arnoldi method with bandpass FIR filter preconditioning.

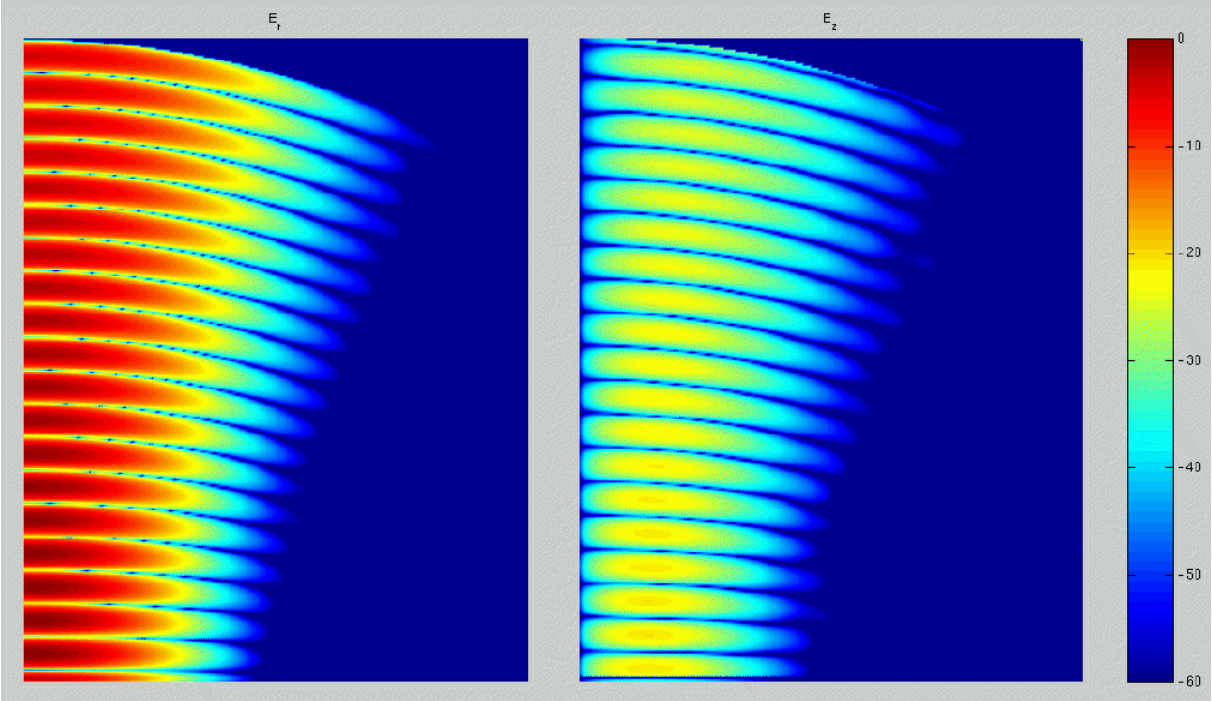


Figure 5.15: Normalized rms [dB] field plots for $\mathrm{TEM}_{0,0,20}$ mode in inhomogeneous resonator.

5.4 Coupled mode method

The test described in this section shows the application of the coupled mode method, a kind of a hybrid method, to the analysis of an inhomogeneous nonuniform waveguide loaded with a ferrite toroid. This inhomogeneous structure of complex shape, partially filled with the material characterized by a complex permeability tensor, is very difficult to analyze by means of classical methods. Any analytical formulation describing this computational problem involves complex operators for at least three field components, what leads to large eigenproblems with complex matrices.

In the initial phase of the coupled mode method, basis functions are computed as a solution for a simplified (basis) structure. The structure is selected so that its analysis is less expensive than the analysis of the original structure. Since the second stage of the coupled mode method relies on the solution of an usually small eigenproblem (see Sec. 3.2.1) the entire method can be much more efficient than any classical approach.

5.4.1 Inhomogeneous nonuniform waveguide loaded with a ferrite toroid

Problem and structure. We are interested in finding the propagation constant of the fundamental mode of the reduced height waveguide shown in Fig. 5.16(a), loaded with a ferrite toroid. Parameters of the structure are: $a = 16.0\,\mathrm{mm}$, $b = 9.4\,\mathrm{mm}$, $d = 6.75\,\mathrm{mm}$, $h = 15.8\,\mathrm{mm}$, $w = 0.5\,\mathrm{mm}$, $h_d = 9.4\,\mathrm{mm}$, $\epsilon_r = 13$, and the frequency range of interest

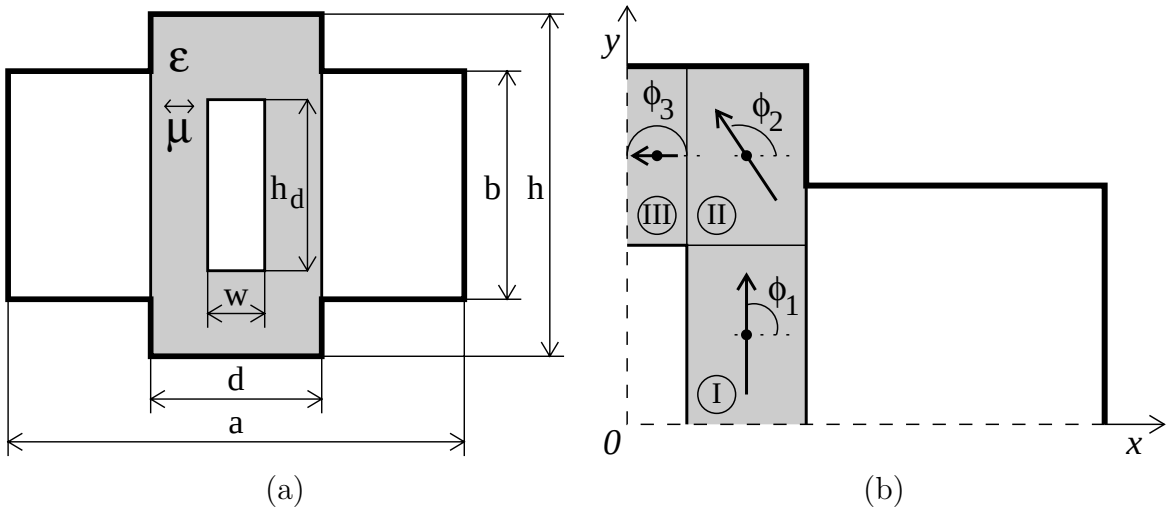


Figure 5.16: Cross-section of the phase shifter structure (a) and division of the toroid to the subregions (b).

ranges from 5 to 6 GHz. Magnetization M_r inside the toroid is influenced by the impulses of electrical current flowing via the wire placed in the central slot of the toroid. Reversing the direction of the current causes the change of M_r sign. For the ferrite toroid at hand, the magnetization reaches the value of $M_r \approx 840$ Gauss. Assuming that the magnetization vector is normal to the z-direction $\underline{\underline{\mu}}$ is defined as

$$\underline{\underline{\mu}} = \begin{bmatrix} \mu \sin^2\phi + \mu_0 \cos^2\phi & (\mu_0 - \mu)\sin\phi\cos\phi & -j\kappa\sin\phi \\ (\mu_0 - \mu)\sin\phi\cos\phi & \mu\cos^2\phi + \mu_0\sin^2\phi & j\kappa\cos\phi \\ j\kappa\sin\phi & -j\kappa\cos\phi & \mu \end{bmatrix} \tag{5.42}$$

where $\mu = \mu_0$, $\kappa = \mu_0\gamma M_r/f$ (gyromagnetic constant $\gamma = 2.8$ [MHz/Oe]) and ϕ is the direction of the magnetization vector (see Fig. 5.16(b)).

Since the structure is symmetric only one of its quarters can be considered in the analysis.

For the purpose of phase shifter application of the analyzed structure (see App. B), we also investigate *nonreciprocal phase shift* $\Delta\Theta = \beta^+ - \beta^-$, where β^+ and β^- are propagation constants of the dominant modes for two opposite directions of the magnetization, i.e. $+M_r$ and $-M_r$ respectively.

Basis functions. Basis transverse electric fields are computed using the FDFD method similar to the one described in Sec. 5.3.1, for the structure filled with dielectric ($\epsilon_r = 13$) and non-gyrotropic ($\underline{\underline{\mu}} = \mu_0$) material. Corresponding transverse magnetic fields are calculated from Maxwell's equations. Due to the symmetry only one quarter of the structure (see Fig. 5.16(b)) is analyzed, using 64×79 grid chosen so that it conforms to the dielectric boundaries.

Formulation of the problem. Taking into consideration only the fundamental mode coupled mode method formalism (3.67) leads to the following simple perturbation formula for calculating propagation constant β in the ferrite [71]

$$\beta = \beta_1 + \frac{\omega \int_{S_f} \vec{h}_1^* \cdot (\underline{\underline{\mu}} - \mu_0) \cdot \vec{h}_1 \, ds}{\int_S \hat{z} \cdot (\vec{e}_1^* \times \vec{h}_1 + \vec{e}_1 \times \vec{h}_1^*) \, ds} \tag{5.43}$$

where ω is the angular frequency, β_1 is the propagation constant of the dominant mode in the basis structure, and $\vec{e}_1$, $\vec{h}_1$ are the corresponding basis electric and magnetic fields. Since the direction ϕ of magnetization vector M_r is a function of position across the ferrite cross-section S_f, so is permeability $\underline{\underline{\mu}}$. In order to account for this fact in the analysis we can divide S_f into subregions, e.g. I, II, and III shown in Fig. 5.16(b), and define ϕ in each of them independently, according to the local direction of M_r.

Results. Field configuration of the fundamental mode in the basis waveguide is similar to the rectangular waveguide TE_{10} mode, which has PMC at $x = 0$ and PEC at $y = 0$. The plot of corresponding propagation constant β_1, computed using the FDFD method described above, is shown in Fig. 5.17(a).

In the calculations of propagation constants β^+ and β^- in the ferrite structure we assume the following directions of magnetization vector M_r in regions I, II and III

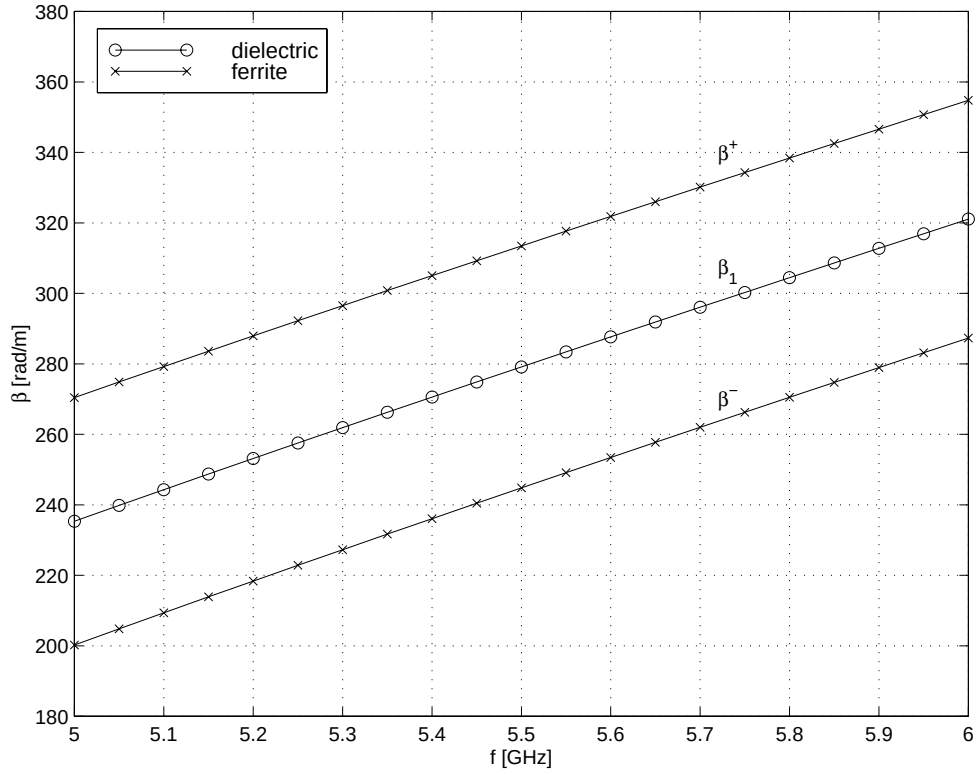
$$\phi_1 = \pi/2 \quad , \quad \phi_2 = \arctan\left(\frac{h - h_d}{w - d}\right) \quad , \quad \phi_3 = \pi \tag{5.44}$$

It was shown in [71] that such an approximation of the magnetization field improves accuracy of computed propagation constants. Plots of β^+ and β^- are shown in Fig. 5.17(a).

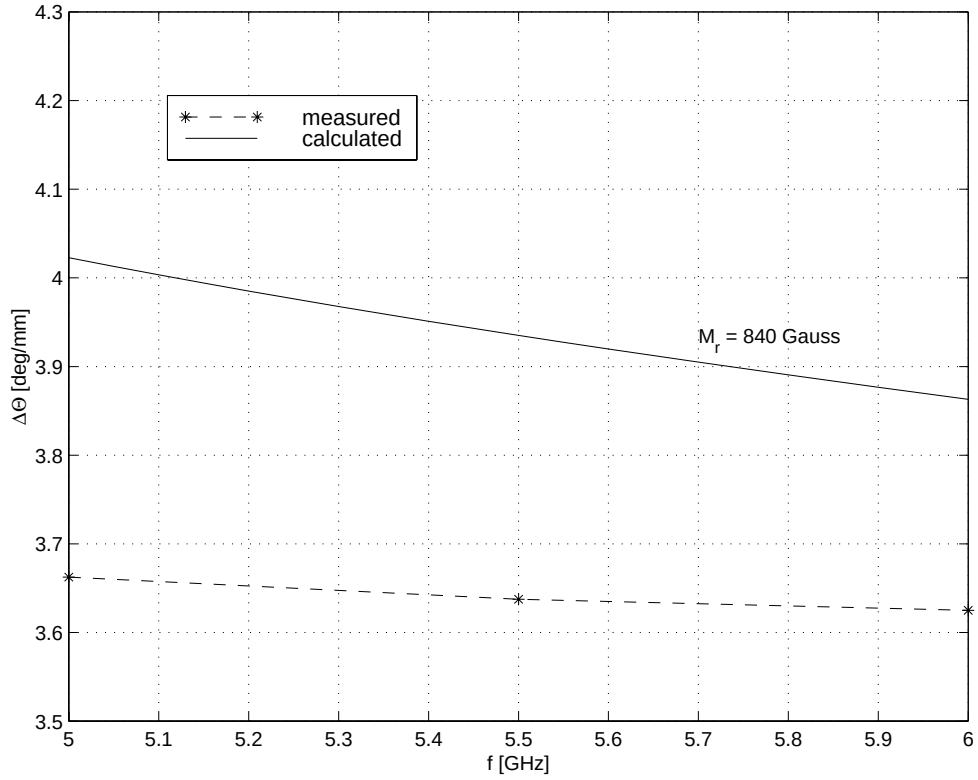
Nonreciprocal phase shift $\Delta\Theta$ calculated for $M_r = 840\,\mathrm{Gauss}$ is shown in Fig. 5.17(b). A validation of the presented results was performed by measurements of a physical ferrite phase shifter structure, described in App. B. For comparison, the measured values are also displayed in Fig. 5.17(b). We can see that computed values of $\Delta\Theta$ are $6.5 \div 10\%$ higher than the measured ones, but the variation of $\Delta\Theta$ has similar character.

There are three main reasons for the observed inaccuracy. One is the application of the coupled mode method, which should only be used when the perturbation of the fields in the basis structure caused by the introduction of the parameters for the original structure is sufficiently small. The second reason is that only the fundamental mode was taken into account (only one term in the expansion). Better accuracy would be expected when using, for example, the Galerkin method similar to the one described in Sec. 5.1.1, involving the expansions for a few higher order modes. The third reason for the discrepancy between the computed and measured results is a very approximate prediction of the magnetization field inside the ferrite toroid (an accurate prediction could be obtained as the solution of a magnetostatic problem for this structure).

It was shown in this test that the perturbation method (resulting from the coupled mode formalism) gives a simple formula for calculating propagation constants of the toroidal phase shifter structures. Application of the perturbation method involving basis



(a)



(b)

Figure 5.17: (a) Propagation constant β_1 of the fundamental mode in the basis dielectric structure and in the ferrite structure (β^+ and β^-) for two opposite directions of magnetization vector M_r. (b) Comparison of calculated nonreciprocal phase shift $\Delta\Theta$ [deg/mm] with measurements.

functions computed by means of the FDFD, enables one to approximate magnetization vector M_r inside the toroid much more accurately than other methods used so far.

5.5 Eigenfunction expansion methods

The eigenfunction expansion (EE) algorithms are another class of hybrid methods which enable one to compute dispersion characteristics of the waveguides, based on the solutions in a few frequency/propagation constant points. In this example, we test the most general eigenfunction expansion approach, described in Sec. 3.2.2, which can employ the eigenfunctions computed at arbitrary (ω, β) points from the dispersion diagram [86–88]. We show how large speedup can be gained if the EE is applied to the analysis of the image line investigated in the previous sections by means of various classical methods. We consider a few particular choices of basis/testing functions and their impact on accuracy and efficiency of the solution. Moreover, we discuss the performance of the EE in the analysis of a circular waveguide loaded with an anisotropic magnetic medium.

5.5.1 Inhomogeneous rectangular waveguide loaded with a dielectric slab

Problem and structure. In the frequency range from 0 to 20 GHz, we investigate dispersion characteristics of the dominant odd modes (with PMC symmetry plane) in the image line shown in Fig. 5.4. The dimensions of the structure are: $a = 15.8\,\text{mm}$, $b = 7.9\,\text{mm}$, $w = 6.32\,\text{mm}$, $h = 3.16\,\text{mm}$, and relative permittivity of the slab $\epsilon_r = 9$. The odd modes can be computed by taking only odd basis functions into expansion. We concentrate on all β-algorithms discussed in Sec. 3.2.2.1 and summarized in Table 3.1 (i.e. β-GS, β-S, β-G) and compare their properties.

Basis functions and reference characteristics. The application of the β-G algorithm requires computation of basis functions for a given frequency f_0. They are computed using FDFD method similar to the one described in Sec. 5.3.1, involving β-formulation (5.31) for transverse electric fields $\vec{E}_t$ and using 40×20 uniform grid that conforms to the boundaries of the slab. For this gird size the FDFD method leads to the sparse matrix eigenproblem of size $N = 1540$. Reference dispersion curves shown in Fig. 5.18 are determined with the same FDFD method.

The application of the β-S algorithm requires computation of basis functions for a given propagation constant β_0. They are determined using an analogous FDFD approach, involving ω-formulation for electric flux $\vec{D}_t$ that can be derived from (2.76) premultiplied with $-\hat{z} \times (\cdot)$.

Basis functions in the β-GS algorithm are evaluated using both, β- and ω-, formulations of the FDFD.

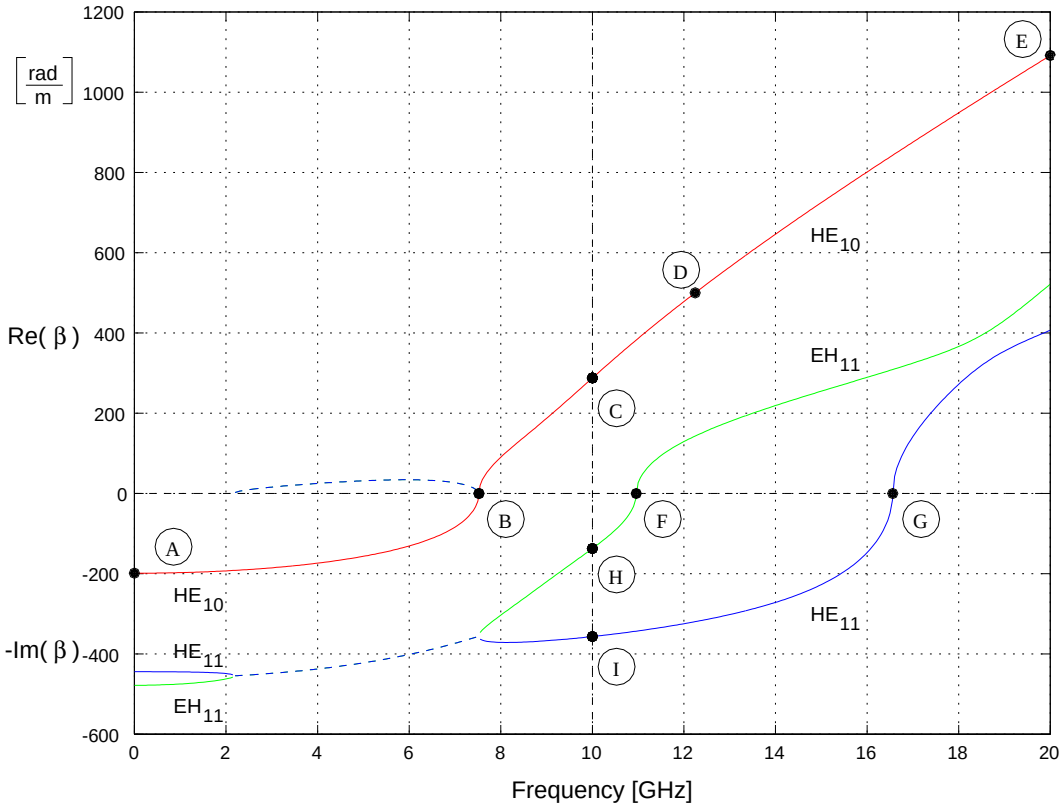
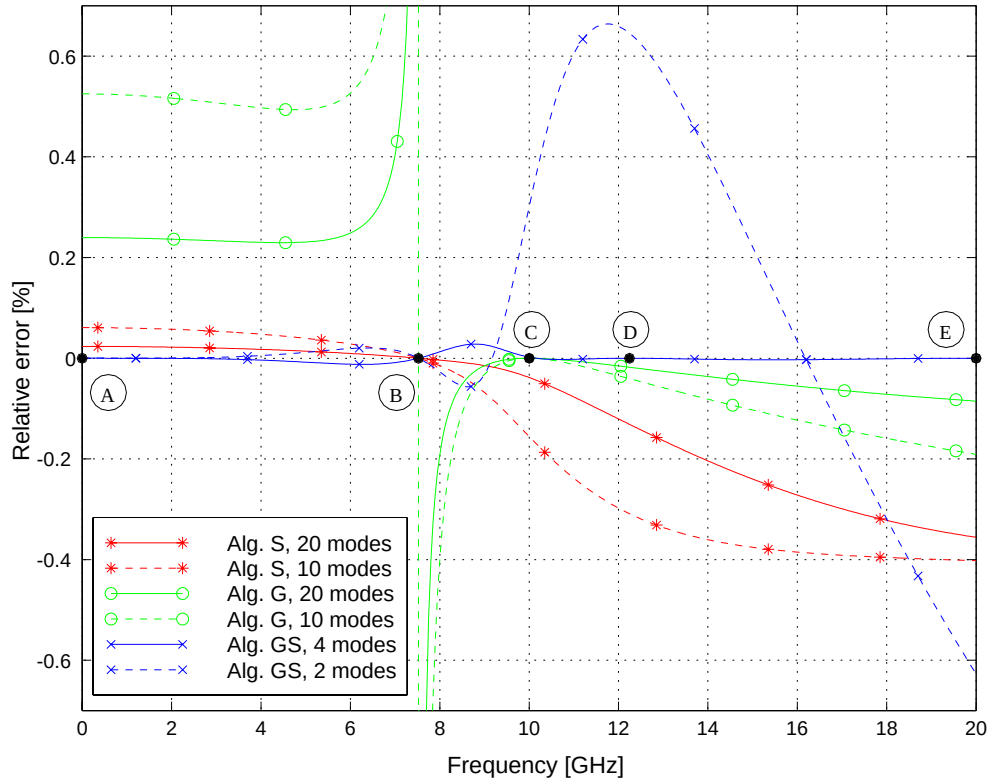


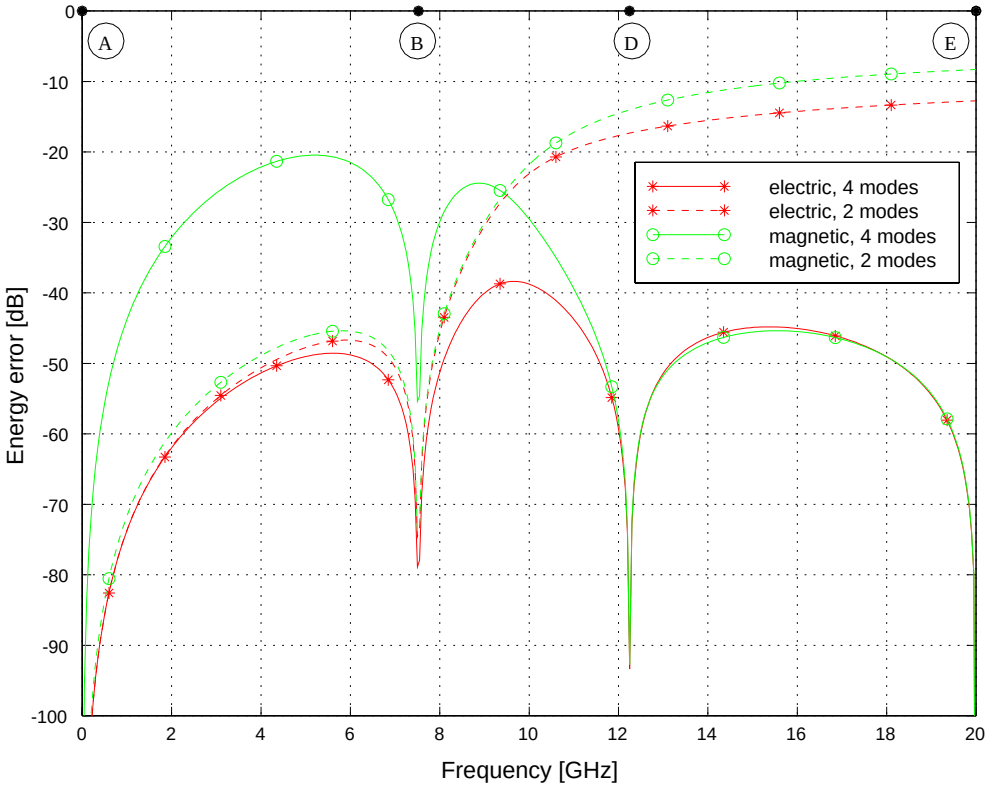
Figure 5.18: Dispersion characteristics for three odd modes in an image line, calculated using the FDFD method. Symbols A–I indicate various points at which modal fields have been calculated and used as a basis in the performed tests. The modes are labelled according to the scheme used in [79].

Results. In the first test [87, 88] the dispersion characteristics of the dominant mode are evaluated using β-GS, β-S and β-G algorithms. Each algorithm involves a different set of eigenfunctions. Several points belonging to three sets, denoted by letters A–I are shown in Fig. 5.18. In particular, for algorithm β-GS which allows eigenfunctions for an arbitrary set of pairs $\{\omega_i, \beta_i\}$, two sets of basis points are used. First set consists of four points corresponding to the same mode, namely A, B, D and E, while the second set contains only the first two of the points mentioned above. Algorithm β-S is implemented for the basis calculated at $\beta_0 = 0$. The first three such points are denoted by B, F, G in Fig. 5.18. Finally, for algorithm β-G, basis functions were calculated at $f_0 = 10\,\mathrm{GHz}$. Again, the first three such points are denoted by C, H, I in Fig. 5.18.

The approximated dispersion characteristics are computed at $m = 401$ points (located in 50 MHz intervals in the 0–20 GHz range) and compared to the FDFD reference solution from Fig. 5.18. According to the relation (3.86) the speedup for β-S and β-G algorithms is $S \approx 400$ and $S \approx 100$ or $S \approx 200$ for β-GS algorithms involving basis functions computed at $n = 4$ or $n = 2$ points, respectively. The relative error in propagation constant of the dominant mode for the three algorithms is shown in Fig. 5.19(a). The error is computed



(a)



(b)

Figure 5.19: (a) Error in propagation constant for the dominant odd mode (relative to FDFD computations shown in Fig. 5.18) in three β-algorithms. The results for algorithms β-G and β-S are plotted for the $N = 10$ and $N = 20$ odd basis modes, while for β-GS algorithm the basis fields are evaluated at points A and B (2 modes) or A, B, D, E (4 modes). (b) Electric and magnetic field energy error (relative to the FDFD field computations).

as

$$\Delta_\beta(f) = \frac{\beta_{\text{ee}}(f) - \beta_{\text{ref}}(f)}{\beta_{\text{ref}}(f)} \cdot 100\% \tag{5.45}$$

where β_{ee} denotes the propagation constant computed by means of the eigenfunction expansion method, and β_{ref} is the one computed using the reference method (in this example the FDFD method). The results for algorithms β-S and β-G are shown for the bases consisting of $N = 10$ or $N = 20$ odd eigenfunctions (evaluated in $n = 1$ frequency or propagation constant point). It is seen that all algorithms are capable of reproducing the dispersion characteristic with a very good accuracy. In spite of the fact that algorithm β-GS uses only two or four points in the basis it gives the best results, because all points (A, B, D, E) used in the basis correspond to the dominant mode. It is seen that even for the smallest basis consisting of just 2 points this algorithm gives very good results below cutoff and may be regarded as quite satisfactory also above cutoff (error in propagation constants below 1%). Adding two more points to the basis pushes the relative error in propagation constant below 0.05% for all points within the region of interest. Fig. 5.19(b) shows the energy errors in reconstructing of electric and magnetic fields computed by β-GS algorithms, relative to the reference FDFD solution. The relative electric energy error Δ_E is computed as the minimization of the following error function

$$\text{err}_E = \frac{\int_S (\alpha \vec{E}_{t\text{ee}} - \vec{E}_{t\text{ref}}) \cdot (\alpha \vec{D}_{t\text{ee}} - \vec{D}_{t\text{ref}})^* \, ds}{\int_S \vec{E}_{t\text{ref}} \cdot \vec{D}^*_{t\text{ref}} \, ds} \tag{5.46}$$

where $\vec{E}_{t\text{ee}}$ and $\vec{D}_{t\text{ee}}$ denote the transverse electric field and flux computed by means of the eigenfunction expansion method, and $\vec{E}_{t\text{ref}}$ and $\vec{D}_{t\text{ref}}$ are the field and flux computed using the reference FDFD method. If scalar α is chosen so that it minimizes the expression on the right side of (5.46) the error Δ_E can be written as [86]

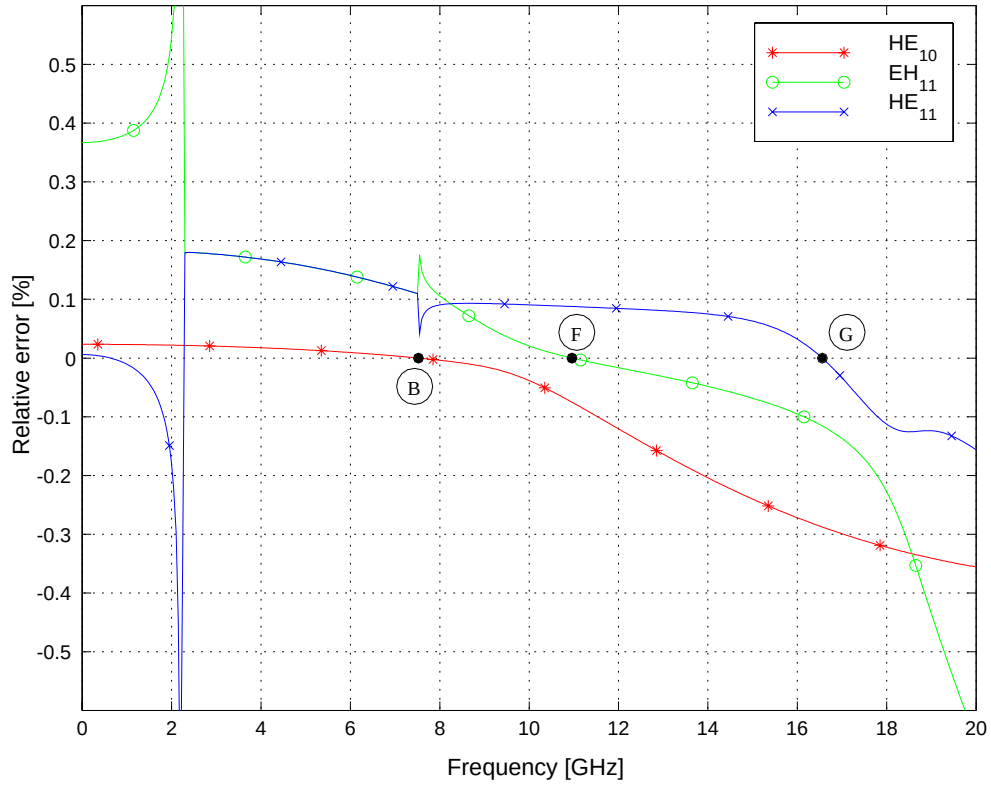
$$\begin{aligned} \Delta_E &\equiv \min_\alpha \text{err}_E \\ &= 1 - \frac{|\int_S \vec{E}_{t\text{ee}} \cdot \vec{D}^*_{t\text{ee}} \, ds|^2}{\int_S \vec{E}_{t\text{ee}} \cdot \vec{D}^*_{t\text{ee}} \, ds \cdot \int_S \vec{E}_{t\text{ref}} \cdot \vec{D}^*_{t\text{ref}} \, ds} \end{aligned} \tag{5.47}$$

The relative magnetic energy error Δ_M can be evaluated in the same way. The resulting expression has the following form

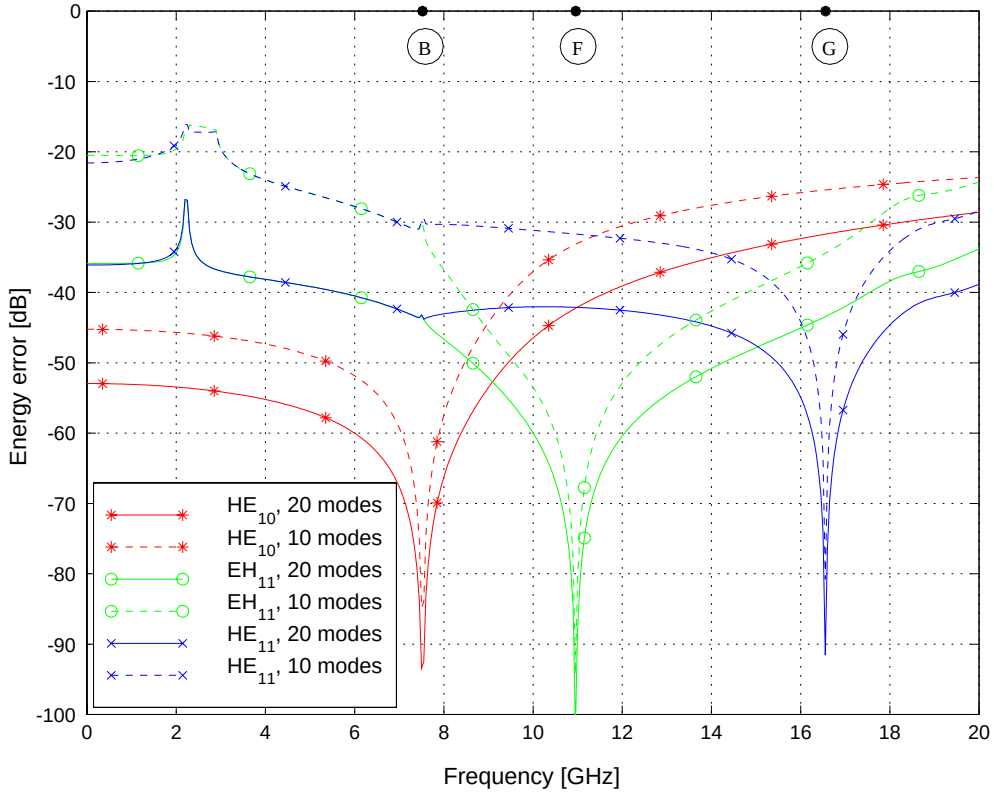
$$\Delta_M = 1 - \frac{|\int_S \vec{H}_{t\text{ee}} \cdot \vec{B}^*_{t\text{ee}} \, ds|^2}{\int_S \vec{H}_{t\text{ee}} \cdot \vec{B}^*_{t\text{ee}} \, ds \cdot \int_S \vec{H}_{t\text{ref}} \cdot \vec{B}^*_{t\text{ref}} \, ds} \tag{5.48}$$

where $\vec{H}_{t\text{ee}}$ and $\vec{B}_{t\text{ee}}$ denote the transverse magnetic field and flux computed by means of the eigenfunction expansion method, and $\vec{H}_{t\text{ref}}$ and $\vec{B}_{t\text{ref}}$ are the field and flux computed using the reference method.

When higher order modes are of interest, algorithm β-S should be used. Fig. 5.20(a) shows the relative error in propagation constant, obtained with this algorithm and basis size $N = 20$, for all modes shown in Fig. 5.18. Note, that two modes become degenerate



(a)



(b)

Figure 5.20: (a) Error in propagation constant for three odd modes (relative to FDFD computations shown in Fig. 5.18) in algorithm β-S using the basis constructed form $N = 20$ odd modes evaluated at $\beta_0 = 0$. (b) Electric field energy error for three modes (relative to the FDFD field computations) in algorithm β-S using the basis constructed from $N = 10$ and $N = 20$ odd modes evaluated at $\beta_0 = 0$.

below cutoff and produce a pair of complex waves which exist over a finite frequency range. Algorithm β-S predicts the propagation constant for this wave with the accuracy better than 0.2%. (Only the real part is shown but the results for the imaginary part are similar). Fig. 5.20(b) shows the electric field energy error for three modes evaluated with algorithm β-S with $N = 10$ and $N = 20$. It is seen that fields of all modes are satisfactorily reproduced. The results for the magnetic field energy error can be found in [88] and they are $\approx 5\,\mathrm{dB}$ worse.

Extensive results for the structure at hand for all β- and ω-algorithms summarized in Table 3.1, including computation of even modes, can be found in [86].

5.5.2 Homogeneous circular waveguide loaded with anisotropic magnetic medium

Problem and structure. In order to test the ability of the eigenfunction expansion algorithms to cope with anisotropic media we investigate dispersion characteristics of the four dominant modes in the circular waveguide shown in Fig. 5.21(a). The structure is homogeneously loaded with an anisotropic magnetic material of relative permittivity $\epsilon_r = 9$ and permeability tensor given by

$$\underline{\underline{\mu}} = \mu_0 \begin{bmatrix} 1 & j0.75 & 0 \\ -j0.75 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{5.49}$$

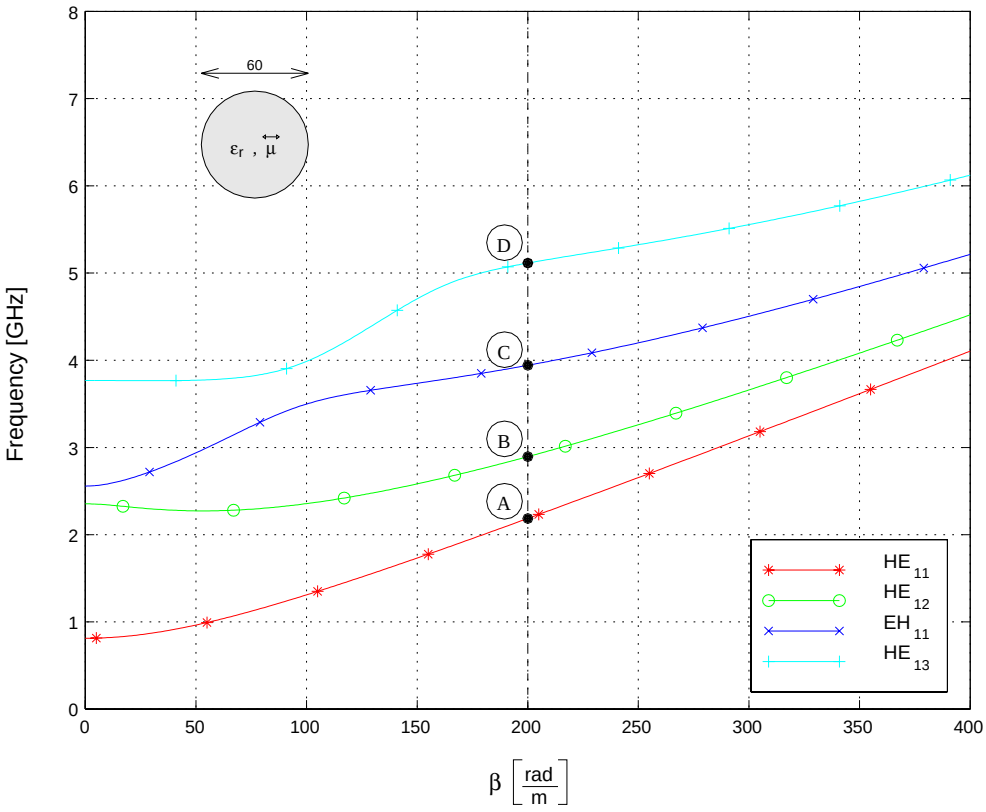
The diameter of the waveguide is $d = 60\,\mathrm{mm}$ and the propagation constant range of interest is from 0 to 400 [rad/m].

Due to gyrotropic properties of the medium the propagation constants of hybrid modes EH_{nm} and HE_{nm} depend on the sign of the index m denoting the angular variation. Here we present the results for modes having the angular dependence $m = -1$.

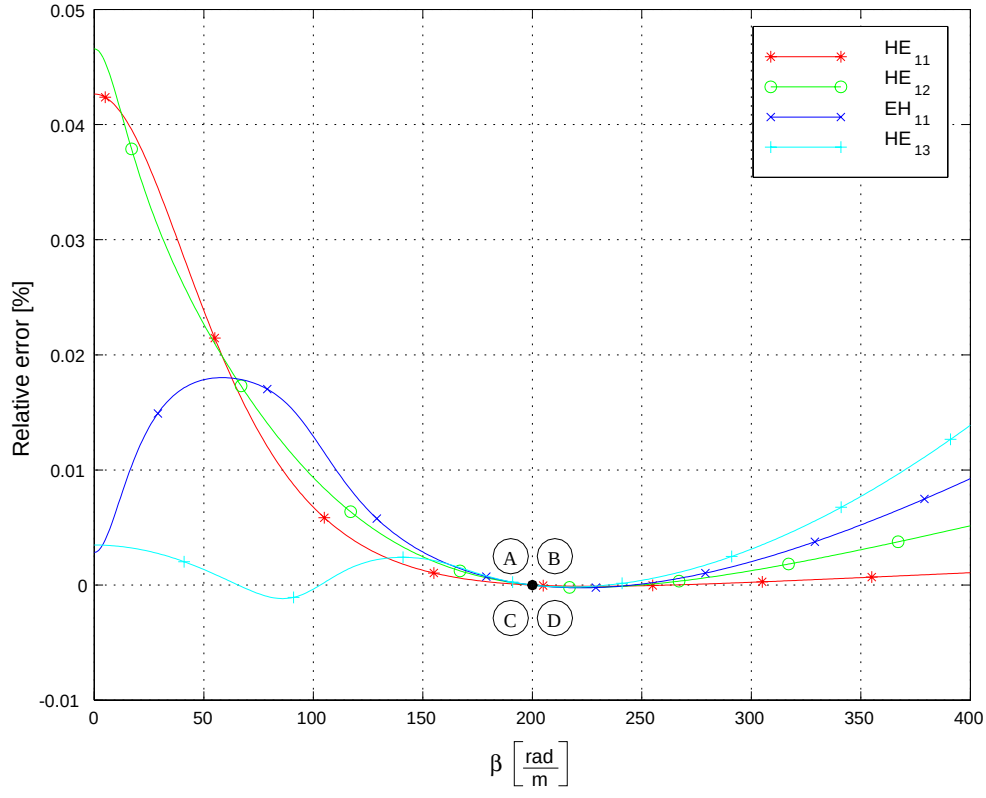
Basis functions and reference characteristics. Basis functions and reference dispersion characteristics $\omega(\beta)$ shown in Fig. 5.21(a) are found analytically by solving nonlinear dispersion equation [36]. This equation is solved for frequency ω being an unknown and propagation constant β being a parameter. For the evaluation of approximated dispersion curves the ω-S algorithm (see Table 3.1) is used, involving basis functions for a given β_0 and resulting in $\omega(\beta)$ characteristics.

Results. Algorithm ω-S is applied with the basis consisting of $N = 20$ eigenfunctions evaluated at $\beta_0 = 200\,[\mathrm{rad/m}]$. Fig. 5.21(b) shows the error in computation of the frequency of the four dominant modes as a function of β, relative to the reference characteristics shown in Fig. 5.21(b). In this case the relative error in frequency is computed as

$$\Delta_f(\beta) = \frac{f_{\mathrm{ee}}(\beta) - f_{\mathrm{ref}}(\beta)}{f_{\mathrm{ref}}(\beta)} \cdot 100\% \tag{5.50}$$



(a)



(b)

Figure 5.21: (a) Dispersion characteristics of modes with the azimuthal index $m = -1$ in a circular waveguide homogeneously loaded with an anisotropic magnetic medium. Letters A–D denote first four points of the basis in ω-S algorithm, evaluated at $\beta_0 = 200\,[\mathrm{rad/m}]$. (b) Relative error in frequency for algorithm ω-S using the basis constructed from $N = 20$ modes with $m = -1$.

It is seen that except for two modes near cutoff, the error is at the level of 0.02%. The largest value of the error, for the HE_{12}, mode occurs at $\beta = 0$ and is lower than 0.05%.

The errors for the modes with the azimuthal index $m = 1$ can be found in [88]. In this case the results were slightly worse and the maximal relative error in the entire region $0 \leq \beta \leq 400\,[\mathrm{rad/m}]$, occurring for HE_{11} mode, is $\approx 0.25\%$.

It was shown, that a new class of eigenfunction expansion hybrid algorithms is very efficient in wideband frequency domain analysis of waveguides. The basis and testing functions used in the algorithms fulfill all internal and external boundary conditions for particular points on the dispersion diagram. Numerical tests show that the accurate solution is obtained even for small number of functions used in the field expansion. It results in small eigenproblems, which can be solved very fast in each frequency or propagation constant point.

Chapter 6

Conclusions

A proper selection of an efficient method for the analysis of a given electromagnetic eigenproblem should take into account many aspects such as choice of an analytical formulation, choice of the method of conversion of the operator problem to the matrix one, application of the hybrid methods, choice of the eigensolver appropriate for a given conversion method, application of preconditioning techniques.

Choice of analytical formulation and conversion method. Classical conversion methods such as the Rayleigh-Ritz method (RR), method of moments (MoM), finite element method (FEM) or finite differences frequency domain method (FDFD) result in the eigenproblems with matrices whose size is directly proportional to the number of field components involved in the analytical formulation. Therefore, the formulations leading to the most efficient approaches are likely to involve as few components as possible. It was also shown that application of the formulation that is spurious-free guarantees the absence of spurious solutions in the spectrum without necessity of using any special approach, such as penalty method or application of the basis/testing functions obeying all Maxwell's equations. Such approach can result in faster convergence of the numerical eigenvalue solver and more accurate solution.

Another aspect of efficiency is related to the choice of the conversion method itself. Among classical conversion methods, the RR and MoM methods involving entire domain basis/testing functions are badly suited for the analysis of structures of complex geometries. In contrast to the RR and MoM, the FEM and FDFD are very well suited for this purpose. However, the disadvantage of both latter methods is that they lead to computationally expensive eigenvalue problems if the structures at hand are filled with media of complex properties (e.g. anisotropic medium, ferrite, chiroferrite). In this case, the coupled mode method (CM) can be applied. This hybrid method leads to a small eigenproblem, what significantly reduces the cost of the analysis. Another hybrid method that was proven to be much faster than any classical method is eigenfunction expansion algorithm (EE). It was used for wideband frequency domain analysis of waveguides. The hybrid methods can also be applied to the efficient analysis of resonators. A comparison of various conversion methods concerning their selected properties is shown in Table 6.1.

Table 6.1: Applicability of various conversion methods to the analysis of the structures with complex geometry and filled with media of complex properties (+ – well suited, ± – might be used, but the problem becomes computationally expensive, − – badly suited).

Conversion methods		Complex geometries	Complex properties
Classical	Rayleigh-Ritz Method of moments	−	+
	Finite element method	+	±
	Finite difference method	+	±
Hybrid	Coupled mode Eigenfunction expansion	+	+

Choice of the eigensolver appropriate for a given conversion method. Efficiency of all numerical algorithms solving matrix eigenvalue problems depends on topological and spectral matrix properties, such as the matrix size, elements' type, symmetry, sparsity pattern, and spectrum configuration. Since these features are characteristic for any particular method of conversion of an operator problem to the matrix one, the performance of the eigensolver is highly dependent on the choice of the conversion method. Suitable choice of a numerical solution method should also take into account some features of the algorithm important in terms of memory and time savings, such as fast convergence and ability to select for computation particular eigenvalues within the entire spectrum. Thus, the performance of the eigensolvers mostly depends on:

Matrix size and sparsity pattern. The size N of an operator matrix and its sparsity pattern are the most important factors influencing the total cost of the solution (in terms of the total computation time and amount of required memory).

In spite of the high computational cost and high memory requirements the methods based on matrix transformations, such as QR and QZ, can be effectively used for *small problems* ($N < 300$) due to high stability and robustness of their numerous and easy available implementations. Such a small problems are generated by the hybrid methods such as the CM or EE.

Since the convergence of the iterative methods highly depends on selection of parameters of the eigensolver, what is not a straightforward task, the QR (QZ) method can also be preferred for *medium size problems* ($300 < N < 1000$). This kind of problems is generated by the methods such as the RR or MoM using entire domain basis/testing functions. However, it should be noted that when parallel computations are performed and the iterative solver is used, additional speedup due to the size of the cache memory can be observed and the QR method becomes substantially slower.

In *large scale problems* ($N > 1000$), generated by the FEM or FDFD, the matrices

are sparse and structured. In this case, the most stable and robust algorithms, such as the QR and QZ methods, are not suitable because:

- they find entire eigenspectrum and eigenspace of the operator, while in electromagnetic problems usually only the eigenvalues located near one of the ends of the operator spectrum are to be found,
- similarity transformations used in QR method or congruence transformations used in QZ method can quickly destroy sparsity of the operator matrix,
- due to high numerical complexity ($\mathcal{O}(N^3)$) and memory requirements ($\mathcal{O}(N^2)$) the size of the problem to be solved can be larger than the maximal admissible size on a given machine,
- they are extremely difficult to parallelize.

In this case, the iterative methods such as the Arnoldi or subspace iteration are preferred. In contrast to the QR method, these algorithms compute only selected eigenvalues and they acquire the information about the matrix from the product of the matrix operator and a vector. Therefore, they can easily take advantage of sparsity and structure of the operator matrix. Suitable choice of the matrix storage format can result in smaller memory requirements ($\mathcal{O}(N)$) and rises the possibility of solving larger problems, while a proper implementation of the matrix-vector product can lead to much faster execution ($\mathcal{O}(N)$). If iterative methods are used for dense matrices their efficiency is degraded, because the cost of the matrix-vector product is ($\mathcal{O}(N^2)$). In spite of this fact, they are much more efficient than the QR method. For efficient application of the iterative methods, their parameters should be optimized for each computer architecture. It is also important to note, that the iterative methods better conform to the multiprocessor super scalar architectures than the QR algorithm. However, only an experienced user can fully exploit capabilities of the modern systems (regarding the quality of software tools, algorithms, cache effects, libraries, parallel computations) to get significant performance improvement. The suggestions concerning the choice of the eigensolver are summarized in Table 6.2.

Matrix symmetry and elements' type. The symmetry of the matrix and the type of the matrix elements are also important for efficiency of computations. The versions of eigensolvers intended for symmetric and real matrices are much faster and less memory consuming than the ones intended for complex and nonsymmetric matrices. However, it should be born in mind, that maintaining of the symmetry of the formulation at the cost of increasing the number of the involved field components or conversion of a standard problem into a generalized one is usually inefficient.

Spectrum configuration and application of preconditioning techniques. Since QR and QZ methods compute all the eigenvalues (and corresponding eigenvectors) of a given eigenproblem, the efficiency of computations does not depend on the eigenvalues of interest and preconditioning techniques cannot be applied.

Table 6.2: Properties of matrices generated with various conversion methods and recommended methods of solution.

Conversion methods		Matrix properties		Suggested methods of solution
		Size	Sparsity pattern	
Classical	Rayleigh-Ritz Method of moments	moderate	dense	Arnoldi/Lanczos subspace iteration QR (QZ)
	Finite element method	large	sparse	Arnoldi/Lanczos subspace iteration
	Finite difference method	very large	sparse, structured diagonally	Arnoldi/Lanczos subspace iteration
Hybrid	Coupled mode Eigenfunction expansion	small	dense	QR (QZ)

The situation is extremely different for iterative methods, such as Arnoldi/Lanczos or subspace iteration, which compute only selected eigenvalues. Since, in the case of iterative methods, the convergence strongly depends on spectral properties of the problem, such as separation of eigenvalues and the spectral radius, the spectral transformation techniques, such as inverse iteration or polynomial filtering using e.g. Chebyshev or digital finite impulse response (FIR) filters can be applied to improve the efficiency of the algorithms. Due to the fact that the solution of generalized problems always requires computation of a matrix decomposition provided that the shift-invert preconditioner can be applied at no additional cost, this strategy was proven to be very efficient in the solution of large generalized eigenproblems. In the standard problems where the eigenvalues of interest are located at the lower end of the real spectrum the preconditioning with Chebyshev polynomials offers considerable memory and time savings. Another preconditioning technique, incorporating bandpass FIR filters, can be used for the efficient computation of the eigenvalues of a standard eigenproblem, which are located far from both spectrum ends. In this case any other alternative technique, such as inverse iteration, cannot be used because the size of the matrix is very large and computation of the LU decomposition of the matrix is not possible. The recommendations on using preconditioning techniques are summarized in Table 6.3.

The most important aspects of this thesis concerned implementation and performance analysis of modern iterative eigensolvers applied to the problems resulting from various classical conversion methods. It was shown, that the Arnoldi method combined with the reverse communication technique was the most efficient in the solution of all considered electromagnetic eigenproblems. The tests concerning the analysis of an image line by means of the Galerkin method showed that the Arnoldi method can be 15 times faster than the QR method in the solution of the eigenproblems involving dense matrices. The

Table 6.3: Recommended preconditioning techniques for standard and generalized eigenproblems in dependence on the location of the eigenvalues of interest in the spectrum.

Location of eigenvalues in the spectrum	Recommended preconditioner	
	Standard eigenproblem	Generalized eigenproblem
higher end	Chebyshev polynomials or none	none
lower end	Chebyshev polynomials	inverse iteration
between the ends	FIR digital filters	shift-invert

results of analysis of the image line by means of the FEM method showed that the Arnoldi method is 10 times faster than the subspace iteration in the solution of generalized eigenproblems involving sparse and structured matrices. In this comparison we did not take into account the relatively long time needed for performing of the LU decomposition. It is worth noting that the total computational time, including the LU time, of the FEM approach is twice as high as the time of the solution of the standard problem resulting from the application of the FDFD method to the same image line structure. Thus, the Arnoldi method combined with the FDFD, producing highly diagonally structured matrices, revealed to be the most efficient approach from all considered ones. Another tests, related to the analysis of a rotationally symmetric dielectric resonators by means of the FDFD, concerned the implementation and performance analysis of various preconditioning techniques incorporating Chebyshev polynomials and FIR digital filters. These techniques could only be implemented in modern iterative eigensolvers, such as the Arnoldi and subspace iteration methods. It was shown that the Arnoldi method using Chebyshev preconditioning could be 4 times faster in the computation of the lowest resonant frequencies of the structure than the method without any preconditioning. Moreover, the Arnoldi method with preconditioning could be 30% faster than the subspace iteration one using the same type of preconditioning. The application of advanced preconditioning involving the FIR filters enabled us to compute the resonant frequencies that are far of both ends of the spectrum, what would not be possible with any other approach.

Another important aspect of this thesis was the application of the hybrid methods such as the coupled mode method and the eigenfunction expansion method to the analysis of waveguide structures. Application of the perturbation method (resulting from the coupled mode formalism) to the analysis of a nonreciprocal ferrite phase shifter led to a simple computational formula. The cost of the analysis of such a complex inhomogeneous structure partially filled with the material of tensor properties by means of any classical method was significantly reduced. The theoretical solution was verified practically and proven to be more accurate the solution by means of the methods used so far. Another new class of hybrid algorithms, i.e. eigenfunction expansion algorithms, was proven to be very efficient in wideband frequency domain analysis of waveguides. Numerical tests showed that the accurate solution was obtained even for small number of functions taken

into the field expansion. It resulted in very small eigenproblems that could be solved very fast in each frequency or propagation constant point. Due to hybrid character of the eigenfunction expansion this approach was much more efficient than any of the standard ones. The maximum available speedup was close to the number of the points swept.

The results of this thesis summarized in previous two paragraphs prove all claims of the thesis. They can open up new possibilities for microwave engineers. The constraints (long computational time) on practical application of the FEM and FDFD methods for effective design of complex devices can be mitigated when the iterative eigensolvers, such as the implicitly restarted Arnoldi method, are used. Another important new possibility is the efficient analysis of structures containing complex materials and fast determination of dispersion characteristics of waveguides.

Acknowledgments/Podziękowania

Przede wszystkim chciałbym wyrazić głębokie i szczere podziękowanie mojemu promotorowi, profesorowi Michałowi Mrozowskiemu, za wiele lat nauki, pomocy, inspiracji i wsparcia, które otrzymałem od niego. Pozwoliło mi to doprowadzić moją pracę do szczęśliwego końca.

I am also grateful to Dr F.A. Fernandez of UCL London for providing FEM analysis program used for calculations presented in Sec. 5.2.1.

Chciałbym również podziękować Komitetowi Badań Naukowych, który sponsorował tę pracę dwoma grantami o numerach 8 T11D 018 11 i 8 T11D 029 15 oraz Centrum Informatycznemu Trójmiejskiej Akademickiej Sieci Komputerowej, które udostępniło mi swoje komputery do wykonania obliczeń.

Serdeczne podziękowania należą się moim przyjaciołom, Michałowi Patrykowi Dębickiemu, Piotrowi Jędrzejewskiemu, Piotrowi Przybyszewskiemu i Michałowi Rewieńskiemu, z którymi przyszło mi dzielić jeden pokój przez te wszystkie lata. Zawsze będę pamiętał ten spędzony wspólnie czas, dyskusje. Piotrowi Przybyszewskiemu pragnę podziękować jeszcze dodatkowo za pomoc w przygotowaniu niektórych ilustracji zawartych w pracy.

Jestem bardzo wdzięczny swoim Rodzicom, którzy nieustannie pomagali mi i wspierali moją pracę dodając mi nowych sił i motywacji.

W końcu chciałbym również podziękować mojej ukochanej żonie Agnieszce za cierpliwość i miłość jaką mi okazywała podczas tego długiego i trudnego dla nas okresu. Zawsze mogłem liczyć na jej pomoc i wsparcie. Nie wiem czy kiedykolwiek uda mi się wynagrodzić jej tak wiele "utraconych" wspólnych chwil, podczas których zmuszona była sama zajmować się naszymi wspaniałymi dziećmi, Martusią i Rafałkiem.

Copyright note

Appendix A

Review of numerical methods for matrix eigenvalue problems

The main part of this review is devoted to algorithms capable of solving standard eigenproblems of the form

$$\underline{\underline{A}}\,\underline{v} = \lambda \underline{v} \tag{A.1}$$

where $\underline{\underline{A}}$ is a matrix of size $n \times n$, $\underline{v}$ is an eigenvector and λ is the corresponding eigenvalue.

In the first section we describe some basic algorithms which can act as independent solvers, but in practice, they are inefficient due to their simplicity. However, they are often involved in more elaborate algorithms. In the next three sections we concentrate on the algorithms, such as *QR method*, *subspace iteration*, and *Arnoldi method*, which are the most general ones, applicable to a broad class of matrix eigenproblems.

It should be mentioned, that many other methods, which are not discussed here, can be found in the literature [39, 98]. The most popular ones are *bisection* and *Jacobi algorithm* intended only for symmetric matrices, or *nonsymmetric Lanczos method*, which is especially intended for large nonsymmetric problems:

Bisection [**39**] is based on slicing the spectrum of a symmetric matrix. In order to apply the method one should transform the matrix into its tridiagonal form, which requires additional effort. The advantage is that the method is able to find only selected eigenvalues. Computation of eigenvectors requires application of inverse iteration, resulting in the additional cost of $\mathcal{O}(n^2)$ per eigenvector. In consequence, the method is less efficient than the QR method if more than $n/4$ eigenvalues is to be computed.

Jacobi algorithm [**39**] is one of the first algorithms conforming well to modern computational techniques. It is based on *Jacobi rotations* (which are special case of *Givens rotations*) applied to the matrix in order to obtain its diagonal form. An advantage of this algorithm is the ease of parallelization of the code, which can be attractive in modern parallel computer systems. However, the serial code is much slower than the QR method and, due to its simplicity, it can be competitive for very small problems or when small accuracy is required.

Nonsymmetric Lanczos method [39, 98] is a kind of Krylov space method, similar to the Lanczos or Arnoldi methods. It is based on an oblique projection, which results in smaller than the Arnoldi method memory requirements. However, its main disadvantage is its potential instability.

A separate, last section is devoted to the methods especially intended for generalized eigenproblems of the form

$$\underline{\underline{A}}\,\underline{v} = \lambda \underline{\underline{B}}\,\underline{v} \tag{A.2}$$

where $\underline{\underline{A}}$ and $\underline{\underline{B}}$ are matrices of size $n \times n$, $\underline{v}$ is an eigenvector and λ is the corresponding eigenvalue.

A.1 Basic algorithms

One of the oldest iterative techniques for solving matrix eigenvalue problems is the power method. The method, in its original form, is intended to find the eigenvalue possessing the largest magnitude and the corresponding eigenvector. Convergence of the method strongly depends on spectral properties of the matrix. In some cases it can be slow or may not happen at all. To overcome these difficulties spectral transformations can be used. In the simplest case, such a transformation can be a linear shift or inverse iteration, while, in general, a polynomial (e.g. Chebyshev) or rational function of the matrix can be used. Spectral transformations can also be used to calculate the eigenvalue from other ends of matrix spectrum than the largest magnitude one. In particular, a linear shift can be used to find the largest or the smallest real part eigenvalue, while the smallest magnitude eigenvalue can be calculated via an inverse iteration. Having the first eigenvalue-eigenvector pair computed, it is possible to extract the next pair with use of deflation techniques. All these aspects are discussed in the next few subsections with reference to the power method. Another algorithm which is useful in approximating the subspace spanned on eigenvectors of interest is projection and is discussed at the end of this section.

A.1.1 The power method

The power method is based on the process of generating the sequence of vectors $\underline{\underline{A}}^{i}\underline{v}^{(0)}$, where $\underline{v}^{(0)}$ is a nonzero starting vector. This sequence converges to the so-called *dominant eigenvector* $\underline{v}_1$ corresponding to the eigenvalue possessing the largest magnitude, called the *dominant eigenvalue* λ_1. Subsequent approximations of λ_1 are calculated with the *Rayleigh quotient* of $\underline{v}_1$

$$\lambda = \frac{\underline{v}^{H}\underline{\underline{A}}\,\underline{v}}{\underline{v}^{H}\underline{v}} \tag{A.3}$$

The iteration is continued until tolerance τ is reached.

The algorithm with necessary normalizations is as follows:

Algorithm 1 THE POWER METHOD

Input: $\underline{\underline{A}}, \underline{v}^{(0)}, \tau$

Output: $\{\lambda, \underline{v}\} : \underline{\underline{A}}\,\underline{v} = \lambda \underline{v}$

(1) $\underline{v} = \frac{\underline{v}^{(0)}}{\|\underline{v}^{(0)}\|}$

(2) **for** $i = 1, 2, \ldots$ **until** $\|\underline{\underline{A}}\,\underline{v} - \lambda \underline{v}\,\|_2 < \tau|\lambda|$

(2.1) $\underline{w} = \underline{\underline{A}}\,\underline{v}$

(2.2) $\lambda = \frac{\underline{v}^H \underline{w}}{\underline{v}^H \underline{v}}$

(2.3) $\underline{v} = \frac{\underline{w}}{\|\underline{w}\|}$

where τ is a given accuracy. Norm $\|.\|$ in Steps (1) and (2.3) can be any of the *Hölder norms*

$$\|\underline{x}\,\|_p = \left(\sum_{i=1}^{n} |\underline{x}_i|^p \right)^{1/p} \tag{A.4}$$

which are also called p-norms. In particular, if $p = \infty$, then the vector is normalized to its maximal element. Such normalization has lower numerical cost than the 2-norm one because it involves comparisons of the vector elements rather than dot product calculation. On the other hand, if the 2-norm normalization is used, eigenvalue λ can be calculated in Step (2.2) using simplified formulae $\lambda = \underline{v}^H \underline{w}$. Note, that calculations in this step are equivalent to Rayleigh quotient calculation (A.3) with the difference that they do not involve any additional matrix-vector product. In order to speed up the algorithm, λ need not to be computed in each iteration.

Convergence and efficiency. Let us assume that the eigenvalues of matrix $\underline{\underline{A}}$ of size n are ordered so that

$$|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n| \tag{A.5}$$

Then, as shown in [39], after i steps of the algorithm, eigenvalue λ approximates dominant eigenvalue λ_1 with accuracy given by

$$|\lambda_1 - \lambda| = \mathcal{O}\left(\left| \frac{\lambda_2}{\lambda_1} \right|^i \right) \tag{A.6}$$

which corresponds to a linear convergence with a *convergence factor*

$$\rho_P = \left| \frac{\lambda_2}{\lambda_1} \right| \tag{A.7}$$

If $|\lambda_2|$ is close to $|\lambda_1|$, the convergence is slow. Moreover, Algorithm 1 do not converges at all if $\lambda_2 \neq \lambda_1$ but $|\lambda_2| = |\lambda_1|$.

A.1.2 Spectral transformations

Spectral transformations are a very powerful tool which can be used for accelerating the process of approximation of eigenvalues and eigenvectors. The idea of spectral transformations is to convert the spectrum of a matrix, so that the convergence factor for the eigenvalues (eigenvectors) of interest is minimized. In this sense spectral transformations are referred to as *preconditioning techniques*.

Consider a matrix $\underline{\underline{A}}$ having an eigenvalue λ with a corresponding eigenvector $\underline{u}$ and a general form of spectral transformation $\phi(\chi)$ being a rational function

$$\phi(\chi) \equiv \frac{p_k(\chi)}{q_l(\chi)} = \frac{\alpha_0 + \alpha_1\chi + \alpha_2\chi^2 + \ldots + \alpha_k\chi^k}{\beta_0 + \beta_1\chi + \beta_2\chi^2 + \ldots + \beta_l\chi^l} \tag{A.8}$$

where $p_k(\chi)$ and $q_l(\chi)$ are polynomials of degree k and l, respectively. If $q(\underline{\underline{A}})$ is nonsingular, we can define $\phi(\underline{\underline{A}}) \equiv q(\underline{\underline{A}})^{-1}p(\underline{\underline{A}})$. Clearly, $\phi(\underline{\underline{A}})$ has an eigenvalue $\theta \equiv \phi(\lambda)$. It can be easily verified that the corresponding eigenvector $\underline{v} = \underline{u}$. Such important property of preserving eigenvectors is representative for spectral transformations. Once eigenpair $\{\theta, \underline{v}\}$ has been computed, the corresponding eigenvalue λ can be calculated from the equivalence $\theta \equiv \phi(\lambda)$ or directly from the Rayleigh quotient (A.3).

Implementation. In terms of the power method we can derive the following algorithm:

Algorithm 2 THE POWER METHOD WITH SPECTRAL TRANSFORMATION

Input: $\underline{\underline{A}}, \phi(\chi), \underline{v}^{(0)}, \tau$

Output: $\{\lambda, \underline{v}\} : \underline{\underline{A}}\,\underline{v} = \lambda\underline{v}$

(1) $\underline{v} = \frac{\underline{v}^{(0)}}{\|\underline{v}^{(0)}\|}$

(2) **for** $i = 1, 2, \ldots$ **until** $\|\underline{\underline{A}}\,\underline{v} - \lambda\underline{v}\,\|_2 < \tau|\lambda|$

(2.1) $\underline{w} = \phi(\underline{\underline{A}})\underline{v}$

(2.2) $\lambda = \frac{\underline{v}^H\underline{\underline{A}}\,\underline{v}}{\underline{v}^H\underline{v}}$ or $\lambda = \frac{\underline{w}^H\underline{\underline{A}}\,\underline{w}}{\underline{w}^H\underline{w}}$

(2.3) $\underline{v} = \frac{\underline{w}}{\|\underline{w}\|}$

Calculation of λ in Step (2.2) should be performed in the manner which minimizes computational cost of the Rayleigh quotient. The choice between the above two expressions depends on transformation $\phi(\chi)$. Moreover, if computation of λ can take advantage of the normalization of $\underline{w}$, Step (2.3) should precede Step (2.2).

Note, that any spectral transformation can also be implemented without affecting of the original algorithm (which do not implement the spectral transformation) by directly preconditioning the matrix operator, i.e., by passing $\phi(\underline{\underline{A}})$ to the program, instead of $\underline{\underline{A}}$.

Convergence and efficiency. If n eigenvalues of $\underline{\underline{A}}$ are ordered so that

$$|\theta_1| \equiv |\phi(\lambda_1)| > |\phi(\lambda_2)| \geq \cdots \geq |\phi(\lambda_n)| \equiv |\theta_n| \tag{A.9}$$

then the Algorithm 2 converges to λ_1 with the convergence factor

$$\rho = \left|\frac{\theta_2}{\theta_1}\right| \equiv \left|\frac{\phi(\lambda_2)}{\phi(\lambda_1)}\right| \tag{A.10}$$

Factor ρ is minimized when $\phi(\chi)$ is constructed so that $|\phi(\lambda_1)| \gg |\phi(\lambda_2)|$.

The simplest spectral transformations a include linear shift and inverse iteration. Other frequently used transformations are polynomial filters, in particular of Chebyshev type. They are briefly discussed below in the context of the power method.

A.1.2.1 Linear shift

A linear shift is the simplest spectral transformation (A.8), which is defined by polynomials $p_1(\chi) = -\sigma + \chi$ and $q_0(\chi) = 1$ with σ representing the *shift*. Thus

$$\phi(\underline{\underline{A}}) = \underline{\underline{A}} - \sigma\underline{\underline{I}} \tag{A.11}$$

$$\theta \equiv \phi(\lambda) = \lambda - \sigma \tag{A.12}$$

where $\underline{\underline{I}}$ is the identity matrix.

Implementation. Implementation of a linear shift in Algorithm 2 is simple. It requires substituting Steps (2.1) and (2.2) by

Algorithm 3 THE POWER METHOD WITH LINEAR SHIFT

Input: $\underline{\underline{A}}, \sigma, \underline{v}^{(0)}, \tau$

Output: $\{\lambda, \underline{v}\} : \underline{\underline{A}}\,\underline{v} = \lambda\underline{v}$

$\vdots$

(2.1) $\underline{w} = (\underline{\underline{A}} - \sigma\underline{\underline{I}})\underline{v}$

(2.2) $\lambda = \sigma + \frac{\underline{v}^H\underline{w}}{\underline{v}^H\underline{v}}$

(2.3) $\underline{v} = \frac{\underline{w}}{\|\underline{w}\|}$

Note that the calculation of the Rayleigh quotient in Step (2.2) is performed without an additional matrix-vector product.

Convergence and efficiency. Linear shift is useful when computing the largest or the smallest real eigenvalue of a matrix with purely real spectrum. Consider such matrix $\underline{\underline{A}}$ with eigenvalues ordered decreasingly

$$\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_{n-1} > \lambda_n \tag{A.13}$$

A suitable choice of a linear shift σ can improve convergence of the algorithm to the eigenvalue $\lambda_1 = \theta_1 + \sigma$ or $\lambda_n = \theta_n + \sigma$. Optimal convergence is obtained when the value of σ is chosen so that it shifts the middle of the "unwanted" part of the spectrum to the origin. Hence, for computing the largest real eigenvalue λ_1, the optimal shift should be chosen as

$$\sigma_{\mathrm{LR}} = \frac{\lambda_2 + \lambda_n}{2} \tag{A.14}$$

while if the smallest real eigenvalue λ_n is of interest

$$\sigma_{\mathrm{SR}} = \frac{\lambda_1 + \lambda_{n-1}}{2} \tag{A.15}$$

Corresponding optimal convergence factors are

$$\rho_{S_{\mathrm{LR}}} = \left| \frac{\lambda_2 - \sigma_{\mathrm{LR}}}{\lambda_1 - \sigma_{\mathrm{LR}}} \right| \tag{A.16}$$

$$\rho_{S_{\mathrm{SR}}} = \left| \frac{\lambda_{n-1} - \sigma_{\mathrm{SR}}}{\lambda_n - \sigma_{\mathrm{SR}}} \right| \tag{A.17}$$

respectively.

Linear shifts can also be used in conjunction with inverse iteration and polynomial (i.e. Chebyshev) acceleration, described below.

A.1.2.2 Inverse iteration

An inverse iteration is a spectral transformation (A.8), which is frequently used concurrently with a linear shift σ. This transformation is defined by polynomials $p_0(\chi) = 1$ and $q_1(\chi) = -\sigma + \chi$. Therefore

$$\phi(\underline{\underline{A}}) = (\underline{\underline{A}} - \sigma \underline{\underline{I}})^{-1} \tag{A.18}$$

$$\theta \equiv \phi(\lambda) = \frac{1}{\lambda - \sigma} \tag{A.19}$$

Implementation. In order to get specific algorithm for inverse iteration in terms of the power method we can substitute Steps (2.1), (2.2) and (2.3) in Algorithm 2 by

Algorithm 4 The power method with inverse iteration

Input: $\underline{\underline{A}}, \sigma, \underline{v}^{(0)}, \tau$

Output: $\{\lambda, \underline{v}\} : \underline{\underline{A}}\,\underline{v} = \lambda \underline{v}$

$\vdots$

(2.1) Solve $(\underline{\underline{A}} - \sigma \underline{\underline{I}})\underline{w} = \underline{v}$

(2.2) $\underline{w} = \frac{\underline{w}}{\|\underline{w}\|}$

(2.3) $\lambda = \sigma + \frac{\underline{w}^H \underline{v}}{\underline{w}^H \underline{w}}$

(2.4) $\underline{v} = \underline{w}$

where $\underline{\underline{I}}$ is the identity matrix. Note that λ is calculated without an additional matrix-vector product.

The iteration in Step (2.1), performed on vector $\underline{v}$, requires an additional comment. Since computation of the matrix inverse is very costly, direct calculation of the transformation (A.18) is, in general, inefficient. Much more practical approach involves computing of the LU decomposition $\underline{\underline{A}} - \sigma \underline{\underline{I}} = \underline{\underline{L}}\,\underline{\underline{U}}$ (or Cholesky for symmetric case) at the beginning of the algorithm and solving linear system $\underline{\underline{L}}\,\underline{\underline{U}}\,\underline{w} = \underline{v}$ instead. The solution $\underline{w}$ of the system is found in two steps: forward substitution $\underline{\underline{L}}\,\underline{y} = \underline{v}$ and back substitution $\underline{\underline{U}}\,\underline{w} = \underline{y}$.

The practical importance of the inverse iteration relies on the fact that its application causes the algorithm to converge to the eigenvalue closest (in magnitude) to the shift σ. In particular, when $\sigma = 0$ it converges toward the eigenvalue of the smallest magnitude. This property allows one to compute any eigenvalue from the matrix spectrum and the corresponding eigenvector.

Convergence and efficiency. If we consider matrix $\underline{\underline{A}}$ with eigenvalues ordered so that

$$|\lambda_1 - \sigma| < |\lambda_2 - \sigma| \leq \cdots \leq |\lambda_n - \sigma| \tag{A.20}$$

then Algorithm 4 converges to λ_1 with convergence factor

$$\rho_I = \left| \frac{\lambda_1 - \sigma}{\lambda_2 - \sigma} \right| \tag{A.21}$$

We can see that convergence is faster if σ is closer to λ_1. This observation conceived an idea of changing the value of σ according to approximations of λ_1 computed in subsequent iterations. The algorithm which uses, in each iteration, Rayleigh quotient (A.3) as a new shift is called *Rayleigh Quotient Iteration* (RQI). It was proposed by Parlett [83], who showed that this algorithm offers quadratic convergence rate, in general case, and even cubic for Hermitian matrices. Despite of that, application of RQI is rather limited due to a high cost of frequent factorizations, which have to be performed in each iteration. In practical algorithm, σ can be "refreshed" after each couple of iterations.

Main inconveniences of using the inverse iteration technique are that it requires the calculation of the costly LU decomposition and that its good convergence properties strongly depend on a priori knowledge of the approximate solution. However, these disadvantages

can be mitigated in some practical cases. In particular, the inverse iteration can be applied at no additional cost for generalized eigenproblems, where such decomposition is always computed (see Sec. A.5.2). It also becomes a very economical tool for calculating eigenvectors once the corresponding eigenvalues have been found. The inverse iteration used in this context is involved by other algorithms (i.e. QR, subspace iteration, Arnoldi method).

A.1.2.3 Chebyshev acceleration

Chebyshev acceleration techniques are based on Chebyshev polynomials.

Chebyshev polynomials. The polynomial of degree k is defined by

$$C_k(t) = \cos\left(k \cos^{-1}(t)\right) \qquad \text{for} \qquad -1 \leq t \leq 1 \tag{A.22}$$

This definition can be easily extended to the case $|t| \geq 1$ using the expansion of cosine function $\cos\theta = (e^{i\theta} + e^{-i\theta})/2$, where $i = \sqrt{-1}$, giving

$$C_k(t) = \frac{1}{2}\left[(t + \sqrt{t^2-1})^k + (t + \sqrt{t^2-1})^{-k}\right] \qquad \text{for} \qquad |t| \geq 1 \tag{A.23}$$

Polynomial character of $C_k(t)$ may not be readily seen yet can be expressed using the recurrence formulae, valid for Chebyshev polynomials

$$\begin{cases} C_0(t) &= 1 \\ C_1(t) &= t \\ C_{k+1}(t) &= 2tC_k(t) - C_{k-1}(t) \qquad \text{for} \qquad k = 1, 2, \ldots \end{cases} \tag{A.24}$$

This three-term recurrence relation is of high practical importance, which will be discussed later in the text.

Chebyshev polynomials oscillate between -1 and 1 for t in the interval $\langle -1, 1\rangle$ while their absolute values increase rapidly outside of the range. The rate of increase is larger for higher polynomial order k.

A very important property of Chebyshev polynomials, which is known from approximation theory, is that they are optimal in the sense of minimizing polynomial function over a certain non-empty interval $\langle \alpha, \beta\rangle$. It is shown in [98] that among all polynomials p_k of degree k such that $p_k(\gamma) = 1$, where $\gamma \geq \beta$ or $\gamma \leq \alpha$, the minimum

$$\min \max_{t \in \langle \alpha, \beta\rangle} |p(t)| \tag{A.25}$$

is reached by the polynomial

$$\hat{C}_k(t) \equiv \frac{C_k\left(1 + 2\frac{t-\beta}{\beta-\alpha}\right)}{C_k\left(1 + 2\frac{\gamma-\beta}{\beta-\alpha}\right)} = \frac{C_k\left(-1 - 2\frac{\alpha-t}{\beta-\alpha}\right)}{C_k\left(-1 - 2\frac{\alpha-\gamma}{\beta-\alpha}\right)} \tag{A.26}$$

The numerator of $\hat{C}_k(t)$ has the form $C_k(M(t))$, where $M(t) = 1 + 2\frac{t-\beta}{\beta-\alpha}$ is a linear transformation which maps the interval $\langle \alpha, \beta\rangle$ into $\langle -1, 1\rangle$. The polynomial in the denominator is a constant function of t used for normalization purposes ($\hat{C}_k(\gamma) = 1$).

Implementation. Let us construct the algorithm based on the power method with spectral transformation (Algorithm 2) defined by the optimal Chebyshev polynomial (A.26). $\hat{C}_k(t)$ can be taken without normalization since Algorithm 5 has its own normalization in Step (2.3). Substituting $p_k(\chi) = C_k(M(\chi))$ and $q_0(\chi) = 1$ we get

Algorithm 5 The power method with Chebyshev acceleration

Input: $\underline{\underline{A}}, c, e, k, \underline{v}^{(0)}, \tau$

Output: $\{\lambda, \underline{v}\} : \underline{\underline{A}}\,\underline{v} = \lambda \underline{v}$

$\vdots$

(2.1) $\underline{w} = C_k\left(\frac{\underline{\underline{A}} - c\underline{\underline{I}}}{e}\right)\underline{v}$

(2.2) $\lambda = \frac{\underline{v}^H \underline{\underline{A}}\,\underline{v}}{\underline{v}^H \underline{v}}$

(2.3) $\underline{v} = \frac{\underline{w}}{\|\underline{w}\|}$

where

$$c = \frac{\alpha + \beta}{2} \qquad \text{and} \qquad e = \frac{\beta - \alpha}{2} \tag{A.27}$$

It should be noted that, in a practical algorithm, calculations in Step (2.1) are performed using three-term recurrence (A.24) in the following way

Algorithm 6 Product of matrix Chebyshev polynomial and a vector via three-term recurrence

Input: $\underline{\underline{A}}, c, e, k, \underline{v}$

Output: $\underline{w} = C_k\left(\frac{\underline{\underline{A}} - c\underline{\underline{I}}}{e}\right)\underline{v}$

(1) $\underline{w}_0 = \underline{v}$

(2) $\underline{w}_1 = \frac{1}{e}(\underline{\underline{A}}\,\underline{v} - c\underline{v})$

(3) **for** $j = 2, \ldots, k$

(3.1) $\underline{w} = \frac{2}{e}(\underline{\underline{A}}\,\underline{w}_1 - c\underline{w}_1) - \underline{w}_0$

(3.2) $\underline{w}_0 = \underline{w}_1$

(3.3) $\underline{w}_1 = \underline{w}$

This approach enables computation of $\underline{w}$ using matrix-vector products only. However, the above algorithm can generate overflow for large k because the generated iteratively product of the Chebyshev polynomial and the vector is not normalized till k iterations are performed. A version of the algorithm which avoids that by suitable normalization can be found in [98] and is called *Chebyshev iteration*.

Convergence and efficiency. Consider a matrix $\underline{\underline{A}}$ whose eigenvalues are real and ordered decreasingly, as in (A.13). If we choose γ to be equal λ_1 or λ_n and interval $\langle\alpha, \beta\rangle$ to include all the remaining eigenvalues, then the Algorithm 5 converges to λ_1 or λ_n respectively. For $k \to \infty$, corresponding convergence factors are defined by

$$\rho_{C_{\mathrm{LR}}} = \frac{a_2 + \sqrt{a_2^2 - e^2}}{a_1 + \sqrt{a_1^2 - e^2}} \tag{A.28}$$

$$\rho_{C_{\mathrm{SR}}} = \frac{a_{n-1} + \sqrt{a_{n-1}^2 - e^2}}{a_n + \sqrt{a_n^2 - e^2}} \tag{A.29}$$

where $a_i = |\lambda_i - c|$. It can be shown that maximal convergence rate for λ_1 is reached if c and e are defined by equations (A.27), where $\alpha = \lambda_n$ and $\beta = \lambda_2$. In this case, the convergence factor (A.28) is equal

$$\rho_{C_{\mathrm{LR}}} = \frac{e}{\lambda_1 - c + \sqrt{(\lambda_1 - c)^2 - e^2}} \tag{A.30}$$

At this point we can compare efficiency of Chebyshev acceleration with the simple linear shift. Considering the same matrix $\underline{\underline{A}}$ and taking the optimal shift for Algorithm 3 defined by (A.14) we get $\sigma_{\mathrm{LR}} = c$. Thus, corresponding convergence factor (A.16) becomes

$$\rho_{S_{\mathrm{LR}}} = \frac{e}{\lambda_1 - c} \tag{A.31}$$

Comparison of (A.30) and (A.31) shows that the Chebyshev acceleration is potentially much more superior to a simple linear shift. Analogous considerations can be performed for convergence of the smallest real eigenvalue λ_n which requires taking $\alpha = \lambda_{n-1}$ and $\beta = \lambda_n$.

Such consistent conclusions cannot be drawn from comparison of the Chebyshev acceleration with the inverse iteration. The former technique is well suited to accelerate the convergence of the eigenvalues from the ends of the spectrum while the latter one can speedup computation of an arbitrary eigenvalue. However, in order to be effective, inverse iteration requires much better approximation of the eigenvalue of interest than the Chebyshev acceleration.

It should be noted, that the above analyzes assumed a matrix with purely real spectrum. Some generalizations are required when the eigenvalues are located in the complex plane. Chebyshev polynomial $C_k\left(\frac{t-c}{e}\right)$ in the complex domain is related to an ellipse of center c and focal distance e. For the largest real eigenvalue λ_1, convergence properties of Algorithm 5 are described by the largest of convergence factors

$$\rho_{C_{\mathrm{LR}}}^{(i)} = \frac{a_i + \sqrt{a_i^2 - e^2}}{a_1 + \sqrt{a_1^2 - e^2}} \qquad \text{for} \qquad 2 \le i \le n \tag{A.32}$$

where a_i are the major semi-axes of confocal ellipses (characterized by the same c and e) passing through λ_i. The problem of convergence optimization is related to finding of the optimal ellipse. This is not a straightforward task, since parameters c and e cannot, in

general, be found from relations (A.27). More details concerning convergence properties of Chebyshev polynomials in the complex plane and computing an optimal ellipse can be found in [98] and [45].

Summarizing, suitably chosen parameters c and a cause that "unwanted" eigenvalues are mapped inside the ellipse while "wanted" are mapped outside. As a consequence eigenvalues of interest become better separated and convergence of the power method to the eigenvalue of the largest or smallest real part can be significantly improved.

A.1.3 Deflation

Consider a matrix $\underline{\underline{A}}$ of the size n with eigenvalues $\{\lambda_i\}_{i=1,\ldots,n}$ ordered as in (A.5) and an iterative solver such as the power method (Alg. 1), which is able to compute the largest magnitude eigenvalue λ_1 and corresponding eigenvector $\underline{v}_1$. In order to find the next largest magnitude eigenvalue λ_2 deflation technique can be used.

A general procedure, called *Wielandt's deflation*, consists of modifying the original matrix $\underline{\underline{A}}$ with the following rank one operation

$$\underline{\underline{A}}_1 = \underline{\underline{A}} - \sigma \underline{v}_1 \underline{w}^H \tag{A.33}$$

where $\underline{w}$ can be arbitrary vector such that $\underline{w}^H \underline{v}_1 = 1$, and σ is a shift. It can be shown that the eigenvalues of $\underline{\underline{A}}_1$ form the following set $\{\lambda_1 - \sigma, \lambda_2, \ldots, \lambda_n\}$. If σ is appropriately chosen, then (A.33) shifts λ_1 toward the origin leaving remaining eigenvalues unchanged. Thus, λ_2 becomes the largest magnitude eigenvalue of $\underline{\underline{A}}_1$ and can be found using the same eigensolver.

One can think that in order to compute some subset of dominant eigenvalues, deflation of form (A.33) can be applied many times producing $\underline{\underline{A}}_1$, $\underline{\underline{A}}_2$ and so on. However, it has to be remembered that matrices $\underline{\underline{A}}_i$ accumulate errors from all previous calculations and deflation should only be used for computing a few dominant eigenvalues.

Let $\underline{u}_i$ be the left eigenvector of $\underline{\underline{A}}$, corresponding to λ_i. In general, deflation (A.33) preserves the right eigenvector $\underline{v}_1$ and the left eigenvectors $\{\underline{u}_i\}_{i=2,\ldots,n}$. If we choose $\underline{w} = \underline{u}_1$ all right and left eigenvectors will be preserved. Such a choice is referred to as *Hotelling's deflation*. Another possibility is to take $\underline{w} = \underline{v}_1$, which has the property of preserving all the Schur vectors of $\underline{\underline{A}}$. This type of deflation has a great practical importance and is frequently used in the algorithms such as QR method, subspace iteration or Arnoldi method, described in the next sections. These algorithms often apply a generalization of the deflation (A.33), which uses several Schur vectors at a time. As a result of the deflation applied to the same matrix $\underline{\underline{A}}$ we get

$$\underline{\underline{A}}_j = \underline{\underline{A}} - \underline{\underline{Q}}_j \Sigma_j \underline{\underline{Q}}_j^H \tag{A.34}$$

where $\underline{\underline{Q}}_j = [\underline{q}_1, \underline{q}_2, \ldots, \underline{q}_j]$ is orthonormal column matrix of j Schur vectors of $\underline{\underline{A}}$ corresponding to $\{\lambda_1, \lambda_2, \ldots, \lambda_j\}$ and $\underline{\underline{\Sigma}}_j = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_j)$ is a diagonal matrix of shifts. It can be shown that the eigenvalues of $\underline{\underline{A}}_j$ become $\{\lambda_1 - \sigma_1, \ldots, \lambda_j - \sigma_j, \lambda_{j+1}, \ldots, \lambda_n\}$. Moreover, as stated above, the associated Schur vectors remain unchanged. More comments on practical implementation of the deflation can be found in [39, 98].

A.1.4 Projection

Projection methods are used in order to approximate the subspace spanned on eigenvectors of the original matrix $\underline{\underline{A}}$ of size n by some m-dimensional subspace $\mathcal{K}$ of $\mathbb{C}^n$, where $m < n$. These methods are based on transformation of the original problem

$$\underline{\underline{A}}\,\underline{u} = \lambda \underline{u} \tag{A.35}$$

to the approximate problem

$$\underline{\underline{P}}_{\mathcal{K}}(\underline{\underline{A}}\,\underline{\tilde{u}} - \tilde{\lambda}\underline{\tilde{u}}) = 0 \tag{A.36}$$

where $\underline{\underline{P}}_{\mathcal{K}}$ is the matrix representation of a *projector*[1] onto subspace $\mathcal{K}$, $\underline{\tilde{u}}$ is an approximation of $\underline{u}$ in this subspace, and $\tilde{\lambda}$ is an approximate eigenvalue λ. Since $\underline{\tilde{u}} \in \mathcal{K}$ we can write

$$\underline{\underline{P}}_{\mathcal{K}}\underline{\underline{A}}\,\underline{\tilde{u}} = \tilde{\lambda}\underline{\tilde{u}} \tag{A.37}$$

The projection is defined in terms of two m-dimensional subspaces: subspace of approximants $\mathcal{K}$ and the so-called left subspace $\mathcal{L}$ which is orthogonal to the residual vector $\underline{\underline{A}}\,\underline{\tilde{u}} - \tilde{\lambda}\underline{\tilde{u}}$. If we denote by V and W matrices whose columns form biorthogonal bases for subspaces $\mathcal{K}$ and $\mathcal{L}$, respectively, i.e.

$$\underline{\underline{W}}^{H}\underline{\underline{V}} = I \tag{A.38}$$

where I is the identity matrix, then the associated projector $\underline{\underline{P}}_{\mathcal{K}}$ can be represented by

$$\underline{\underline{P}}_{\mathcal{K}} = \underline{\underline{V}}\,\underline{\underline{W}}^{H} \tag{A.39}$$

This is referred to as *oblique projection*.

Using (A.39) and letting

$$\underline{\tilde{u}} = \underline{\underline{V}}\,\underline{y} \tag{A.40}$$

we can reformulate the approximate eigenproblem (A.37) to the form of

$$\underline{\underline{B}}_{m}\underline{y} = \tilde{\lambda}\underline{y} \tag{A.41}$$

where

$$\underline{\underline{B}}_{m} = \underline{\underline{W}}^{H}\underline{\underline{A}}\,\underline{\underline{V}} \tag{A.42}$$

Note, that $\underline{\underline{B}}_{m}$ is a matrix of size m and $m \ll n$ in practice. Therefore, computation of the eigenvalues $\tilde{\lambda}$ and the eigenvectors $\underline{y}$ of eigenproblem (A.41) is not expensive. The associated approximate eigenvectors $\underline{\tilde{u}}$ can be found using transformation (A.40). It was shown in [98] that the quality of the approximation depends on the angle between the exact eigenvector $\underline{u}$ and the subspace $\mathcal{K}$. If $\underline{u} \in \mathcal{K}$ then the approximation $\underline{\tilde{u}}$ is exact.

Most of the practical algorithms making use of projection methods, such as subspace iteration or Arnoldi method, apply the so-called *orthogonal projection*. This is a special

[1] A projector P is a linear transformation from $\mathbb{C}^n$ to itself which is idempotent, i.e., such that $P^2 = P$.

case, where $\mathcal{L} = \mathcal{K}$. As a consequence of condition (A.38) $\underline{\underline{W}} = \underline{\underline{V}}$ and columns of $\underline{\underline{V}}$ form an orthonormal basis for $\mathcal{K}$. Thus projector (A.39) takes the form

$$\underline{\underline{P}}_{\mathcal{K}} = \underline{\underline{V}}\,\underline{\underline{V}}^{H} \tag{A.43}$$

It results in

$$\underline{\underline{B}}_{m} = \underline{\underline{V}}^{H}\underline{\underline{A}}\,\underline{\underline{V}} \tag{A.44}$$

The procedure consisting of computing orthonormal basis $\underline{\underline{V}}$ and matrix $\underline{\underline{B}}_m$, solving approximate eigenproblem (A.41) for eigenpairs $\{\tilde{\lambda}, \underline{y}\}$ and determining corresponding approximate eigenvalues $\tilde{\lambda}$ (*Ritz values*) and eigenvectors $\underline{\tilde{u}} = \underline{\underline{V}}\,\underline{y}$ (*Ritz vectors*) of $\underline{\underline{A}}$, is called the *Rayleigh-Ritz procedure* (RR). The pairs of the approximate eigensolutions $\{\tilde{\lambda}, \underline{\tilde{u}}\}$ are called *Ritz pairs*.

A frequently used variation of the RR procedure computes Schur vectors of $\underline{\underline{B}}_m$ and, using transformation (A.40), calculates corresponding approximate Schur vectors of $\underline{\underline{A}}$. Such an approach is computationally more stable and more natural for algorithms which internally operate on the Schur vectors.

Dealing with orthogonal projections is numerically safer than dealing with oblique ones. However, methods based on oblique projections can offer some advantages. In particular, they may allow computing approximations to left as well as right eigenvectors simultaneously or (and) require far less storage than similar orthogonal projection methods.

A.2 QR method

QR method is the most popular technique for finding all the eigenvalues (and eigenvectors) of a standard dense eigenproblem (A.1). The method was developed in 1961 by Francis [35] and Kublanovskaya [62] concurrently. The algorithm consists in transforming of the operator matrix $\underline{\underline{A}}$ into the upper triangular Schur form via the process called *QR iteration*. It consists of a series of *similarity transformations* of form $\underline{\underline{A}}^{(i+1)} = \underline{\underline{Q}}^{(i)^{-1}}\underline{\underline{A}}^{(i)}\underline{\underline{Q}}^{(i)}$ called *QR steps*. Since $\underline{\underline{Q}}^{(i)}$ is orthogonal they are equivalent to $\underline{\underline{A}}^{(i+1)} = \underline{\underline{Q}}^{(i)^{H}}\underline{\underline{A}}^{(i)}\underline{\underline{Q}}^{(i)}$.

The algorithm can be summarized as follows:

Algorithm 7 QR METHOD

Input: $\underline{\underline{A}}, \tau$

Output: $\{\underline{\underline{H}}, \underline{\underline{V}}\} : \underline{\underline{A}}\,\underline{\underline{V}} = \underline{\underline{V}}\,\underline{\underline{H}}$; $\underline{\underline{V}}^{H}\underline{\underline{V}} = \underline{\underline{I}}$; $\underline{\underline{H}}$ is upper triangular

(1) $\underline{\underline{H}}^{(1)} = \underline{\underline{V}}^{(1)^{H}}\underline{\underline{A}}\,\underline{\underline{V}}^{(1)}$ (Hessenberg reduction)

(2) **for** $i = 1, 2, \ldots$ **until** $\{|h_{j+1,j}| < \tau(|h_{j,j}| + |h_{j+1,j+1}|)\}_{j=1,\ldots,n-1}$ (QR iteration)

(2.1) $\underline{\underline{H}}^{(i)} = \underline{\underline{Q}}^{(i)}\underline{\underline{R}}^{(i)}$ (QR factorization)

(2.2) $\underline{\underline{H}}^{(i+1)} = \underline{\underline{R}}^{(i)}\underline{\underline{Q}}^{(i)}$

(2.3) $\underline{\underline{V}}^{(i+1)} = \underline{\underline{V}}^{(i)} \underline{\underline{Q}}^{(i)}$

In order to reduce computational cost of subsequent QR factorizations (2.1) the algorithm starts with a transformation of $\underline{\underline{A}}$ to the form which is easy to factorize. Initial Hessenberg reduction (1) is usually performed with using orthogonal *Householder reflections* which can zero the elements below the first subdiagonal, column by column. Therefore after $n-2$ basic transformations we get Hessenberg form. Each QR factorization (2.1) usually involves orthogonal *Givens rotations* which are capable of zeroing single elements, so $n-1$ basic transformations are needed to triangularize the Hessenberg matrix. The QR step ends with multiplication (2.2) which turns back the matrix to the Hessenberg form.

Eigenvalues and eigenvectors. It can be shown [39] that successive QR iterations (2) tend the main subdiagonal elements $h_{j+1,j}$ to vanish. The algorithm stops when they become sufficiently small provided that Schur decomposition $\underline{\underline{V}}^H \underline{\underline{A}} \underline{\underline{V}} = \underline{\underline{H}}$ is found. The Schur form $\underline{\underline{H}}$ is desirable, because its eigenvalues are located on the diagonal. Since similarity transformations do not change the eigenvalues of the original matrix $\underline{\underline{A}}$, they are equal to the eigenvalues of $\underline{\underline{H}}$.

Step (2.3) accumulates successive transformations and is performed when all the eigenvectors $\{\underline{u}_j\}_{j=1,\ldots,n}$ of $\underline{\underline{A}}$ are requested. They are computed via the process called diagonalization (see [39] for details) once the Schur decomposition has been found. If only a certain subset of the eigenvectors is desired, an alternative method, involving inverse iteration technique (described in Sec. A.1.2.2), can be used. A widely followed rule of thumb advises to use this method if fewer than 25% of the eigenvectors are requested. Each eigenvector $\underline{u}_j$ corresponding to the eigenvalue λ_j is found as $\underline{u}_j = \underline{\underline{V}}_1 \underline{v}_j$, where $\underline{v}_j$ is produced by the Algorithm 4 with $\underline{\underline{A}} = \underline{\underline{H}}_1$ and $\sigma = \lambda_j$. It can be a very economical process since

- transformations in Step (2.3) have not to be accumulated (it gives savings of computational cost of order $\mathcal{O}(n^3)$ flops),
- Hessenberg matrices $\underline{\underline{H}}_1 - \lambda \underline{\underline{I}}$ can be factored very efficiently (in $\mathcal{O}(n^2)$ flops),
- only one iteration is typically required to produce an adequate approximate eigenvector.

Therefore each eigenvector can be found in $\mathcal{O}(n^2)$ flops.

Convergence and efficiency. Consider that the eigenvalues $\{\lambda_j\}_{j=1,\ldots,n}$ of $\underline{\underline{A}}$ are ordered decreasingly as in (A.13). In practice, it usually corresponds to the order of their approximations along the diagonal of $\underline{\underline{H}}$. It can be shown [39] that subdiagonal $h_{j+1,j}$ entry converges to zero with rate

$$\rho_{\mathrm{QR}} = \left| \frac{\lambda_{j+1}}{\lambda_j} \right| \tag{A.45}$$

The convergence can be, therefore, extremely slow, if the modules of any neighboring eigenvalues are not sufficiently different (e.g. complex pair). Algorithm 7 can be, however, accelerated by incorporating of shifting techniques described in Sec. A.1.2.

The following algorithm modifies Steps (2.1) and (2.2) of QR iteration in Algorithm 7 in order to implement shifts

Algorithm 8 EXPLICITLY SHIFTED QR METHOD

Input: $\underline{\underline{A}}, \tau$

Output: $\{\underline{\underline{H}}, \underline{\underline{V}}\} : \underline{\underline{A}}\,\underline{\underline{V}} = \underline{\underline{V}}\,\underline{\underline{H}}$; $\underline{\underline{V}}^H\underline{\underline{V}} = \underline{\underline{I}}$; $\underline{\underline{H}}$ is upper triangular

$\vdots$

(2.1) Select shift $\sigma^{(i)}$

(2.2) $\underline{\underline{H}}^{(i)} - \sigma^{(i)}\underline{\underline{I}} = \underline{\underline{Q}}^{(i)}\underline{\underline{R}}^{(i)}$ (QR factorization)

(2.3) $\underline{\underline{H}}^{(i+1)} = \underline{\underline{R}}^{(i)}\underline{\underline{Q}}^{(i)} + \sigma^{(i)}\underline{\underline{I}}$

(2.4) $\underline{\underline{V}}^{(i+1)} = \underline{\underline{V}}^{(i)}\underline{\underline{Q}}^{(i)}$

Convergence rate for the above algorithm is strongly dependent on the shift σ and for the eigenvalues ordered so that

$$|\lambda_1 - \sigma| \geq |\lambda_2 - \sigma| \geq \cdots \geq |\lambda_{n-1} - \sigma| \geq |\lambda_n - \sigma| \tag{A.46}$$

it is defined as

$$\rho_{S\mathrm{QR}} = \left| \frac{\lambda_{j+1} - \sigma}{\lambda_j - \sigma} \right| \tag{A.47}$$

It is evident that choosing σ close to λ_n, one can thoroughly improve the convergence of appropriate entries of Hessenberg form $\underline{\underline{H}}^{(i)}$. Fortunately, $\underline{\underline{H}}^{(i)}$ offers quite good estimates of the eigenvalues along the diagonal. The h_{nn} entry is regarded as the best one and is usually selected in each iteration (Step 2.1) as the shift $\sigma^{(i)}$. It results in rapid zeroing of the $h_{n,n-1}$ entry and fast convergence of h_{nn} to λ_n.

Such a single shift approach is inefficient when the matrix is real and h_{nn} is to approximate a complex eigenvalue. In this case real h_{nn} is no longer a good approximation. Taking a complex shift is not advised because it results in computations in complex arithmetic. In order to avoid these problems, the so-called *double shift strategy* can be used, based on performing two successive single shift QR steps at once. It is usually implemented without explicitly formulating $\underline{\underline{H}}^{(i)} - \sigma^{(i)}\underline{\underline{I}} = \underline{\underline{Q}}^{(i)}\underline{\underline{R}}^{(i)}$ factorizations and is called *implicitly shifted QR method*.

Another technique involved in practical QR algorithm is deflation (mentioned in Sec. A.1.3) which is performed in each QR iteration for eigenvalues regarded as the exact ones. Each deflation results in reducing of the size of the eigenproblem, which accelerates the whole algorithm.

Problem specific considerations. The version of the QR method described above is suited for nonsymmetric eigenproblems. The algorithm applied to a symmetric matrix is considerably simplified because:

- symmetric Hessenberg matrix $\underline{\underline{H}}$ is tridiagonal, so similarity transformations are less costly,
- symmetry and tridiagonal band structure are preserved when a single shift QR step is performed,
- consideration of complex shifts can be abandoned since the spectrum of the matrix is purely real,
- symmetric Schur matrix is diagonal and Schur vectors (columns of $\underline{\underline{V}}$) are identical to the eigenvectors of $\underline{\underline{A}}$, so their computation does not require any additional cost.

Empirical observations show that implicitly shifted QR algorithm for nonsymmetric real matrices requires $\approx 25n^3$ flops, if $\underline{\underline{H}}$ and $\underline{\underline{V}}$ are computed, while $\approx 10n^3$ flops is needed, if only the eigenvalues are desired. The algorithms for symmetric real matrices require $\approx 9n^3$ flops and $\approx 4/3n^3$ flops respectively.

Algorithms for complex matrices are similar to those for real eigenproblems with the difference that they operate on complex numbers. In practice, they are written so that all operations are performed in real arithmetic. The complex algorithms are ≈ 4 times more expensive than their corresponding real analogs (the complex analog of a real symmetric matrix is a Hermitian matrix).

More details on the techniques used in QR algorithms (e.g. balancing), as well as references to numerous papers devoted to the QR method, can be found in [39].

A.3 Subspace iteration

Subspace iteration, called also *simultaneous iteration*, is one of the simplest and the most popular methods for solving large, sparse and in general, nonsymmetric eigenproblems. The method was originally introduced in 1957 by Bauer [10] as *Treppeniteration* (staircase iteration). The subspace iteration computes the eigenvalues of largest modulus and can be viewed as a generalization of the power method described in Sec. A.1.1. The algorithm generates the sequence of matrices $\underline{\underline{A}}^i \underline{\underline{Q}}_m^{(0)}$, where $\underline{\underline{Q}}_m^{(0)}$ is an $n \times m$ column subspace formed from $m < n$ linearly independent vectors. Their linear independence is progressively lost during the iteration and should be reestablished in order to get convergence to the eigenvectors corresponding to different largest magnitude eigenvalues. For that purpose, the original Bauer's version uses LU decomposition, nevertheless usually QR decomposition is more frequently performed. Such a version of the method is sometimes called *orthogonal iteration*, since QR factorization additionally orthonormalize the

vectors. As a consequence, the columns of $\underline{\underline{Q}}_m$ converge to appropriate Schur vectors, in place of the eigenvectors.

The following basic subspace iteration algorithm simultaneously computes m Schur vectors of $\underline{\underline{A}}$ corresponding to the largest magnitude eigenvalues.

Algorithm 9 SUBSPACE ITERATION

Input: $\underline{\underline{A}}, \{\underline{q}_j\}_{j=1,\ldots,m}$ linearly independent, τ

Output: $\{\underline{\underline{T}}_m, \underline{\underline{Q}}_m\} : \underline{\underline{A}}\,\underline{\underline{Q}}_m = \underline{\underline{Q}}_m\underline{\underline{T}}_m$; $\underline{\underline{Q}}_m^H\underline{\underline{Q}}_m = \underline{\underline{I}}$; $\underline{\underline{T}}_m$ is upper triangular

(1) $\underline{\underline{Q}}_m^{(0)} = \left[\frac{\underline{q}_1}{\|\underline{q}_1\|}, \ldots, \frac{\underline{q}_m}{\|\underline{q}_m\|}\right]$

(2) **for** $i = 1, 2, \ldots$ **until** $\left\{\|(\underline{\underline{Q}}_m^{(i)} - \underline{\underline{Q}}_m^{(i-1)})_j\|_2 < \tau\right\}_{j=1,\ldots,m}$

(2.1) $\underline{\underline{V}}_m^{(i)} = \underline{\underline{A}}\,\underline{\underline{Q}}_m^{(i-1)}$

(2.2) $\underline{\underline{V}}_m^{(i)} = \underline{\underline{Q}}_m^{(i)}\underline{\underline{R}}_m^{(i)}$ (QR factorization)

(3) $\underline{\underline{T}}_m = \underline{\underline{Q}}_m^{(i)H}\underline{\underline{A}}\,\underline{\underline{Q}}_m^{(i)}$

The above algorithm is similar to the power method presented in the Algorithm 1. The main difference is that the iteration is performed on m vectors at once, thus $\underline{\underline{V}}_m \in \mathbb{C}^{n\times m}$, $\underline{\underline{T}}_m \in \mathbb{C}^{m\times m}$ and usually $m \ll n$. Step (2.2) can be viewed as a normalization process that is similar to the normalization used in the power method (Step (2.3) the in Algorithm 1). As a stopping criterion a simpler stationarity condition can be used instead of costly residual criterion used in the power method. It is based on a comparison of the columns of $\underline{\underline{Q}}_m$ calculated in succeeding iterations. Such criterion, however, can fail if $\underline{\underline{A}}$ has equal or nearly equal eigenvalues.

Eigenvalues and eigenvectors. Assuming that the eigenvalues $\{\lambda_j\}_{j=1,\ldots,n}$ of $\underline{\underline{A}}$ are ordered decreasingly as in (A.5), it can be shown that the above algorithm generates column matrix $\underline{\underline{Q}}_m = [\underline{q}_1, \ldots, \underline{q}_m]$, where each Schur vector $\underline{q}_j$ corresponds to λ_j. Since $\underline{\underline{T}}_m$ is a Schur form and its eigenvalues are equal to the diagonal entries, the eigenvalues of $\underline{\underline{A}}$ can be obtained as $\lambda_j = t_{jj}$. Corresponding eigenvectors $\underline{u}_j$ can be calculated by means of the inverse iteration in the following way: $\underline{u}_j = \underline{\underline{Q}}_m\underline{v}_j$, where $\underline{v}_j$ is produced by the Algorithm 4 with $\underline{\underline{A}} = \underline{\underline{T}}_m$ and $\sigma = \lambda_j$.

Convergence and efficiency. As in the power method, the convergence of the algorithm depends on the separation between neighboring eigenvalues. For eigenvalues ordered as above the convergence of the j-th column of $\underline{\underline{Q}}_m$ to the Schur vector of $\underline{\underline{A}}$ is proportional to the ratio

$$\rho_{\mathrm{SI}} = \left|\frac{\lambda_{j+1}}{\lambda_j}\right| \tag{A.48}$$

Convergence of the eigenvalues is governed by the same ratio and can be extremely slow if λ_j is close to λ_{j+1}. In order to improve the convergence, projection technique can be incorporated. In context of the subspace iteration, this technique was introduced for the first time by Stewart [110].

The additional enhancement in efficiency can be expected if matrix $\underline{\underline{Q}}_m$ is multiplied by $\underline{\underline{A}}$ several (say l) times before the relatively expensive orthonormalization Step (2.2) is performed.

The following algorithm takes advantage of the ideas mentioned above.

Algorithm 10 SUBSPACE ITERATION WITH PROJECTION

Input: $\underline{\underline{A}}, \{\underline{q}_j\}_{j=1,\ldots,m}$ linearly independent, k, τ

Output: $\{\underline{\underline{T}}_m, \underline{\underline{Q}}_m\} : \underline{\underline{A}}\,\underline{\underline{Q}}_m = \underline{\underline{Q}}_m \underline{\underline{T}}_m$; $\underline{\underline{Q}}_m^H \underline{\underline{Q}}_m = \underline{\underline{I}}$; $\underline{\underline{T}}_m$ is upper triangular

(1) $\underline{\underline{Q}}_m^{(0)} = \left[\frac{\underline{q}_1}{\|\underline{q}_1\|}, \ldots, \frac{\underline{q}_m}{\|\underline{q}_m\|}\right]$

(2) **for** $i = 1, 2, \ldots$ **until** $\left\{\|(\underline{\underline{A}}\,\underline{\underline{Q}}_m^{(i)} - \underline{\underline{Q}}_m^{(i)}\underline{\underline{T}}_m^{(i)})_j\|_2 < |t_{jj}|\tau\right\}_{j=1,\ldots,k}$

(2.1) Select l

(2.2) $\underline{\underline{V}}_m^{(i)} = \underline{\underline{A}}^l \underline{\underline{Q}}_m^{(i-1)}$

(2.3) $\underline{\underline{V}}_m^{(i)} = \underline{\underline{Q}}_m^{(i)} \underline{\underline{R}}_m^{(i)}$ (QR factorization)

(2.4) $\underline{\underline{B}}_m = \underline{\underline{Q}}_m^{(i)H} \underline{\underline{A}}\,\underline{\underline{Q}}_m^{(i)}$ (projection)

(2.5) $\underline{\underline{T}}_m^{(i)} = \underline{\underline{X}}_m^{(i)H} \underline{\underline{B}}_m \underline{\underline{X}}_m^{(i)}$ (Schur form via QR algorithm)

(2.6) $\underline{\underline{Q}}_m^{(i)} = \underline{\underline{Q}}_m^{(i)} \underline{\underline{X}}_m^{(i)}$

where $k \leq m$ is the number of required eigenvalues.

Proper selection of l is very important for general performance of the algorithm. For larger values of l, the subspace spanned on the dominant eigenvectors is better approximated by column subspace $\underline{\underline{V}}_m^{(i)}$. On the other hand, too large l can cause problems in orthogonalization step because the columns of $\underline{\underline{V}}_m^{(i)}$ may have become (numerically) linearly dependent. In practice, l can be determined based on observations of convergence rate of particular Schur vectors.

It can be shown [110] that the application of orthogonal projection (2.4) accelerates convergence of the j-th eigenvalue to the rate

$$\rho_{\text{SIP}} = \left|\frac{\lambda_{m+1}}{\lambda_j}\right| \tag{A.49}$$

In this case even the convergence of the last required k-th eigenvalue can be substantially accelerated by choosing large enough m.

Symmetric case. In the case of Hermitian matrix, Algorithm 10 can be simplified because projection (2.4) results in a symmetric matrix $\underline{\underline{B}}_m$. Therefore, a more efficient symmetric QR algorithm can be applied in Step (2.5), leading to a symmetric Schur (diagonal) form $\underline{\underline{T}}_m$. Moreover, computation of $\underline{\underline{A}}$'s eigenvectors requires no additional cost because the eigenvectors are equal to the Schur vectors $\underline{\underline{Q}}_m$, computed in Step (2.6).

Some variations of the subspace iteration (e.g., [111]) use an oblique projection instead of the orthogonal one. Then, the QR decomposition is not performed and so-called *lop-sided* oblique projection $\underline{\underline{B}}_m = (\underline{\underline{V}}_m^H \underline{\underline{V}}_m)^{-1} \underline{\underline{V}}_m^H \underline{\underline{A}}\, \underline{\underline{V}}_m$ is computed in place of Step (2.4).

The subspace iteration is the first of the algorithms discussed here which does not operate on matrix elements but uses only the information obtained from the product of a matrix operator and a vector. This is a very important feature, since considerable time savings can be gained, if the matrix is sparse and the matrix-vector product is adequately implemented.

Practical subspace iteration codes often use other basic algorithms discussed in Sec. A.1, which can additionally enhance the performance of the method. The most frequently used ones are *locking* (a form of deflation) and various spectral transformations. Details can be found in [98].

A.4 Arnoldi method

Arnoldi method is an orthogonal projection method for searching a few eigenvalues and eigenvectors of a general non-Hermitian eigenproblem (A.1). It was developed in 1951 by Arnoldi [4], who introduced the method as a means of reducing a dense matrix $\underline{\underline{A}}$ into the upper Hessenberg form $\underline{\underline{H}} = \underline{\underline{V}}^H \underline{\underline{A}}\, \underline{\underline{V}}$. He observed that the reduction truncated before completion can give good approximations of some eigenvalues.

Comparison of j-th columns in $\underline{\underline{A}}\, \underline{\underline{V}} = \underline{\underline{V}}\, \underline{\underline{H}}$ leads to

$$\underline{\underline{A}}\, \underline{v}_j = \sum_{i=1}^{n} h_{ij} \underline{v}_i = \sum_{i=1}^{j+1} h_{ij} \underline{v}_i \tag{A.50}$$

Separating the last term in the summation we get

$$\underline{w}_j \equiv h_{j+1,j} \underline{v}_{j+1} = \underline{\underline{A}}\, \underline{v}_j - \sum_{i=1}^{j} h_{ij} \underline{v}_i \;;\; h_{ij} = \underline{v}_i^H \underline{\underline{A}}\, \underline{v}_j \tag{A.51}$$

If $\underline{w}_j \neq 0$ then $\underline{q}_{j+1}$ can be expressed by

$$\underline{v}_{j+1} = \frac{\underline{w}_j}{h_{j+1,j}} \;;\; h_{j+1,j} = \|\underline{w}_j\|_2 \tag{A.52}$$

Equations (A.51) and (A.52) define iterative *Arnoldi process*. After m steps, this process generates the following *m-step Arnoldi factorization*, starting from an initial vector $\underline{v}_1$.

$$\underline{\underline{A}}\, \underline{\underline{V}}_m = \underline{\underline{V}}_m \underline{\underline{H}}_m + \underline{w}_m \underline{e}_m^H \tag{A.53}$$

where $\underline{\underline{V}}_m = [\underline{v}_1, \ldots, \underline{v}_m]$ is column orthogonal and $\underline{\underline{H}}_m \in \mathbb{C}^{m\times m}$ is upper Hessenberg with non-negative subdiagonal elements. Since $\underline{\underline{V}}_m \underline{w}_m = 0$ by construction, then

$$\underline{\underline{H}}_m = \underline{\underline{V}}_m^H \underline{\underline{A}}\, \underline{\underline{V}}_m \tag{A.54}$$

It should be noted that the columns of $\underline{\underline{V}}_m$, called *Arnoldi vectors*, form an orthonormal basis for a *Krylov subspace* $\mathcal{K}_m$

$$\mathcal{K}_m \equiv \mathrm{span}\{\underline{v}_1, \underline{\underline{A}}\,\underline{v}_1, \underline{\underline{A}}^2 \underline{v}_1, \ldots, \underline{\underline{A}}^{m-1} \underline{v}_1\} \tag{A.55}$$

Therefore $\underline{\underline{H}}_m$ is an orthogonal projection of $\underline{\underline{A}}$ onto $\mathcal{K}_m$.

The following algorithm computes m-step Arnoldi factorization.

Algorithm 11 ARNOLDI METHOD

Input: $\underline{\underline{A}}, \underline{v}_0, m$

Output: $\{\underline{\underline{H}}_m, \underline{\underline{V}}_m, \underline{w}_m\} : \underline{\underline{A}}\,\underline{\underline{V}}_m = \underline{\underline{V}}_m \underline{\underline{H}}_m + \underline{w}_m \underline{e}_m^T$; $\underline{\underline{V}}_m^H \underline{\underline{V}}_m = \underline{\underline{I}}$; $\underline{\underline{V}}_m \underline{w}_m = 0$; $\underline{\underline{H}}_m$ is upper Hessenberg

(1) $\underline{v}_1 = \frac{\underline{v}_0}{\|\underline{v}_0\|}$; $\underline{\underline{V}}_1 = \underline{v}_1$; $\hat{\underline{\underline{H}}}_1 = 0$

(2) **for** $j = 1, \ldots, m$

(2.1) $\underline{w}_j = \underline{\underline{A}}\,\underline{v}_j$

(2.2) $\underline{h} = \underline{\underline{V}}_j^H \underline{w}_j$; $\underline{w}_j = \underline{w}_j - \underline{\underline{V}}_j \underline{h}$

(2.3) $\underline{\underline{H}}_j = [\hat{\underline{\underline{H}}}_j, \underline{h}]$

(2.4) $\beta_j = \|\underline{w}_j\|_2$; $\underline{v}_{j+1} = \frac{\underline{w}_j}{\beta_j}$

(2.5) $\hat{\underline{\underline{H}}}_j = \begin{bmatrix} \underline{\underline{H}}_j \\ \beta_j \underline{e}_j^T \end{bmatrix}$; $\underline{\underline{V}}_{j+1} = [\underline{\underline{V}}_j, \underline{v}_{j+1}]$

With the successive iterations of the Arnoldi method, $\underline{\underline{V}}_m$ converges to an invariant subspace of $\underline{\underline{A}}$ and $\underline{w}_j$, computed in Step (2.2), becomes smaller. On the one hand, the decreasing of $\underline{w}_j$ is desirable because it indicates that the approximate eigenvalues become more exact (the rationale for that is discussed below in the paragraph devoted to eigenvalues and eigenvectors). On the other hand, however, significant digits of $\underline{w}_j$ can be easily lost due to numerical cancellation, leading to loss of numerical orthogonality of the iteratively generated Arnoldi vectors $\{\underline{v}_j\}_{j=1,\ldots,m}$. It should be noted that the orthogonality of $\underline{\underline{V}}_m$ is crucial for the algorithm since non-orthogonal basis vectors may cause numerical difficulties, such as spurious eigenvalues.

Great effort has been made in the last years to overcome the difficulties in the orthogonalization process. Different orthogonalization schemes have been proposed, i.e., *Gram-Schmidt* (GS), *Modified Gram-Schmidt* (MGS) or *Householder* approach (see [98] for details). Algorithm 11 uses the GS orthogonalization in Step (2.2). However, in order to obtain orthogonal to working precision vectors, a reorthogonalization step is sometimes

required. A simple and efficient method, called *DGKS correction*, has been proposed in [22] by Daniel et al. It consists in computing of the following iterative refinement steps

$$\underline{s} = \underline{\underline{V}}_j^H \underline{w}_j \ ; \ \underline{w}_j = \underline{w}_j - \underline{\underline{V}}_j \underline{s} \ ; \ \underline{h} = \underline{h} + \underline{s} \tag{A.56}$$

just after Step (2.2). If the DGKS correction is necessary, it is sufficient, in practice, to perform only one step of the refinement (A.56).

It should be noted that both, Step (2.2) and correction (A.56), can be accomplished with using Level 2 BLAS operations. This is important for performance on vector and parallel computers. Other alternative orthogonalization types are not recommended because they cannot be formulated in terms of Level 2 BLAS.

Eigenvalues and eigenvectors. In the projection method, such as the Arnoldi method, Ritz approximations $\{\tilde{\lambda}_j, \underline{\tilde{u}}_j\}$ of the eigensolutions of $\underline{\underline{A}}$ can be determined from eigenpairs of the approximate eigenproblem $\underline{\underline{H}}_m \underline{y} = \tilde{\lambda} \underline{y}$ (see Section A.1.4). Each Ritz vector $\underline{\tilde{u}}_j$, corresponding to the Ritz value $\tilde{\lambda}_j$, is calculated as $\underline{\tilde{u}}_j = \underline{\underline{V}}_m \underline{y}_j$. Quality of the approximation is represented by the residual norm r, called *Ritz estimate*

$$r_j = \|\underline{\underline{A}} \underline{\tilde{u}}_j - \underline{\tilde{u}}_j \tilde{\lambda}_j\|_2 = \|(\underline{\underline{A}} \underline{\underline{V}}_m - \underline{\underline{V}}_m \underline{\underline{H}}_m) \underline{y}_j\|_2 = h_{m+1,m} |\underline{e}_m^H \underline{y}_j| = \beta_m |\underline{e}_m^H \underline{y}_j| \tag{A.57}$$

If $\underline{w}_m = 0$ then $r_j = 0$ for $j = 1, \ldots, m$, $\underline{\underline{V}}_m$ spans an invariant subspace of $\underline{\underline{A}}$ and the Ritz approximations $\{\tilde{\lambda}_j, \underline{\tilde{u}}_j\}$ are exact.

Convergence and efficiency. Convergence theory of orthogonal projection methods onto a Krylov subspace for a general nonsymmetric matrix is very complex and will not be discussed here on the whole. We only mention some convergence properties important for comparison with other methods. More details can be found in [98].

As m increases, approximate Ritz pairs $\{\tilde{\lambda}_j, \underline{\tilde{u}}_j\}$, computed from the m-step Arnoldi factorization, converge to the eigenpairs of $\underline{\underline{A}}$. The subspace spanned on eigenvectors $\{\underline{\tilde{u}}_j\}_{j=1,\ldots,m}$ converges the most rapidly to the subspace spanned on the eigenvectors of $\underline{\underline{A}}$ corresponding to the largest magnitude eigenvalues. Therefore approximations of these vectors seem to be the most accurate. However, the accuracy also depends on the choice of the initial vector $\underline{v}_0$. In consequence, the other eigenvectors, better approximated by $\underline{v}_0$, may become more exact.

As an example, describe the convergence of the subspace $\mathcal{K}_m$ to an eigenvector $\underline{u}_1$ of $\underline{\underline{A}}$, corresponding to the largest eigenvalue λ_1. Assume that the initial vector $\underline{v}_1$ has the expansion $\underline{v}_1 = \sum_{j=1}^n \alpha_j \underline{u}_j$ with respect to the eigenbasis $\{\underline{u}_j\}_{j=1,\ldots,n}$, in which $\|\underline{u}_j\|_2 = 1$ for $j = 1, \ldots, n$ and $\alpha_1 \neq 0$. Define the distance between $\underline{u}_1$ and the subspace $\mathcal{K}_m$ as $\|(\underline{\underline{I}} - \underline{\underline{P}}_{\mathcal{K}}) \underline{u}_1\|_2$, where $\underline{\underline{P}}_{\mathcal{K}} = \underline{\underline{V}}_m \underline{\underline{V}}_m^H$ is a projection matrix. It is shown in [98], that the distance is bounded as follows

$$\|(\underline{\underline{I}} - \underline{\underline{P}}_{\mathcal{K}}) \underline{u}_1\|_2 \leq \xi_1 \frac{C_{m-1}\left(\frac{a}{e}\right)}{|C_{m-1}\left(\frac{\lambda_1 - c}{e}\right)|} \tag{A.58}$$

where $\xi_1 = \sum_{j=2}^{n} |\alpha_j|/|\alpha_1|$ and C_{m-1} is Chebyshev polynomial of degree $m-1$. Symbols c, e, a denote respectively the center, focal distance and the major semi-axis of the ellipse, containing all the eigenvalues of $\underline{\underline{A}}$ except λ_1.

It can be seen that the upper bound in (A.58) becomes small if $\underline{v}_0$ ($\underline{v}_1 = \underline{v}_0/\|\underline{v}_0\|$) is "reach" in the direction of $\underline{u}_1$ and λ_1 is well separated from other eigenvalues. These observations are obviously coincident with the convergence properties of the methods discussed in the previous sections, however, potential convergence of the Arnoldi method can be much superior due to the properties of Chebyshev polynomials, described in Sec. A.1.2.3.

Another observation is that the appropriate selection of the initial vector $\underline{v}_0$ can lead to the convergence to the eigenvalues other than the largest ones. In consequence, if $\underline{v}_0$ is "reach" in the direction of some arbitrary eigenvectors of interest, it may result in a small number of iterations m, which should be performed in order to calculate these eigenvectors with a required accuracy.

Symmetric case. Symmetric version of the Arnoldi method is called the *Lanczos method*. It can be viewed as a simplification of Arnoldi's method for the particular Hermitian matrix case. In fact, this method was developed earlier. It was introduced by Lanczos [63] in 1950, i.e., a year before the Arnoldi method was developed. Due to the symmetry of the operator not only can numerous simplifications be done in the algorithm, but also the convergence of the method is superior.

The simplifications are considerable because Hessenberg matrix $\underline{\underline{H}}_m$, generated in the Lanczos process, is real, tridiagonal and symmetric, and therefore the solution of the approximate eigenproblem is much simpler and less time consuming. Moreover, the calculations in Step (`2.2`) are radically simplified since they involve only a single dot product and a single scalar-vector multiplication (see [39, 98] for details).

In a symmetric case the eigenspectrum of $\underline{\underline{A}}$ is purely real. Assume that the eigenvalues are ordered decreasingly as in (A.13). The largest and the smallest eigenvalues are then denoted by λ_1 and λ_n respectively. It was shown in [39] that the extremal (the largest and the smallest) eigenvalues of $\underline{\underline{H}}_m$, generated in the Lanczos method, become progressively better approximates of $\underline{\underline{A}}$'s extremal eigenvalues as m increases. This behavior is the consequence of the Lanczos method's convergence properties described by the so-called *Kaniel-Paige theory*. Samples of this theory along with references can be found in [39]. It was therein shown that the choice of the initial vector is not so important for the convergence properties of the algorithm as in the nonsymmetric case because the convergence in the Kaniel-Paige style is very fast. A comparison of the Lanczos method and the power method (Algorithm 1) in context of the largest eigenvalue convergence demonstrated that the Lanczos method is much more superior.

Similarly to the subspace iteration, discussed in the previous section, the Arnoldi method represents a class of iterative algorithms which are independent on matrix storage scheme and problem type. The information about the operator matrix is passed to the algorithm as a result of a matrix-vector product in Step (`2.1`).

A.4.1 Restarting of the Arnoldi method

There are two important problems with the above simple Arnoldi method. The first one is acquisition of "wanted" eigenvalues. The reason is, that convergence properties of the algorithm favor the eigenspace corresponding to the eigenvalues with the largest absolute values while they may not be the ones of interest. The second problem is, that since the required eigenvectors are not a priori known (even approximately), the number of Arnoldi iterations m, which need to be taken in order to obtain expected accuracy, can be considerable. It may result in high requirements for memory and computation time due to increasing, in each iteration, size of generated matrices m.

One of the possibilities, which can resolve the above problems, is to restart the algorithm after each m iterations with appropriately updated starting vector $\underline{v}_1$. In this manner interesting information from a very large Krylov subspace is iteratively extracted and compressed in the fixed size m-dimensional subspace. The update has the form

$$\underline{v}_1^{(i+1)} = \phi(\underline{\underline{A}})\underline{v}_1^{(i)} \tag{A.59}$$

where ϕ is a filtering polynomial constructed so that "unwanted" part of the eigenspectrum is filtered out from $\underline{v}_1$. Such update determines the amount of required memory and causes that the algorithm converges to the "wanted" eigenvalues. The information about the spectrum, used for the construction of ϕ, may be given a priori from the spectral properties of $\underline{\underline{A}}$ or may be derived from $\underline{\underline{A}}$'s approximate eigenvalues just before the restart. The techniques based on the concept of restarting with use of polynomial filtering are referred to as *polynomial acceleration techniques*.

In general, the subspace size m can be selected independently before each restart. Assuming that m never exceeds some fixed maximal value, the amount of required memory is, however, still strictly determined.

Two alternative approaches to restarting procedure has been proposed. The first one is called *explicit restarting* and the second one is called *implicit restarting*. They are briefly discussed below.

A.4.1.1 Explicit restarting

Explicit restarting for symmetric problems (the Lanczos method) was developed by Cullum and Donath [20]. The explicitly restarted Arnoldi method (for nonsymmetric case) was proposed by Saad [96–98]. The explicitly restarted algorithm, computing k eigenvalues of interest, has the following form:

Algorithm 12 Explicitly restarted Arnoldi method

Input: $\underline{\underline{A}}, \underline{v}_0, k, m, \tau$

Output: $\left\{\{\lambda_j, \underline{u}_j\} : \underline{\underline{A}}\,\underline{u}_j = \lambda_j \underline{u}_j\right\}_{j=1,\ldots,k}$

(1) $\underline{v}_1 = \frac{\underline{v}_0}{\|\underline{v}_0\|}$

(2) **for** $i = 1, 2, \ldots$ **until** $\left\{ r_j < |\tilde{\lambda}_j| \tau \right\}_{j=1,\ldots,k}$

(2.1) $\underline{\underline{A}}\,\underline{\underline{V}}_m = \underline{\underline{V}}_m \underline{\underline{H}}_m + \underline{w}_m \underline{e}_m^T$ (m-step Arnoldi factorization)

(2.2) Compute $\{\tilde{\lambda}_j, \underline{\tilde{u}}_j, r_j\}_{j=1,\ldots,m}$

(2.3) Construct ϕ

(2.4) $\underline{v}_0 = \phi(\underline{\underline{A}})\underline{v}_1$; $\underline{v}_1 = \frac{\underline{v}_0}{\|\underline{v}_0\|}$

In each iteration, the Ritz approximations $\{\tilde{\lambda}_j, \underline{\tilde{u}}_j\}$ are calculated at Step (2.2) from eigensolutions of $\underline{\underline{H}}_m$. The QR method is usually used for this purpose. Nonexpensive stopping criterion in Step (2) is based on the corresponding Ritz estimates r_j (A.57), computed for k "wanted" eigenpairs. A main difference between various explicitly restarted Arnoldi/Lanczos algorithms relies on the construction of the filtering polynomial ϕ.

The original restarting strategy, proposed by Saad [96], was to replace the starting vector with a linear combination of Ritz vectors, corresponding to wanted Ritz values. Assuming that the first k Ritz pairs $\{\tilde{\lambda}_j, \underline{\tilde{u}}_j\}$ are "wanted", the replacement has the form

$$\underline{v}_1^{(i+1)} = \sum_{j=1}^{k} \alpha_j \underline{\tilde{u}}_j \tag{A.60}$$

Since $\underline{\tilde{u}}_j \in \mathcal{K}_m$ then $\underline{\tilde{u}}_j = \phi_j(\underline{\underline{A}})\underline{v}_1^{(i)}$ and (A.60) becomes

$$\underline{v}_1^{(i+1)} = \sum_{j=1}^{k} \alpha_j \phi_j(\underline{\underline{A}})\underline{v}_1^{(i)} \tag{A.61}$$

Comparing this equation to (A.59) we can see that (A.61) evidently defines the filtering polynomial $\phi = \sum_{j=1}^{k} \alpha_j \phi_j$.

In each iteration convergence of the algorithm depends on the choice of the coefficients α_j. In order to equalize the convergence rates of "wanted" Ritz values Saad [96] suggested a simple heuristic choice of the coefficients, favoring the Ritz vectors that have least converged. In this choice, j-th Ritz vector $\underline{\tilde{u}}_j$ is weighted with the value of its Ritz estimate ($\alpha_j = r_j$).

Another restarting strategy is to sort the spectrum of $\underline{\underline{A}}$ into two disjoint sets Ω_w and Ω_u, corresponding respectively to "wanted" and "unwanted" eigenvalues and construct the polynomial ϕ, which is minimized on an open convex set $\mathcal{C}_u$ containing Ω_u with $\Omega_w \cap \mathcal{C}_u = \varnothing$. One of Saad's proposals [97] was to choose ϕ to be a Chebyshev polynomial. Such approach is called the *Arnoldi-Chebyshev method*. In this case $\mathcal{C}_u$ is an ellipse. The problem of constructing the optimal ellipse has been studied in [45] and [98].

An alternative, which can be more appropriate in some cases, is to enclose the set Ω_u in a polygonal region $\mathcal{C}_u$ and find such polynomial ϕ that has minimum on $\mathcal{C}_u$ in a least squares sense. This is referred to as *least squares - Arnoldi method*. The details can be found in [98].

In order to examine some convergence properties of the Arnoldi-Chebyshev method and the least squares - Arnoldi method we expand the starting vector $\underline{v}_1$ into a series of $\underline{\underline{A}}$'s eigenvectors $\underline{u}_j$

$$\underline{v}_1^{(i)} = \sum_{j=1}^{n} \gamma_j \underline{u}_j \tag{A.62}$$

According to (A.59) we get the updated starting vector

$$\underline{v}_1^{(i+1)} = \sum_{j=1}^{n} \gamma_j \phi(\lambda_j) \underline{u}_j \tag{A.63}$$

If the eigenvalues $\{\lambda_j\}_{j=1,\ldots,n}$ are ordered as in (A.9), then the j-th original expansion coefficient is essentially attenuated, in each iteration, by the factor

$$\rho = \left| \frac{\phi(\lambda_j)}{\phi(\lambda_1)} \right| \tag{A.64}$$

In consequence, the iteration converges faster to the eigenvalues of interest, located outside the region $\mathcal{C}_u$.

A.4.1.2 Implicit restarting

Implicit restarting was proposed by Sorensen [108, 109]. This approach is more efficient and numerically stable than the explicit one. It combines the implicitly shifted QR mechanism (see Section A.2) with the k-step Arnoldi/Lanczos factorization. The implicitly restarted algorithm can be presented in the following steps:

Algorithm 13 IMPLICITLY RESTARTED ARNOLDI METHOD

Input: $\underline{\underline{A}}, \underline{v}_0, k, m, \tau$

Output: $\left\{\{\lambda_j, \underline{u}_j\} : \underline{\underline{A}}\,\underline{u}_j = \lambda_j \underline{u}_j\right\}_{j=1,\ldots,k}$

(1) $\underline{v}_1 = \frac{\underline{v}_0}{\|\underline{v}_0\|}$

(2) $\underline{\underline{A}}\,\underline{\underline{V}}_k = \underline{\underline{V}}_k \underline{\underline{H}}_k + \underline{w}_k \underline{e}_k^T$ (k-step Arnoldi factorization)

(3) **for** $i = 1, 2, \ldots$ **until** $\left\{r_j < |\tilde{\lambda}_j|\tau\right\}_{j=1,\ldots,k}$

(3.1) $\underline{\underline{A}}\,\underline{\underline{V}}_m = \underline{\underline{V}}_m \underline{\underline{H}}_m + \underline{w}_m \underline{e}_m^T$ (perform p additional steps ($k + p = m$))

(3.2) Compute $\{\tilde{\lambda}_j, \tilde{\underline{u}}_j, r_j\}_{j=1,\ldots,m}$

(3.3) Select shifts $\{\mu_j\}_{j=1,\ldots,p}$

(3.4) $\underline{\underline{A}}\,\underline{\underline{V}}_m^+ = \underline{\underline{V}}_m^+ \underline{\underline{H}}_m^+ + \underline{w}_m \underline{e}_m^T \underline{\underline{Q}}$ (apply p shifts via implicit QR)

(3.5) $\underline{\underline{A}}\,\underline{\underline{V}}_k^+ = \underline{\underline{V}}_k^+ \underline{\underline{H}}_k^+ + \underline{w}_k^+ \underline{e}_k^T$ (discard the last p columns)

Main differences between this algorithm and the previous explicitly restarted one rely on the way of applying filtering polynomial ϕ and on the fact that successive iterations in the implicit algorithm need not to be restarted from scratch but starts from the k-step factorization.

Applying the shifts $\{\mu_j\}_{j=1,\ldots,p}$ to the Arnoldi factorization is mathematically equivalent to explicitly restarting the algorithm with the initial vector (A.59), where

$$\phi(\lambda) = \prod_{j=1}^{p}(\lambda - \mu_j) \tag{A.65}$$

is a polynomial ϕ of degree p, whose roots are the shifts. This process is accomplished in Step (3.4) and can be implemented in the following way:

Algorithm 14 Application of shifts into m-step Arnoldi factorization

Input: $\underline{\underline{A}}\,\underline{\underline{V}}_m = \underline{\underline{V}}_m\underline{\underline{H}}_m + \underline{w}_m\underline{e}_m^T$, $\{\mu_j\}_{j=1,\ldots,p}$

Output: $\underline{\underline{A}}\,\underline{\underline{V}}_m^+ = \underline{\underline{V}}_m^+\underline{\underline{H}}_m^+ + \underline{w}_m\underline{q}^H : \underline{q}^H = \underline{e}_m^T\underline{\underline{Q}}$

(1) $\underline{q} = \underline{e}_m$

(2) **for** $j = 1, \ldots, p$

(2.1) $\underline{\underline{H}}_m - \mu_j\underline{\underline{I}} = \underline{\underline{Q}}_j\underline{\underline{R}}_j$ (QR factorization)

(2.2) $\underline{\underline{H}}_m = \underline{\underline{Q}}_j^H\underline{\underline{H}}_m\underline{\underline{Q}}_j$

(2.3) $\underline{\underline{V}}_m = \underline{\underline{V}}_m\underline{\underline{Q}}_j$

(2.4) $\underline{q}^H = \underline{q}^H\underline{\underline{Q}}_j$

It is worthwhile to note that each iteration (2) corresponds to a single iteration of the explicitly shifted QR Algorithm 8 with $\underline{\underline{H}} = \underline{\underline{H}}_m$ and $\sigma = \mu_j$. The implicitly shifted version of Algorithm 14 is mathematically equivalent, yet the QR factorization at Step (2.1) is not explicitly computed (see [108] for details).

A k-step Arnoldi factorization, which is used as a starting point for successive iterations of Algorithm 13, is obtained in Step (3.5) by equating the first k columns in the "shifted" m-step Arnoldi factorization, resulting from Step (3.4). In consequence, each iteration involves p matrix-vector multiplications (in Step (3.1)) in contrast to $m = k+p$ products, required by the explicit Algorithm 12.

Any polynomial acceleration technique, described in Section A.4.1.1, can be used in this implicit scheme on condition, that the filtering polynomial is of degree p. If the roots of the polynomial can be easily determined, e.g. in the case of Chebyshev polynomials, it is obvious, that ϕ can be applied by selecting shifts μ_j equal to its p roots. In order to apply ϕ which is specified by the coefficients, e.g. in the case of least squares polynomials, an alternative technique is also proposed in Sorensen's [108].

Another Sorensen's shift selection strategy that has proved successful in practice is called *exact shifts*. In this strategy, the shifts μ_j are selected to be equal to Ritz values $\tilde{\lambda}_j$, corresponding to p "unwanted" eigenvalues. This is mathematically equivalent to Saad's explicitly updating $\underline{v}_1$ with a particular linear combination of k "wanted" Ritz vectors $\underline{\tilde{u}}_j$.

The details of the implicitly restarted Arnoldi algorithm can be found in [108, 109].

Compared to the explicit restarting techniques, the implicit ones are more stable, due to implementation of the implicitly shifted QR restarting mechanism, and more efficient, approximately by a factor of $(k+p)/p$.

The restarted Arnoldi/Lanczos algorithms are very efficient tools for finding eigenvalues located at any end of the spectrum. The performance of these methods can be additionally enhanced by the application of deflation and preconditioning techniques, which can be used independently on the restarting techniques described in Section A.4.1. The most effective preconditioners are inverse iteration and Chebyshev polynomials. More implementation details can be found in [98].

A.5 Generalized eigenproblems

In practice, there are two general approaches of solving generalized eigenproblems of form (A.2). The first one is the QZ method and the second one is the reduction of the eigenproblem to a standard one and exploiting the methods for standard eigenproblems discussed in the previous sections. These approaches are shortly discussed in the succeeding sections.

A.5.1 QZ method

QZ method can be regarded as a version of the QR method intended for generalized eigenproblems of the form (A.2). It is based on a series of orthogonal transformations $\underline{\underline{A}}^{(i+1)} = \underline{\underline{Q}}^{(i)H}\underline{\underline{A}}^{(i)}\underline{\underline{Z}}^{(i)}$ and $\underline{\underline{B}}^{(i+1)} = \underline{\underline{Q}}^{(i)H}\underline{\underline{B}}^{(i)}\underline{\underline{Z}}^{(i)}$, which lead the pencil $\{\underline{\underline{A}}, \underline{\underline{B}}\}$ to the generalized Schur decomposition $\{\underline{\underline{H}}, \underline{\underline{T}}\}$, where $\underline{\underline{H}}$ and $\underline{\underline{T}}$ are upper triangular. The algorithm can be summarized as follows:

Algorithm 15 QZ METHOD

Input: $\underline{\underline{A}}, \underline{\underline{B}}, \tau$

Output: $\{\underline{\underline{H}}, \underline{\underline{V}}, \underline{\underline{T}}, \underline{\underline{W}}\} : \underline{\underline{A}}\,\underline{\underline{W}} = \underline{\underline{V}}\,\underline{\underline{H}}$; $\underline{\underline{B}}\,\underline{\underline{W}} = \underline{\underline{V}}\,\underline{\underline{T}}$; $\underline{\underline{V}}^H\underline{\underline{V}} = \underline{\underline{I}}$; $\underline{\underline{W}}^H\underline{\underline{W}} = \underline{\underline{I}}$; $\underline{\underline{H}}, \underline{\underline{T}}$ are upper triangular

(1) $\underline{\underline{H}}^{(1)} = \underline{\underline{V}}^{(1)H}\underline{\underline{A}}\,\underline{\underline{W}}^{(1)}$; $\underline{\underline{T}}^{(1)} = \underline{\underline{V}}^{(1)H}\underline{\underline{B}}\,\underline{\underline{W}}^{(1)}$ (Hessenberg-triangular reduction)

(2) **for** $i = 1, 2, \ldots$ **until** $\{|h_{j+1,j}| < \tau(|h_{j,j}| + |h_{j+1,j+1}|)\}_{j=1,\ldots,n-1}$ (QZ iteration)

(2.1) Compute $\{\underline{\underline{Q}}^{(i)}, \underline{\underline{Z}}^{(i)}\}$: $\underline{\underline{Q}}^{(i)^H} \underline{\underline{H}}^{(i)} \underline{\underline{Z}}^{(i)}$ is upper Hessenberg;
$\underline{\underline{Q}}^{(i)^H} \underline{\underline{T}}^{(i)} \underline{\underline{Z}}^{(i)}$ is upper triangular;
$\underline{\underline{H}}^{(i)} \underline{\underline{T}}^{(i)^{-1}} \stackrel{\text{QR}}{=} (\underline{\underline{Q}}^{(i)} \underline{\underline{D}}^{(i)}) \underline{\underline{R}}^{(i)}$;
$\underline{\underline{D}}^{(i)} = \text{diag}(1, \pm 1, \ldots, \pm 1)$; $\underline{\underline{Z}}^{(i)^H} \underline{\underline{Z}}^{(i)} = \underline{\underline{I}}$

(2.2) $\underline{\underline{H}}^{(i+1)} = \underline{\underline{Q}}^{(i)^H} \underline{\underline{H}}^{(i)} \underline{\underline{Z}}^{(i)}$

(2.3) $\underline{\underline{T}}^{(i+1)} = \underline{\underline{Q}}^{(i)^H} \underline{\underline{T}}^{(i)} \underline{\underline{Z}}^{(i)}$

(2.4) $\underline{\underline{V}}^{(i+1)} = \underline{\underline{V}}^{(i)} \underline{\underline{Q}}^{(i)}$

(2.5) $\underline{\underline{W}}^{(i+1)} = \underline{\underline{W}}^{(i)} \underline{\underline{Z}}^{(i)}$

General ideas incorporated in the above algorithm are similar to the concepts exploited by the Algorithm 7. In order to reduce computational cost of *QZ iterations* the algorithm starts with the reduction of $\underline{\underline{A}}$ and $\underline{\underline{B}}$ to Hessenberg form $\underline{\underline{H}}^{(1)}$ and upper triangular form $\underline{\underline{T}}^{(1)}$, respectively. This is performed using orthogonal Householder and Givens transformations. Same transformations are used in each QZ iteration, determining orthogonal matrices $\underline{\underline{Q}}^{(i)}$ and $\underline{\underline{Z}}^{(i)}$. It should be noted, that $\underline{\underline{Q}}^{(i)}$ has the same first column as would be obtained in the QR iteration applied to $\underline{\underline{H}}^{(i)} \underline{\underline{T}}^{(i)^{-1}}$, while the next columns can only differ in sign. The construction of $\underline{\underline{Z}}^{(i)}$ is dictated by the conservation of the Hessenberg-triangular structure of the pencil $\{\underline{\underline{H}}, \underline{\underline{T}}\}$. The iteration carries on until the stopping criterion of the algorithm is fulfilled, i.e., until the magnitudes of the main subdiagonal elements $h_{j+1,j}$ of $\underline{\underline{H}}^{(i)}$ become sufficiently small. Finally, the algorithm results in the generalized Schur decomposition $\{\underline{\underline{V}}^H \underline{\underline{A}} \underline{\underline{W}} = \underline{\underline{H}}, \underline{\underline{V}}^H \underline{\underline{B}} \underline{\underline{W}} = \underline{\underline{T}}\}$. The overall process is mathematically equivalent to the QR method Algorithm 7 performed for $\underline{\underline{A}}^{(i)} \underline{\underline{B}}^{(i)^{-1}}$ matrix.

The eigenvalues λ of the original pencil $\{\underline{\underline{A}}, \underline{\underline{B}}\}$ are equal to the eigenvalues of the pencil $\{\underline{\underline{H}}, \underline{\underline{T}}\}$ and can be calculated as $\{\lambda_j = h_{jj}/t_{jj} : t_{jj} \neq 0\}_{j=1,\ldots,n}$. The transformations $\underline{\underline{Q}}^{(i)}$ and $\underline{\underline{Z}}^{(i)}$ are accumulated in Steps (2.4) and (2.5) only if all the eigenvectors $\{\underline{u}_j\}_{j=1,\ldots,n}$ of $\underline{\underline{A}}$ are desired. If just some eigenvectors are requested and the inverse iteration technique, described in Sec. A.5.2 is used, these costly operations can be avoided. In this case, the eigenvector $\underline{u}_j$, corresponding to the eigenvalue λ_j, is found as $\underline{u}_j = \underline{\underline{W}}^{(1)} \underline{v}_j$, where $\underline{v}_j$ is generated by Algorithm 17 with $\underline{\underline{A}} = \underline{\underline{H}}^{(1)}$, $\underline{\underline{B}} = \underline{\underline{T}}^{(1)}$ and $\sigma = \lambda_j$.

Efficiency of the QZ method can be improved in practical algorithms by application of the techniques such as implicit shift or deflation. Implementation details can be found in [39]. Computational cost of such optimized QZ method is $\approx 66n^3$ flops if $\underline{\underline{H}}$, $\underline{\underline{T}}$, $\underline{\underline{V}}$ and $\underline{\underline{W}}$ are desired, while it needs $\approx 30n^3$ flops if only the eigenvalues are computed.

In reality the QZ method is applicable only to nonsymmetric eigenproblems because the transformations of the form $\underline{\underline{A}}^{(i+1)} = \underline{\underline{Q}}^{(i)^H} \underline{\underline{A}}^{(i)} \underline{\underline{Z}}^{(i)}$ and $\underline{\underline{B}}^{(i+1)} = \underline{\underline{Q}}^{(i)^H} \underline{\underline{B}}^{(i)} \underline{\underline{Z}}^{(i)}$ destroy symmetry of the pencil.

A.5.2 Reduction to standard form

The simplest transformation method of generalized eigenproblem (A.2) into a standard one is to compute, for nonsingular $\underline{\underline{B}}$, the operator

$$\underline{\underline{C}} = \underline{\underline{B}}^{-1}\underline{\underline{A}} \tag{A.66}$$

and to solve standard $\underline{\underline{C}}\,\underline{v} = \lambda\underline{v}$ problem. This approach is explored by iterative methods (i.e., the power method, subspace iteration or the Arnoldi method) in the case of nonsymmetric $\underline{\underline{B}}$. However, $\underline{\underline{B}}^{-1}\underline{\underline{A}}$ matrix is never explicitly formulated since the matrix inversion is very costly, ill conditioned and destroys matrix structure and sparsity. In practice, at the beginning of the iterative algorithm, LU decomposition $\underline{\underline{B}} = \underline{\underline{L}}\,\underline{\underline{U}}$ is computed, factoring $\underline{\underline{B}}$ into the lower and upper triangular matrices, respectively. Then appropriate triangular systems are solved each time, when the product of $\underline{\underline{B}}^{-1}$ and a vector is to be computed. Therefore, in order to compute $\underline{w} = \underline{\underline{B}}^{-1}\underline{\underline{A}}\,\underline{v}$ the following steps are performed:

Algorithm 16 $\underline{\underline{B}}^{-1}\underline{\underline{A}}$ MATRIX-VECTOR PRODUCT VIA LU DECOMPOSITION OF $\underline{\underline{B}}$

Input: $\underline{\underline{A}}, \{\underline{\underline{L}}, \underline{\underline{U}}\} : \underline{\underline{B}} = \underline{\underline{L}}\,\underline{\underline{U}}, \underline{v}$

Output: $\underline{w} = \underline{\underline{B}}^{-1}\underline{\underline{A}}\,\underline{v}$

(1) $\underline{x} = \underline{\underline{A}}\,\underline{v}$

(2) Solve $\underline{\underline{L}}\,\underline{y} = \underline{x}$ (forward substitution)

(3) Solve $\underline{\underline{U}}\,\underline{w} = \underline{y}$ (back substitution)

For dense matrices, computation of the LU requires $2n^3/3$ flops, while each forward or back substitution needs n^2 flops. This cost is relatively high but in the case of structured sparse matrices, e.g. banded, it can be substantially reduced. Denote by p and q the lower and the upper bandwidth of $\underline{\underline{B}}$, where $p, q \ll n$. The cost of the factorization is reduced to $2npq$ flops while forward and back substitutions cost $2np$ and $2nq$, respectively.

In order to transform a nonsymmetric generalized eigenproblem into a standard one the decomposition (e.g. LU) is always computed. In this case, a shift-invert preconditioner, which also requires the factorization (see Sec. A.1.2.2), can be applied at no additional cost. An example algorithm exploring the shift-invert technique in the context of the power method has the following form

Algorithm 17 INVERSE ITERATION FOR GENERALIZED EIGENPROBLEM

Input: $\underline{\underline{A}}, \underline{\underline{B}}, \sigma, \underline{v}_0, \tau$

Output: $\{\lambda, \underline{v}\} : \underline{\underline{A}}\,\underline{v} = \lambda\underline{\underline{B}}\,\underline{v}$

(1) $\underline{v} = \frac{\underline{v}_0}{\|\underline{v}_0\|}$

(2) $(\underline{\underline{A}} - \sigma\underline{\underline{B}}) = \underline{\underline{L}}\,\underline{\underline{U}}$ (LU decomposition)

(3) **for** $i = 1, 2, \ldots$ **until** $\|\underline{\underline{A}}\,\underline{v} - \lambda\underline{\underline{B}}\,\underline{v}\|_2 < \tau|\lambda|$

(3.1) $\underline{x} = \underline{\underline{B}}\,\underline{v}$

(3.2) Solve $\underline{\underline{L}}\,\underline{y} = \underline{x}$ (forward substitution)

(3.3) Solve $\underline{\underline{U}}\,\underline{w} = \underline{y}$ (back substitution)

(3.4) $\underline{w} = \frac{\underline{w}}{\|\underline{w}\|}$

(3.5) $\lambda = \sigma + \frac{\underline{w}^H\underline{\underline{B}}\,\underline{v}}{\underline{w}^H\underline{\underline{B}}\,\underline{w}}$

(3.6) $\underline{v} = \underline{w}$

It should be noted, that Steps (3.1) to (3.3) implicitly solve $(\underline{\underline{A}} - \sigma\underline{\underline{B}})\underline{w} = \underline{\underline{B}}\,\underline{v}$ system. Subsequent approximations of λ are calculated with the Rayleigh quotient

$$\lambda = \frac{\underline{v}^H\underline{\underline{A}}\,\underline{v}}{\underline{v}^H\underline{\underline{B}}\,\underline{v}} \tag{A.67}$$

If the normalization in Step (3.4) has the form $\|\underline{w}\|_B = \underline{w}^H\underline{\underline{B}}\,\underline{w}$, the Rayleigh quotient calculation in Step (3.5) can be simplified so that $\lambda = \sigma + \underline{w}^H\underline{\underline{B}}\,\underline{v}$.

The algorithm converges to the eigenvalue λ closest to the shift σ. In order to improve convergence, the shift can be occasionally changed during the iteration. However, the change of σ implies expensive recomputation of the LU decomposition and should not be performed too frequently.

Symmetric case. In the case of symmetric generalized eigenproblem, the approach based on involving the LU decomposition in implicit formulation of $\underline{\underline{B}}^{-1}\underline{\underline{A}}$ matrix is not advised since it destroys the operator symmetry. Assume that $\underline{\underline{B}}$ is positive definite. In order to transform the problem into a symmetric standard one, Cholesky factorization $\underline{\underline{B}} = \underline{\underline{G}}\,\underline{\underline{G}}^H$ can be computed and the following symmetric operator can be formulated

$$\underline{\underline{C}} = \underline{\underline{G}}^{-1}\underline{\underline{A}}\,\underline{\underline{G}}^{-H} \tag{A.68}$$

Its eigenvalues correspond to the eigenvalues of the pencil $\{\underline{\underline{A}}, \underline{\underline{B}}\}$. The eigenvectors $\underline{v}$ of the pencil can be computed as $\underline{v} = \underline{\underline{G}}^{-H}\underline{q}$, where $\underline{q}$ are the eigenvectors of $\underline{\underline{C}}$. If the size of the matrices is small, then $\underline{\underline{C}}$ can be explicitly formulated and QR method can be used for determination of the eigenvalues and eigenvectors. It leads to the following algorithm

Algorithm 18 Solution of symmetric generalized eigenproblem by means of Cholesky factorization and QR method

Input: $\{\underline{\underline{A}}, \underline{\underline{B}}\} : \underline{\underline{A}} = \underline{\underline{A}}^H$; $\underline{\underline{B}} = \underline{\underline{B}}^H$; $\underline{\underline{B}}$ is positive definite

Output: $\{\mathrm{diag}(\lambda_1, \ldots, \lambda_n), \underline{\underline{V}}\} : \underline{\underline{V}}^H\underline{\underline{A}}\,\underline{\underline{V}} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$; $\underline{\underline{V}}^H\underline{\underline{B}}\,\underline{\underline{V}} = \underline{\underline{I}}$

(1) $\underline{\underline{B}} = \underline{\underline{G}}\,\underline{\underline{G}}^H$ (Cholesky factorization)

(2) $\underline{\underline{C}} = \underline{\underline{G}}^{-1}\underline{\underline{A}}\,\underline{\underline{G}}^{-H}$

(3) $\mathrm{diag}(\lambda_1, \ldots, \lambda_n) = \underline{\underline{Q}}^H\underline{\underline{C}}\,\underline{\underline{Q}}$ (Schur form via QR algorithm)

(4) $\underline{\underline{V}} = \underline{\underline{G}}^{-H}\underline{\underline{Q}}$

This algorithm requires $\approx 14n^3$ flops.

If $\underline{\underline{A}}$ and $\underline{\underline{B}}$ are large and sparse, then $\underline{\underline{C}}$ should not be explicitly formulated due to high cost of the process and probable loss of sparsity and structure. In this case, some iterative eigensolver is used, which implements a product of the operator $\underline{\underline{C}}$ and a vector in the way similar to shown in Algorithm 16.

The cost of Cholesky factorization is half as high as the cost of the LU decomposition, while forward and back substitutions for $\underline{\underline{G}}$ and $\underline{\underline{G}}^H$ matrices are same expensive as for $\underline{\underline{L}}$ and $\underline{\underline{U}}$. Similarly to the LU decomposition, the cost of the factorization and the substitutions can be reduced for banded $\underline{\underline{B}}$.

Appendix B

Nonreciprocal ferrite phase shifter

This appendix is intended as an illustration of practical significance of some computational techniques discussed in this thesis and presents design considerations for a nonreciprocal latching ferrite phase shifter that was manufactured and measured in Telecommunications Research Institute, Gdańsk Division [15]. The measurement results of this structure validate the analysis performed in Sec. 5.4.1.

B.1 Structure considerations

Waveguide ferrite phase shifters are often used in the feeding systems of high-power electronically scanned array antennas. Among a variety of phase shifter configurations latching ferrite phase shifters with toroidal ferrite section are the most frequently used for that purpose. The main advantages of such structures are elimination of external magnets and reduction of switching power [115]. A cross-section of a typical latching ferrite structure is shown in Fig. B.1(a). A ferrite toroid can be permanently magnetized by the impulses of electrical current flowing via a wire placed in the central slot of the toroid. Depending on the direction of the current, induced remanence magnetization is $+M_r$ or $-M_r$, which correspond to the states where propagation constant of the dominant mode is, respectively, β^+ and β^-. These values define one of the most important phase shifter parameters, i.e. *nonreciprocal phase shift*

$$\Delta\Theta = \beta^+ - \beta^- \tag{B.1}$$

which is the phase change per unit length between two states of opposite magnetization.

In order to improve the figure of merit of the phase shifter, which is the phase change divided by the attenuation, various modifications of the basic structure shown in Fig. B.1(a) have been proposed. Clark [17] investigated the effect of chamfering the corners of the ferrite toroid. He found that the modified structure resulted in 20% increase in the nonreciprocal phase shift and a reduction of the microwave attenuation. Ince et al. have verified these results in [47] and they showed that the increase is in fact 10%, while another 10% was probably caused by the elimination of mechanical stresses occurring in the non-chamfered toroid. Another modification, contributing nothing to the phase

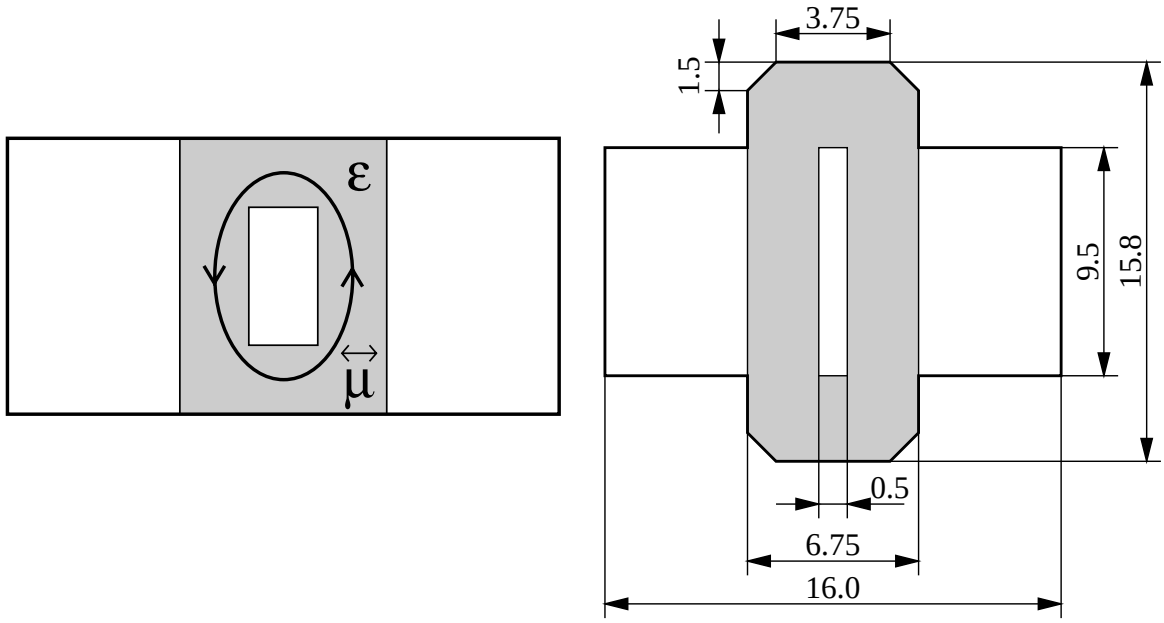


Figure B.1: Cross-sections of a simple phase shifter structure (a) and the physical structure designed in Telecommunications Research Institute, Gdańsk Division (b).

change but only decreasing the microwave loss, has been proposed by Mizobuchi and Kurebayashi [78]. They used reduced height grooved waveguide to obtain 20% improvement of the figure of merit compared to the phase changer in a rectangular waveguide. Another problem arising in the design of the phase shifters are higher order modes degrading the reflection and transmission characteristics of the device. The methods for suppressing the parasitic modes, consisting on the application of a thin resistive sheet in the central y-z plane of the guide, are described in [116].

Taking into account the described above improvements of the basic structure, the phase shifter shown in Fig. B.1(b) was designed and manufactured in Telecommunications Research Institute, Gdańsk Division [15].

B.2 Analysis

The simplified analysis of the reduced height grooved waveguide structure has been performed in [78] by means of the transmission matrix technique. However, in order to apply this method, the toroid is approximated by three vertical slabs homogeneous in the y-direction. Main disadvantage of this approach is that the nonuniform magnetization vector inside the ferrite cannot be accurately modeled.

In our approach, we suggest the application of the perturbation method (which is a special case of the coupled mode method for the only one fundamental mode taken into expansion) with basis fields computed by means of the finite difference frequency domain

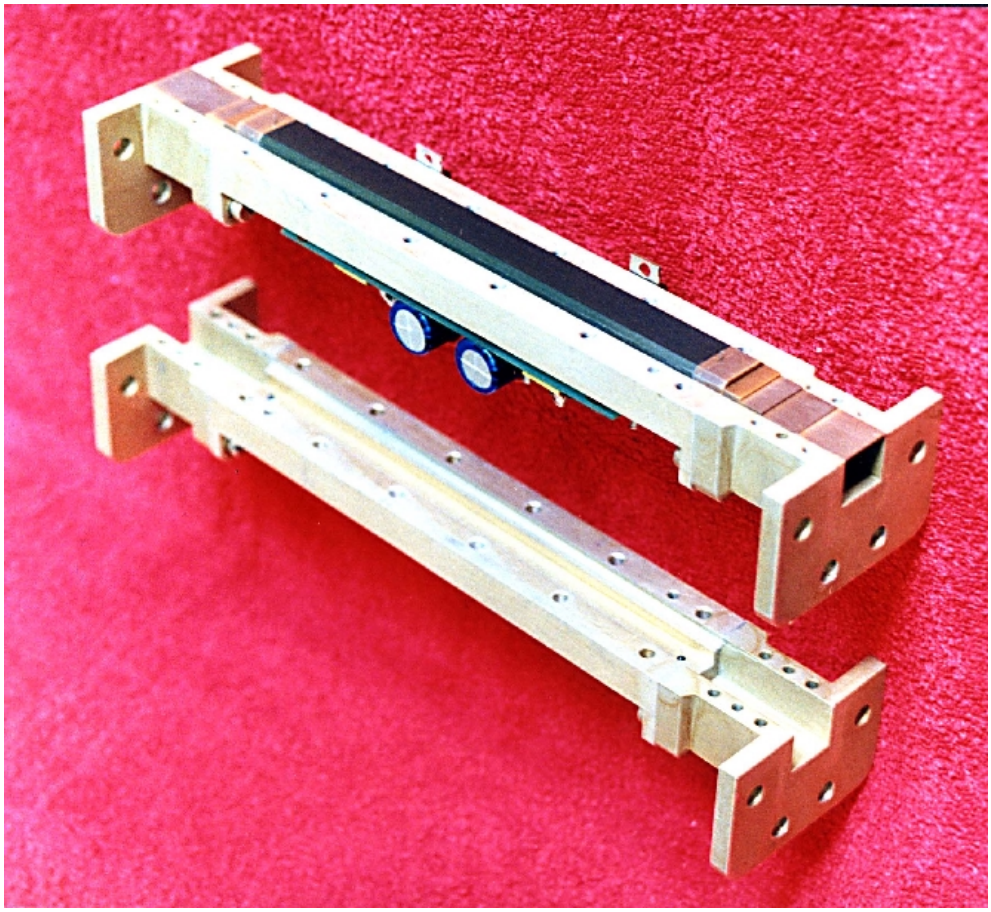
method. The details of the analysis are described in Sec. 5.4.1.

The application of the perturbation method allows one to divide the structure into any subregions with locally defined magnetization vector and thus exact modeling of nonuniform distribution of the magnetization vector is possible.

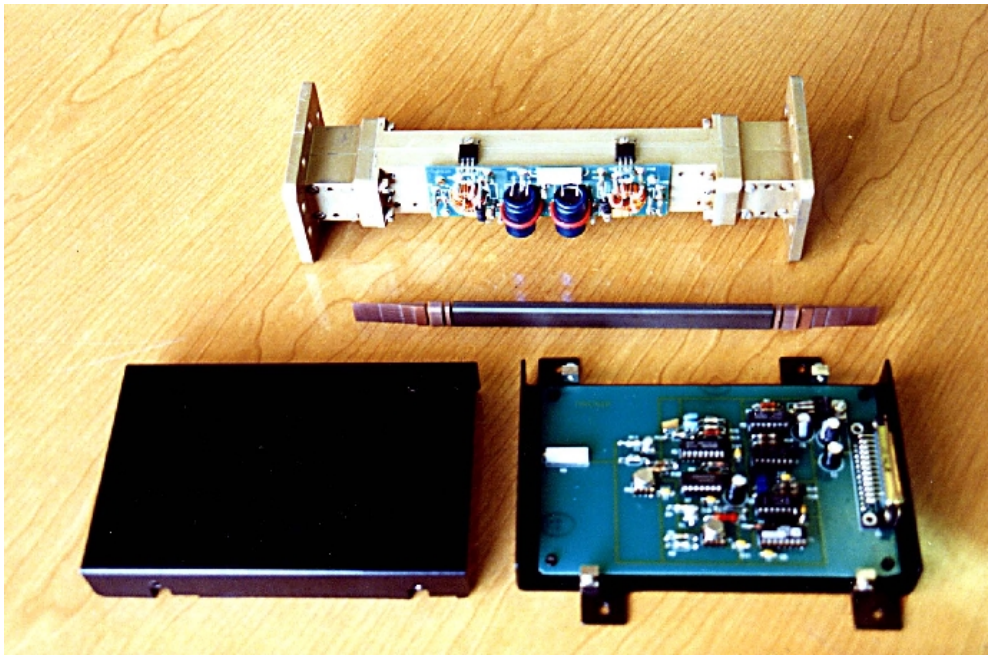
B.3 Realization and measurements

The manufactured phase shifter structure shown in Fig. B.2 included a 16 mm long yttrium-gadolinium garnet G-120 ferrite toroid fabricated by POLFER Ltd. The ferrite section was matched to R-58 waveguide using dielectric loaded four-section transformers providing maximally flat impedance matching in the operation frequency range from 5 to 5.6 GHz.

Measurement results of the nonreciprocal phase shift $\Delta\Theta$ along with the analytical values are compared in Fig. 5.17(b) and discussed in Sec. 5.4.1. Measurement results of the insertion loss and the return loss of the total structure can be found in [15].



(a)



(b)

Figure B.2: The ferrite phase shifter structure manufactured and measured in Telecommunications Research Institute, Gdańsk Division. Two halves of the waveguide structure (a) and a view of the entire structure with driver part and long ferrite toroid (b).

Bibliography

[1] M. M. Afande, K. Wu, M. Giroux, and R. G. Bosisio. A finite-difference frequency-domain method that introduces condensed nodes and image principle. *IEEE Trans. Microwave Theory Tech.*, 43:838–846, Apr. 1995.

[2] S. Ahmed and P. Daly. Finite-element method for inhomogeneous waveguides. *Proc. IEE*, 116:1661–1664, Oct. 1969.

[3] P. L. Arlett, A. K. Bahrani, and O. C. Zienkiewicz. Applications of finite elements to the solution of Helmholtz's equation. *Proc. IEE*, 115:1762–1766, Dec. 1968.

[4] W. E. Arnoldi. The principle of minimized iterations in the solution of the matrix eigenvalue problem. *Quarterly of Applied Mathematics*, 9:17–29, 1951.

[5] I. Awai and T. Itoh. Coupled-mode theory analysis of distributed nonreciprocal structures. *IEEE Trans. Microwave Theory Tech.*, 29:1077–1086, Oct. 1981.

[6] C. A. Balanis. *Advanced Engineering Electromagnetics*. John Wiley & Sons, New York, 1989.

[7] I. Bardi, O. Biro, K. Preis, G. Vrisk, and K. R. Richter. Nodal and edge element analysis of inhomogeneously loaded waveguides. *IEEE Trans. Magnetics*, 29:1466–1469, Mar. 1993.

[8] E. Barrett et al. *TEMPLATES for Solution of Linear Systems: Building Blocks for Iterative Methods*. SIAM, Philadelphia, 1994.

[9] T. Barts et al. Maxwell's grid equations. *Frequenz*, 44:9–16, Jan. 1990.

[10] F. L. Bauer. Das Verfahren der Treppeniteration und verwandte Verfahren zur Losung algebraischer Eigenwertproblem. *Z. Angew. Mat. Phys.*, 8:214–235, 1957.

[11] M. J. Beaubien and A. Wexler. An accurate finite-difference method for higher order waveguide modes. *IEEE Trans. Microwave Theory Tech.*, 16:1007–1017, Dec. 1968.

[12] M. J. Beaubien and A. Wexler. Unequal-arm finite-difference operators in the positive-definite successive overrelaxation (PDSOR) algorithm. *IEEE Trans. Microwave Theory Tech.*, 18:1132–1149, Dec. 1970.

[13] K. Beilenhoff, W. Heinrich, and H. L. Hartnagel. Improved finite-difference formulation in frequency domain for three-dimensional scattering problems. *IEEE Trans. Microwave Theory Tech.*, 40:540–546, Mar. 1992.

[14] K. Bierwirth, N. Schulz, and F. Arndt. Finite-difference analysis of rectangular dielectric waveguide structures. *IEEE Trans. Microwave Theory Tech.*, 34:1104–1113, Nov. 1986.

[15] A. Buda and Z. Sawicki. Single-toroid nonreciprocal latching ferrite phase shifter in modified waveguide. In *Proc. 12-th International Conference on Microwaves & Radar (MIKON-98)*, pages 557–561, Kraków, Poland, May 20-22 1998.

[16] M. Celuch-Marcysiak and W. Gwarek. Higher order modeling of media interfaces for enhanced FDTD analysis of microwave circuits. In *Proc. 24th European Microwave Conference*, pages 1530–1535, Cannes, France, 1994.

[17] W. P. Clark. A technique for improving the figure-of-merit of a twin-slab nonreciprocal ferrite phase shifter. *IEEE Trans. Microwave Theory Tech.*, 16:974–975, Nov. 1968.

[18] D. G. Corr and J. B. Davies. Computer analysis of the fundamental and higher order modes in single and coupled microstrip. *IEEE Trans. Microwave Theory Tech.*, 20:669–678, Oct. 1972.

[19] Z. J. Csendes and P. Silvester. Numerical solution of dielectric loaded waveguides: I–finite-element analysis. *IEEE Trans. Microwave Theory Tech.*, 18:1124–1131, Dec. 1970.

[20] J. Cullum and W. E. Donath. A block Lanczos algorithm for computing the q algebraically largest eigenvalues and a corresponding eigenspace for large, sparse symmetric matrices. In *Proc. IEEE Conference on Decision and Control*, pages 505–509, New York, 1974. IEEE Press.

[21] P. Daly. Hybrid-mode analysis of microstrip by finite-element methods. *IEEE Trans. Microwave Theory Tech.*, 19:19–25, Jan. 1971.

[22] J. Daniel, W. B. Gragg, L. Kaufman, and G. W. Stewart. Reorthogonalization and stable algorithms for updating the Gram-Schmidt QR factorization. *Math. Comp.*, 30:772–795, 1976.

[23] J. B. Davies. Finite element analysis of waveguides and cavities - a review. *IEEE Trans. Magnetics*, 29:1578–1583, Mar. 1993.

[24] J. B. Davies, F. A. Fernandez, and G. Y. Philippou. Finite element analysis of all modes in cavities with circular symmetry. *IEEE Trans. Microwave Theory Tech.*, 30:1975–1980, Nov. 1982.

[25] J. B. Davies and C. A. Muilwyk. Numerical solution of uniform hollow waveguides with boundaries of arbitrary shape. *Proc. IEE*, 113:277–284, Feb. 1966.

[26] M. P. Dębicki, P. Jędrzejewski, J. Mielewski, P. Przybyszewski, and M. Mrozowski. Application of the Arnoldi method to the solution of electromagnetic eigenproblems on the multiprocessor power challenge architecture. Technical Report 19/95, Technical University of Gdańsk, Faculty of Electronics, Gdańsk, Poland, 1995.

[27] M. P. Dębicki, P. Jędrzejewski, J. Mielewski, P. Przybyszewski, and M. Mrozowski. Zastosowanie systemów wieloprocesorowych o architekturze superskalarnej do rozwiązywania macierzowych zagadnień własnych. In *Materiały konf. Zaawansowane Technologie Informatyczne w Nauce Polskiej*, pages 225–234, Kraków, Poland, Oct. 21-23 1995.

[28] M. P. Dębicki, P. Jędrzejewski, J. Mielewski, P. Przybyszewski, and M. Mrozowski. Solution of electromagnetic eigenproblems on multiprocessor superscalar computers. In *Proc. Applied Computational Electromagnetics Symposium (ACES 96)*, pages 214–220, Monterey, CA, Mar. 18-22 1996.

[29] M. P. Dębicki, P. Jędrzejewski, J. Mielewski, P. Przybyszewski, and M. Mrozowski. Zastosowanie metody Arnoldiego do rozwiązywania elektromagnetycznych zagadnień własnych z wykorzystaniem systemów wieloprocesorowych o architekturze superskalarnej. In *Materiały konf. VIII Krajowe Sympozjum Nauk Radiowych (URSI'96)*, pages 103–106, Wrocław, Poland, Feb. 15-16 1996.

[30] F. A. Fernandez, J. B. Davies, S. Zhu, and Y. Lu. Sparse matrix eigenvalue solver for finite element solution of dielectric waveguides. *Electronics Lett.*, 27:1824–1826, 26th Sep. 1991.

[31] F. A. Fernandez and Y. Lu. Variational finite element analysis of dielectric waveguides with no spurious solutions. *Electronics Lett.*, 26:2125–2126, 6th Dec. 1990.

[32] F. A. Fernandez and Y. Lu. A variational finite element formulation for dielectric waveguides in terms of transverse magnetic fields. *IEEE Trans. Magnetics*, 27:3864–3867, Sep. 1991.

[33] F. A. Fernandez and Y. Lu. *Microwave and Optical Waveguide Analysis by the Finite Element Method*. Research Studies Press, Taunton, Somerset, England, 1996.

[34] F. A. Fernandez, Y. Lu, J. B. Davies, and S. Zhu. Finite element analysis of complex modes in inhomogeneous waveguides. *IEEE Trans. Magnetics*, 29:1601–1604, Mar. 1993.

[35] J. G. F. Francis. The QR transformation: A unitary analogue to the LR transformation, Parts I and II. *Comp. J.*, 4:265–272, 332–345, 1961.

[36] A. J. Baden Fuller. *Ferrites at Microwave Frequencies*. Peter Peregrinus Ltd., London, England, 1987.

[37] A. T. Galick, T. Kerkhoven, and U. Ravaioli. Iterative solution of the eigenvalue problem for a dielectric waveguide. *IEEE Trans. Microwave Theory Tech.*, 40:699–705, Apr. 1992.

[38] F. H. Gil and J. P. Martinez. Analysis of dielectric resonators with tuning screw and supporting structure. *IEEE Trans. Microwave Theory Tech.*, 33:1453–1457, Dec. 1985.

[39] G. H. Golub and C. F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, Maryland, 1989.

[40] J. M. Guan and C. C. Su. Resonant frequencies and field distributions for the shielded uniaxially anisotropic dielectric resonator by the FD–SIC method. *IEEE Trans. Microwave Theory Tech.*, 45:1767–1777, Oct. 1997.

[41] A. G. Gurewicz. *Ferryty w zakresie mikrofal*. PWN, Warszawa, Poland, 1964.

[42] R. F. Harrington. *Time-Harmonic Electromagnetic Fields*. McGraw-Hill, New York, 1961.

[43] H. A. Haus and W. Huang. Coupled-mode theory. *Proceedings of the IEEE*, 79:1505–1518, Oct. 1991.

[44] K. Hayata, K. Miura, and M. Koshiba. Finite-element formulation for lossy waveguides. *IEEE Trans. Microwave Theory Tech.*, 36:268–276, Feb. 1988.

[45] D. Ho, F. Chatelin, and M. Bennani. Arnoldi-Tchebychev procedure for large scale nonsymmetric matrices. *Math. Modeling and Num. Analysis*, 24:53–65, 1990.

[46] J. S. Hornsby and A. Gopinath. Numerical analysis of a dielectric-loaded waveguide with a microstrip line – finite-difference methods. *IEEE Trans. Microwave Theory Tech.*, 17:684–690, Sep. 1969.

[47] W. J. Ince, D. H. Temme, and F. G. Willwerth. Toroid corner chamfering as a method of improving the figure of merit of latching ferrite phasers. *IEEE Trans. Microwave Theory Tech.*, 19:563–564, June 1971.

[48] J. Jin. *The Finite Element Method in Electromagnetics*. John Wiley & Sons, New York, 1993.

[49] W. Junding, Y. Z. Xiong, M. J. Shi, G. F. Chen, and M. D. Yu. Analysis of twin ferrite toroidal phase shifter in grooved waveguide. *IEEE Trans. Microwave Theory Tech.*, 42:616–621, Apr. 1994.

[50] N. Kaneda, B. Houshmand, and T. Itoh. FDTD analysis of dielectric resonators with curved surfaces. *IEEE Trans. Microwave Theory Tech.*, 45:1645–1649, Sep. 1997.

[51] A. Kędzior and J. Krupka. Application of the Galerkin method for determination of quasi TE_{i0k} mode frequencies of a rectangular cavity containing a dielectric sample. *IEEE Trans. Microwave Theory Tech.*, 30:196–198, Feb. 1982.

[52] H. Klingbeil, K. Beilenhoff, and H. L. Hartnagel. A local mesh refinement algorithm for the FDFD method using a polygonal grid. *IEEE Microwave Guided Wave Lett.*, 6:52–54, Jan. 1996.

[53] H. Klingbeil, K. Beilenhoff, and H. L. Hartnagel. Finite-difference analysis of structures consisting of roundly and rectangularly shaped domains. *IEEE Microwave Guided Wave Lett.*, 8:295–297, Aug. 1998.

[54] Y. Kobayashi and T. Senju. Resonant modes in shielded uniaxial-anisotropic dielectric rod resonators. *IEEE Trans. Microwave Theory Tech.*, 41:2198–2205, Dec. 1993.

[55] M. Koshiba, K. Hayata, and M. Suzuki. Finite-element formulation in terms of the electric-field vector for electromagnetic waveguide problems. *IEEE Trans. Microwave Theory Tech.*, 33:900–905, Oct. 1985.

[56] M. Koshiba, K. Hayata, and M. Suzuki. Improved finite-element formulation in terms of the magnetic field vector for dielectric waveguides. *IEEE Trans. Microwave Theory Tech.*, 33:227–233, Mar. 1985.

[57] J. Krupka. Optimization of an electrodynamic basis for determination of the resonant frequencies of microwave cavities partially filled with a dielectric. *IEEE Trans. Microwave Theory Tech.*, 31:302–305, Mar. 1983.

[58] J. Krupka. Computations of frequencies and intrinsic q factors of TE_{0nm} modes of dielectric resonators. *IEEE Trans. Microwave Theory Tech.*, 33:274–277, Mar. 1985.

[59] J. Krupka. *Metody analizy i wybrane własności mikrofalowych struktur rezonansowych.* Wydawnictwa Politechniki Warszawskiej, Warszawa, Poland, 1989.

[60] J. Krupka. Resonant modes in shielded cylindrical ferrite and single-crystal dielectric resonators. *IEEE Trans. Microwave Theory Tech.*, 37:691–697, Apr. 1989.

[61] J. Krupka, D. Cros, M. Aubourg, and P. Guillon. Study of whispering gallery modes in anisotropic single-crystal dielectric resonators. *IEEE Trans. Microwave Theory Tech.*, 42:56–61, Jan. 1994.

[62] V. N. Kublanovskaya. On some algorithms for the solution of the complete eigenvalue problem. *USSR Comp. Math. Phys.*, 3:637–657, 1961.

[63] C. Lanczos. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *J. Res. Nat. Bur. Stand.*, 45:255–282, 1950.

[64] J. E. Lebaric and D. Kajfez. Analysis of dielectric resonator cavities using the finite integration technique. *IEEE Trans. Microwave Theory Tech.*, 37:1740–1747, Nov. 1989.

[65] R. B. Lehoucq and J. A. Scott. An evaluation of software for computing eigenvalues of sparse nonsymmetric matrices. Technical Report MCS-P547-1195, Argonne National Laboratory, Jan. 1996.

[66] Y. Lu and F. A. Fernandez. An efficient finite element solution of inhomogeneous anisotropic and lossy dielectric waveguides. *IEEE Trans. Microwave Theory Tech.*, 41:1215–1223, June/Jul. 1993.

[67] Y. Lu and F. A. Fernandez. Finite element analysis of lossy dielectric waveguides. *IEEE Trans. Magnetics*, 29:1609–1612, Mar. 1993.

[68] Y. Lu and F. A. Fernandez. Vector finite element analysis of integrated optical waveguides. *IEEE Trans. Magnetics*, 30:3116–3119, Sep. 1994.

[69] D. R. Lynch and K. D. Paulsen. Origin of vector parasities in numerical Maxwell solutions. *IEEE Trans. Microwave Theory Tech.*, 39:383–394, Mar. 1991.

[70] J. Mazur, J. Mielewski, and J. Popik. Coupled mode analysis of the waveguiding structures containing chiroferrite media. In *Proc. 3-rd International Workshop on Chiral, Bi-isotropic and Bi-anisotropic Media (CHIRAL'94)*, pages 185–190, Périgueux, France, May 18-20 1994.

[71] J. Mielewski and A. Buda. Analysis of the nonreciprocal ferrite phase shifter with nonuniform cross-section. In *Proc. 12-th International Conference on Microwaves & Radar (MIKON-98)*, pages 509–513, Kraków, Poland, May 20-22 1998.

[72] J. Mielewski and M. Mrozowski. Finite-difference analysis of dielectric waveguides using a fast non-symmetric eigensolver. In *Proc. 11-th International Microwave Conference (MIKON-96)*, pages 566–570, Warsaw, Poland, May 27-30 1996.

[73] J. Mielewski and M. Mrozowski. Application of Arnoldi method in FEM analysis of dielectric waveguides. In *Proc. XX-th National Conference Circuit Theory and Electronic Networks (KKTOiUE)*, pages 565–569, Kołobrzeg, Poland, Oct. 21-23 1997.

[74] J. Mielewski and M. Mrozowski. Application of the Arnoldi method in FEM analysis of waveguides. *IEEE Microwave Guided Wave Lett.*, 8:7–9, Jan. 1998.

[75] J. Mielewski, A. Ćwikła, and M. Mrozowski. Accelerated FD analysis of dielectric resonators. *IEEE Microwave Guided Wave Lett.*, 8:375–377, Nov. 1998.

[76] J. Mielewski, A. Ćwikła, and M. Mrozowski. Analysis of shielded anisotropic dielectric resonators using FDFD and the Arnoldi method. In *Proc. 12-th International Conference on Microwaves & Radar (MIKON-98)*, pages 335–339, Kraków, Poland, May 20-22 1998.

[77] A. R. Mitchell and D. F. Griffiths. *The Finite Difference Methods in Partial Differential Equations*. John Wiley & Sons, Chichester, 1980.

[78] A. Mizobuchi and H. Kurebayashi. Nonreciprocal remanence ferrite phase shifters using the grooved waveguide. *IEEE Trans. Microwave Theory Tech.*, 26:1012–1016, Dec. 1978.

[79] M. Mrozowski. *Guided Electromagnetic Waves - Properties and Analysis*. Research Studies Press, Taunton, Somerset, England, 1997.

[80] M. M. Ney. Method of moments as applied to electromagnetic problems. *IEEE Trans. Microwave Theory Tech.*, 33:972–980, Oct. 1985.

[81] V. Nikolski. *Variational Methods for Internal Problems of Electrodynamics*. Science, Moscow, 1967.

[82] T. Okada and D. Kajfez. Fit formulation for cylindrical cavities resulting in a symmetric matrix. In *Proc. URSI Int. Symposium on Electromagnetic Theory*, pages 332–334, Stockholm, Sweden, Aug. 14-17 1989. The Royal Institute of Technology.

[83] B. N. Parlett. The Rayleigh quotient iteration and some generalizations for nonnormal matrices. *Math. Comp.*, 28:679–693, 1974.

[84] S. S. Patrick and K. J. Webb. Behaviour of a magnetic field vector-based finite difference analysis for optical waveguides. *IEEE Trans. Magnetics*, 27:3883–3885, Sep. 1991.

[85] S. V. Polstyanko, R. Diczij-Edlinger, and J. F. Lee. Fast frequency sweep technique for efficient analysis of dielectric waveguides. *IEEE Trans. Microwave Theory Tech.*, 45:1118–1126, July 1997.

[86] P. Przybyszewski, J. Mielewski, and M. Mrozowski. Efficient eigenfunction expansion algorithms for analysis of waveguides. Technical Report 88/96, Technical University of Gdańsk, Faculty of Electronics, Telecomunications and Informatics, Gdańsk, Poland, 1996.

[87] P. Przybyszewski, J. Mielewski, and M. Mrozowski. A fast technique for analysis of waveguides. *IEEE Microwave Guided Wave Lett.*, 8:109–111, Mar. 1998.

[88] P. Przybyszewski, J. Mielewski, and M. Mrozowski. A new class of eigenfunction expansion methods for fast frequency domain analysis of waveguides. *IEEE Trans. Microwave Theory Tech.*, 2000 (in the 2-nd review).

[89] B. M. A. Rahman and J. B. Davies. Finite-element analysis of optical and microwave waveguide problems. *IEEE Trans. Microwave Theory Tech.*, 32:20–28, Jan. 1984.

[90] B. M. A. Rahman and J. B. Davies. Finite-element solution of integrated optical waveguides. *J. Lightwave Techol.*, 2:682–688, Oct. 1984.

[91] B. M. A. Rahman and J. B. Davies. Penalty function improvement of waveguide solution by finite elements. *IEEE Trans. Microwave Theory Tech.*, 32:922–928, Aug. 1984.

[92] B. M. A. Rahman, F. A. Fernandez, and J. B. Davies. Review of finite element methods for microwave and optical waveguides. *Proc. IEEE*, 79:1442–1448, Oct. 1991.

[93] M. Rewieński. *High Performance Algorithms for Large Scale Electromagnetic Modelling*. PhD thesis, Technical University of Gdańsk, Gdańsk, Poland, 1999.

[94] U. Rienen and T. Weiland. Triangular discretization method for the evaluation of rf-fields in cylindrically symmetric cavities. *IEEE Trans. Magnetics*, 21:2317–2320, Nov. 1985.

[95] T. Rozzi, L. Pierantoni, and M. Farina. Eigenvalue approach to the efficient determination of the hybrid and complex spectrum of inhomogeneous, closed waveguide. *IEEE Trans. Microwave Theory Tech.*, 45:345–353, Mar. 1997.

[96] Y. Saad. Variations on Arnoldi's method for computing eigenelements of large unsymmetric matrices. *Lin. Alg. Appl.*, 34:269–295, 1980.

[97] Y. Saad. Chebyshev acceleration techniques for solving nonsymmetric eigenvalue problems. *Math. Comp.*, 42:567–588, 1984.

[98] Y. Saad. *Numerical methods for large eigenvalue problems*. Halsted Press-John Wiley & Sons Inc., New York, 1992.

[99] M. N. O. Sadiku. *Numerical Techniques in Electromagnetics*. CRC Press, Boca Raton, 1992.

[100] R. S. Schechter. *The Variational Method in Engineering*. McGraw-Hill, New York, 1967.

[101] D. Schmitt, R. Schuhmann, and T. Weiland. The complex subspace iteration for the computation of eigenmodes in lossy cavities. *Int. Journal of Numerical Modelling*, 8:385–398, Sep. 1995.

[102] D. Schmitt, B. Steffen, and T. Weiland. 2D and 3D computations of lossy eigenvalue problems. *IEEE Trans. Magnetics*, 30:3578–3581, Sep. 1994.

[103] D. Schmitt and T. Weiland. 2D and 3D computations of eigenvalue problems. *IEEE Trans. Magnetics*, 28:1793–1796, Mar. 1992.

[104] N. Schulz, K. Bierwirth, and F. Arndt. Finite-difference method without spurious solutions for the hybrid-mode analysis of diffused channel waveguides. *IEEE Trans. Microwave Theory Tech.*, 38:722–729, Jun. 1990.

[105] N. Schulz, K. Bierwirth, F. Arndt, and U. Köster. Rigorous finite-difference analysis of coupled channel waveguides with arbitrarily varying index profile. *J. Lightwave Techol.*, 9:1244–1253, Oct. 1991.

[106] E. Schweig and W. B. Bridges. Computer analysis of dielectric waveguides: a finite-difference method. *IEEE Trans. Microwave Theory Tech.*, 32:531–541, May 1984.

[107] P. Silvester. A general high-order finite-element waveguide analysis program. *IEEE Trans. Microwave Theory Tech.*, 17:204–210, Apr. 1969.

[108] D. C. Sorensen. Implicit application of polynomial filters in k-step Arnoldi method. Technical Report TR90-27, Rice University, Department of Mathematical Sciences, Houston, TX, 1990.

[109] D. C. Sorensen. Implicitly restarted Arnoldi/Lanczos methods for large scale eigenvalue calculations. In *Proc. ICASE/LaRC Workshop on Parallel Numerical Algorithms*, Kluwer, Norfolk, Va, May 23-25 1994.

[110] G. W. Stewart. Simultaneous iteration for computing invariant subspaces of non-Hermitian matrices. *Numerische Mathematik*, 25:123–136, 1976.

[111] W. J. Stewart and A. Jennings. A simultaneous iteration algorithm for real matrices. *ACM Transactions on Mathematical Software*, pages 184–198, June 1981.

[112] C. C. Su. An efficient numerical procedure using the shifted power method for analyzing dielectric waveguides without inverting matrices. *IEEE Trans. Microwave Theory Tech.*, 41:539–542, Mar. 1993.

[113] C. C. Su and J. M. Guan. Finite-difference analysis of dielectric-loaded cavities using the simultaneous iteration of the power method with the Chebyshev acceleration technique. *IEEE Trans. Microwave Theory Tech.*, 42:1998–2006, Oct. 1994.

[114] M. M. Taheri and D. Mirshekar-Syahkal. Accurate determination of modes in dielectric-loaded cylindrical cavities using a one-dimensional finite element method. *IEEE Trans. Microwave Theory Tech.*, 37:1536–1541, Oct. 1989.

[115] M. A. Treuhaft and L. M. Silber. Use of microwave ferrite toroids to eliminate external magnets and reduce switching power. *Proc. IRE*, 46:1538, 1958.

[116] G. N. Tsandoulas, D. H. Temme, and F. G. Willwerth. Longitudinal section mode analysis of dielectrically loaded rectangular waveguides with application to phase shifter design. *IEEE Trans. Microwave Theory Tech.*, 18:88–95, Feb. 1970.

[117] L. Valor and J. Zapata. Efficient finite element analysis of waveguides with lossy inhomogeneous anisotropic materials characterized by arbitrary permittivity and permeability tensors. *IEEE Trans. Microwave Theory Tech.*, 43:2452–2459, Oct. 1995.

[118] C. Vasallo. *Théorie des Guides d'Ondes Èlectromagnétiques*, volume 1. Eyrolles, Paris, 1985.

[119] V. Vemuri and W. J. Karplus. *Digital Computer Treatment of Partial Differential Equations*. Prentice-Hall Series in Computational Mathematics. Prentice-Hall, Englewood Cliffs, New Jersey, 1981.

[120] J. S. Wang and N. Ida. Eigenvalue analysis in electromagnetic cavities using divergence free finite elements. *IEEE Trans. Magnetics*, 27:3978–3981, Sep. 1991.

[121] J. P. Webb. Edge elements and what they can do for you. *IEEE Trans. Magnetics*, 29:1460–1465, Mar. 1993.

[122] T. Weiland. Eine Methode zur Lösung der Maxwellschen Gleichungen für sechskomponentige Felder auf diskreter Basis. *AEÜ*, 31:116–120, Mar. 1977.

[123] T. Weiland. Eine numerische Methode zur Lösung des Eigenwellenproblems längshomogener Wellenleiter. *AEÜ*, 31:308–314, Jul./Aug. 1977.

[124] T. Weiland. Verlustbehaftete Wellenleiter mit beliebiger Randkontur und Materialbelegung. *AEÜ*, 33:170–174, Apr. 1979.

[125] T. Weiland. Three dimensional resonator mode computation by finite difference method. *IEEE Trans. Magnetics*, 21:2340–2343, Nov. 1985.

[126] A. Ćwikła, J. Mielewski, M. Mrozowski, and J. Wosik. Accurate full wave analysis of open hemispherical resonators loaded with dielectric layers. In *Proc. IEEE MTT-S International Microwave Symposium*, pages 1265–1268, Anaheim, CA, USA, June 13-19 1999.

[127] K. S. Yee. Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media. *IEEE Trans. Antennas Propag.*, 14:302–307, May 1966.

[128] K. A. Zaki and C. Chen. New results in dielectric-loaded resonators. *IEEE Trans. Microwave Theory Tech.*, 34:815–824, July 1986.