Improving accuracy of the numerical solution of Maxwell's equations by processing edge singularities of the electromagnetic field

Igor Semenikhin
Valiev Institute of Physics and Technology of the Russian Academy of Sciences,
Nakhimovsky prosp. 34, 117218 Moscow, Russia
isemenihin@gmail.com

Abstract. In this paper we present a methodology for increasing the accuracy and accelerating the convergence of numerical methods for solving Maxwell's equations in the frequency domain by taking into account the behavior of the electromagnetic field near the geometric edges of wedge-shaped structures. Several algorithms for incorporating treatment of singularities into methods for solving Maxwell's equations in two-dimensional structures by the examples of the analytical modal method and the spectral element method are discussed. In test calculations, for which we use diffraction gratings, the significant accuracy improvement and convergence acceleration were demonstrated. In the considered cases of spectral methods an enhancement of convergence from algebraic to exponential or close to exponential is observed. Diffraction efficiencies of the gratings, for which the conventional methods fail to converge due to the special values of permittivities, were calculated.

Keywords: Maxwell's equations, edge singularity, convergence acceleration, diffractive optics

1. Introduction

It is well known that the electromagnetic field near geometric edges, both conductors and dielectrics, may possess singularities [1]. This leads to a slow convergence and sometimes even to the inability to obtain a solution in the seemingly simplest cases by conventional methods. For example, not so long ago the authors of [2] found that the calculation of the scattering of a TM polarized electromagnetic field on a lamellar diffraction grating (Fig. 1a) with certain values of the permittivities is not possible even by means of such proven methods as the Fourier Modal Method (FMM, RCWA) [3] and the Analytical Modal Method (AMM) [4,5,6].

Fig. 1. Geometric parameters and permittivities of lamellar a) and triangular b) diffraction gratings. One period is depicted. The medium permittivity is $\varepsilon_0$, the incidence angle of the external electromagnetic field equals $\theta$. Dotted circles with a radius $\rho_0$ indicate the areas where the singularities arise (wedge assemblies); c) a wedge assembly in the general case.

It was shown in [7] that the same problem arises under certain conditions in the case of triangular diffraction gratings and generally in the case of any objects where external electromagnetic radiation is scattered at the intersection of faces with certain values of dielectric permittivity and with a certain value of the refractive index of the external environment. In [7,8] the reason of this was investigated and the conditions under which conventional calculation methods cease to converge were deduced. The singularities primarily occur in the components of the electromagnetic field lying in a plane
perpendicular to the edge of wedge [1]. In particular, for TM polarization the components of the electric field near the edge of diffraction gratings asymptotically change proportionally to $\rho^{\tau-1}$, while the magnetic component parallel to the edge changes as $a + b(\varphi) \cdot \rho^{\tau}$, where $\rho$ is the distance to the edge in the polar coordinate system $(\rho, \varphi)$, the constants $a$, the function $b(\varphi)$ and the value of $\tau$ characterizing the singularity depend on the grating parameters and the refractive index of the external medium. In number of simple cases the value of $\tau$ can be obtained analytically [9,10], in the rest it is necessary to solve the transcendental equation numerically. In general, $\tau$ is a complex value and $\rho^{\tau-1} = \rho^{\text{Re}(\tau)-1} \exp[i \text{Im}(\tau) \ln \rho]$. For $0 < \text{Re}(\tau) < 1$ and a finite imaginary part of $\tau$, the oscillations of electric components with decreasing $\rho$ tend to infinity both in amplitude and frequency. In case of $\text{Re}(\tau) \to 0$ and $\rho \to 0$ simultaneously, the magnetic component also has no limiting value and oscillates with an increasing frequency around a certain constant. The closer the numerical methods converge, moreover, at $\text{Re}(\tau) = +0$ the conventional calculation methods such as FMM cease to work. Exactly such $\tau$ was considered in [2]. Hereinafter, by the term $+0$ we denote an infinitesimal positive value. Some modifications of FMM were proposed in [2,11] to solve the problem in the case of lamellar gratings. However, the authors of [8] expressed doubt about the reliability of the obtained results; besides, that did not solve the problem in the whole. The authors of [8] also noted that they do not know the methods currently in use that could give the correct result for $\text{Re}(\tau) = +0$. Of course, the case under consideration is specific, however, very often even in ordinary calculations it is necessary to obtain greater speed and accuracy than existing methods allow, which as a rule do not take into account edge singularities of the field explicitly.

To date, various approaches have been developed for processing edge singularities, including: Adaptive Spatial Resolution (ASR) in modal methods [12,13], mesh refinement toward the singularities and various adaptive strategies in the Finite Element Method [14,15], decomposition of the solution into a regular part and a singular part as in the Singular Field Method [16] and in the Singular Complement Method [17,18], the Weighted Regularization Method [19] and the method using of Perfectly Matched Layers at corners [20]. The simplest strategies, such as ASR, do not use additional information about the behavior of the field near the singularity. While more advanced ones, such as the Singular Complement Method, use the explicit knowledge of the expression of the singularities near the edge.

It must be said that the behavior of the electromagnetic field near the edge of wedge as in Fig. 1b) and near the edge of wedge assembly in Fig. 1a) and Fig. 1c) was studied quite a long time ago [21,22], since understanding the law of the field behavior, we can take it into account in the numerical calculation. For example, the convergence of spectral methods which are based on the expansion of the solution in smooth functions such as orthogonal polynomials or trigonometric functions, to which FMM and AMM belong, directly depends on the smoothness of the solution [23]. If in the region under consideration the solution and all its derivatives are continuous, a very fast, so-called exponential or spectral convergence is realized, when the higher expansion coefficients decrease exponentially. If the solution has singularities in the expansion domain, then the convergence will be, at best, algebraic, when the coefficients decrease according to a power law. If the solution in the region near the singularity is known, one can use the expansion in smooth functions outside this region, and then join these two solutions at the boundary, thus very fast convergence can be achieved.

In the original work [24] the field near the singularity was obtained in the form of a power series in $\rho$, so-called Meixner series, with coefficients calculated from a recurrent system of differential equations. It was later clarified [25] that in the case of angles that are multiples of $\pi/2$ this series also contains the logarithms of $\rho$, and that the logarithms appear in a more general case when angles are rational numbers multiple of $\pi$ [25,26].

Based on the results obtained earlier by other authors, we somewhat reformulate the method of calculating the field near the singularity for more convenient integration with solution methods used in the rest part of the domain. We implement the transition from the static case, where the problem re-
duces to solving the eigenvalue problem similar to that arising in modal methods, to the case of finite wavelength. As examples of practical application the solution near the edges is added to several well-known methods. In the first case it is the AMM, where a system of linear equations for the coefficients of the transmitted and the reflected modes is modified as a result of taking into account the field singularity. In the second case it is the spectral element method, where the region containing the singularity acts as one of the elements. In addition, an algorithm for constructing the matrix of Dirichlet-to-Neumann map operator (DtN) is presented for the region containing the singularity. This allows adding singularity processing to any multi-domain method for solving Maxwell’s equations in the two-dimensional case when the DtN algorithm can be applied to join computational domains [27]. In particular, we briefly touch on the boundary integral method [28] by which a verification of reference values is carried out for the examples presented in this work. In all cases considered, the singularity handling allows to significantly increase the accuracy and the computation speed, as well as achieve exponential or close to exponential convergence. Additionally, we performed calculations of the electromagnetic field in the cases $\text{Re}(\tau) = +0$ analyzed in [7,8], for which it was not possible to obtain the result by conventional methods and the author knows only one method described in Ref. [20] which have allowed to tackle such problems.

2. Solution near the edge of a wedge assembly in polar coordinates

The general solution of Maxwell’s equations for nonconical diffraction in the two-dimensional (2-D) case can be represented as a linear combination of two fundamental polarizations: the transverse electric polarization (TE) and the transverse magnetic (TM). In this paper, we shall focus on the case of TM polarization as on a more complex one; in the case of TE polarization, everything can be done in a similar way. For TM polarization only the $H_z$ component of magnetic field is not zero and the equation in the frequency domain for this component in polar coordinates is:

$$\rho^{-2} \partial_\rho \left( \rho \partial_\rho \frac{\varepsilon}{\rho} H_z \right) + \rho^{-2} \varepsilon(\varphi) \partial_\varphi \left[ \varepsilon^{-1}(\varphi) \partial_\varphi H_z \right] + \varepsilon(\varphi) \kappa_0^2 H_z = 0, \quad H_\rho = H_\varphi = 0, \quad (\text{TM})$$

where $\varepsilon(\varphi)$ is the relative dielectric constant, $k_0 = \frac{2\pi}{\lambda}$, $\lambda$ denotes the wavelength of the incident radiation. Hereafter, for simplicity we set the relative magnetic permeability to unity. We shall solve Eq. (1) in a circle $\rho \leq \rho_0$, see Fig. 1). For convenience we introduce a new variable: $r = \rho / \rho_0$ and equation (1) takes the form:

$$r \partial_r \left( r \partial_r H_z \right) + \varepsilon(\varphi) \partial_\varphi \left[ \varepsilon^{-1}(\varphi) \partial_\varphi H_z \right] + \varepsilon(\varphi) \gamma^2 r^2 H_z = 0,$$

where $\gamma = 2\pi (\rho_0 / \lambda)$ is the dimensionless constant much less than unity.

We shall seek the field $H_z$ in the form of a series:

$$H_z(r, \varphi) = \sum_{j=0}^{\infty} h_j G_j(r, \varphi),$$

where the set of functions $\{G_j(r, \varphi)\}$ is the solutions of Eq. (2) for the boundary conditions that form the complete set of linearly independent functions (basis) $\{G_j(r = 1, \varphi)\}$ on the boundary $r = 1^1$. Then the coefficients $h_j$ can be found from the boundary conditions $b(\varphi) \equiv H_z(r = 1, \varphi) = 0$. We shall use as such a basis a linear combination of angular eigenfunctions, which is obtained naturally when considering a static case.

When passing to the static case ($\lambda \to \infty$, $\gamma \to 0$), Eq. (2) for each function from $\{G_j(r, \varphi)\}$ turns into:

$^1$ Note that the system of functions $\{G_j(r, \varphi)\}$ itself is not a basis in a circle $r \leq 1$, since it can be used to represent only solutions of equation (2).
\[ \rho \partial_r \left( \rho \partial_r G_j \right) = -\varepsilon(\phi) \partial_\phi \left[ \varepsilon^{-1}(\phi) \partial_\phi G_j \right]. \]  

(4)

It is solved by the separation of variables:

\[ G_j(r, \phi) = \Phi_j(\phi) R_j(r), \]

(5)

where \( \Phi_j(\phi) \) are the eigenfunctions of the one-dimensional Helmholtz equation:

\[ -\varepsilon(\phi) \partial_\phi \left[ \varepsilon^{-1}(\phi) \partial_\phi \Phi_j(\phi) \right] = \tau_j^2 \Phi_j(\phi), \]  

(6)

corresponding to the eigenvalues \( \tau_j^2 \). For definiteness, we shall chose the sign of \( \tau_j \) so that \( \text{Re}(\tau_j) \geq 0 \).

Equation (6) is completely analogous to the eigenvalue problem that arises in modal methods, for example, in case of a lamellar grating with replacing the \( x \) coordinate by the azimuth angle \( \phi \) and the wavelength \( \lambda \to \infty \) (\( k_0 = 0 \)), see the next paragraph. The solutions of Eq. (6) in some simple cases can be found analytically (see Appendix 1). The main properties of the eigenfunctions \( \Phi_j(\phi) \), including their linear independence and completeness, which is important for representing through them an arbitrary condition on the boundary \( r = 1 \), as well as the orthogonality with weight \( \varepsilon^{-1}(\phi) \) to the adjoint eigenfunctions \( \Phi_j^+(\phi) \) corresponding to the differential operator adjoint to the operator in Eq. (6) are given in [6], and the general theory of such non-self-adjoint boundary eigenvalue problems is presented, for example, in [29].

Substituting Eq. (5) into Eq. (4) and taking into account Eq. (6), we obtain the equation for \( R_j^0(r) \):

\[ \left[ \rho \partial_r \left( \rho \partial_r \right) - \tau_j^2 \right] R_j^0(r) = 0, \]

(7)

with solution:

\[ R_j^0(r) = a_j r^{\tau_j} + b_j r^{-\tau_j}. \]

(8)

For \( \text{Re}(\tau_j) > 0 \) the term with a negative sign in exponent is rejected due to physical reason (the requirement of a finite energy in a finite volume)\(^1\). We set the coefficients \( a_j = 1 \) so that on the boundary \( r = 1 \) the functions \( R_j^0(r) \) are equal to unity.

In case of \( \gamma > 0 \) we shall seek the functions \( G_j(r, \phi) \) in the form:

\[ G_j(r, \phi) = \sum_{k=0}^{\infty} \Phi_k(\phi) R_{k,j}(r), \]

(9)

with the condition that upon passing to the static case \( \gamma = 0 \), there must be

\[ \lim_{\gamma \to 0} R_{k,j}(r) = \delta_{k,j} R_j^0(r), \]

(10)

where \( \delta_{k,j} \) is the Kronecker delta. Further, we find the functions \( R_{k,j}(r) \) in the form of an absolutely convergent series that continuously depends on \( \gamma \), when the wedge angles are the product of \( \pi \) and rational numbers (see Appendix 2.2, the case of irrational multipliers is also briefly discussed there); therefore, for sufficiently small \( \gamma \) the linear independence of \( G_j(r = 1, \phi) \) Eq. (9) will be preserved. To get an equation for the functions \( R_{k,j}(r) \), we substitute Eq. (9) into Eq. (2), taking into account Eq. (6):

\[ \sum_{k=0}^{\infty} \left[ \rho \partial_r \left( \rho \partial_r \right) - \tau_k^2 + r^2 \gamma^2 \varepsilon(\phi) \right] \Phi_k(\phi) R_{k,j}(r) = 0. \]

(11)

\(^1\) This requirement can be written as: \( \int_V \varepsilon |E|^2 + \mu |H|^2 \, dv < \infty \), where \( V \) is a finite cylindrical volume containing the edge, and the components of the electric field are \( E_\rho \sim \rho^{-2} \partial_\phi H_\phi \), \( E_\phi \sim \partial_\phi H_\rho \).
Left multiplying Eq. (11) by the complex conjugate adjoint eigenfunctions \( \Phi_j^*(\varphi) \) and integrating over \( \varphi \) on the interval \([0, 2\pi]\) with the weight function \( \varepsilon^{-1}(\varphi) \), we obtain
\[
\left[r\partial_r (r\partial_r) - \tau_k^2\right]R_{k,j}(r) + r^2\gamma^2\sum_{l=0}^{\infty} c_{k,l}R_{l,j}(r) = 0, \tag{12}
\]
where matrix elements \( c_{k,l} \) are
\[
c_{k,l} = \int_0^{2\pi} \Phi_j^*(\varphi)\Phi_l(\varphi) d\varphi, \tag{13}
\]
and we take into account the orthogonality of \( \Phi_j^*(\varphi) \) and \( \Phi_l(\varphi) \) with weight \( \varepsilon^{-1}(\varphi) \) [5, 6].

The functions \( R_{k,j}(r) \) can be found by the iteration method, taking the static case \( R_0^j(r) \) as the zeroth-order approximation (see Appendix 2.1). The solution is obtained as a series in even powers of \( \gamma \):
\[
R_{k,j}(r) = R_0^j(r)\left[\delta_{k,j} + \gamma^2 a_{k,j,1}(r)\cdot r^2 + \gamma^4 a_{k,j,2}(r)\cdot r^4 + \ldots\right], \tag{14}
\]
where \( a_{k,j,i}(r) \) are either constants independent of \( r \) or if \( \tau_k = \tau_j + 2l \) holds, where \( l \) is a positive integer, then \( a_{k,j,i}(r), i \geq 1 \) consist of the sum of the constants and the logarithms to an integer power less than the power of \( r \). The second option is realized when the angles of the wedges are the product of \( \pi \) and rational numbers. For small \( r \), as well as for small \( \gamma \) (large \( \lambda \)), \( R_{k,j}(r) \) tends to the static case (10). In Appendix 2.2 we show that the series (14) absolutely converges for any finite values of \( \gamma \) when the wedge angles are the product of \( \pi \) and rational numbers, and we briefly discuss the case of angles that are multiples of an irrational factor of \( \pi \); there are also formulas for practical calculations of the coefficients of the series, see Eq.(110)-(112). Since we know the behavior of the functions \( R_{k,j}(r) \) for \( r \to 0 \), this functions can also be found by a direct solution of the Cauchy problem for a system of ordinary differential equations (Appendix 2.3).

Summarizing all the above, the solution near the edge of a wedge assembly inside the circle \( \rho_0 \) can be written as follows:
\[
H_z(\rho, \varphi) = \sum_{j=0}^{\infty} h_j \hat{G}_j(\rho/\rho_0, \varphi) = \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \Phi_k(\varphi)R_{k,j}(\rho/\rho_0) =
= \sum_{j,k=0}^{\infty} h_j \Phi_k(\varphi)\cdot (\rho/\rho_0)^{\tau_j}\left[\delta_{k,j} + a_{k,j,1}(\rho/\rho_0)\cdot (k_0\rho)^2 + a_{k,j,2}(\rho/\rho_0)\cdot (k_0\rho)^4 + \ldots\right], \tag{15}
\]
where \( \tau_j \) are the square roots of the eigenvalues such that \( \text{Re}(\tau_j) \geq 0 \), and \( \Phi(\varphi) \) are corresponding eigenfunctions of the problem (6) arising in a static case (\( \lambda = \infty \)), \( R_{k,j}(\rho/\rho_0) \) are the solutions of the system of ordinary differential equations (12), and the coefficients \( h_j \) are found from the boundary conditions at \( \rho = \rho_0 \).

3. Incorporating the solution near the edge of a wedge assembly into the Analytical Modal Method

In this section we shall apply the above-described technique for calculating the electromagnetic field near edges to accelerate the convergence of the Analytical Modal Method which is the base representative of the class of modal methods. Within these methods the simulation domain, for example, a diffraction grating, is divided into parallel layers, in each of which the permittivity changes only in the plane of the layer and is assumed to be constant along the perpendicular direction allowing the separation of spatial variables, see Fig.1a). Within each layer the eigenmodes of the electromagnetic field are calculated and the general solution is then expressed by means of an eigenmode expansion. The expansion coefficients can be found by applying the proper boundary conditions. In the case of TM polariza-
tion the magnetic field within each $l$-th layer ($l=0..L$, where $l=0$ corresponds to the superstrate layer and $l=L$ to the substrate) with relative permittivity $\varepsilon(l)(x)$ satisfies the following equation:

$$\partial^2_{xx}H_z = -\varepsilon(l)(x)\partial_x \left[ \varepsilon^{-1}(x)\partial_x H_z \right] - k^2 H_z, \quad H_z = H_y = 0. \quad (16)$$

Along $x$ axis the quasiperiodic Floquet boundary conditions are applied:

$$H_z(x+p, y) = \exp(i k_0 n_0 \sin \theta) H_z(x, y), \quad (17)$$

where $p$ is the period of grating, $n_0$ is the refractive index of the superstrate and $\theta$ is the incidence angle of external radiation in the form of a plane wave (Fig. 1):

$$H_z^{\text{inc}} = \exp[i k_0 n_0 (x \sin \theta + y \cos \theta)]. \quad (18)$$

After separating the variables in Eq. (16), the solution within each $l$-th layer can be written as [4,6]:

$$H_z(x,y) = \sum_n \left[ c_{l,n} \exp(ik_0 \kappa_{l,n}y) + c_{l,-n} \exp(-ik_0 \kappa_{l,n}y) \right] \psi_{l,n}(x), \quad (19)$$

where $c_{l,n}$ and $c_{l,-n}$ are the amplitudes of forward and backward propagating modes, $\kappa_{l,n}$ and $\psi_{l,n}(x)$ are the square roots of eigenvalues and the eigenfunctions of the one-dimensional Helmholtz equation with the quasiperiodic Floquet boundary conditions (17):

$$\varepsilon(l) \partial_x^2 + \kappa^2 = 0. \quad (20)$$

The sign of square root in (19) is chosen so that its imaginary part would be positive, whereas for real positive $\kappa^2_{l,n}$ the positive square root should be used. Thus, the directions of propagation and decaying of modes coincide. Various methods for solving Eq. (20) can be found in [4,6,30,31] and the same methods are applicable for solving Eq. (6).

In the edges regions bounded by circles of radius $\rho_0$ and marked in Fig. 1a) by a dashed line the solution is written in the form (15). In numerical calculations, we restrict the sum (15) over $j$ to the first $N_w$ terms. The number of terms in the sum over $k$ is chosen to be larger than $N_w$, so that with their further increase the final calculation results do not change within the required accuracy. For simplicity the numbers $N_w$ will be taken the same for all wedge assemblies $w$.

The equations for the coefficients $c$ in Eq. (19) for each layer and the coefficients $h$ in Eq. (15) for each region near the edge are obtained by equating the values of the field and its normal derivative multiplied by $\varepsilon^{-1}$ on both sides of the interface $\Gamma$ indicated in Fig. 1a) by the dashed line:

$$H_z^- = H_z^+, \quad \varepsilon^{-1} \partial_n H_z^- = -\varepsilon^{-1} \partial_n H_z^+, \quad (21)$$

Fig. 2. a) Integration of conditions on the boundary (21) with test functions; b) the choice of radius $\rho_0$ for a given number of functions $N_0$ used to represent the solution near the edge of a wedge assembly.
with a subsequent integration of these equations over a system of test functions. At the boundary of the layers on the segments \([x_{j-1}, x_j]\) between the edges Fig. 2a), indicated by the blue dashed line, we integrate Eq. (21) over the variable \(x\) with the test functions for which we use the complex conjugate eigenfunctions of the superstrate \(\psi_{0,m}^*\) (Fourier matching scheme). At the boundary of the region around the edge, indicated in Fig. 2a) by the red dash-dotted line, we integrate Eq. (21) along the circle \(\rho_0\) with the test functions \(\Phi^*(\varphi)\) and the weight \(e^{-1}(\varphi)\). More details on the preparation of equations for the coefficients can be found in Appendix 3.

The radius \(\rho_0\) of the circle around the edge of a wedge assembly is chosen so that the ratio of the number of functions (9) to the length of the projection of the circle boundary onto the \(x\) axis, see Fig. 2b), was approximately equal to the ratio of the number of modes to the period. The radius in this case is \(\rho_0 = pN_w/(4M)\). At lower values of \(\rho_0\), the convergence slows down; at large values, in addition, the condition number of the matrix of the resulting system of linear equations grows rapidly. The condition number increases for the same reason as in the Rayleigh method [32], due to the fact that exponentially decaying modes in Eq. (19), when integrated over the radius of the circle, give a very small contribution, which, for a finite length of the mantissa of numbers, is practically not distinguishable for different modes against the background of the remaining coefficients of the linear system.

| № | Relative permittivities | Geometric dimensions in units of wavelength \(\lambda\) | \(\theta\) | Tested characteristic and its reference value | References |
|---|---|---|---|---|---|
| 1 | \(\varepsilon_0 = \varepsilon_1 = 1\), \(\varepsilon_i = (0.22 + 6.7i)^2\) | \(p = h = 1, d = p/2\) | 30° | \(R_0 = 0.8484816789046437\) | 31, 33, 34, 35 |
| 2 | \(\varepsilon_0 = \varepsilon_1 = 1, \varepsilon_i = 2.3^2, \varepsilon_i = 1.5^2\) | \(p = 2, h = 1, d = 0.468\) | 30° | \(T_1 = 0.5105923632002064\) | 31, 34, 35, 36 |
| 3 | \(\varepsilon_0 = 1.5^2, \varepsilon_i = -2.5 + 0i\), \(\varepsilon_i = 1.3^2\) | \(p = 10/31, h = 14/31, d = p/2\) | 0° | \(R = 0.5942180852105747\) | 2.8, 11 |

Table 1. Parameters of the lamellar diffraction gratings in Fig. 1a) used as examples. The TM-polarized electromagnetic wave is incident at an angle \(\theta\).

As practical examples we calculated the diffraction efficiencies of several lamellar gratings in Fig. 1a). Table 1 shows the parameters of these gratings and references to some works in which they were previously used for such tests. In this paper we do not touch on hypersingularity [37] when \(\text{Re}(\tau_1) = 0\) therefore, in instance, for the dielectric constant of grating No. 3 in table 1 we write \(\varepsilon_i = -2.5 + 0i\), which leads to the corresponding choice of the sign of the complex number \(\tau_1\), such that \(\text{Re}(\tau_1) = +0\).

Fig. 3 presents a comparison of the convergence of the modal methods often used in the calculation of such structures in the case of a metal diffraction grating No. 1 of Table 1. Relative error of the reflected zero-order diffraction efficiency \(R_0\) calculated by the Fourier modal method (FMM), by the parametric Fourier modal method (PFMM) [12], and by the Analytical Modal Method (AMM), both with and without \((N_w = 0)\) implementing expansion (15) near the edges, are shown. The value of the adaptive spatial resolution parameter \(\eta = 0.995\) in the PFMM method is selected from the condition of best convergence. The value \(R_0 = 0.8484816789046437\), calculated using multiple-precision arithmetic and rounded to 16 significant digits, was used as a reference. The reference values used in the work were verified by several methods, including the boundary integral method with incorporated edge singularity processing; for more detail, see the Section 4 of the article.

From Fig. 3 it can be seen that the FMM, PFMM methods and the regular version of AMM \((N_w = 0)\) significantly lose in convergence rate to the modified AMM method, which explicitly takes into account edge singularities. In addition, in the case of a proportional increasing the number of functions \(N_w\) in expansion (15)
simultaneously with the number of modes $M$ the modified AMM method shows exponential or close to exponential convergence (Fig. 3a), while the rest of the presented spectral methods are limited by algebraic convergence due to the presence of edge singularities.

Fig. 3. a) Relative error of the reflected zero-order diffraction efficiency $R_0$ of the lamellar diffraction grating No. 1 of Table 1 versus the number of modes $M$ calculated by various modal methods; b) the results of calculation using the modified AMM method with a fixed number of functions $N_w$ used for representing the solution near the edges.

In the case of a fixed number of functions $N_w$ in the expansion (15) of the solution near the edges the convergence of the method is algebraic, Fig. 3b). Besides, with increasing $N_w$ the rate of convergence grows significantly. It can be noted that if earlier, in order to achieve relative accuracy $10^{-12}$ by the usual AMM method $M=3.2 \times 10^5$ modes were required in this example [31], now, to obtain the same accuracy using the modified AMM method, only a few hundred modes are enough.

Fig. 4 shows the calculation of the first-order transmitted diffraction efficiency $T_1$ of a dielectric lamellar grating No. 2 of Table 1. As in the case of the metal diffraction grating the processing of edge singularities significantly accelerates convergence. And with a simultaneous proportional increase in the number of functions $N_w$ together with the number of modes $M$, an exponential or close to exponential convergence is also observed in Fig. 4a).

Fig. 4. a) Relative error of the first order transmitted diffraction efficiency $T_1$ of the dielectric lamellar diffraction grating No. 2 of Table 1 versus the number of modes $M$ calculated by various modal methods; b) the results of calculation using the modified AMM method with a fixed number of functions $N_w$ used for representing the solution near the edges.
In the following example, Fig. 5, the reflectivity of a lamellar diffraction grating No 3 of Table 1, featured the real part $\tau_i$ of all four edge singularities equal to $+0$, is calculated. It was previously studied in the papers [2,8,11], which we talked about in the introduction. Conventional methods such as FMM, PFMM, and AMM do not work in this case, and only the results of the modified AMM method are shown in the figures. If the singularity is explicitly taken into account, the value $\text{Re}(\tau_i) = +0$ no longer presents a problem, and the convergence of calculating the reflectance of the grating is as fast as in the previous examples.

![Fig. 5](image)

**Fig. 5.** a) Relative error of the reflectivity of the lamellar diffraction grating No. 3 of Table 1, where all edge singularities have $\text{Re}(\tau_i) = +0$, versus the number of modes $M$ calculated by the modified AMM; b) the same calculation but with a fixed number of functions $N_w$ used for representing the solution near the edges.

It should also be noted that the accuracy curves, Fig. 5b), calculated with a fixed number of functions $N_w$, tend practically to a constant with an increase in the number of modes $M$. This expresses the fact that in the case of $\text{Re}(\tau_i) = +0$ the achievable accuracy is limited not only by the number of used modes $M$, but also by the number of functions $N_w$ in the regions containing the edge.

![Fig. 6](image)

**Fig. 6.** a) Reflectivity (solid blue line), extrapolation by Thiele formula of reflectivity using four points (dashed red line) and the $\text{Re}(\tau_i)$ of the singularities at the upper edges (thin dashed line) and at the lower edges (thin dash-dotted line), see Fig 1a), of the diffraction grating No. 3 of table 1, depending on the imaginary part of the dielectric constant $\epsilon_i$, while the real part $\epsilon_r$ remains the same; b) the relative error in the calculation of reflectivity by the PFMM method for the four values of the imaginary part of the dielectric constant $\epsilon_i$, which were used in the extrapolation procedure at the panel a).
Fig. 6a) shows how the reflectivity and the real part of $\tau_{\text{l}}$ vary with a decrease in the imaginary part of the dielectric constant $\epsilon_{\text{i}}$ of the central part of the diffraction grating No. 3 of Table 1 from 2.5 to zero while the real part $\epsilon_{\text{r}}$ remains the same. The curves of reflectivity and of $\text{Re}(\tau_{\text{l}})$ are quite smooth, and their derivatives remain finite at $\text{Re}(\tau_{\text{l}}) = 0$.

From fig. 6b) a reducing convergence rate of the conventional PFMM method is seen, with a decrease of the imaginary part of $\epsilon_{\text{i}}$ and a corresponding decrease of $\text{Re}(\tau_{\text{l}})$. Due to the presence of the edge singularities, the convergence of the PFMM is algebraic and the slope of the convergence line is relaxed in logarithmic coordinates, while the number of modes required to obtain the result with the same accuracy increases exponentially. For $\text{Re}(\tau_{\text{l}}) = 0$ the PFMM method ceases to work, but using extrapolation we can approximately estimate the value of the reflectivity at the point $\text{Im}(\epsilon_{\text{i}}) = 0$. Thiele extrapolation, using the four points marked in Fig. 6a) in circles, gives a rounded to two significant digits value of $R = 0.58$, which differs from the exact one by 2.4%. It is clear that such an extrapolation can only be carried out \textit{a posteriori}, knowing in advance about the sufficient smoothness of the curve in the direction of the extrapolated point. If the location of points is not close enough to $\text{Im}(\epsilon_{\text{i}}) = 0$, then, as will be shown in the last example of the next section, such an extrapolation procedure can give a physically incorrect result.

Fig. 7. The modified AMM method with the ratio $\frac{N_{\text{w}}}{M} = \frac{M}{4}$ between the number of functions $N_{\text{w}}$ and the number of modes $M$ is compared in terms of time consumption with the PFMM method with $\eta = 0.99$ in the case of diffraction grating No. 3 of Table 1, depending on the required calculation accuracy and the value of $\text{Im}(\epsilon_{\text{i}})$. In the area highlighted in blue it is more advantageous to use the PFMM method, in the green area the modified AMM method is preferable.

Until now, we have compared the convergence of the methods depending on the number of modes $M$. However, in real applications the time of calculation of the result with the required accuracy is primarily important. The calculation of the field near the corners in the form (15) takes a certain amount of time and, on the whole, the modified AMM method, which explicitly takes into account the edges singularities, works slower than the usual AMM method with the same $M$. The number of operations when calculating by formula (15) is proportional to $N_{w}^4$ and in the case when $N_{w} \sim M$ the total number of operations in the modified AMM method grows as $M^4$. While in most modal methods the number of operations is proportional to $M^3$, moreover, in the iterative implementation of the AMM method it is proportional to $M^2$ [31]. However, the convergence of the conventional methods is only algebraic, that is, the calculation errors is proportional to $M^{-q}$ with a certain constant $q$. At the
same time, the convergence of the modified AMM method is exponential or close to exponential, and the calculation error is proportional to $e^{-cM}$ with some constant $c$. Therefore, with an increase in the result accuracy requirements, the modified method will ultimately win in terms of time consumption. Also, with decreasing of the values of $\text{Re}(\tau_1)$, conventional methods need an exponential increase in the number of modes to achieve the required accuracy, whereas the convergence of the modified method practically does not change. Thus, conventional methods, which do not take into account the edges singularities explicitly, can gain in computation time only if the required accuracy is moderate and the value of $\text{Re}(\tau_1)$ is not too small. As an example, we provide a diagram approximately showing when one method or another is preferable in the case of diffraction grating No. 3 of Table 1, depending on the required calculation accuracy and the value of $\text{Im}(\varepsilon)$. The modified AMM method is compared with the PFMM method in Fig. 7). This test was carried out using only one processor core without any parallelization.

4. Incorporating the solution near the edges into the Spectral Element Method

In this section, we shall incorporate the solution near the edges into the Spectral Element Method (SEM) [38], which is a universal method and which is very close to the finite element method. In this method the computational domain is divided into subdomains, in each of which the permittivity remains uniform or vary smoothly. The solution in the subdomains is represented in the form of an expansion over known basis functions, for which the orthogonal polynomials are usually used. In a subdomain containing an edge we shall represent the solution in the form of the series (15). The equations for the expansion coefficients in SEM are wrote in such a way that within each subdomain the solution satisfies Maxwell's equations, and at the boundaries with other subdomains it satisfies the corresponding conditions for the electromagnetic field. The coefficients can be calculated all at once by solving one large system of linear equations, which is usually done by some iterative method or, by implementing the Dirichlet-to-Neumann mapping (DtN) technique, where firstly the boundary conditions for the subdomains are found, and after that the Dirichlet problem is solved separately for each subdomain [27,39,40].

Speaking about the spectral element method, we adhere to the terminology of the book [23], referring to methods that use the spectral approach inside the subdomains, not only limiting by the Galerkin method in a weak form, as in the original proposal [38]. A formalism based on the application of the Dirichlet-to-Neumann (DtN) map, which we shall use, was previously discussed by us in [41], here we only briefly outline the main points.

In the case of TM polarization the magnetic field in each subdomain with uniform permittivity $\varepsilon$ satisfies the following equation:

$$
\nabla_z^2 H_z + \nabla_z^2 H_z + \varepsilon k_0^2 H_z = 0, \quad H_z = H_y = 0.
$$

(22)

It can be solved both by the Galerkin method and by the pseudo-spectral method [23,39,40]. If a subdomain has not the rectangular form, then mapping into a rectangular area is applied by introducing new coordinates. For joining solutions in neighboring regions, it is convenient to use the Dirichlet-to-Neumann map [27,39,40]. To obtain the DtN map operator matrix for a subdomain, on the boundary $\Gamma$ of this subdomain a set of $N$ linearly independent boundary conditions at $N$ points of the Gauss or Gauss-Lobatto quadrature is defined and, by solving the Dirichlet problem for each of them, the corresponding normal derivatives at these points of the boundary are calculated. After that we compose two square matrices $N \times N$: in the first matrix $F$ there will be $N$ boundary conditions as columns, and in the second matrix $\nabla_n F$ there will be columns of the corresponding normal derivatives at the boundary. Since the boundary condition vectors are chosen linearly independent, an arbitrary boundary condition $f_\Gamma$ defined at these points can be represented as their linear combination with some coefficients $c$, and the corresponding normal derivative $\nabla_n f_\Gamma$ can be easily calculated:

$$
f_\Gamma = F \cdot c \Rightarrow \nabla_n F \cdot c = \nabla_n f_\Gamma \Rightarrow \nabla_n f_\Gamma = (\nabla_n F \cdot F^{-1}) f_\Gamma = Df_\Gamma,
$$

(23)
where $D$ is the DtN map operator matrix. Boundary conditions, as well as normal derivatives, can be specified not only in the form of values at $N$ points, but also in the form of $N$ expansion coefficients for some fixed set of basis functions at the boundary, which is convenient if we are calculating a solution inside the subdomain using the Galerkin method. In this case, we shall say that DtN map acts in the space of these basis functions. The transition from one representation to another, as well as the calculation of the DtN map matrix for the subdomain containing the edge in various representations is given in Appendix 4.

Knowing the DtN map operator matrices $D_1$ and $D_2$ for two subdomains with a shared border $\Gamma'$:

$$\begin{align*}
(D_1)_{1,1} &+ (D_1)_{1,2} (D_1')_{2,1} & (D_1)_{1,2} (D_1')_{2,1} & f_{1,1} \\
(D_2)_{1,1} &+ (D_2)_{1,2} (D_2')_{2,1} & (D_2)_{1,2} (D_2')_{2,1} & f_{2,1}
\end{align*}$$

and taking into account boundary conditions on $\Gamma'$, which for TM polarization are:

$$f_{1,1} = (D_1')_{2,1} f_{1,2} + (D_1')_{2,1} f_{2,2},$$

where $\beta = \varepsilon_1 / \varepsilon_2$ is the ratio of the permittivities of the materials at points on both sides of the shared boundary $\Gamma'$, the DtN map operator matrix of these two elements combining can be straightforwardly calculated:

$$\begin{align*}
(D_1)_{2,1} + (D_1)_{2,2} (D_1')_{2,1} & (D_1)_{2,2} (D_1')_{2,1} & f_{1,1} \\
(D_2)_{2,1} + (D_2)_{2,2} (D_2')_{2,1} & (D_2)_{2,2} (D_2')_{2,1} & f_{2,1}
\end{align*}$$

If we want to connect two adjacent boundaries $\Gamma_2$ and $\Gamma_3$ of the region characterized by a matrix of DtN map operator:

$$\begin{align*}
(D_1)_{1,1} &+ (D_1)_{1,2} (D_1')_{2,1} & (D_1)_{1,2} (D_1')_{2,1} & f_{1,1} \\
(D_2)_{1,1} &+ (D_2)_{1,2} (D_2')_{2,1} & (D_2)_{1,2} (D_2')_{2,1} & f_{2,1}
\end{align*}$$

with conditions on the boundaries:

$$f_{1,2} = a f_{1,1}, \quad \partial_n f_{1,1} = -\beta \partial_n f_{1,2},$$

where $a, \beta$ are diagonal matrices as, for example, in the case of quasiperiodic boundary conditions (17), then, substituting Eq. (28) into Eq. (27), we obtain the DtN map for the remaining part of the boundary $\Gamma_1$:

$$\begin{align*}
(D_1)_{1,1} + (D_1)_{1,2} a (D_1')_{2,1} + (D_1)_{1,2} (D_1')_{2,1} & f_{1,1} \\
(D_2)_{1,1} + (D_2)_{1,2} a (D_2')_{2,1} + (D_2)_{1,2} (D_2')_{2,1} & f_{2,1}
\end{align*}$$

And vice versa, after the conditions on the boundary $\Gamma_1$ are found, the conditions on the boundaries $\Gamma_2, \Gamma_3$ will be:

$$f_{1,2} = D' f_{1,1}, \quad f_{1,2} = a f_{1,2}.$$
\( c_i^+ \) of the transmitted wave in substrate can be found using expansion (19) of the field in the superstrate and in the substrate, and the definition of the DtN map at the upper and lower boundaries of the domain:

\[
\begin{pmatrix}
D_{1,1} & D_{1,2} \\
D_{2,1} & D_{2,2}
\end{pmatrix}
\begin{pmatrix}
\Psi_1 \\
0
\end{pmatrix}
\begin{pmatrix}
c_i^+ + c_1^+ \\
c_2^+ + c_2^+
\end{pmatrix}
= ik_0
\begin{pmatrix}
\Psi_1 \kappa_1 & 0 \\
0 & -\Psi_2 \kappa_2
\end{pmatrix}
\begin{pmatrix}
-c_i^+ + c_1^- \\
-c_2^- + c_2^-
\end{pmatrix},
\tag{31}
\]

where the square matrices \( \Psi_i \) consist of the columns \( \psi_{i,j}(x_i) \) at the grid nodes.

The coefficients \( c_{1,j}^+ = \delta_{j,0} \), and \( c_2^+ \equiv 0 \) are known and specify the zero order incident wave in the superstrate and the upward going wave in the substrate, respectively. Instead of equating the field at grid points, the Galerkin method can be implemented by equating field integrals with test functions on both sides of the boundary, for more details see [41]. After the magnetic field at the boundary of the entire domain is obtained, we can, using Eq. (26) and Eq. (30), and acting in the opposite order, find the boundary conditions for each of the subdomains. Then, the field inside the subdomains can be calculated by solving the Dirichlet problem. In particular, the calculation of the coefficients \( h_i \) in the expansion (15) of the magnetic field in the region near the edge is given in Appendix 4. It may be noticed that the DtN mapping approach is very similar to the S-matrix technique used in modal methods [42], with the difference that instead of reflection and transition coefficients, the field values at the subdomains boundaries are searched.

It should be noted that in the considered approach the DtN map matrices of individual subdomains can be calculated by using an arbitrary method, not necessarily by the spectral method. In the case of a uniform dielectric constant inside each element, it is very convenient to calculate the DtN map matrix by using the boundary integral method [28,36]. Since in this method the calculated subdomain does not need to be mapped into a rectangle, we get considerable freedom in choosing the partition into subdomains. In particular, we can represent regions of uniform permittivity as these subdomains. The subdomains containing the edges, as before, are calculated separately, and the whole technique of combining the subdomains remains the same. The boundary integral method scales much better with the number of collocation points increasing, and we employed its multiple-precision arithmetic implementation with incorporated edge singularity processing to verify the reference values used in the article.

As practical calculations shown, the value of the radius \( \rho_0 \) of the region containing the edge, where the solution is represented in the form of series (15), can be chosen in a first approximation so that the ratio of the number of functions \( N_w \) to the length of the boundary of the region is equal to the average density of grid points per period, in accordance with the expression: \( N_w / 2\pi \rho_0 = N_x / p \), where \( N_x \) is the number of grid points along the \( x \) axis over the grating period \( p \). If the high accuracy result is required, then for faster convergence it is convenient to use the formula:

\[
\rho_0 / p = c_1 N_w / N_x + c_2 (N_w / N_x)^2 + \ldots,
\tag{32}
\]

where the coefficients \( c_i \) can be approximately chosen from the condition of the best convergence of the method to the result of intermediate accuracy already obtained. In the work we used this formula with two coefficients \( c_i \) in the series.

As examples, several triangular diffraction gratings Figs. 1b), the parameters of which are listed in Table 2, were used. Gratings with the same or similar characteristics were considered earlier in [7,8].
Table 2. Parameters of the triangular diffraction gratings Fig. 1b) used as examples. A TM-polarized electromagnetic wave is incident at an angle $\theta$.

| №  | Relative permittivities | Geometric characteristics | $\theta$ | Reference value |
|----|------------------------|---------------------------|--------|----------------|
| 1  | $\varepsilon_e = 1, \varepsilon_s = -1.5 + 0i$ | $p = \lambda, \alpha = 30^0, \vartheta = 120^0$ | $0^0$ | $R_0 = 0.9983547969216146$ |
| 2  | $\varepsilon_e = 1, \varepsilon_s = -4 + 0i$ | $p = \lambda / 1.3, \alpha = 30^0, \vartheta = 60^0$ | $30^0$ | $R_0 = 0.7006512718026266$ |
| 3  | $\varepsilon_e = 1, \varepsilon_s = -1.5 + 1.5i$ | $p = \lambda / 0.9, \alpha = 30^0, \vartheta = 90^0$ | $60^0$ | $R_0 = 0.1161837518324407$ |

Fig. 8a) shows the relative error of the zero-order reflected diffraction efficiency $R_0$ of a grating No. 1 of Table 2 versus the number of grid points $N_x$ along the $x$ axis over the grating period $p$, calculated by the SEM with various numbers of functions $N_w$ in the subdomains containing edges. Previously this example was considered in [7], where the absence of the C-method (or Chandezon method [43]) convergence was observed due to the purely imaginary value of $\tau_1$. Note that in the case of a triangular grating Fig. 1b), the value of $\tau_1$ for the edge at the top vertex of the profile is the same as at the bottom; see the remark following the Eq. (41) in Appendix 1. It can be seen that the situation is the same as in the case of the analytical modal method: when the number of functions $N_w$ is constant, the convergence of the spectral method is algebraic, and is exponentially fast or close to it if $N_w$ is increasing proportionally with the number of nodes $N_x$.

In Fig. 8b) the spatial distribution of the squared magnitude of the magnetic field $|H(x, y)|^2$ is displayed. The color scale, shown on the right, is linear. The calculation was carried out at $N_x = 160$ grid points and at the number of functions $N_w = 18$. A thick white line marks the grating profile; thin lines indicate the boundaries of the subdomains. For better visibility of the boundaries of subdomains containing singularities, the radii $\rho_0$ of these subdomains in Fig. 8b) were taken an order of magnitude greater than the optimal ones, Eq. (32).

Fig. 8. a) Relative error of the reflected zero-order diffraction efficiency $R_0$ of a triangular diffraction grating No. 1 of Table 2 versus the number of grid points $N_x$ along the $x$ axis over the grating period $p$, calculated by the spectral element method with various numbers of functions $N_w$ in the subdomains containing edges; b) the squared magnitude of the magnetic field $|H(x, y)|^2$. The thin white lines indicate the boundaries of the subdomains, and the thick one marks the grating profile.

It is seen that near the grating boundary, there are regions in which the magnetic field is much larger than the amplitude of the incident plane wave Eq. (18), which is equal to unity.
Fig. 9a) shows the relative error of the reflected minus first-order diffraction efficiency $R_1$ of a triangular grating No. 2 of Table 2 versus the number of grid points $N_x$, calculated by the SEM with various numbers of functions $N_w$ in the subdomains containing edges. In this example, as in the previous one $\text{Re}(\tau) = +0$, therefore, the conventional numerical methods also do not work here, which was pointed out in [7].

In Fig. 9b) the space distribution of the squared magnetic field is depicted. The optimal radii $\rho_0$, calculated by Eq. (32), of the subdomains containing singularities are very small, and the corresponding boundaries are not visible in the figure. The entire computational domain is inclined by an angle $(\alpha-(180°-\vartheta-\alpha))/2$, which is equal to the difference in angles at the base of the triangle divided by 2, see Fig 1b). So that the angles of the subdomains are as close to $90°$ as possible, this reduces the condition number of the linear system matrix when we solve the Dirichlet problem inside these subdomains. The magnitude of the magnetic field reaches its maximum near the lower angle of the grating, in the region containing the edge that is shown beside.

In Fig. 10a) the convergence of several methods was compared using the example of a triangular grating No. 3 of Table 2. We examined:

a) the conventional spectral element method without increasing the accuracy of the solution in the edge regions;

b) the parametric C-method with adaptive transformation of coordinates near edges (or Adaptive Spatial Resolution (ASR) as it is called in the original paper [13]) and a transformation parameter $\eta = 0.99$;

c) the spectral element method with adaptive transformation of coordinates (or adaptive spatial resolution) in subdomains near edges according to the law $\rho = \rho_0(1-\sin t)$, where the parameter $t$ changes from zero to $\pi/2$ at the edge point;

d) the spectral element method utilizing expansion (15) in the subdomains containing the edges and using the proportional increasing of the number of functions $N_w$ with the number of collocation points $N_x$. 

Fig. 9. a) Relative error of the reflected minus first-order diffraction efficiency $R_1$ of a triangular diffraction grating No. 2 of Table 2 versus the number of grid points $N_x$ along the $x$ axis over the grating period $p$, calculated by the spectral element method with various numbers of functions $N_w$ in the subdomains containing edges; b) the squared magnitude of the magnetic field $|H(x,y)|^2$. The thin white lines indicate the boundaries of the subdomains, and the thick one marks the grating profile.
Fig. 10. a) Relative error of the reflected zero-order diffraction efficiency $R_0$ of a triangular diffraction grating No. 3 of Table 2 versus the number of grid points $N_x$ over the grating period $p$ for the spectral element method (SEM) and versus the number of modes $M$ for the parametric C-method with Adaptive Spatial Resolution (ASR); b) the results of calculation by the spectral element method with a fixed number of functions $N_w$ used for representing the solution near the edges.

The spectral element method employing expansion (15) demonstrates the best convergence. In Fig. 10b) the convergence of this method is shown for a different number of functions $N_w$ in the subdomains containing the edges.

Fig. 11. a) Reflectivity, reflected zero-order diffraction efficiency $R_0$ and the real part of the $\tau_1$ versus the imaginary part of the dielectric constant $\varepsilon_s$ of a triangular diffraction grating No. 3 of Table 2, the real part $\varepsilon_s$ remains the same: $\text{Re}(\varepsilon_s) = -1.5$; b) relative error of the reflected zero-order diffraction efficiency $R_0$ calculated by the parametric C-method for several values of the imaginary part of the dielectric constant $\varepsilon_s$.

Fig. 11 demonstrates how the reflected zero-order diffraction efficiency $R_0$, the reflectivity, and the real part of the $\tau_1$ change when the imaginary part of the dielectric constant $\varepsilon_s$ of grating No. 3 of Table 2 varies while the real part $\varepsilon_s$ remains the same. Just as in the last example of the previous section, the change in all examined quantities occurs smoothly and the derivatives stay finite at $\text{Re}(\tau_1) = +0$. However, if we try to extrapolate $R_0$ by using the four values of $\varepsilon_s$, for which in Fig. 11b) graphs of convergence are presented, to $\text{Im}(\varepsilon_s) = 0$ by utilizing the Thiele formula, similar to how it was done in the last example of the previous section, we get a physically incorrect result:
The polynomial extrapolation of $R_0$ using the same points gives a relative error of about 30%. To improve the result, it is necessary to shift the extrapolation points closer to $\Im(e_\varepsilon)=0$, which requires an exponential increase of the computation time in the case of conventional methods. Fig. 11b) demonstrates how the convergence rate of the parametric C-method diminishes, with a decrease in the imaginary part of $e_\varepsilon$ and a corresponding decrease in the real part of $\tau_1$.

For $\Re(\tau_1)=+0$ the parametric C-method, like the conventional SEM, does not converge to a certain value. The reason for the divergence, consisting in the behavior of the field near the edges, has already been discussed in the papers [7,8]. Having the ability to calculate the field near the edges, we shall illustrate this using the example of the triangular grating No. 3 of table 2. In Fig. 12 the left panel shows the distribution of the squared magnitude of the magnetic field $|H(x,y)|^2$ for the dielectric constant of the grating material $e_\varepsilon = -1.5+1.5i$, and the right panel shows $|H(x,y)|^2$ for $e_\varepsilon = -1.5+0i$.

![Fig. 12. The space distribution of the squared magnitude of the magnetic field $|H(x,y)|^2$ in the case of a diffraction grating No. 3 of Table 2 for two values of the dielectric constant of the grating material. The thin white lines indicate the boundaries of the subdomains, and the thick one marks the grating profile.](image)

In the presented resolution the change in the field in both pictures seems quite smooth, and although in the second case its maximum value is several times larger, the behavior of the $|H(x,y)|^2$ distribution cannot say why the conventional numerical methods work in one case and not in the other. However, if we sequentially increase the scale of the region near the upper edge of Fig. 12b), we obtain a picture of the distribution of the squared magnitude of the magnetic field shown in Fig. 13.

![Fig. 13. The squared magnitude of the magnetic field $|H(x,y)|^2$ at the apex of the triangular diffraction grating No. 3 of Table 2 with $e_\varepsilon = -1.5+0i$ ($\Re(\tau_1)=+0$), for several regions embedded in each other; the radius of the first region is $k_0\rho_0 = 0.3$, the divider $d^{1/2}=23.3$. It can be seen that if the first and the third images are still different, then the second and the fourth are almost identical.](image)
The radius of the region near the edge \( k_0 \rho_0 = 0.3 \) in the first image of Fig. 13 is the same as in the Fig. 12. In each subsequent image of Fig. 13 the radius decreases by about \( d^{1/2} = 23.3 \) times, the scale on the right is the same for all images. Despite of the reduction in scale, the field continues to change without decreasing the amplitude, without reaching a constant value. And if we wanted to represent such a distribution in the form of a complex Fourier series as in the C-method, or by using an orthogonal polynomials expansion as in the spectral element method, we would need more and more functions to take into account field changes.

To approximately represent the magnetic field near the top of the profile for small values of \( \rho \), we restrict ourselves to the first terms of series (15) containing only \( \tau_0 \) and \( \tau_1 \):

\[
H_z(\rho, \varphi) = \sum_{j,k=0,1} h_j \Phi_k(\varphi) \cdot \left( \rho / \rho_0 \right)^{\tau_j} \delta_{k,j} = h_0 + h_1 \Phi_1(\varphi) \cdot \left( \rho / \rho_0 \right)^{\tau_1},
\]

where we consider that \( \tau_0 = 0 \) and \( \Phi_0(\varphi) = 1 \).

Fig. 14 demonstrates the graphs of the relative error of such approximation Eq. (33) inside the region \( k_0 \rho_0 = 0.3 \) near the edge at the apex of the triangular profile.

The value of \( \tau_1 \) in the general case is the imaginary quantity \( \tau_1 = \alpha + i \beta \), and the squared magnetic field can be written:

\[
|H(\rho, \varphi)|^2 = |H_z(\rho, \varphi)|^2 \approx |h_0|^2 + \left( \rho / \rho_0 \right)^{2\alpha} |h_1 \Phi_1(\varphi)|^2 + 2\left( \rho / \rho_0 \right)^{\alpha} \text{Re}[h_0 \ast h_1 \Phi_1(\varphi)e^{i h_1 \ln(\rho / \rho_0)}].
\]

For a finite value of \( \alpha > 0 \), the squared magnitude of the magnetic field for small \( \rho \) converges to a constant \( |h_0|^2 \). If \( \alpha = 0, \beta \neq 0 \), then for small but finite values of \( \rho \neq 0 \) we have:

\[
|H(\rho, \varphi)|^2 \approx |h_0|^2 + |h_1 \Phi_1(\varphi)|^2 + 2 \text{Re}[h_0 \ast h_1 \Phi_1(\varphi)e^{i h_1 \ln(\rho / \rho_0)}] = a(\varphi) + b(\varphi) \cos(\beta_1 \ln \rho + \delta),
\]

where \( a(\varphi) \) and \( b(\varphi) \) do not depend on \( \rho \). The squared magnitude of the magnetic field in this case will oscillate with an increasing frequency when \( \rho \) goes to zero. And for the radii \( \rho / \rho_0 = \exp(-2\pi / \beta) \approx d \) the picture will begin to repeat, as we saw in Fig. 13. The components of the electric field \( E_\rho \sim \rho^5 \partial_\rho H_z \sim \rho^5, E_\varphi \sim \partial_\varphi H_z \sim \rho^5 \) will oscillate with increasing amplitude.

Fig. 15 shows the squared magnitude of the magnetic field depending on the distance to the apex of the triangular grating (see Fig. 12) when moving along the profile from the left side to the top. The distance is plotted on a logarithmic scale. The dashed line shows the asymptotic approximation...
calculated by Eq. (34). In the case of $\varepsilon_r = -1.5+1.5i$, see Fig. 15a), when approaching the peak ($\rho \to 0$) the oscillations decay quickly, and the magnetic field converges to a constant.

Fig. 15. The squared magnitude of the magnetic field depending on the distance to the apex of the triangular grating No. 3 of Table 2 (see Fig. 12), when moving along the profile line from the left side to the top for two values of the dielectric constant of the grating material: a) $\varepsilon_r = -1.5+1.5i$, b) $\varepsilon_r = -1.5+0i$. The dashed line shows the asymptotic approximation calculated by the Eq. (34).

In this case the field can be represented with any predetermined accuracy by using a finite number of basis functions, and the usual calculation methods work well here. However, the smaller the value of $\alpha_\tau \equiv \text{Re}(\tau_i)$ in Eq. (34), the slower the oscillations damping will be, and the more functions will be needed to represent the solution at a given accuracy. This behavior is seen in Fig. 11b). For $\varepsilon_r = -1.5+0i$ the oscillations, in the case when the quantity $\rho$ decreases, persist with a constant amplitude and with an increasing frequency, therefore, it was not possible to obtain the result by conventional methods [7,8]. Nevertheless, the explicit calculation of the field near the edges, as shown in the present paper, copes with this problem.

**Conclusion**

In this work we studied an approach which improves the accuracy and accelerates the convergence of numerical methods for solving the Maxwell’s equations by means of taking into consideration the singularities of the electromagnetic field that arise near the geometric edges of scattering objects. Several algorithms to incorporate the singularities processing into methods for solving the Maxwell’s equations in two-dimensional structures were analyzed. As implementation examples we used the analytical modal method and the spectral element method. In the first case, the resulting system of equations for the coefficients of the eigenmodes underwent changes. In the second case, for each region containing a singularity a Dirichlet-to-Neumann map operator was constructed and subsequently these regions were processed, like all the others.

In test examples, for which we used the calculation of the lamellar and the triangular diffraction gratings characteristics, a significant increase in accuracy and an acceleration of convergence was achieved. In the examples considered, the possibility to increase the type of convergence of spectral methods from algebraic to exponential or close to exponential was demonstrated. Diffraction efficiencies of the gratings, for which the conventional methods fail to converge due to the special values of permittivities, were calculated. In this work we confined ourselves to the two-dimensional case only; however, the method presented in the article can be extended to the three-dimensional case as well.

In conclusion we note, that the approach presented in the fourth paragraph of the paper, which employs the Dirichlet-to-Neumann map for joining subdomains, is universal and allows to easily em-
bed the processing of singularities in various 2D multi-domain methods, regardless of what the numerical algorithm for solving Maxwell’s equations is used inside individual domains. In addition to the spectral methods analyzed in the article, it can be the boundary integral method, the finite-difference method, the finite-element method, and etc.

Appendix 1. Solution of the eigenvalue problem for angular functions \( \Phi_k(\varphi) \) in some simple cases

The examples considered here have already been investigated previously in [9,10,24]. For completeness, we present the main results for simple cases, writing out the whole series of roots of the eigenvalues \( \tau_k \), \( \text{Re}(\tau_k) \geq 0 \), taking into consideration the degree of degeneracy, which is important for the correct writing of system (12). Note that when we write the final formulae for \( \tau_k \) in this appendix, we do not sort \( \tau_k \) by the value of the real part.

Let us first solve the problem Eq. (6) in the case of a homogeneous infinite wedge with an angle \( \vartheta \) made of a material with a permittivity \( \varepsilon_1 \) in a medium with a permittivity \( \varepsilon_2 \), Fig. 16a. We position the \( x \) axis on the axis of symmetry, and we shall seek for the function \( \Phi_k(\varphi) \) in the upper half-plane \( 0 \leq \varphi \leq \pi \), in the lower half it will be either antisymmetric or symmetric. Accordingly, at points \( \varphi = 0, \pi \) either the function itself or its derivative must vanish. In the first case, the solution can be written as:

\[
\Phi_k(\varphi) = \begin{cases} 
   a_{1,k} \sin \tau_k \varphi, & 0 \leq \varphi < \vartheta / 2 \\
   a_{2,k} \sin \tau_k (\varphi - \pi), & \vartheta / 2 \leq \varphi \leq \pi
\end{cases} \quad (36)
\]

and in the second:

\[
\Phi_k(\varphi) = \begin{cases} 
   a_{1,k} \cos \tau_k \varphi, & 0 \leq \varphi < \vartheta / 2 \\
   a_{2,k} \cos \tau_k (\varphi - \pi), & \vartheta / 2 \leq \varphi \leq \pi
\end{cases} \quad (37)
\]

Using the boundary conditions Eq. (21) for the antisymmetric case Eq. (36) we have:

\[
\begin{aligned}
   a_{1,k} \sin \tau_k \vartheta / 2 - a_{2,k} \sin \tau_k (\vartheta / 2 - \pi) &= 0 \\
   \frac{a_{1,k} \tau_k}{\varepsilon_1} \cos \tau_k \vartheta / 2 - \frac{a_{2,k} \tau_k}{\varepsilon_2} \cos \tau_k (\vartheta / 2 - \pi) &= 0
\end{aligned} \quad (38)
\]

Dividing the second line by the multipliers \( \tau_k \), since the constant solution in the case of \( \tau_0 = 0 \) we already know, and equating the determinant of Eq. (38) to zero, we obtain the equation for \( \tau_k \):

\[
\varepsilon_1 \sin \tau_k \vartheta / 2 \cdot \cos \tau_k (\vartheta / 2 - \pi) - \varepsilon_2 \sin \tau_k (\vartheta / 2 - \pi) \cdot \cos \tau_k \vartheta / 2 = 0. \quad (39)
\]

Similarly for the symmetric case:
\[ \varepsilon_1 \cos \tau_k \vartheta / 2 \cdot \sin \tau_k (\vartheta / 2 - \pi) - \varepsilon_2 \cos \tau_k (\vartheta / 2 - \pi) \cdot \sin \tau_k \vartheta / 2 = 0. \]  

(40)

It is convenient to rewrite equations (39), (40) in the form:
\[ \pm \delta \cdot \sin \tau_k (\pi - \vartheta) = \sin \tau_k \pi, \quad \delta = (\varepsilon_1 - \varepsilon_2) / (\varepsilon_1 + \varepsilon_2), \]

(41)

where the “+” sign in front of \( \delta \) corresponds to the antisymmetric case and “−” sign to the symmetric one. Note that replacing \( \varepsilon_1 \leftrightarrow \varepsilon_2 \) is identical to changing the sign “±” before \( \delta \) in Eq. (41) by “∓”, thus the set \( \{ \tau_j \} \) in a simple triangular grating in Fig. 1b) is the same for the upper and for the lower edges, only the symmetry of the functions \( \Phi_j(\varphi) \) for the corresponding \( \tau_j \) will be different.

When \( \vartheta = \pi / 2 \), the Eq. (41) will be:
\[ \pm \delta \sin(\tau_k \pi / 2) = \sin \tau_k \pi. \]

(42)

It can be noted that if \( \tau_k \) is a solution of Eq. (42), then \( \tau_k + 2n \), where \( n \) is an integer, is also a solution and, considering the zero solution \( \tau_0 = 0 \), it is enough to find \( \tau_k \) in the interval \( 0 < \text{Re}(\tau_k) \leq 2 \), all the others are obtained by adding \( 2n, n = 1, 2, 3, \ldots \). Substituting \( y = \tau_k \pi / 2 \), we obtain the equation
\[ \pm \delta \sin y = 2 \sin y \cos y, \]
whose solutions will be:

\[ a) \cos y = \pm \delta / 2 \Rightarrow \tau_1 = 4n \pm 2\pi^{-1} \arccos(+\delta / 2), \quad \tau_2 = 4n \pm 2\pi^{-1} \arccos(-\delta / 2) \]

\[ b) \sin y = 0 \Rightarrow \tau_{3,4} = 2n \]

(43)

The solutions \( \tau = 4n \pm 2\pi^{-1} \arccos(+\delta / 2) \) correspond to the antisymmetric case and \( \tau = 4n \pm 2\pi^{-1} \arccos(-\delta / 2) \) to the symmetric one. And there are two solutions \( \tau = 2n \) in Eq. (43) b): one for the symmetric and one for the antisymmetric case. Thus, the whole set of solutions \( \tau_k, \text{Re}(\tau_k) \geq 0 \) can be written as:
\[ \tau = \{ 0, 2\pi^{-1} \arccos(\delta / 2), 2\pi^{-1} \arccos(-\delta / 2), 2 \} + 2n, \quad n = 0, 1, 2, \ldots, \]

(44)

Such a notation explicitly takes into account that the eigenvalues corresponding to \( \tau = 2, 4, 6, \ldots \) are doubly degenerate. Note that the middle terms in the brackets of expression (44) are not sorted by the value of the real part, as well as in the other formulae in this appendix, and to determine the \( \tau_k \) characterizing the singularity it is necessary to choose one of them whose real part is minimal. In numerical calculation only the first \( N_w \) terms of the series \( \{ \tau_j \} \) sorted in ascending order of the real parts of \( \tau_k \) are used. In general case of arbitrary \( \vartheta \) when \( \tau_k \) is calculated numerically, for example, by method represented in Ref. [31], the symmetry of the eigenvectors corresponding to nondegenerate \( \tau_k \) can be readily identified by substituting \( \tau_k \) into Eq. (39) and Eq. (40) and comparing the absolute values of the right-hand sides. The values of the corresponding \( \tau_k \) coefficients \( a_{1,k}, a_{2,k} \) can be obtained from linear equation (38) for the symmetric and from similar equation for the antisymmetric case.

When \( \vartheta = \pi \cdot 1 / m \), where \( l, m = 1, 2, 3, \ldots \), we have:
\[ \pm \delta \cdot \sin[\tau_k \pi(m-l) / m] = \sin \tau_k \pi. \]

(45)

It can be seen that it is enough to find solutions of Eq. (45) in the interval \( 0 < \text{Re}(\tau_k) \leq m \), all the others are obtained by adding \( m \cdot n, n = 1, 2, 3, \ldots \). Moreover, between solutions \( \tau = 0 \) and \( \tau = m \) there will be \( 2(m-1) \) solutions, and solutions \( \tau = m \cdot n \) are twice degenerate, since they satisfy both the symmetric and the antisymmetric cases. For example, if \( \vartheta = \pi / 3 \) then by substituting \( y = \tau_k \pi / 3 \), we obtain the equations:
\[ \pm \delta \sin 2y = \sin 3y \Leftrightarrow \pm 2\delta \sin(y) \cos(y) = 4\sin(y) \cos(y)^2 - \sin(y), \]

(46)
whose solutions can be written in the form of six values $\tau_0, \ldots, \tau_5$ with a shift of $3n$:
\[
\tau = \{0, 3\pi^{-1} \arccos((\pm \delta - \sqrt{\delta^2 + 4})/4), 3\pi^{-1} \arccos((\pm \delta + \sqrt{\delta^2 + 4})/4), 3\} + 3n, \quad n = 0, 1, 2, \ldots \quad (47)
\]
And for $\vartheta = \pi \cdot 2/3$:
\[
\tau = \{0, 3\pi^{-1} \arccos(-\sqrt{\pm \delta / 2}), 3\pi^{-1} \arccos(\sqrt{\pm \delta / 2}), 3\} + 3n, \quad n = 0, 1, 2, \ldots \quad (48)
\]

The next simple case is a combination of four right-angled wedges with permittivities $\varepsilon_l$, $l = 1..4$, Fig. 16b). The solution of Eq. (6) in each $l$-th section is:
\[
\Phi_l(\varphi) = a_l e^{i\varphi_0} + b_l e^{-i\varphi_0}. \quad (49)
\]
Using the conditions for the function and for the derivative at the four boundaries $\varphi_l = l\pi/2$, $l = 1..4$, we obtain a homogeneous system of 8 equations:
\[
a_1 e^{i\pi/2} + b_1 e^{-i\pi/2} = a_2 e^{i\pi/2} + b_2 e^{-i\pi/2}, (a_1 e^{i\pi/2} - b_1 e^{-i\pi/2})/\varepsilon_1 = (a_2 e^{i\pi/2} - b_2 e^{-i\pi/2})/\varepsilon_2 \\
\quad \cdots \\
\quad a_4 e^{i2\pi} + b_4 e^{-i2\pi} = a_1 + b_1, (a_4 e^{i2\pi} - b_4 e^{-i2\pi})/\varepsilon_4 = (a_1 - b_1)/\varepsilon_1
\]
where we have divided the equations for derivatives by the factor $i \cdot \tau$, since we consider the constant solution at $\tau = 0$ already known. In order that the linear system (50) has a nonzero solution, it is necessary that its determinant must be equal to zero, from this condition the values of $\tau_k$ can be found. From Eq. (50) it follows that if $\tau_k$ satisfies the condition that the determinant of this equation is equal to zero, then the number $\tau_k + 2n$, where $n$ is an integer, also satisfy this condition. Thus, it is enough to find solutions in the interval $0 < \Re(\tau_k) \leq 1$, all the others can be obtained by adding $2n, n = 1, 2, 3, \ldots$ Besides, if $\tau_k$ is a solution, then $2 - \tau_k$ is also solution, therefore, if there are solutions other than 0 and $2n$, then such $\tau_k$ exists that $0 < \Re(\tau_k) \leq 1$, and if $\tau_k \neq 1$ then the singularity of the electric field components exists, which is proportional to $\rho^{\pi - 1}$. By substituting $x = e^{i\tau \pi}$ and by equating determinant of Eq. (50) to zero we obtain:
\[
e \cdot x^9 \left[ c \cdot (x^4 + 1) - 4d \cdot (x^3 + x) - 2(c - 4d) \cdot x^2 \right] = 0 \\
c = (\varepsilon_1 + \varepsilon_2)(\varepsilon_2 + \varepsilon_3)(\varepsilon_3 + \varepsilon_4)(\varepsilon_4 + \varepsilon_1), \quad d = (\varepsilon_1 \varepsilon_3 - \varepsilon_2 \varepsilon_4)^2, \quad e = (\varepsilon_1 \varepsilon_2 \varepsilon_3 \varepsilon_4)^2
\]
Or, after simple conversions:
\[
x^9 (x - 1)^2 (x^2 - 2x + 1) = 0, \quad \alpha = (2d/c) - 1.
\]
Zero solutions $x = 0$ have no physical meaning, two solutions with $x = 1$ correspond to $\tau_{1,4} = 2$, and solutions corresponding to the last bracket vanishing:
\[
\alpha = (x^3 + 1)/2x = (x + x^{-1})/2 = \cos(\tau_k \pi),
\]
are:
\[
\tau_1 = \pi^{-1} \arccos(2d/c - 1) = 2\pi^{-1} \arccos(\sqrt{d/c}), \quad \tau_2 = 2 - 2\pi^{-1} \arccos(\sqrt{d/c})
\]

Thus, the whole set of numbers $\tau$ can be represented as:
\[
\tau = \{0, 2\pi^{-1} \arccos(\sqrt{d/c}), 2 - 2\pi^{-1} \arccos(\sqrt{d/c}), 2\} + 2n, \quad n = 0, 1, 2, \ldots
\]

In the general case when all wedges in a wedge assembly have angles that are multiples of $\pi/m$, where $m$ is a positive integer, then it follows from the system of $4m$ equations for the continuity
of functions $\Phi_k(\phi)$ Eq. (49) and conditions on the derivative on $2m$ boundaries $\phi_l = l\pi/m$, $l = 1..2m$:

$$a_1 e^{i\pi/m} + b_1 e^{-i\pi/m} = a_2 e^{i\pi/m} + b_2 e^{-i\pi/m}, (a_1 e^{i\pi/m} - b_1 e^{-i\pi/m})/\varepsilon_1 = (a_2 e^{i\pi/m} - b_2 e^{-i\pi/m})/\varepsilon_2$$

$$\ldots$$

$$a_{2m} e^{i2\pi} + b_{2m} e^{-i2\pi} = a_1 + b_1, (a_{2m} e^{i2\pi} - b_{2m} e^{-i2\pi})/\varepsilon_{2m} = (a_1 - b_1)/\varepsilon_1$$

(56)

that if $\tau_k$ satisfies the condition that the determinant of Eq. (56) is equal to zero, then the numbers $\tau_k + n \cdot m$, where $n$ is an integer, also satisfy this condition. Thus, it is enough to find solutions in the interval $0 < \Re(\tau_k) \leq m$, all the rest are obtained by adding $n \cdot m$, $n = 1,2,3,...$; besides, if $\tau_k$ is a solution then $m - \tau_k$ is also solution. It also follows that the number of distinct roots sorted in ascending order of the real part between the values $\tau_k + (n-1) \cdot m$ and $\tau_k + n \cdot m$ is the same for all $k$ and $n$. Assuming that the eigenvalues as quantities that have a physical meaning in modal methods continuously depend on the permittivity $^1$, it is easy to show that there will be exactly $2(m-1)$ solutions between the values $\tau = (n-1) \cdot m$ and $\tau = n \cdot m$. Indeed, in the case of uniform permittivity on the closed interval $[0,2\pi]$ the unnormalized eigenfunctions of the problem:

$$\partial^2 \Phi_k(\phi) = -\tau_k^2 \Phi_k(\phi)$$

(57)

are $\{\cos(\tau \phi), \sin(\tau \phi)\}$ and all eigenvalues $\tau_k$, except for $\tau_0 = 0$, are doubly degenerate and represent a sequence of natural numbers: $\tau = 0,1,2,3,4,...$. And there are $2(m-1)$ solutions between the values $\tau = (n-1) \cdot m$ and $\tau = n \cdot m$ in this case. By continuously changing the dielectric constant inside the wedges with angles that are multiples of $\pi/m$, one can see that the doubly degenerate eigenvalues $\tau_{2m+1} = \tau_{2m} = n \cdot m$ remain unchanged, the first corresponds to the eigenfunction $\cos(\tau_{2m+1} \phi)$ $(a_j = b_j = 1/2$ in Eq. (56) ), the second corresponds to eigenfunction $\varepsilon(\phi)\sin(\tau_{2m} \phi)$ $(a_j = -b_j = \varepsilon_j/(2i)$ ). Under the condition $\tau'_k = \tau_k + m$ between $\tau'_k$ and $\tau_k$ the number of roots, sorted by the real part, lying between the values $\tau = (n-1) \cdot m$ and $\tau = n \cdot m$, $n = 1,2,...$ must be the same. If the roots $\tau_k$ change continuously, that is they do not disappear and new ones do not arise, then the number of solutions with a real part between the values $(n-1) \cdot m$ and $n \cdot m$ will remain unchanged. Thus, we can write: $\tau_{k+2m} = \tau_k + n \cdot m$.

In Appendix 2 we shall show that the fulfillment of the relation $\tau'_k = \tau_k + n \cdot m$ between the roots leads to the presence of the logarithms of $r$ in the series $R_{k,j}(r)$ Eq.(14), when the angles of the wedges are the product of rational numbers and $\pi$.

Appendix 2. Calculation of radial functions $R$

2.1 Iterative calculation of radial functions $R$

We have to solve the system of differential equations (12) in the domain $0 \leq r \leq 1$, to obtain a set of functions $R_{k,j}(r)$ satisfying condition (10). In the numerical solution we restrict ourselves to the first functions $R_{k,j}(r)$, $k = 0..N_w - 1$, $j = 0..N_w - 1$ corresponding to the roots $\tau_k$, sorted in ascending order of the real part. We shall solve the system of equations (12) iteratively, finding the new $l+1$ approximation $R_{k,j}^{(l+1)}$ by substituting of the previous $R_{k,j}^{(l)}$ into the right side of the system:

---

1 The property of the continuous dependence of the roots $\tau_k$ on the dielectric constant is used in some numerical implementations of the analytical modal method to calculate the values of $\tau_k$, see, for example, [30].
\[
\left[ r\partial_r \left( r\partial_r \right) - \tau_k^2 \right] R^{(i+1)}_{k,j}(r) = -r^2 \gamma^2 \sum_{k=0}^{N_w-1} c_{k,j} R^{(i)}_{k,j}(r). \tag{58}
\]

As a zeroth approximation we take the static solution:
\[
R^{(0)}_{k,j}(r) = \delta_{k,j} R^{0}_0(r) = \delta_{k,j} r^{\tau_j}. \tag{59}
\]

The approximations \( R^{(i)}_{k,j}(r) \) can be calculated independently for each \( j = 0..N_w - 1 \), the difference is only in the initial approximation (59).

Solution of Eq. (58) is the sum of the general solution of the corresponding homogeneous system and the particular solution of the inhomogeneous one. As a solution of the homogeneous system that satisfies the finiteness condition at zero and Eq. (10), we take the static solution (59) at each iteration. At the first iteration the solution of equation (58) is
\[
\left[ r\partial_r \left( r\partial_r \right) - \tau_k^2 \right] R^{(1)}_{k,j}(r) = -r^2 \gamma^2 c_{k,j} r^{\tau_j} \Rightarrow R^{(1)}_{k,j}(r) = R^{(0)}_{k,j}(r) + \begin{cases} 
\gamma^2 c_{k,j} \frac{r^{\tau_j+2}}{\tau_k^2 - \left( \tau_j + 2 \right)^2}, & \tau_k \neq \tau_j + 2 \\
\frac{r^{\tau_j+2}}{4\tau_k^2} \ln r, & \tau_k = \tau_j + 2
\end{cases} \tag{60}
\]

In Eq. (60) we obtained the sum of the zeroth approximation with a term containing the powers and logarithms of \( r \) multiplied by \( \gamma^2 \). If we continue this process the result will be a series in even powers of \( \gamma \). Besides,
\[
R^{(i)}_{k,j}(r) = R^{(i-1)}_{k,j}(r) + \gamma^2 r^{\tau_j+2} a_{k,j,i}(r), \tag{61}
\]

where \( a_{k,j,i}(r) \) consists of constants and possibly also logarithms of \( r \). It can be seen from Eq. (60) that the power of \( r \) on the right-hand side after each iteration increases by 2 and if there are \( \tau_k = \tau_j + 2l \), where \( l \) is some positive integer, then starting from the \( l \)-th approximation the term \( R^{(i)}_{k,j}(r) \) will contain logarithms of \( r \) and in the next iteration the logarithms will be already in all \( R^{(i+1)}_{k,j}(r), i = 0,1,2,\ldots \). As follows from Appendix 1, if the wedge assembly consists of wedges with angles that are multiples of \( \pi / m \), where \( m \) is a positive integer, or in other words, the angles are products of rational numbers by \( \pi \), then there are roots \( \tau_k = \tau_j + n \cdot m, n = 1,2,\ldots \). Thus, logarithms appear in the approximation \( R^{(i)}_{k,j}(r) \) as soon as \( 2l = n \cdot m \) is satisfied. And the functions \( a_{k,j,i}(r) \) in Eq. (14) and Eq. (15) will contain the logarithms for \( i \geq l \). When the angles are multiples of \( \pi / 2 \), the logarithms in \( R^{(i)}_{k,j}(r) \) appear starting from the first approximation, when the angles are multiples of \( \pi / 3 \) starting from the third, etc.

At the next iterations we have to solve equations containing the right-hand side of the form:
\[
\left[ r\partial_r \left( r\partial_r \right) - \tau_k^2 \right] u(r) = -r^2 \gamma^2 c \cdot r^{\alpha} (\ln r)^\beta, \tag{62}
\]

where the degree of the logarithm \( \beta \) is an integer, \( \beta < \text{Re}(\alpha) \). A particular solution to the equation with such a right-hand side can be calculated using formal integral representation:
\[
u_{\text{particular}}(r) = -\gamma^2 c \cdot \begin{cases}
\int_0^r \left( r^\alpha \ln r \right)^\beta \left( r^\alpha \ln r \right)^\beta dr^\gamma, & \tau_k = \tau_0 = 0 \\
\frac{1}{2\tau_k} \left[ r^\gamma \int_0^{r^\gamma} (r^\gamma)^{\beta+1} (\ln r^\gamma)^\beta dr^\gamma - r^{-\gamma} \int_0^r (r^\gamma)^{\beta+1} (\ln r^\gamma)^\beta dr^\gamma \right], & \tau_k = \alpha + 2 \\
\frac{1}{2\tau_k} \left[ r^\gamma \int_0^{r^\gamma} (r^\gamma)^{\beta+1} (\ln r^\gamma)^\beta dr^\gamma - r^{-\gamma} \int_0^r (r^\gamma)^{\beta+1} (\ln r^\gamma)^\beta dr^\gamma \right], & \tau_k = \alpha + 2
\end{cases} \tag{63}
\]
where in the third line for \( \tau_k \neq \alpha + 2 \) the lower limit of the first integral must be assigned as \( \sigma = 0 \) if \( \text{Re}(\tau_k) < \text{Re}(\alpha) + 2 \) and as \( \sigma = +\infty \) otherwise. Integrals with a logarithm are calculated using the recurrence formula:

\[
\int r^\alpha (\ln r)^bdr = \frac{r^{\alpha+1}(\ln r)^b}{a+1} - \frac{b}{a+1}\int r^\alpha (\ln r)^{b-1}dr = r^\alpha \sum_{q=0}^{b} \frac{(-1)^q b!}{(a+1)^{q+1}(b-q)!} (\ln r)^{b-q} + \text{const}. \tag{64}
\]

So, Eq. (63) can be rewritten in expanded form:

\[
\sum_{q=0}^{\beta} \frac{(-1)^q (q+1)\beta!}{(\alpha+2)^{q+2}(\beta-q)!} (\ln r)^{\beta-q}, \tau_k = \tau_0 = 0
\]

\[
\sum_{q=0}^{\beta+1} \frac{(-1)^q \beta!}{2^{q+1}(\alpha+2)^{q+1}(\beta+1-q)!} (\ln r)^{\beta+1-q}, \tau_k = \alpha + 2
\]

\[
\sum_{q=0}^{\beta} \frac{[(\alpha+2+\tau_k)^{q+1}-(\alpha+2-\tau_k)^{q+1}]\beta!}{2\tau_k[\tau_k^2-(\alpha+2)^2]^{q+1}(\beta-q)!} (\ln r)^{\beta-q}, \text{otherwise}
\]

Thus, the solution will look like the series:

\[
R_{k,j}(r) = R_0^j(r)\left[\delta_{k,j} + \gamma^2 a_{k,j,1}(r) \cdot r^2 + \gamma^4 a_{k,j,2}(r) \cdot r^4 + \ldots\right], \tag{66}
\]

where \( a_{k,j}(r) \) are either constants independent of \( r \), or if \( \tau_k = \tau_j + 2l \) holds, where \( l \) is a positive integer, then \( a_{k,j,l}(r), \ i \geq l \) consist of the sum of the constants and the logarithms in an integer power which is less than the power of multiplier \( r \) in Eq. (66), so that \( R_{k,j}(r \to 0) \to \delta_{k,j} R_0^j(r) \). The second option is realized when the angles of the wedges are the product of \( \pi \) and rational numbers. As practical calculations show, for better accuracy the number \( N_\phi \) should be taken much greater than \( N_w \), and on finishing the iterative process \( R_{k,j}(r) \) can be truncated to a square matrix if necessary. The required number of functions \( N_\phi \) can be determined by taking the value several times larger than \( N_w \) as the initial approximation, and then double \( N_\phi \) until the changes in the final calculation results become less than the specified error.

### 2.2 Convergence of the series \( R_{k,j}(r) \)

In this appendix we shall show that the series (66) converges absolutely for any finite values of \( \gamma \), when the domain under consideration consists of wedges with angles multiples of \( \pi/m \), where \( m \) is a positive integer. We also briefly discuss the case of angles that are multiples of an irrational factor of \( \pi \).

If the iterative process (61) is briefly written in the form:

\[
R_{k,j}^{[l]}(r) = R_{k,j}^{[l-1]}(r) + u_{k,j,l}(r), \quad l = 1, 2, \ldots
\]

then the series (66) will look like:

\[
R_{k,j}(r) = \delta_{k,j} R_0^j(r) + \sum_{l=1}^{\infty} u_{k,j,l}(r) = \sum_{l=0}^{\infty} u_{k,j,l}(r). \tag{67}
\]

To prove the convergence of this series, below we will construct a majorizing series and show its convergence.

To write down analytical formulas for the terms of series (67), let us trace their evolution. During the first \( L_4 - 1 \) iterations, until the equality \( \tau_k = \tau_j + 2L_4 \) for some \( \tau_k \) is satisfied, there are no logarithms in \( R_{k,j}^{[l]}(r) \) and

\[
R_{k,j}^{[l]}(r) = R_{k,j}^{[l-1]}(r) + \gamma^{2l} r^{t_{[l]}(r)} a_{k,j,l}(r) = R_{k,j}^{[l-1]}(r) + \gamma^{2l} r^{t_{[l]}(r)} A_{k,j}^{[l]}, \tag{68}
\]

where
where we have introduced the notation \( t_j(l) = \tau_j + 2l \) and \( A_{k,j}^{[l,0]} \) are constants independent of \( r \):

\[
A_{k,j}^{[1,0]} = \frac{c_{k,j}}{\tau_k^2 - t_j(l)^2}, \
A_{k,j}^{[2,0]} = \frac{1}{\tau_k^2 - t_j(l)^2} \sum_{s=0}^{\infty} c_{k,s} A_{k,j}^{[1,0]}, \ldots, 
A_{k,j}^{[l,0]} = \frac{1}{\tau_k^2 - t_j(l)^2} \sum_{s=0}^{\infty} c_{k,s} A_{k,j}^{[l-1,0]}.
\] (69)

The first upper index in \( A_{k,j}^{[l,0]} \) corresponds to the iteration number and the second to power of the logarithm by which this coefficient will be multiplied. Note that \( \tau_j(l)^2 \geq 0 \) since \( \Re(\tau_j) \geq 0 \) and the equality holds only for \( j = 0 \). For \( 1 < l < L \) the inequality \( \tau_j \neq t_j(l) \) holds for all \( k = 0, \infty \) and in the case under consideration, when the angles are proportional to the rational factor of the number \( \pi \) it implies that \( \tau_{k+2nm} = \tau_k + n \cdot m \) and, for example, the first series in Eq. (69) can be written as:

\[
\left| \sum_{s=0}^{\infty} c_{k,s} c_{s,j} \right| \leq C^2 \sum_{s=0}^{\infty} \frac{1}{\tau_s^2 - t_j^2} = C^2 \sum_{q=0}^{2m-1} \left[ \sum_{n=0}^{\infty} \frac{1}{(\tau_q + n \cdot m)^2 - t_j^2} \right],
\] (70)

where \( C = \max_{k,j} |c_{k,j}| \) is some finite number, since the functions \( \Phi_k(\varphi) \) in Eq. (13) are continuous on a closed interval \([0, 2\pi]\) and, therefore, are bounded by the Weierstrass extreme value theorem and consequently the integral of their product on this interval is also bounded. Because the series in square brackets in Eq. (70) converges, the entire series (70) also converges. Similarly for other series in Eq. (69).

Since \( \tau_{k+2nm} = \tau_k + n \cdot m \), the function

\[
\hat{d}_j(l) = \min_{k=0, \ldots, \infty, \tau_k = t_j(l)} |\tau_k - t_j(l)|
\] (71)

is periodic with period \( \Delta l = m/2 \) for even \( m \) and \( \Delta l = m \) otherwise. Let us define:

\[
d_j = \min_{l=1, \ldots, \infty} \hat{d}_j(l) = \min_{l=1, \Delta l} \hat{d}_j(l)
\] (72)

The finite length of the period \( \Delta l \) is a fundamental feature of the case of angles that are multiples of the rational factor of \( \pi \). This allows us to determine finite \( d_j \) Eq. (72) and prove the convergence of the series (66) simply enough. Also in this case we can choose a continuous function \( D_j(l) \) that grows with increasing \( l \) no faster than logarithmic function, so that:

\[
\sum_{k=0}^{\infty} \frac{1}{\tau_k^2 - t_j(l)^2} \leq \frac{\hat{D}_j(l)}{\hat{\tau}_j(l)} \leq \frac{D_j \cdot \ln(l+1)}{d_j \cdot |t_j(l)|},
\] (73)

where \( D_j > 0 \) is some constant. This statement is easily proved by dividing (73) into the inner and outer sums as in Eq. (70) and replacing the inner sum over \( n \) by a majorizing integral with appropriately chosen limits.

Note that if the condition \( \tau_k \neq t_j(l), k = 0, \infty \) was satisfied for any \( l \) and there was a fixed value of \( d_j \), then the convergence of series (66) for any finite \( \gamma \) would be proved elementary, since the sequence \( A_{k,j}^{[l,0]} \) for \( l \gg 1 \) asymptotically behaves like \( A_{k,j}^{[l,0]} \sim C^l \prod_{q=1}^{l} \hat{D}_j(q) / (d_j 2!^l) \), which follows from (69), (73). And series (67) for \( l \gg 1 \) will be dominated by a converging series:

---

1 If the number of roots between \( \tau_j \) and \( \tau_j + n \cdot m \), \( n = 1, 2, \ldots \) differed from \( 2(m-1) \), the further proof would remain the same; it is only important for us that this number is permanent and does not depend on \( n \) or \( k \).
\[\tilde{u}_l = \gamma^{2l}C^{l} \left[ \prod_{q=1}^{l} \tilde{D}_q(q) \right] f \left( d_q^{l} 2^l \right). \]  

(74)

As is known, series of the form \( \sum_{l=0}^{\infty} \tilde{u}_l \) converge absolutely if starting from some index \( l \), the ratio \( |\tilde{u}_{l+1} / \tilde{u}_l| \) is less than some constant \( q < 1 \) (d’Alembert ratio test), since the series is dominated by a converging geometric progression with the common ratio \( q \). For series (74), this criterion is satisfied for indexes \( l > \gamma^2 C \cdot D_j \ln(l+2)/(2d_j) \). In our case, due to the fact that there are roots \( \tau_k = t_j(L_\beta) \), \( \beta = 1, 2, \ldots \), logarithms will appear in series (66), however, as we will see below, the asymptotic form of the terms of the series will remain approximately the same, and series (66) will converge for an arbitrary fixed \( \gamma \).

When the equality \( \tau_{k_i} = t_j(L_1) \), i.e. \( \tau_{k_i} = \tau_j + 2L_1 \) is valid for the first time for some \( k_i \), then we get in the solution for the \( k_i \)-th row of system (58) (or in more rows, if there is a degeneracy) the logarithm in Eq. (75) so that on the next iteration the coefficients \( A_{k-i,j}^{[l+1]} \) will have in the solution for the \( l \)-th iteration, equation (58) will have the form:

\[ R_{k-i,j}^{[l+1]}(r) = R_{k-i,j}^{[l]}(r) + \gamma^{2l+1} \left[ A_{k-i,j}^{[l]} d_j \ln r + A_{k-i,j}^{[l+1]} \right], \quad \tau_{k_i} = t_j(L_1), \]

\[ A_{k-i,j}^{[l]} = -\frac{1}{d_j \cdot 2t_j(L_1)} \sum_{r=0}^{\infty} c_{k_i,j} A_{k-i,j}^{[l-1]} \]  

(75)

for all other equations of the infinite system (58), the solution remains in the form (68) with the coefficients:

\[ A_{k-i,j}^{[l+0]} = \frac{1}{\tau_k^2 - t_j(L_1)^2} \sum_{r=0}^{\infty} c_{k_i,j} A_{k-i,j}^{[l-1]} = \frac{1}{[\tau_k - t_j(L_1)][\tau_k + t_j(L_1)]} \sum_{r=0}^{\infty} c_{k_i,j} A_{k-i,j}^{[l-1]}, \quad k \neq k_i \]  

(76)

We specially added a factor \( d_j \) to the denominator of the expression for the coefficient \( A_{k-i,j}^{[l+1]} \) at the logarithm in Eq. (75) so that on the next iteration the coefficients \( A_{k-i,j}^{[l+1]} \) and \( A_{k-i,j}^{[l+1]} \) would be of the same order of magnitude in terms of \( d_j \). Indeed, at the \( L_1 + 1 \)-th iteration, equation (58) will have the form:

\[ \left[ r \partial_j, (r \partial_j) - \tau_k^2 \right] R_{k-i,j}^{[l+1]}(r) = -r^2 \gamma^2 \sum_{r=0}^{\infty} c_{k_i,j} R_{k-i,j}^{[l]}(r) = -r^2 \gamma^2 \left[ \sum_{r=0}^{\infty} c_{k_i,j} R_{k-i,j}^{[l]}(r) - \gamma^{2l} \left[ A_{k-i,j}^{[l+1]} d_j \ln r + \sum_{r=0}^{\infty} c_{k_i,j} A_{k-i,j}^{[l]} \right] \right]. \]  

(77)

The solution to Eq. (77) for \( \tau_k \neq t_j(L_1 + 1) \) is

\[ R_{k-i,j}^{[l+1]}(r) = R_{k-i,j}^{[l]}(r) + \gamma^{2l+1} \left[ A_{k-i,j}^{[l+1]} d_j \ln r + A_{k-i,j}^{[l+1]} \right], \]  

(78)

where

\[ A_{k-i,j}^{[l+1]} = -\frac{1}{t_j(L_1 + 1)} \left[ \sum_{r=0}^{\infty} c_{0,r} A_{k-i,j}^{[l]} - c_{0,0} A_{k-i,j}^{[l]} \right] 2d_j, \quad k = 0 \]

\[ A_{k-i,j}^{[l+1]} = \frac{1}{\tau_k^2 - t_j(L_1 + 1)^2} \left[ \sum_{r=0}^{\infty} c_{k_i,r} A_{k-i,j}^{[l]} + c_{k_i,0} A_{k-i,j}^{[l]} \right] \left[ \frac{d_j t_j(L_1 + 1)}{\tau_k^2 - t_j(L_1 + 1)^2} \right], \quad k > 0 \]  

(79)

\[ A_{k-i,j}^{[l+1]} = \frac{1}{\tau_k^2 - t_j(L_1 + 1)^2} c_{k_i,0} A_{k-i,j}^{[l]} \]
The sum in the expression (79) for the coefficient $A_{k,j}^{[1,0]}$ is
\[
\sum_{s=0}^{\infty} c_{k,s} A_{k,j}^{[1,0]} = \sum_{s=0}^{\infty} \frac{c_{k,s}}{t_j(L_s)} U_{k,s} + \frac{c_{k,j}}{2^2 t_j(L_s)^2} U_{k,j},
\]  
and the product in the case of coefficient $A_{k,j}^{[1,1]}$ is
\[
c_{k,j} A_{k,j}^{[1,1]} = \frac{1}{d_j \cdot 2 t_j(L_s)} c_{k,j} U_{k,j},
\]
where for brevity we introduced the notation $U_{k,s} = \sum_{s=0}^{\infty} c_{s,0} A_{k,j}^{[1,s,0]}$. Taking into account definitions (71), (72), it can be seen that expressions (80) and (81) are of the same order of magnitude in terms of $d_j$.

If at this iteration the equality $\tau_{k_2} = t_j(L_2)$ is satisfied for some index $k_2$, the logarithm to the second power will appear in the row $k_2$. Otherwise, until the equality $\tau_{k_2} = t_j(L_2)$ for some $k_2$ is satisfied at the next iterations, the logarithm of the first degree will be present on the right side of each line of Eq. (58), and for iteration $L_l < l < L_2$ the solution will be written as:
\[
R_{k,j}^{[l]}(r) = R_{k,j}^{[l-1]}(r) + \gamma^{2(k_l-1)} t_j^{k_l} \left[ A_{k,j}^{[l]} d_j \ln r + A_{k,j}^{[l,0]} \right],
\]
where
\[
A_{k,j}^{[l,0]} = \begin{cases}
\frac{-1}{t_j(L_s)^2} \sum_{s=0}^{\infty} c_{k,s} \left[ A_{k,j}^{[l-1,0]} - A_{k,j}^{[l-1,1]} d_j \frac{2}{t_j(L_s)} \right], & k = 0 \\
\frac{1}{\tau_{k_2} - t_j(L_s)^2} \sum_{s=0}^{\infty} c_{k,s} \left[ A_{k,j}^{[l-1,0]} + A_{k,j}^{[l-1,1]} d_j \frac{t_j(L_s)}{\tau_{k_2} - t_j(L_s)^2} \right], & k > 0.
\end{cases}
\]
\[
A_{k,j}^{[l,1]} = \frac{1}{\tau_{k_2} - t_j(L_s)^2} \sum_{s=0}^{\infty} c_{k,s} A_{k,j}^{[l-1,1]}
\]
When the equality $\tau_{k_2} = t_j(L_2)$ is realized for some $k_2$, we get in the solution for the $k_2$-th row of system (58) (or in more rows, if there is a degeneracy) the squared logarithm:
\[
R_{k_2,j}^{[l_2]}(r) = R_{k_2,j}^{[l_2-1]}(r) + \gamma^{2(k_2_l-1)} t_j^{k_2_l} \sum_{s=0}^{\infty} c_{k_2,s} \left[ A_{k_2,j}^{[l_2-1,0]} \ln r + A_{k_2,j}^{[l_2-1,1]} \right] - \frac{1}{2 t_j(L_2)^2} + \frac{1}{2^2 t_j(L_2)^3} + \frac{1}{2^3 t_j(L_2)^4} + \ldots
\]
\[
= \gamma^{2(k_2_l-1)} t_j^{k_2_l} \left[ A_{k_2,j}^{[l_2-1,0]} \ln r + A_{k_2,j}^{[l_2-1,1]} \right],
\]
where
\[
A_{k_2,j}^{[l_2,0]} = -\frac{1}{d_j 2 t_j(L_2)^2} A_{k_2,j}^{[l_2-1,0]},
\]
\[
A_{k_2,j}^{[l_2,1]} = \frac{1}{d_j 2 t_j(L_2)^2} \sum_{s=0}^{\infty} c_{k_2,s} A_{k_2,j}^{[l_2-1,1]} d_j + \frac{A_{k_2,j}^{[l_2-1,1]} d_j}{2 t_j(L_2)^2},
\]

28
For all other lines of the solution everything will be as before, in accordance with Eq. (82). Here we also added \( d_j \) to the denominator of the coefficient \( A_{k,j}^{[2]} \) in Eq. (86), so that \( A_{k,j}^{[2,1]} \), \( A_{k,j}^{[2,1,1]} \) and \( A_{k,j}^{[2,1,2]} \) would be of the same order of magnitude in terms of \( d_j \). In addition, in Eq. (85) we transfer the factor 1/2 from the coefficient \( A_{k,j}^{[2,2]} \) to the logarithm.

If we further continue this process, then in each iteration \( L_\beta \), when \( \tau_k = t_j(L_\beta) \) is satisfied for some \( k_\beta \), in the solution for the \( k_\beta \)-th row of system (58) (or in more rows, if there is a degeneracy), the behavior of the coefficients \( A_{k,j}^{[q]} \) is no singularity in Eq. (90). Thus, the convergence of series (88) is completely determined by the behavior of the coefficients \( A_{k,j}^{[q]} \). If with an increase in the iteration number \( l \) they change according to the power law: \( A_{k,j}^{[q]} \sim \alpha^l \), where \( \alpha \) is a constant, then the parameter \( \gamma = 2\pi(\rho_0 / \lambda) \) characterizing the radius of convergence \( \rho_0 \) is bounded above by some value \( \gamma_{\max} \sim \alpha^{-1/2} \). If with an increase in \( l \)
the coefficients $A_{k,j}^{[l,q]}$ decrease faster than according to a power law, for example, inversely proportional to the factorial $l$, then the series will converge for any finite $\gamma$. Before embarking on an analytical study of series (89), let us carry out a numerical experiment. If the first case $|A_{k,j}^{[l,q]}| \sim \alpha^l$ is realized, then the value of $\ln |A_{k,j}^{[l+1,q]}| / |A_{k,j}^{[l,q]}|$ with an increase in $l$ will come to a constant $\ln \alpha$. In the second case, this value will decrease indefinitely; so, for example, for $|A_{k,j}^{[l,q]}| \sim \alpha^l / l!$, it will be proportional to $\ln(\alpha / (l+1))$. In Fig. 17, the solid blue line shows the dependence of $\ln \left( \max_{k,q} \left| A_{k,j}^{[l+1,q]} \right| / \max_{k,q} \left| A_{k,j}^{[l,q]} \right| \right)$ versus the iteration number $l$ at $j = 0$ for the triangular diffraction grating No. 3 of Table 2, next to it the function $\ln(\alpha / (l+1))$ is depicted by the green dashed line. A similar decreasing dependence can be obtained for other examples and numbers $j$, indicating that the second option is actually realized.

![Fig. 17. The solid blue line shows the dependence of $\ln(A^{[l+1]}/A^{[l]})$, where $A^{[l]} = \max_{k,q} |A_{k,j}^{[l,q]}|$ versus the iteration number $l$ at $j = 0$ for the triangular diffraction grating No. 3 of Table 2, next to it the function $\ln(\alpha / (l+1))$, $\alpha = 0.5$ is depicted by the green dashed line.](image)

Let us at first estimate the sum in the third line Eq. (89):

$$
\left| A_{k,j}^{[L,q]} \right| \leq C \cdot F_j(L_{\beta}) \sum_{p=0}^{\beta-1-q} d_j^p \left[ t_j(L_{\beta}) + \tau_k \right]^{\rho+1} - \left[ t_j(L_{\beta}) - \tau_k \right]^{\rho+1} \leq C \cdot F_j(L_{\beta}) \sum_{p=0}^{\beta-1-q} d_j^p \left| t_{k,j}^{[L,q,p]} \right|, \quad (91)
$$

where we denote $F_j(L_{\beta}) \equiv \max_{q} \sum_{s=0}^{\infty} \left| A_{k,j}^{[L,s]} \right|$. The absolute value of $t_{k,j}^{[L,q,p]}$ can be estimated as:

$$
\left| t_{k,j}^{[L,q,p]} \right| = \frac{1}{2 |\tau_k|} \left[ \frac{1}{\left| \tau_k - t_j(L_{\beta}) \right|^{\rho+1}} - (-1)^{\rho+1} \right] \leq \frac{1}{|\tau_k|} \left[ t_j(L_{\beta}) - \tau_k \right]^{\rho+1}, \quad \text{otherwise}
$$

(92)

In the case of $|\tau_k - t_j(L_{\beta})| \leq |\tau_k + t_j(L_{\beta})|$, taking into account that $d_j \leq |t_j(L_{\beta}) - \tau_k|$, for the sum in Eq. (91) we obtain:

$$
\sum_{p=0}^{\beta-1-q} d_j^p \left| t_{k,j}^{[L,q,p]} \right| \leq \sum_{p=0}^{\beta-1-q} \frac{d_j^p}{|\tau_k|} \left[ t_j(L_{\beta}) - \tau_k \right]^{\rho+1} = \frac{1}{|\tau_k|} \left[ t_j(L_{\beta}) - \tau_k \right]^\rho \sum_{p=0}^{\beta-1-q} d_j^p \left[ t_j(L_{\beta}) - \tau_k \right]^p \leq \frac{\beta - q}{|\tau_k|} \left[ t_j(L_{\beta}) - \tau_k \right]. \quad (93)
$$
And it can be written:
\[ |A_{k,j}^{[p,q]}| \leq C \cdot F_j(L_\beta) \frac{\beta - q}{|\tau_k - t_j(L_\beta)|} \cdot \frac{|\tau_k - t_j(L_\beta)| - |\tau_k + t_j(L_\beta)|}{|\tau_k + t_j(L_\beta)|} = \frac{C \cdot F_j(L_\beta)}{|t_j(L_\beta)|}. \] (94)

If \(|\tau_k - t_j(L_\beta)| > |\tau_k + t_j(L_\beta)|\), then since \(d_j \leq |t_j(L_\beta) + \tau_k|\), we obtain a relation that is completely analogous to equation (93):
\[ \frac{1 - \sum_{p=0}^{\beta-q-1} d_j^{p \cdot (p+1)} |t_j(L_\beta)|}{|t_j(L_\beta)|} \leq \left( \frac{C \cdot F_j(L_\beta)}{|t_j(L_\beta)|} \right)^q, q < \beta. \] (95)

For the case \(k = 0, (\tau_0 = 0)\), taking into account that \(d_j \leq |t_j(1) - \tau| \leq 2\) and \(|t_j(L_\beta)| > 2\beta\), we have:
\[ \sum_{p=0}^{\beta-q} d_j^p |t_j(L_\beta)|^{p+1} \leq \frac{1}{|t_j(L_\beta)|} \sum_{p=0}^{\beta-q} d_j^{p \cdot (p+1)} \leq \left( \frac{C \cdot F_j(L_\beta)}{|t_j(L_\beta)|} \right)^q, q < \beta. \] (96)

For definiteness, let us analyze the case, when in each iteration \(l\) the equality \(\tau_k = t_j(l)\) is realized for some index \(k\) and the maximum power of the logarithm in Eq. (88) increases by unity (this is the case when the angles of wedges are proportional to \(\pi/2\)). For other cases everything is done in the same way, we will discuss the difference below. For the second line of Eq. (89), when \(\tau_k = t_j(L_\beta)\), the upper bound of the absolute value of \(A_{k,j}^{[p,q]}\) will be
\[ \left| A_{k,j}^{[p,q]} \right| \leq C \cdot F_j(L_\beta) \frac{\beta - q}{|t_j(L_\beta)|} \cdot \frac{|t_j(L_\beta)| - |t_j(L_\beta) + \tau_k|}{|t_j(L_\beta) + \tau_k|} = \frac{C \cdot F_j(L_\beta)}{|t_j(L_\beta)|}. \] (97)

Passing to the next \((L_\beta + 1)-th\) iteration, let us estimate the sums over \(s\) in Eq. (89), considering Eq. (94)-(96):
\[ \sum_{s=0}^{\infty} c_{k,s} A_{k,j}^{[p,q]} \leq C \cdot F_j(L_\beta) \cdot C \sum_{s=0}^{\infty} A_{k,j}^{[p,q]} = C^2 F_j(L_\beta) \tilde{D}_j(L_\beta) \left( \frac{\beta - q}{|t_j(L_\beta)|} \right)^q, q < \beta. \] (98)

where the function \(\tilde{D}_j(l) > 0\) increases no faster than logarithmic function. For the case \(\tau_k = t_j(L_\beta)\), there will be one term (or more, if there is a degeneracy) for \(s = k_\beta\) from Eq. (97):
\[ c_{k,k_\beta} A_{k,j}^{[p,q]} \leq C^2 F_j(L_\beta) \left[ |t_j(L_\beta)| \right]^{-1}, q < \beta. \] (99)

Substituting Eq. (98) and Eq. (99) into the analogue of the expression in the third line of Eq. (89) but for the iteration \((L_\beta + 1)\), under the condition \(|\tau_k - t_j(L_\beta + 1)| \leq |\tau_k + t_j(L_\beta + 1)|\) we obtain:
\[ A_{k,j}^{[p,q+1]} \leq \frac{C^2 F_j(L_\beta)}{|t_j(L_\beta)|} \sum_{p=0}^{\beta-q} d_j^{p \cdot (p+1)} \left[ \frac{\tilde{D}_j(L_\beta)(\beta - (p + q) + \eta)}{|t_j(L_\beta) + \tau_k|} \right]^q \leq \frac{C^2 F_j(L_\beta)}{|t_j(L_\beta)|} \tilde{D}_j(L_\beta) \left[ \frac{\beta - q}{|t_j(L_\beta)|} \left( \frac{(\beta + q) - 1}{2} \right) + \eta \cdot (\beta + 1 - q) \right], \] (100)

where \(\eta\) is the order of degeneracy for \(\tau_k = t_j(L_\beta)\). For the case: \(|\tau_k - t_j(L_\beta + 1)| > |\tau_k + t_j(L_\beta + 1)|\), the expression \(|t_j(L_\beta + 1) + \tau_k|\) would be contained in the denominator of Eq. (100) instead of \(|t_j(L_\beta + 1) - \tau_k|\). A similar expression can be obtained for the case \(\tau_k = \tau_0 = 0\), with the only difference
that instead of the expression \(|\tau_k - t_j(L_\beta)\) the denominator will contain \(|t_j(L_\beta + 1)\) instead. The same differences will be observed in the next iterations, so we will not focus further on the cases \(|\tau_k - t_j(l)| > |\tau_k + t_j(l)|\) and \(\tau_k = \tau_0 = 0\).

Note that the term with the factor \((\beta - q)(\beta + 1 - q)\) in Eq. (100) contributes to \(A_{k,j}^{L_{\beta + 1,q}}\) only for \(q < \beta\), in accordance with the remark immediately after Eq. (89).

For the analogue of the expression in the second line of Eq. (89) corresponding to \(1\) \((\beta - q)(\beta + 1 - q)\) we have:

\[
\begin{aligned}
\sum_{s=0}^\infty c_{k,s}A_{k,s}^{L_{\beta + 1,q}} &\leq \frac{C^3 F_j(L_\beta)}{t_j(L_\beta) t_j(L_\beta + 1)} \left[\frac{\bar{D}_j(L_\beta)(\beta - q)(\beta + 1 - q)}{2} + \eta(\beta + 1 - q)\right],
\end{aligned}
\]

where we have included the term proportional to \(\bar{D}_j(L_\beta)\) from Eq. (101) for \(A_{k,s}^{L_{\beta + 1,q}}\).

For the case \(\tau_{k,s} = t_j(L_\beta + 1)\) and \(q = \beta + 1\), there will be one term (or more, if there is a degeneracy) for \(s = k_{\beta + 1}\) from Eq. (101):

\[
\begin{aligned}
|c_{k_{\beta + 1},s}A_{k_{\beta + 1},s}^{L_{\beta + 1,q}}| &\leq \frac{C^3 F_j(L_\beta)}{t_j(L_\beta) t_j(L_\beta + 1)} \frac{\bar{D}_j(L_\beta)(\beta - q)(\beta + 1 - q)}{2} + \eta(\beta + 1 - q).
\end{aligned}
\]

Substituting Eq. (102) and Eq. (103) into counterpart of the third line of Eq. (89) for the iteration \((L_\beta + 2)\), we obtain:

\[
\begin{aligned}
\sum_{p=0}^{\beta+q-1} \frac{d_j^p}{t_j(L_\beta + 1)} \left[\frac{\bar{D}_j(L_\beta + 1)\bar{D}_j(L_\beta)(\beta - q)(\beta + 1 - q) + \eta(\beta + 1 - q)}{2} + \eta(\beta + 1 - q)\right]
\end{aligned}
\]

When calculating the sum over index \(p\), we used the formula:

\[
\sum_{p=0}^{\beta+q-1} \frac{[\beta - (p + q)][\beta + 1 - (p + q)] \ldots \times [\beta + n - (p + q)]}{n + 2} = \frac{[\beta - q][\beta + 1 - q] \ldots \times [\beta + n + 1 - q]}{n + 2},
\]

which can be obtained from the relation for the sums of binomial coefficients: \(\sum_{k=0}^m \binom{n+k}{n} = \binom{n+m+1}{n+1}\) [44]. And we assume that the order of degeneracy \(\eta\) is the same for \(\tau_k = t_j(L_\beta + 1)\), if this is not the
case, then we choose as $\eta$ the maximum order of degeneracy. Continuing this process further, after $L_\beta + N$ iterations for the case $\tau_{k,x} \neq t_j(L_\beta + N)$, $|\tau_k - t_j(L_\beta + N)| > |\tau_i + t_j(L_\beta + N)|$, $\tau_k \neq 0$ we get:

$$
|A_{k,j}^{(L_\beta + N,q)}| \leq \frac{1}{|\tau_i||t_j(L_\beta + N) - \tau_i|} \cdot \frac{C^{N+1} F_j(L_\beta)}{\prod_{n=0}^{N} t_j(L_\beta + n)} \times \sum_{n=1}^{N+1} \prod_{p=1}^{n-1} \tilde{D}_j(L_\beta + N - p) \frac{\eta^{N+1-q}(\beta + N - q)!}{n!(\beta + N - n)!}.
$$

(106)

In the cases $|\tau_k - t_j(L_\beta + N)| > |\tau_k + t_j(L_\beta + N)|$ and $\tau_k = \tau_0 = 0$ expressions $|\tau_i||t_j(L_\beta + N) + \tau_i|$ and $|t_j(L_\beta + N)d_j|$ will be in the denominator of Eq. (106), respectively, instead of the expression $|\tau_i||t_j(L_\beta + N) - \tau_i|$. Considering that $[[|\tau_i||t_j(L_\beta + N) \pm \tau_i|^{-l} < \tilde{D}_j(L_\beta + N)](t_j(L_\beta + N)|d_j|)^{-l}$ and $\tilde{D}_j(l) \leq \tilde{D}_j \cdot \ln(l + 1)$, where $\tilde{D}_j > 0$ is a constant, we have:

$$
|A_{k,j}^{(L_\beta + N,q)}| \leq \frac{C^{N+1} F_j(L_\beta)|\text{Max}(1, \tilde{D}_j) \cdot \ln(L_\beta + N + 1)|^{N+1}\eta^N}{d_j^{N+1} \prod_{n=0}^{N} t_j(L_\beta + n)} \cdot \sum_{n=1}^{N+1} \frac{\eta^n(\beta + N)!}{n!(\beta + N - n)!}.
$$

(107)

By using the relation: $\sum_{n=0}^{N+1}(\beta + N) \leq \sum_{n=0}^{N} (\beta + N) = 2^{\beta + N}$ [44], expression (107) can be simplified:

$$
|A_{k,j}^{(L_\beta + N,q)}| \leq \frac{2^{\beta + N} C^{N+1} F_j(L_\beta)|\text{Max}(1, \tilde{D}_j) \cdot \ln(L_\beta + N + 1)|^{N+1}\eta^N}{d_j^{N+1} \prod_{n=0}^{N} t_j(L_\beta + n)}.
$$

(108)

A similar expression is obtained in the case of $\tau_{k,x} = t_j(L_\beta + N)$. Thus, taking into account that $|t_j(l)| \geq 2l$ and setting $l' = L_\beta + N$, we find that the d’Alembert ratio test for the majorizing series (90) will be fulfilled for indexes:

$$
I' > \gamma^2 C \cdot \eta \cdot \text{Max}(1, \tilde{D}_j) \cdot \ln(l' + 2) \cdot d_j^{-l}.
$$

(109)

This implies that series (67) converges absolutely for any finite $\gamma$, which completes the proof.

In the case when the equality $\tau_k = t_j(l)$ is realized every $\Delta l$ iterations, the step in the sum over $n$ in Eq. (106) will be equal to $\Delta l$ and the degree of the constant $\eta$ will decrease $\Delta l$ times. Thus, in this case, we obtain a criterion similar to expression (109).

Let us now turn to the case of angles that are multiples of irrational factors of $\pi$. The complexity of the proof here lies in the fact that in this case we cannot define a finite $d_j$, since for any $\hat{d}_j(l)$ there can be an $l' > l$ such that $\hat{d}_j(l') < \hat{d}_j(l)$. As a result it is not so easy to choose the majorizing series of the form Eq. (74) with fixed $d_j$, and it is necessary to carefully investigate how the numbers $\tau_k$ behave in this case, which is beyond the scope of this paper. On the other hand, the case of angles that are multiples of irrational factors of $\pi$ is rather of theoretical interest. In practical calculations we use numbers with a finite number of digits in the mantissa, thus rounding irrational numbers to rational ones. In addition, in real physical applications the angles are specified with some tolerance, which cannot be less than a certain physical limit, since material objects consist of vibrating atoms and molecules, and the exact values of the angles are just an abstraction.

In the conclusion of this section, we note that in numerical calculations it is more convenient to make the replacement $\hat{A}_{k,j}^{(l,q)} = d_j^q A_{k,j}^{(l,q)}$, so, instead of Eq. (88) we have:
\[ R_{k,j}^{(l)}(r) = R_{k,j}^{(l-1)}(r) + \gamma^{2l} r^{l/(l)} \sum_{q=0}^{Q} \frac{\tilde{A}_{k,j}^{[l,q]}}{q!} \left( \ln r \right)^q, \quad Q = \begin{cases} \beta_{i-1} + 1, & \tau_k = t_j(l) \\ \beta_{i-1}, & \text{otherwise} \end{cases}, \]  

where \( t_j(l) \equiv \tau_j + 2l \), the coefficients \( \tilde{A}_{k,j}^{[l,q]} \) are defined recurrently as nested series:

\[
\tilde{A}_{k,j}^{[l,q]} = \left\{ \begin{array}{l}
\sum_{j=0}^{\infty} \sum_{p=0}^{\beta_{i-1}-q} \frac{(-1)^{p+1} (p+1)}{t_j(l)^{p+2}} A_{k,j}^{(l-1,p+q)}, \quad \tau_k = 0
\\
\sum_{j=0}^{\infty} \sum_{p=0}^{\beta_{i-1}+q} \frac{(-1)^{p+1}}{2 t_j(l)^{p+1}} \tilde{A}_{k,j}^{(l-1,p+q-1)}, \quad \tau_k = t_j(l)
\\
\sum_{j=0}^{\infty} \sum_{p=0}^{\beta_{i-1}-q} \frac{[t_j(l) + \tau_k]^{p+1} - [t_j(l) - \tau_k]^{p+1}}{2 \tau_k \left( \tau_k^2 - t_j(l)^2 \right)^{p+1}} A_{k,j}^{(l-1,p+q)} \text{, otherwise}
\end{array} \right.
\]

and at the beginning of the iterative process \( R_{k,j}^{(0)}(r) = \delta_{k,j} r^{\tau_j}, \tilde{A}_{k,j}^{(0,q)} = \delta_{k,j} \cdot \delta_{q,0}, \beta_0 = 0 \). If in the process of calculating \( \tilde{A}_{k,j}^{[l,q]} \) at least once the equality \( \tau_k = t_j(l) \) occurs, then the number \( \beta \) as a result of this iteration increases by 1:

\[
\beta_t = \begin{cases} \beta_{i-1} + 1, & \text{if } \exists k, \tau_k = t_j(l) \\ \beta_{i-1}, & \text{otherwise} \end{cases}.
\]

Note also that in formulae (88)-(89) the number \( \beta \) was used as \( \beta_t \) in Eq. (111).

2.3 Calculation of radial functions \( R \) by direct solution of a system of differential equations

The first few radial functions can be also found by simpler but less accurate method, solving numerically the Cauchy problem for a system of differential equations. Let us represent the functions \( R_{k,j}(r) \)

\[
R_{k,j}(r) = R_{j}(r) \cdot f_{k,j}(r),
\]

From Eq. (66) it follows that \( f_{k,j}(r) \) is written as

\[
f_{k,j}(r) = \delta_{k,j} + \gamma^2 a_{k,j,1}(r) \cdot r^2 + \gamma^4 a_{k,j,2}(r) \cdot r^4 + \ldots,
\]

where the functions \( a_{k,j}(r) \) consist of the sum of the constants or the constants and the logarithms in an integer power. Therefore, since the power of the logarithms in any of the terms is less than that of \( r \), the initial conditions for the function \( f_{k,j}(r) \) and its derivative \( f'_{k,j}(r) \) will be

\[
f_{k,j}(0) = \delta_{k,j}, \quad f'_{k,j}(0) = 0.
\]

Substituting Eq. (113) into Eq. (12) we obtain for each \( j = 0..N_w - 1 \) a system of \( N_\Phi \) differential equations:

\[
\left[ r \partial_j \left( r \partial_j \right) + 2 \tau_j \partial_j \tau_j + \left( \tau_j^2 - \tau_k^2 \right) \right] f_{k,j}(r) + r^2 \gamma^2 \sum_{p=0}^{N_{k,j}-1} c_{k,j}^{(l)} f_{k,j}(r) = 0, \quad k = 0..N_\Phi - 1.
\]

Appendix 3. The system of linear equations for the coefficients in the analytical modal method taking into consideration the edge singularities

In numerical calculations we restrict ourselves in expansion (19) to the first \( M \) eigenfunctions, and in expansions (3) near each \( w \)-th of the total number of \( W_i \) angular singularities, situated at the
boundary with the coordinate $y_j$ between layers $l-1$ and $l$ in Fig. 2a, to the first $N_w$ functions. Substituting Eq. (19) for the field on the boundary between the layers $l-1$ and $l$ directly into Eq. (21), multiplying by complex conjugated test functions $\phi^*_j(x)$, which can be, for example, the first $M - \sum_{w=1}^{N_w} N_w / 2$ terms of the complex Fourier series and integrating over segments $[x_{2j}(y), x_{2j+1}(y))]$, $j = 0, 1, \ldots$ between regions with angular singularities (in Fig. 2a) these are segments $[x_0 \ldots x_1]$, $[x_2 \ldots x_3]$ and $[x_4 \ldots x_5]$), we obtain a system of $2M - \sum_{w=1}^{N_w} N_w$ linear equations:

$$
\begin{align*}
A_{l-1,y} & \left[ e_{l-1} e^*_l + \tilde{c}_{l-1} \right] = A_{l,y} \left[ e^*_l + e_{l} \tilde{c}^*_l \right] , \\
B_{l-1,y} & \left[ e_{l-1} e^*_l - \tilde{c}_{l-1} \right] = B_{l,y} \left[ e^*_l - e_{l} \tilde{c}^*_l \right] ,
\end{align*}
$$

(117)

where the diagonal matrix $e_{l} = \exp(ik_e\kappa \Delta y_j)$, $\Delta y_j = y_{j+1} - y_j$, the matrices $A_{l,y}$, $B_{l,y}$ have the elements:

$$
\begin{align*}
\left( A_{l,y} \right)_{p,m} &= \sum_{j=1}^{M} \int_{x_{2j-1}(y_j)}^{x_{2j+1}(y_j)} \phi^*_j(x) \psi_{l,m}(x) dx , \\
\left( B_{l,y} \right)_{p,m} &= \kappa_{l,m} \sum_{j=1}^{M} \int_{x_{2j-1}(y_j)}^{x_{2j+1}(y_j)} \phi^*_j(x) \psi_{l,m}(x) dx ,
\end{align*}
$$

(118)

and for the stability of the numerical solution of the linear system we made the substitution $c_{l,m} = \tilde{c}_{l,m} e_{l,m}$, see [31]. Physically this means that the expansion coefficients of upward and downward waves are defined at opposite sides of the layer. In layer $l-1$, at the boundary surrounding edge $w$ with center at point $(x_w, y_j)$ and radius $\rho_w$, conditions (21) are written:

$$
\begin{align*}
\sum_{m=1}^{M} \left[ e^*_l e_{l-1,m} e^*_{l-1,m}(\varphi) + \tilde{c}_{l-1,m} e^*_{l-1,m}(\varphi) \right] \psi_{l-1,m}(\varphi) &= \sum_{j=0}^{N_w} h_j \sum_{k=0}^{N_w} \Phi_{n}(\varphi) R_{k,j} , \\
i k_0 \sin \left( \sum_{m=1}^{M} \kappa_{l,m} \left[ e^*_l e_{l-1,m} e^*_{l-1,m}(\varphi) - \tilde{c}_{l-1,m} e^*_{l-1,m}(\varphi) \right] \psi_{l-1,m}(\varphi) \right) + \\
+ \cos \varphi \left( \sum_{m=1}^{M} \left[ e^*_l e_{l-1,m} e^*_{l-1,m}(\varphi) + \tilde{c}_{l-1,m} e^*_{l-1,m}(\varphi) \right] \psi_{l-1,m}(\varphi) \right) &= \sum_{j=0}^{N_w} h_j \sum_{k=0}^{N_w} \Phi_{n}(\varphi) \rho^0_{w,k} R'_{k,j} ,
\end{align*}
$$

(119)

where $e^*_l(\varphi) = \exp(\pm i k_0 \kappa_{l,m} \rho_w \sin \varphi)$ and the signs were chosen taking into account the fact that the $y$ axis is directed downward, the functions $\Phi_{n}(\varphi)$, $R_{k,j}(1)$ and the coefficients $h_j$ refer to the $w$-th edge, for brevity we do not denote this as an additional index, we also used the abbreviated notation $\psi_{l,m}(\varphi) = \psi_{l,m}(x_{w} + \rho_w \cos \varphi)$, $\psi'_{l,m}(\varphi) = \partial_\varphi \psi_{l,m}(x_{w} + \rho_w \cos \varphi)$. When writing Eq. (119) we took into account that $\partial_\varphi R_{k,j}(\varphi / \rho_w) |_{\rho_w=\rho^0} = \rho^0_{w,k} R'_{k,j}(1)$ and truncate $R_{k,j}$ to a square matrix by setting $N_\varphi = N_w$. Multiplying Eq. (119), and similar equations written for a part of the boundary around an edge in the layer $l$, by complex conjugate adjoint eigenfunctions $\Phi^*_n(\varphi)$ with weight $\epsilon^{-1}(\varphi)$ and by integrating over the angle, taking into account orthogonality, we obtain a system of $2N_w$ equations:

$$
\begin{align*}
A^+_l e_{l-1} c_{l-1} + A^+_l e^*_l c^*_{l-1} + A^+_l e^*_l c^*_{l-1} + A^+_l e^*_l c^*_{l-1} = R \cdot h, \\
B^+_l e_{l-1} c_{l-1} - B^+_l e^*_l c^*_{l-1} + B^+_l e^*_l c^*_{l-1} - B^+_l e^*_l c^*_{l-1} = \rho^0 \epsilon^{-1} R' \cdot h ,
\end{align*}
$$

(120)

where the matrix elements $A^+_l e_{l-1,n,m}$, $B^+_l e_{l-1,n,m}$, $A^+_l e^*_l n,m$ and $B^+_l e^*_l n,m$ are:

$$
\begin{align*}
A^+_l e_{l-1,n,m} &= \int_{\pi}^{2\pi} \Phi^*_n e^*_l e^*_{l-1,m}(\varphi) e^*_{l-1,m}(\varphi) d\varphi, \\
B^+_l e_{l-1,n,m} &= \int_{\pi}^{2\pi} \Phi^*_n e^*_l e^*_l e^*_{l-1,m}(\varphi) e^*_{l-1,m}(\varphi) d\varphi, \\
A^+_l e^*_l n,m &= \int_{\pi}^{2\pi} \Phi^*_n e^*_l e^*_l e^*_{l-1,m}(\varphi) e^*_{l-1,m}(\varphi) d\varphi, \\
B^+_l e^*_l n,m &= \int_{\pi}^{2\pi} \Phi^*_n e^*_l e^*_{l-1,m}(\varphi) e^*_{l-1,m}(\varphi) + \cos \varphi \psi'_{l-1,m}(\varphi) e^*_{l-1,m}(\varphi) d\varphi
\end{align*}
$$

(121)
By expressing the coefficients $h$ of the expansion inside the region containing the edge in terms of the coefficients $c$, we obtain $N_u$ equations:

$$\left( A^{-1}_{n-1} - D^{-1}B_{n-1}^{-1}\right)e_{n-1}c_{n-1} + \left( A^{-1}_{n-1} + D^{-1}B_{n-1}^{-1}\right)c_{n-1} + \left( A^{-1}_{n} - D^{-1}B_{n}^{-1}\right)c_{n} + \left( A^{-1}_{n} + D^{-1}B_{n}^{-1}\right)e_{n}c_{n} = 0 , \quad (122)$$

where the matrix $D = \rho_0^{-1}R'(1)R^{-1}(1)$ is the Dirichlet-to-Neumann map operator matrix in the representation of the functions $\Phi_i(\varphi)$, see Appendix 4. After the same is done for the remaining edges between the layers $l-1$ and $l$ we obtain together with Eq. (117) $2M$ equations in total. In the same way the equations for the coefficients of all layers are written. The right-hand side is obtained by substituting the known coefficients $c_{0,m}^+$ for the incident monochrome wave, where $\delta(0,m)$ is the Kronecker symbol, and the zero value for the coefficients $c_{l}^-$ of the last layer in Eq. (117) and Eq. (122), and transferring the result to the right-hand side.

Appendix 4. Calculation of the Dirichlet-to-Neumann map operator matrix and the solution of the Dirichlet problem for a domain containing an edge

To obtain the DtN map operator matrix $D_s$ of a region containing an edge and restricted by boundary $\rho = \rho_0$, it is convenient to use the functions $G_j(\rho / \rho_0 = 1, \varphi), \quad j = 0..N_u - 1$, Eq. (9), as a set of linearly independent boundary conditions. Matrices of these linearly independent vectors of boundary conditions and the corresponding normal derivatives at $N_u$ points $\varphi_i$, for which we can take the points of zeros of orthogonal polynomials, for example, Legendre or Chebyshev in each sector of uniform permittivity, will have the elements:

$$F_{i,j} = G_j(1, \varphi_i) = \sum_{k=0}^{N_u-1} \Phi_k(\varphi_i)R_{k,j}(1), \quad \partial_n F_{i,j} = \partial_n G_j(1, \varphi_i) = \sum_{k=0}^{N_u-1} \Phi_k(\varphi_i)\rho_0^{-1}R'_{k,j}(1). \quad (123)$$

Whence, using Eq. (23) we obtain the DtN map operator matrix $D_s$ acting on the functions defined at the collocation points $\varphi_i$.

It is also easy to obtain the DtN map operator matrix $D_s^{\Phi_i}$ in the representation of functions $\Phi_i(\varphi)$, when this matrix acts on the expansion coefficients $c_i$ of the conditions $f_i(\varphi)$ at the boundary $\rho = \rho_0$ in the functions $\Phi_i(\varphi)$, and as a result we obtain the expansion coefficients $c_i'$ of the corresponding normal derivative $\partial_n f_i(\varphi)$ in these functions:

$$f_i(\varphi) = \sum_{j=0}^{N_u-1} \Phi_i(\varphi)c_j, \quad \partial_n f_i(\varphi) = \sum_{j=0}^{N_u-1} N_j(\varphi)c_j = \sum_{j=0}^{N_u-1} \Phi_i(\varphi)c_j', \quad (124)$$

where $N_j(\varphi)$ are normal derivatives of the solutions of Eq. (1) corresponding to the boundary conditions $\Phi_i(\varphi)$. Multiplying the last equality in Eq. (124) by the complex conjugate adjoint eigenfunctions $\Phi_i^*(\varphi)$ and integrating over $\varphi$ with weight $\varepsilon^{-1}(\varphi)$, taking into account orthogonality, we find the expression for the coefficients of the matrix $D_s^{\Phi_i}$:

$$c_j' = \sum_{j=0}^{N_u-1} \left\{ \int_0^{2\pi} \Phi_i^*(\varphi)\varepsilon^{-1}(\varphi)N_j(\varphi)\varphi d\varphi \right\} c_j = \sum_{j=0}^{N_u-1} D_s^{\Phi_i}(\varphi) c_j'. \quad (125)$$

To find the normal derivatives $N_j(\varphi)$, we formally decompose $\Phi_i(\varphi)$ in functions $G_j(1, \varphi)$ (9):

$$\Phi_i(\varphi) = \sum_{j=0}^{N_u-1} h_j G_j(1, \varphi) = \sum_{j=0}^{N_u-1} h_j \sum_{k=0}^{N_u-1} \Phi_k(\varphi)R_{k,j}(1) \quad \text{(126)}$$
Multiplying Eq. (126) by \( \Phi_i^*(\varphi) \) and integrating over \( \varphi \) with weight \( \varepsilon^{-1}(\varphi) \), gives us the coefficient matrix: \( h = R^{-1}(1) \). Using the found coefficients \( h_{i,j} \), we can write the expression for the normal derivatives \( N_i(\varphi) \) corresponding to \( \Phi_i(\varphi) \) on the boundary \( \rho = \rho_0 \):

\[
N_i(\varphi) = \sum_{j=0}^{N_{i}-1} h_{j,i} \partial_n G_j(1,\varphi) = \sum_{j=0}^{N_{i}-1} R_{j,i}^{-1}(1) \sum_{k=0}^{N_{i}-1} \Phi_k(\varphi) \rho_0^{-1} R_{k,j}(1).
\]

(127)

Thus, substituting \( N_i(\varphi) \) into Eq. (125) we obtain the DtN map operator matrix \( D_{s}^{(\Phi)} \) in the representation of functions \( \Phi_i(\varphi) \):

\[
D_{s}^{(\Phi)} = \rho_0^{-1} R'(1) R^{-1}(1).
\]

(128)

It is more convenient to work in the representation of coefficients, since it does not require equality of the number of collocation points at adjacent boundaries. In this work we used this very representation, connecting firstly all the elements around a region containing an edge into the single element with the DtN map operator matrix \( D \) by means of (25) and after that changing representation on the boundary of this region from the values at collocation points \( \varphi_i \) to the values of expansion coefficients in the eigenfunctions \( \{\Phi_k(\varphi_i)\} \). Then we join it to the region containing an edge characterized by DtN map operator matrix \( D_{s}^{(\Phi)} \).

To change the representation in DtN map operator matrix \( D \) at part \( \Gamma_s \) of the boundary to the basis functions \( \{\Phi_i(\varphi)\} \) representation without changing the representation on the other part \( \Gamma \) of the boundary, let us to represent an arbitrary boundary condition \( f_{\Gamma_s}(\varphi) \) and its corresponding normal derivative \( \partial_n f_{\Gamma_s}(\varphi) \) in the basis functions \( \{\Phi_k(\varphi)\} \) at Gauss or Gauss-Lobatto quadrature points \( \varphi_i \). In matrix form this can be written:

\[
f_{\Gamma_s} = \Phi \cdot c_{\Gamma_s}, \quad \partial_n f_{\Gamma_s} = \Phi \cdot c'_{\Gamma_s},
\]

(129)

where the matrix \( \Phi \) consists of the columns of the basis functions \( \Phi_i(\varphi) \), and \( c_{\Gamma_s}, c'_{\Gamma_s} \) are the coefficients columns. The action of the DtN map operator matrix \( D \) on the function at the border, which includes the parts \( \Gamma_s \) and \( \Gamma \), will be

\[
\begin{pmatrix}
D_{1,1} & D_{1,2} \\
D_{2,1} & D_{2,2}
\end{pmatrix}
\begin{pmatrix}
f_{\Gamma_s} \\
\partial_n f_{\Gamma_s}
\end{pmatrix} =
\begin{pmatrix}
\Phi & 0 \\
0 & I
\end{pmatrix}
\begin{pmatrix}
c_{\Gamma_s} \\
\partial_n c_{\Gamma_s}
\end{pmatrix},
\]

(130)

where \( f_{\Gamma} \) and \( \partial_n f_{\Gamma} \) are the columns of the function values and the corresponding normal derivative on the boundary part \( \Gamma \), \( I \) is the identity matrix. Multiplying Eq. (130) by the transposed matrix of complex conjugate adjoint eigenfunctions \( \tilde{\Phi} \) with the diagonal matrix of the corresponding integration weights \( \mathbf{w} \), we obtain the DtN map operator matrix \( D \) in the mixed representation:

\[
\begin{pmatrix}
\tilde{\Phi}w & 0 \\
0 & I
\end{pmatrix}
\begin{pmatrix}
\Phi & 0 \\
0 & I
\end{pmatrix}
= 
\begin{pmatrix}
I & 0 \\
0 & I
\end{pmatrix}
\Rightarrow
\begin{pmatrix}
\tilde{\Phi}wD_{1,1} & \tilde{\Phi}wD_{1,2} \\
D_{2,1} & D_{2,2}
\end{pmatrix}
\begin{pmatrix}
c_{\Gamma_s} \\
\partial_n c_{\Gamma_s}
\end{pmatrix} =
\begin{pmatrix}
c'_{\Gamma_s} \\
\partial_n c'_{\Gamma_s}
\end{pmatrix}.
\]

(131)

If the matrix \( \Phi \) is square, then for \( \tilde{\Phi}w \) we can simply take the inverse matrix \( \Phi^{-1} \). If the number of quadrature points \( \varphi_i \) is greater than the number of basis functions \( \{\Phi_k(\varphi_i)\} \), then equalities in (131) will be fulfilled with spectral accuracy inherent to Gauss quadratures.

After the conditions \( b(\varphi_i) \) in the points \( \varphi_i \) of the boundary \( \rho = \rho_0 \) of the region containing the edge were received, the coefficients \( h_j \) of the expansion of the field (15) in the functions \( G_j(\rho / \rho_0 = 1, \varphi_i) \) can be found from the equations:

\[
b(\varphi_i) = \sum_{j=0}^{N_{i}-1} h_j G_j(\rho / \rho_0 = 1, \varphi_i) \Leftrightarrow G \cdot h = b.
\]

(132)
If the boundary conditions are represented in the form of expansion coefficients in the functions $\Phi_i(\varphi)$:

$$b(\varphi) = \sum_{j=0}^{N_w-1} \Phi_i(\varphi)c_j = \sum_{j=0}^{N_w-1} h_j G_j(\rho / \rho_0 = 1, \varphi) = \sum_{j=0}^{N_w-1} h_j \sum_{k=0}^{N_w-1} \Phi_k(\varphi)R_{k,j}(\rho / \rho_0 = 1)$$  \hfill (133)

then, multiplying on the left side by $\Phi_i^*(\varphi)$, and integrating with the weight $\epsilon^{-1}(\varphi)$, we obtain a similar equation for the coefficients $h_j$ in the matrix notation:

$$c_i = \sum_{j=0}^{N_w-1} R_{i,j}(\rho / \rho_0 = 1)h_j \iff \mathbf{R} \cdot \mathbf{h} = \mathbf{c}.$$  \hfill (134)

Acknowledgments

The author is grateful to Prof. Lifeng Li for his valuable comments on the paper.

References

[1] J. G. Van Bladel, Singular Electromagnetic Fields and Sources (IEEE Press Series on Electromagnetic Wave Theory), 1st ed. Wiley-IEEE Press, 1996.

[2] K. M. Gundu and A. Mafi, Reliable computation of scattering from metallic binary gratings using Fourier-based modal methods, J. Opt. Soc. Am. A 27 (7) (2010) 1694-1700, https://doi.org/10.1364/JOSAA.27.001694.

[3] L. Li, Fourier modal method, in: E. Popov, ed., Gratings: Theory and Numeric Applications, 2nd rev. ed., Institute Fresnel, AMU 2014, Chap. 13, pp. 537-578 (www.fresnel.fr/numerical-grating-book-2).

[4] R. E. Collin, Reflection and transmission at a slotted dielectric interface. Canadian Journal of Physics 34 (4) (1956) 398-411, https://doi.org/10.1139/p56-047

[5] L. C. Botten, M. S. Craig, R. C. McPhedran, J. L. Adams & J. R. Andrewartha, The Dielectric Lamellar Diffraction Grating, Opt. Acta 28 (3) (1981) 413-428, http://dx.doi.org/10.1080/713820571.

[6] L. C. Botten, M. S. Craig, R. C. McPhedran, J. L. Adams, and J. R. Andrewartha, The finitely conducting lamellar diffraction grating, Opt. Acta 28 (8) (1981) 1087-1102, http://dx.doi.org/10.1080/713820680.

[7] L. Li, Field singularities at lossless metal–dielectric arbitrary-angle edges and their ramifications to the numerical modeling of gratings, J. Opt. Soc. Am. A 29 (4) (2012) 593-604, https://doi.org/10.1364/JOSAA.29.000593.

[8] L. Li and G. Granet, Field singularities at lossless metal-dielectric right-angle edges and their ramifications to the numerical modeling of gratings, J. Opt. Soc. Am. A 28 (5) (2011) 738-746, http://dx.doi.org/10.1364/JOSAA.28.000738.

[9] M. S. Bobrovnikov and V. P. Zamarayeva, On the singularity of the field near a dielectric wedge, Sov. Phys. J. 16 (1973) 1230–1232, https://doi.org/10.1007/BF00890884.

[10] M. Bressan and P. Gamba, Analytical expressions of field singularities at the edge of four right wedges, IEEE Microw. Guided Wave Lett. 4 (1) (1994) 3-5, http://dx.doi.org/10.1109/75.267676.

[11] K. M. Gundu and A. Mafi, Constrained least squares Fourier modal method for computing scattering from metallic binary gratings, J. Opt. Soc. Am. A 27 (11) (2010) 2375-2380, https://doi.org/10.1364/JOSAA.27.002375.
[12] G. Granet, Reformulation of the lamellar grating problem through the concept of adaptive spatial resolution, J. Opt. Soc. Am. A 16 (10) (1999) 2510-2516. 
http://dx.doi.org/10.1364/JOSAA.16.002510.

[13] G. Granet, J. Chandezon, J.-P. Plumey, and K. Raniriharinosy, Reformulation of the coordinate transformation method through the concept of adaptive spatial resolution. Application to trapezoidal gratings, J. Opt. Soc. Am. A 18 (9) (2001) 2102-2108, 
https://doi.org/10.1364/JOSAA.18.002102.

[14] G. Bao, Z. Chen, and H. Wu, Adaptive finite-element method for diffraction gratings, J. Opt. Soc. Am. A 22 (6) (2005) 1106-1114, https://doi.org/10.1364/JOSAA.22.001106

[15] S.C. Brenner, J. Gedicke and LY. Sung, An Adaptive $P_1$ Finite Element Method for Two-Dimensional Transverse Magnetic Time Harmonic Maxwell's Equations with General Material Properties and General Boundary Conditions, J Sci Comput 68 (2016) 848–863
https://doi.org/10.1007/s10915-015-0161-x

[16] A.-S. Bonnet-Ben Dhia, C. Hazard, and S. Lohrengel, A singular field method for the solution of maxwell's equations in polyhedral domains, SIAM Journal on Applied Mathematics, 59 (6) (1999) 2028–2044, https://doi.org/10.1137/S0036139997323383

[17] F. Assous, P. Ciarlet and E. Sonnendrücker, Resolution of the Maxwell equations in a domain with reentrant corners, ESAIM:M2AN 32(3) (1998) 359-389, 
https://doi.org/10.1051/m2an/1998320303591

[18] F. Assous, P. Ciarlet, S. Labrunie, J. Segré, Numerical solution to the time-dependent Maxwell equations in axisymmetric singular domains: the singular complement method, J. of Comp. Phys., 191(1) (2003) 147-176, https://doi.org/10.1016/S0021-9991(03)00309-7

[19] M. Costabel and M. Dauge, Weighted regularization of Maxwell equations in polyhedral domains, Numerische Mathematik, 93(2) (2002) 239–277, https://doi.org/10.1007/s0021100100388

[20] A.-S. Bonnet-Ben Dhia, C. Carvalho, L. Chesnel, P. Ciarlet, On the use of Perfectly Matched Layers at corners for scattering problems with sign-changing coefficients, J. of Comp. Phys, 322, (2016) 224-247, https://doi.org/10.1016/j.jcp.2016.06.037

[21] G. A. Grinberg, Selected Problems in the Mathematical Theory of Electric and Magnetic Phenomena [in Russian], Moscow-Leningrad, Izd-vo AN SSSR, 1948.

[22] J. Meixner, Die Kantenbedingung in der Theorie der Beugung elektromagnetischer Wellen an vollkommen leitenden ebenen Schirmen. Ann. Phys. 441 (1) (1950) 1-9. 
https://doi.org/10.1002/andp.19494410103.

[23] J. P. Boyd, Chebyshev and Fourier spectral methods, 2nd ed. Dover, 2001.

[24] J. Meixner, The behavior of electromagnetic fields at edges, IEEE Trans. Antennas Propagat., 20 (4) (1972) 442–446, https://doi.org/10.1109/TAP.1972.1140243.

[25] G.I. Makarov, A.V. Osipov, Structure of Meixner's series. Radiophys. Quantum Electron. 29 (6) (1986) 544–549, https://doi.org/10.1007/BF01035092.

[26] E. Marx, Logarithmic terms in fields near the edge of a dielectric wedge, International Symposium on Antennas and Propagation Society, Merging Technologies for the 90's, Dallas, TX, USA, vol.3 (1990), pp. 1083-1086, https://doi.org/10.1109/APS.1990.115299.

[27] A. Quarteroni and A. Valli, Domain Decomposition Methods for Partial Differential Equations, Oxford Science Publications, 1999.

[28] Y. Wu and Y. Y. Lu, Analyzing diffraction gratings by a boundary integral equation Neumann-to-Dirichlet map method, J. Opt. Soc. Am. A 26 (11) (2009) 2444-2451, 
https://doi.org/10.1364/JOSAA.26.002444.
[29] R. Mennicken, M. Möller, Non-Self-Adjoint Boundary Eigenvalue Problems, Volume 192, 1st ed. North-Holland, 2003

[30] G. Tayeb & R. Petit, On the Numerical Study of Deep Conducting Lamellar Diffraction Gratings, Opt. Acta 31 (12) (1984), 1361-1365, http://dx.doi.org/10.1080/713821455.

[31] I. Semenikhin, M. Zanuccoli, Application of the iterative approach to modal methods for the solution of Maxwell’s equations, J. of Comp. Phys. 300 (2015) 438-454, https://doi.org/10.1016/j.jcp.2015.07.052.

[32] Lord Rayleigh, On the Dynamical Theory of Gratings, Proc. R. Soc. Lond. A 79 (532) (1907) 399-416, https://doi.org/10.1098/rspl.1907.0051.

[33] L. Li and C. W. Haggans, Convergence of the coupled-wave method for metallic lamellar diffraction gratings, J. Opt. Soc. Am. A 10(6) (1993) 1184-1189, https://doi.org/10.1364/JOSAA.10.001184.

[34] P. Lalanne and J.-P. Hugonin, Numerical performance of finite-difference modal methods for the electromagnetic analysis of one-dimensional lamellar gratings, J. Opt. Soc. Am. A 17 (6) (2000) 1033-1042, http://dx.doi.org/10.1364/JOSAA.17.001033.

[35] K. Edee, Modal method based on subsectional Gegenbauer polynomial expansion for lamellar gratings, J. Opt. Soc. Am. A 28 (10) (2011) 2006–2013, http://dx.doi.org/10.1364/JOSAA.28.002006.

[36] W. Lu and Y. Y. Lu, High order integral equation method for diffraction gratings, J. Opt. Soc. Am. A 29 (5) (2012) 734-740, https://doi.org/10.1364/JOSAA.29.000734.

[37] L. Li, Hypersingularity, electromagnetic edge condition, and an analytic hyperbolic wedge model, J. Opt. Soc. Am. A 31 (4) (2014) 808-817, https://doi.org/10.1364/JOSAA.31.000808.

[38] A. T. Patera, A spectral element method for fluid dynamics: Laminar flow in a channel expansion, J. of Comp. Phys. 54 (3) (1984) 468-488, https://doi.org/10.1016/0021-9991(84)90128-1.

[39] R. Peyret, Spectral Methods for Incompressible Viscous Flow, Springer, New York, (2002), https://doi.org/10.1007/978-1-4757-6557-1.

[40] C. Canuto, M. Y. Hussaini, A. Quarteroni, Th. A. Zang, Spectral Methods Evolution to Complex Geometries and Applications to Fluid Dynamics., Springer, Berlin, (2007), https://doi.org/10.1007/978-3-540-30728-0.

[41] I. Semenikhin, Efficient method for the solution of Maxwell’s equations for nanostructured materials, ITM Web Conf. 30 08007 (2019), https://doi.org/10.1051/itmconf/20193008007.

[42] Y. K. Ko and J. C. Inkson, Matrix method for tunneling in heterostructures: Resonant tunneling in multilayer systems, Phys. Rev. B 38 (14) (1988) 9945-9951, https://link.aps.org/doi/10.1103/PhysRevB.38.9945.

[43] R J. Chandezon, D. Maystre, and G. Raoult, A new theoretical method for diffraction gratings and its numerical application, J. Opt. (Paris) 11(4) (1980) 235–241 https://doi.org/10.1088/0150-536x/11/4/005.

[44] I. S. Gradshteyn and I. M. Ryzhik, Table of Integrals, Series, and Products, Daniel Zwillinger and Victor Moll (eds.), 8th ed., Academic Press, 2014, p.4.