Numerische Mathematik manuscript No.
(will be inserted by the editor)

Accuracy of staircase approximations in finite-difference methods for wave propagation

Jon Häggblad · **Olof Runborg**

Received: date / Accepted: date

Abstract While a number of increasingly sophisticated numerical methods have been developed for time-dependent problems in electromagnetics, the Yee scheme is still widely used in the applied fields, mainly due to its simplicity and computational efficiency. A fundamental drawback of the method is the use of staircase boundary approximations, giving inconsistent results. Usually experience of numerical experiments provides guidance of the impact of these errors on the final simulation result. In this paper, we derive exact discrete solutions to the Yee scheme close to the staircase approximated boundary, enabling a detailed theoretical study of the amplitude, phase and frequency errors created. Furthermore, we show how evanescent waves of amplitude $O(1)$ occur along the boundary. These characterize the inconsistencies observed in electromagnetic simulations and the locality of the waves explain why, in practice, the Yee scheme works as well as it does. The analysis is supported by detailed proofs and numerical examples.

Keywords FDTD · Yee · staircasing

Mathematics Subject Classification (2000) 35L05 · 65N22 · 78M20

1 Introduction

One of the more popular methods for numerically solving wave propagation problems is the Yee scheme, also sometimes referred to as the Finite-Difference Time-Domain (FDTD) method. Originally devised for Maxwell's equations in electromagnetics [18], it has since also found many uses in acoustic simulations [1, 11, 15]. It is a simple algorithm based on compact centered finite difference approximations on

J. Häggblad · O. Runborg
Department of Numerical Analysis, KTH Royal Institute of Technology, SE-10044 Stockholm, Sweden,
E-mail: olofr@kth.se

O. Runborg
Swedish e-Science Research Centre (SeRC), Stockholm, Sweden.

a uniform staggered grid. The scheme is explicit, making it highly efficient, as well as having a small memory footprint since the field values are not stored on all discretization points. A drawback of this approach is that boundaries not aligned with the grid are difficult to model, and hence are usually approximated by staircasing. Not only is this a poor approach, it is inconsistent [16]. Hence, in general the scheme will produce $O(1)$ errors. Despite this, it is still common in applications, due to its simplicity.

Previous work on the analysis of the numerical errors caused by staircase approximations include the oft-cited paper by Cangellaris and Wright [2], where they show that staircasing of a boundary with $\pi/4$ inclination compared to the grid admit surface waves. Holland [8] shows numerically the large errors generated in practical problems. The issue is also discussed in a number of publications regarding the development of more accurate boundary approximations [4,7,9,14,17].

Since staircase approximations of boundaries are still common in applications, a more detailed theoretical understanding is worthwhile rather then to rely on numerical experiments. In this paper we study a two-dimensional model problem for numerical solution of electromagnetic waves by the classical Yee scheme with a boundary approximated by staircasing. This model problem includes boundaries of all rational inclinations.

Despite the fact that the most popular use of the Yee scheme is in electromagnetics, we shall instead use the acoustic wave equation formulation, as this gives a simpler notation. We will mainly focus on perfect electric conductor (PEC) boundaries for both the TM and TE modes of Maxwell's equations. These correspond to stress release and perfectly rigid boundaries in acoustics.

In two dimensions the acoustic equations are given by

$$p_t = a(u_x + v_y), \qquad u_t = bp_x, \qquad v_t = bp_y, \tag{1.1}$$

where p denotes the pressure and u, v are the x- and y-components of the velocity field. We will assume that a and b are constant. We also introduce the velocity $c = \sqrt{ab}$. The equations (1.1) are equivalent to the two-dimensional transverse magnetic (TM) and transverse electric (TE) vacuum modes of Maxwell's equations under the set of substitutions $p = E_z$, $u = H_y$, $v = -H_x$ and $p = H_z$, $u = -E_y$, $v = -E_x$, respectively, together with $a = 1/\varepsilon$, $b = 1/\mu$. Note that in three dimensions there is no such equivalence.

We will consider (1.1) set in a semi infinite open domain $\Omega = \{(x,y) \in \mathbb{R}^2 : y > \alpha x\}$, where $0 < \alpha < 1$. On the boundary $\Gamma = \{(x,y) \in \mathbb{R}^2 : y = \alpha x\}$, we will prescribe boundary conditions. We consider homogeneous Dirichlet conditions both in the form $p = 0$, referred to as *soft boundaries*, and $\hat{\mathbf{n}} \cdot (u,v) = 0$, referred to as *hard boundaries*, where $\hat{\mathbf{n}} \perp (1,\alpha)$ is the unit normal vector. These two forms of boundary conditions correspond to PEC boundaries for the TM and TE mode, respectively, in two-dimensional electromagnetics. The equations are complemented with initial data for p, u and v.

In the analysis we will restrict ourselves to *rational* slopes α which lie between zero and one, i.e., we write $\alpha = \mu/\nu$, where μ and ν are two positive, relatively prime integers with $0 < \mu < \nu$. The cases $\alpha = 0$ and $\alpha = 1$ are quite special, since the staircase approximation of the boundary becomes of higher order than in the other cases [16]. The errors in the Yee scheme are then much more benign, see Remark 4.1.

Our aim is to analyze the errors generated in staircase approximations by deriving exact discrete modal solutions. As far as the authors know, for staircase boundaries this has not been done before.

A number of boundary modeling techniques have been created to improve the accuracy of the Yee scheme for curved boundaries without introducing non-orthogonal coordinates or unstructured grids. One of the earliest is the contour-path FDTD method (CP-FDTD) [9], which in its initial form was plagued by late-time instabilities, regardles of the timestep used [10]. This can be fixed by e.g. adding a term to the update equations [12]. Later another scheme was introduced by Dey and Mittra [3], usually dubbed locally conformal FDTD (CFDTD), which is much simpler. This class of methods involve weighting the update stencil according to the fraction of the sides of the unit cube which are inside the domain. See [13] for an overview. Another approach is to remove the highest order term in the Taylor expansion of the local truncation error [16,5,6]. A big motivation for studying the boundary errors in detail in this paper is to give further insights and hopefully open the door to additional improvements to the above techniques.

Note that, although commonly used to analyze numerical stability, we here use the modal solutions to study the typical errors in the Yee scheme. That the Yee scheme is stable for a staircase boundary is already known, see e.g. [5].

Our main results of the analysis is that away from boundaries the total error is dominated by a first order error stemming from an error in the effective (discrete) reflection coefficient of the staircased boundary. Close to the boundaries, on the other hand, an $O(1)$ error is present. It comes from evanescent waves that concentrate on the boundary, but die off exponentially with the number of grid points away from the boundary. Hence, the $O(1)$ errors produced by the inconsistent discretization will be localized at the boundary. Our conclusion is that this explains why, in practice, the Yee scheme works as well as it does. We note also that the convergence rate in L^2 norm is formally reduced to $O(\sqrt{h})$ due to the large errors at the boundary.

While the model problem might at first seem oversimplified, we argue that asymptotically it is still applicable to general boundaries and domains, since the critical effect of the boundary on the accuracy is independent of the grid size h. We also show a numerical example with a more general domain where we see that the numerical results are consistent with the conclusions of the analysis.

The article is organized as follows. After stating the Yee scheme in Section 2, we derive the necessary conditions for modal solutions in Section 3, giving explicit expressions first for soft boundaries, and then hard boundaries. We then proceed to analyze the accuracy by looking at the asymptotic behavior at fine discretizations in Section 4. In Section 5 we collect the proofs of the analysis. We finish by verifying the analysis with numerical tests in Section 6.

2 The Yee scheme

We introduce the Yee staggered grid as follows. Let $x_m = mh$, $y_j = jh$ and $t_n = n\Delta t$, where h is the grid spacing and Δt is the time step. Denote by I_{mj} the Yee cell $[x_m, x_{m+1}] \times [y_j, y_{j+1}]$. The unknowns are approximated in different spatial locations in the cell: u, v on the cell boundary and p in the center of the cell. Moreover, p is approximated half a time step off from u and v. More precisely,

$$p^{n+\frac{1}{2}}_{m+\frac{1}{2},j+\frac{1}{2}} \approx p(t_n + \Delta t/2, x_m + h/2, y_j + h/2),$$
$$u^n_{m,j+\frac{1}{2}} \approx u(t_n, x_m, y_j + h/2), \qquad v^n_{m+\frac{1}{2},j} \approx v(t_n, x_m + h/2, y_j).$$

Discretizing (1.1) on this staggered grid gives the Yee scheme, which for interior cells reads

$$p^{n+\frac{1}{2}}_{m+\frac{1}{2},j+\frac{1}{2}} = p^{n-\frac{1}{2}}_{m+\frac{1}{2},j+\frac{1}{2}} + a\lambda \left(u^n_{m+1,j+\frac{1}{2}} - u^n_{m,j+\frac{1}{2}} + v^n_{m+\frac{1}{2},j+1} - v^n_{m+\frac{1}{2},j} \right), \tag{2.1}$$

$$u^{n+1}_{m,j+\frac{1}{2}} = u^n_{m,j+\frac{1}{2}} + b\lambda \left(p^{n+\frac{1}{2}}_{m+\frac{1}{2},j+\frac{1}{2}} - p^{n+\frac{1}{2}}_{m-\frac{1}{2},j+\frac{1}{2}} \right), \tag{2.2}$$

$$v^{n+1}_{m+\frac{1}{2},j} = v^n_{m+\frac{1}{2},j} + b\lambda \left(p^{n+\frac{1}{2}}_{m+\frac{1}{2},j+\frac{1}{2}} - p^{n+\frac{1}{2}}_{m+\frac{1}{2},j-\frac{1}{2}} \right), \tag{2.3}$$

where $\lambda = \Delta t/h$.

The boundary is approximated by staircasing where the boundary cells are those whose centers are just inside the domain, i.e. the set of boundary cells Γ_C is defined as $\Gamma_C = \left\{ I_{mj} : \alpha(x_m + h/2) \leq y_j < \alpha(x_m + h/2) + h \right\}$.

We furthermore define the indices for these cells as (m, j_m) so that

$$\Gamma_C = \bigcup_m I_{m,j_m}, \qquad j_m = \lceil (m + 1/2)\alpha - 1/2 \rceil . \tag{2.4}$$

The boundary cells are illustrated in Fig. 2.1. We note also that since $\alpha = \mu/\nu$ the indices satisfy

$$j_{m+\nu} = j_m + \mu. \tag{2.5}$$

The discretization in the boundary cells depends on the type of boundary conditions chosen. For all cases we consider, the stencil is only altered in the boundary cells, not in any other cells. The precise discretization will be detailed later on.

3 Modal solutions of the Yee scheme

The aim is to study the behavior of numerical solutions of (2.1), (2.2) and (2.3) around staircase approximations of boundaries in the Yee scheme. To this end we want to find exact discrete modal solutions for a given time frequency ω.

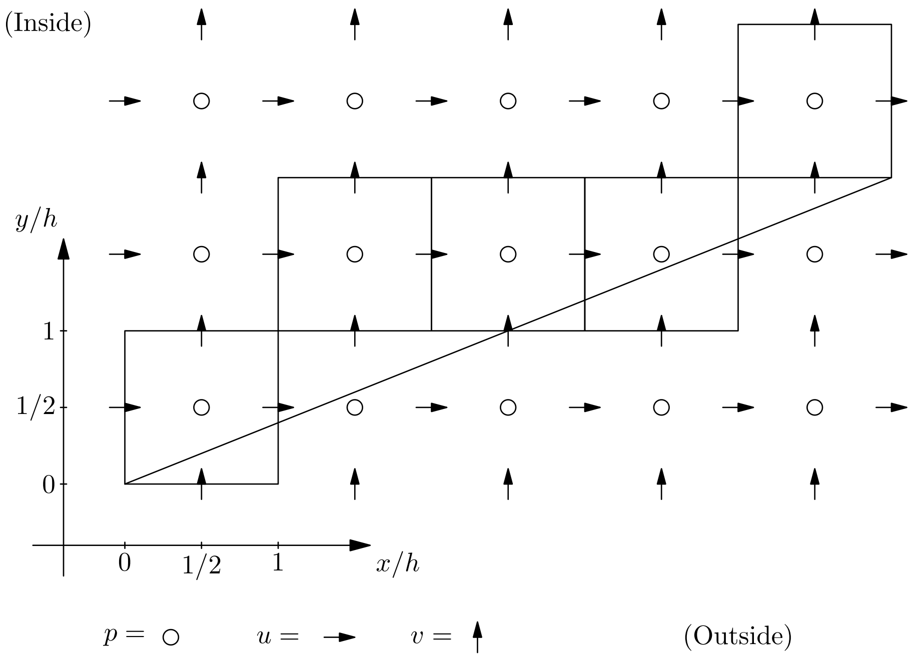

Fig. 2.1 Illustrating the boundary cells. The cells marked by squares are the boundary cells Γ_C and thus are adjacent to the staircased boundary.

3.1 Continuous case

To motivate our derivation of the discrete modes we consider first the continuous case. We fix a time frequency ω and seek solutions of the form

$$\bar{\mathbf{w}}(t,x,y) := \begin{pmatrix} p(t,x,y) \\ u(t,x,y) \\ v(t,x,y) \end{pmatrix} = e^{i\omega t} \begin{pmatrix} \bar{P}(x,y) \\ \bar{U}(x,y) \\ \bar{V}(x,y) \end{pmatrix} =: e^{i\omega t} \bar{\mathbf{W}}(x,y),$$

where $\bar{\mathbf{W}}(x,y)$ is the modal solution. This corresponds to the frequency space solutions given by Helmholtz equation. We can also think of $\bar{\mathbf{W}}$ as the Fourier transform in time of $\bar{\mathbf{w}}(t,x,y)$ at ω. Propagating modes are given by sums of exponentials,

$$\bar{\mathbf{W}}(x,y) = \bar{\mathbf{w}}_{\text{in}} e^{ik_x x + ik_y y} + \beta \bar{\mathbf{w}}_{\text{ref}} e^{ik'_x x + ik'_y y}, \tag{3.1}$$

where (recall $c^2 = ab$)

$$\bar{\mathbf{w}}_{\text{in}} = \begin{pmatrix} 1 \\ bk_x/\omega \\ bk_y/\omega \end{pmatrix}, \qquad \bar{\mathbf{w}}_{\text{ref}} = \begin{pmatrix} 1 \\ bk'_x/\omega \\ bk'_y/\omega \end{pmatrix},$$

which represents an incoming and a reflected wave. The wave numbers k_x, k_y, k'_x and k'_y should satisfy the dispersion relation

$$c^2(k_x^2 + k_y^2) = c^2({k'_x}^2 + {k'_y}^2) = \omega^2, \tag{3.2}$$

and the exponentials should be equal when evaluated on the boundary,

$$k_x + \alpha k_y = k_x' + \alpha k_y' \mod 2\pi. \tag{3.3}$$

The value of β is determined by the chosen boundary condition. For soft boundaries $\beta = -1$ and for hard boundaries $\beta = 1$. We note that these modes can be parametrized by the temporal frequency ω and the value of $K = k_x + \alpha k_y$, which corresponds to the angle of incidence with the boundary. The same turns out to be true for the discrete modes. For these parameters we assume that there is a real parameter $\eta > 0$ such that

$$|K| \leq \eta \frac{|\omega|}{\sqrt{1+\alpha^2}}, \qquad |\omega| \geq \eta, \tag{3.4}$$

which essentially means that we only consider waves hitting the boundary at an angle. These conditions also mean that (3.2)–(3.3) has two distinct solutions.

Remark 3.1 There are also non-propagating modal solutions with zero ω. These correspond to the non-zero rotational part of the solution, which is stationary. They are of the form

$$\bar{\mathbf{W}}(x,y) = \bar{\mathbf{w}}_{\text{stat}} e^{ik_x x + ik_y y}, \qquad \bar{\mathbf{w}}_{\text{stat}} = \begin{pmatrix} 0 \\ -k_y \\ k_x \end{pmatrix},$$

For soft boundaries all such modes are valid. For hard boundaries they are restricted to the case $k_x + \alpha k_y = 0$.

3.2 Admissible discrete modes

As in the continuous case we fix a temporal frequency ω and seek a discrete modal solution $\mathbf{W}$ such that

$$\mathbf{w}_{m,j}^n := \begin{pmatrix} p_{m,j}^n \\ u_{m,j}^n \\ v_{m,j}^n \end{pmatrix} = e^{i\omega n\Delta t}\mathbf{W}(m,j) := e^{i\omega n\Delta t} \begin{pmatrix} P(m,j) \\ U(m,j) \\ V(m,j) \end{pmatrix}.$$

The modal solutions will be parameterized by ω and a value K, which corresponds to the angle of incidence with the boundary, as above. Note that in this definition the discrete solution can be evaluated at any real values of (m,j); in the Yee scheme only certain discrete values are used, and they are different for P, U and V.

To find admissible discrete modal solutions we will first consider the free space case and find what restrictions are imposed. Motivated by the continuous case we look for a solution of the form

$$\mathbf{W}(m,j) := \hat{\mathbf{w}}E(m,j), \qquad \hat{\mathbf{w}} = \begin{pmatrix} \hat{p} \\ \hat{u} \\ \hat{v} \end{pmatrix}, \qquad E(m,j) = e^{ik_x mh + ik_y jh},$$

where (k_x, k_y) are unknown wave numbers. For the actual grid functions used in the scheme, this means

$$
\begin{aligned}
p^{n+\frac{1}{2}}_{m+\frac{1}{2},j+\frac{1}{2}} &= e^{i\omega(n+\frac{1}{2})\Delta t}\hat{p}E\left(m+1/2,j+1/2\right),\\
u^{n+1}_{m,j+\frac{1}{2}} &= e^{i\omega(n+1)\Delta t}\hat{u}E\left(m,j+1/2\right),\\
v^{n+1}_{m+\frac{1}{2},j} &= e^{i\omega(n+1)\Delta t}\hat{v}E\left(m+1/2,j\right).
\end{aligned}
$$

Then for the time derivative

$$
\frac{e^{i\omega(n+1)\Delta t}-e^{i\omega n\Delta t}}{\Delta t} = i\frac{\sin\left(\omega\Delta t/2\right)}{\Delta t/2}e^{i\omega(n+\frac{1}{2})\Delta t} =: i\tilde{\omega}(\Delta t)e^{i\omega(n+\frac{1}{2})\Delta t},
$$

where tilde (˜) denotes the mapping

$$
\tilde{x}(y) = \frac{\sin xy/2}{y/2}. \tag{3.5}
$$

Similarly for the space derivatives,

$$
\begin{aligned}
\frac{E(m+1,j)-E(m,j)}{h} &= i\tilde{k}_x(h)E\left(m+1/2,j\right),\\
\frac{E(m,j+1)-E(m,j)}{h} &= i\tilde{k}_y(h)E\left(m,j+1/2\right).
\end{aligned}
$$

Entering this into the scheme (2.1)–(2.3) we obtain

$$
\begin{aligned}
i\tilde{\omega}\hat{p}E\left(m+1/2,j+1/2\right)e^{i\omega n\Delta t} &= ia\left(\tilde{k}_x\hat{u}+\tilde{k}_y\hat{v}\right)E\left(m+1/2,j+1/2\right)e^{i\omega n\Delta t},\\
i\tilde{\omega}\hat{u}E\left(m,j+1/2\right)e^{i\omega(n+\frac{1}{2})\Delta t} &= ib\tilde{k}_x\hat{p}E\left(m,j+1/2\right)e^{i\omega(n+\frac{1}{2})\Delta t},\\
i\tilde{\omega}\hat{v}E\left(m+1/2,j\right)e^{i\omega(n+\frac{1}{2})\Delta t} &= ib\tilde{k}_y\hat{p}E\left(m+1/2,j\right)e^{i\omega(n+\frac{1}{2})\Delta t},
\end{aligned}
$$

which simplifies to the linear equations for $\hat{p}$, $\hat{u}$ and $\hat{v}$,

$$
\tilde{\omega}\hat{p} = a\left(\tilde{k}_x\hat{u}+\tilde{k}_y\hat{v}\right), \qquad \tilde{\omega}\hat{u} = b\tilde{k}_x\hat{p}, \qquad \tilde{\omega}\hat{v} = b\tilde{k}_y\hat{p}.
$$

We have a nontrivial solution when the system matrix of these equations is singular, that is when

$$
\det\begin{pmatrix}\tilde{\omega} & -a\tilde{k}_x & -a\tilde{k}_y\\ -b\tilde{k}_x & \tilde{\omega} & 0\\ -b\tilde{k}_y & 0 & \tilde{\omega}\end{pmatrix} = \tilde{\omega}\left(c^2(\tilde{k}_x^2+\tilde{k}_y^2)-\tilde{\omega}^2\right) = 0.
$$

When $\tilde{\omega}\neq 0$ we have propagating modes and the null space is spanned by the vector

$$
\hat{\mathbf{w}} = \begin{pmatrix}1\\ b\tilde{k}_x/\tilde{\omega}\\ b\tilde{k}_y/\tilde{\omega}\end{pmatrix}.
$$

The corresponding $\mathbf{W}$ can be written as

$$
\mathbf{W}(m,j) = \begin{pmatrix}1\\ b\tilde{k}_x/\tilde{\omega}\\ b\tilde{k}_y/\tilde{\omega}\end{pmatrix}E(m,j).
$$

Based on the analysis above we will call a discrete propagating modal solution of this type *admissible* for ω and K if (k_x,k_y) satisfy the discrete dispersion relation

$$c^2(\tilde{k}_x^2+\tilde{k}_y^2)=\tilde{\omega}^2, \tag{3.6}$$

and the boundary relation

$$k_x+\alpha k_y = K \mod \frac{2\pi}{h}. \tag{3.7}$$

We will also simply say that (k_x,k_y) are admissible for the real numbers ω and K if (3.6) and (3.7) hold. Note that (3.6) can be expanded to

$$c^2\lambda^2\left(\sin^2\frac{hk_x}{2}+\sin^2\frac{hk_y}{2}\right)=\sin^2\frac{\omega\Delta t}{2}, \qquad \lambda=\frac{\Delta t}{h}.$$

Since the solution does not change if we add an integer multiple $2\pi/h$ to the real part of k_x or k_y we will further assume that

$$0\le \text{Re } k_x < \frac{2\pi}{h}, \qquad 0\le \text{Re } k_y < \frac{2\pi}{h}. \tag{3.8}$$

In Section 5, Theorem 5.1, we shall see that the system (3.6) and (3.7) always has 2ν solutions when (k_x,k_y) are restricted to this set. Moreover, for small enough h there are $\nu-1$ solutions with a negative imaginary part in k_x or k_y, which corresponds to exponentially growing waves. These we discard. We denote the remaining solutions by (k_x^r,k_y^r), $r=0,\dots,\nu$.

With this we are now ready to consider some explicit forms of staircasing.

3.3 Soft boundaries

First we consider homogeneous Dirichlet conditions in the p variable, which are sometimes referred to as soft or stress release boundaries in the acoustic community. In electromagnetics they correspond to PEC boundaries for the TM mode. Thus the boundary condition in the continuous case is given by $p(t,x,y)=0, y=\alpha x, x\in\mathbb{R}$. The corresponding numerical boundary condition is

$$p^{n+\frac{1}{2}}_{m+\frac{1}{2},j_m+\frac{1}{2}}=0, \qquad \forall m\in\mathbb{Z}.$$

We will derive modal solutions which satisfy this for a fixed pair ω and K. As mentioned in the previous section there are $\nu+1$ admissible pairs (k_x^r,k_y^r) which satisfy (3.6) and (3.7) and are bounded, for each choice of ω and K, when h is small enough. We denote the P part of these modes by $P_r(m,j)=e^{ik_x^r mh+ik_y^r jh}$, $r=0,\dots,\nu$. From these we take a linear combination and set

$$P(m,j)=\sum_{r=0}^{\nu}\alpha_r P_r(m,j),$$

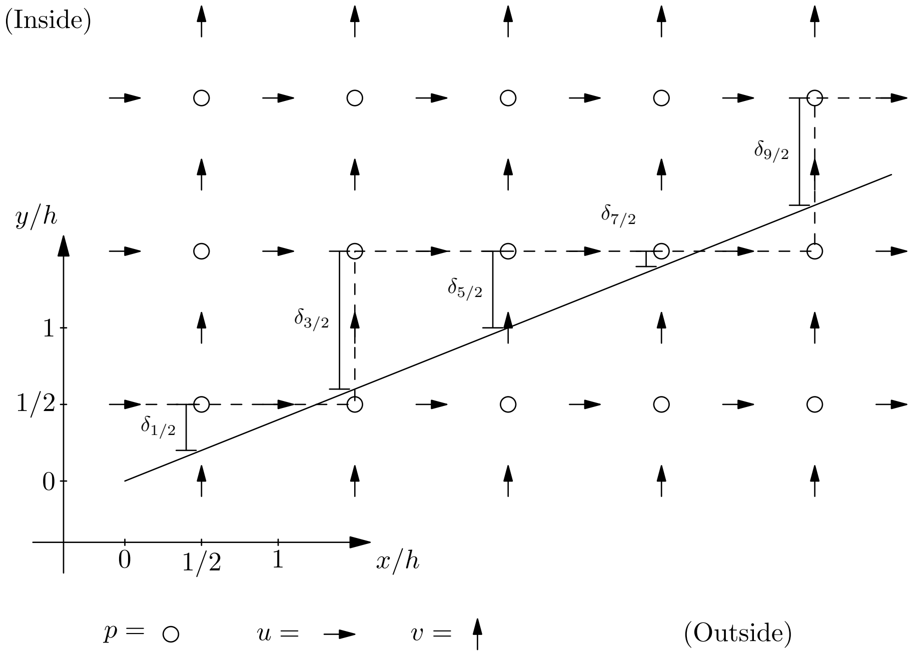

Fig. 3.1 Staircasing of soft boundaries and $\mu = 2, \nu = 5$. The boundary is defined by the pressure points at $\left(\frac{1}{2},\frac{1}{2}\right)$, $\left(\frac{3}{2},\frac{3}{2}\right)$, $\left(\frac{5}{2},\frac{3}{2}\right)$, $\left(\frac{7}{2},\frac{3}{2}\right)$, $\left(\frac{9}{2},\frac{5}{2}\right)$.

where α_r are to be determined such that

$$P\left(m+\frac{1}{2}, j_m+\frac{1}{2}\right) = e^{-i\omega(n+\frac{1}{2})\Delta t} p^{n+\frac{1}{2}}_{m+\frac{1}{2},j_m+\frac{1}{2}} = 0, \qquad \forall m \in \mathbb{Z}.$$

Let us now define the offsets $\delta_{m+\frac{1}{2}}$ between the exact boundary and the staircase approximation at these points, $\delta_{m+\frac{1}{2}} = j_m + 1/2 - \alpha(m+1/2)$. See Fig. 3.1. Then $0 \le \delta_{m+\frac{1}{2}} < 1$ and $\delta_{m+\nu+\frac{1}{2}} = \delta_{m+\frac{1}{2}}$ by the definition of j_m in (2.4) and by (2.5). We obtain

$$\begin{aligned}
P\left(m+\frac{1}{2}, j_m+\frac{1}{2}\right) &= \sum_{r=0}^{\nu} \alpha_r e^{ik_x^r\left(m+\frac{1}{2}\right)h + ik_y^r\left(j_m+\frac{1}{2}\right)h} \\
&= \sum_{r=0}^{\nu} \alpha_r e^{i\left(k_x^r+\alpha k_y^r\right)\left(m+\frac{1}{2}\right)h + ik_y^r \delta_{m+\frac{1}{2}} h} \\
&= e^{iK\left(m+\frac{1}{2}\right)h} \sum_{r=0}^{\nu} \alpha_r e^{ik_y^r \delta_{m+\frac{1}{2}} h}.
\end{aligned}$$

We note that the function within the sum is ν-periodic in m and it is therefore sufficient to enforce the zero condition for $m = 0, \ldots, \nu-1$. To further simplify, we note that $\delta_{m+\frac{1}{2}}\nu = (j_m + 1/2 - \alpha(m+1/2))\,\nu = j_m\nu - m\mu + (\nu-\mu)/2 := d_m + \bar{d}$, where

$$d_m = j_m\nu - m\mu + \left\lfloor \frac{\nu-\mu}{2} \right\rfloor \in \mathbb{N}, \qquad \bar{d} = \left\{ \frac{\nu-\mu}{2} \right\}. \tag{3.9}$$

Here $\{x\}$ denotes the fractional part of x. Note that $\bar{d}$ is independent of m. Thus if we define $z_r := \exp\left(ik_y^r h/\nu\right)$, we get the condition

$$P\left(m+\frac{1}{2}, j_m+\frac{1}{2}\right) = e^{iK\left(m+\frac{1}{2}\right)h} \sum_{r=0}^{\nu} \alpha_r z_r^{d_m+\bar{d}} = 0, \qquad m = 0,\ldots,\nu-1.$$

We can write this in matrix form as $A\boldsymbol{\alpha} = 0$, where

$$A = ZS \in \mathbb{C}^{\nu\times(\nu+1)}, \qquad \boldsymbol{\alpha} = (\alpha_0,\ldots,\alpha_\nu)^T \in \mathbb{C}^{\nu+1}, \tag{3.10}$$

and

$$Z = \begin{pmatrix} z_0^{d_0} & \cdots & z_\nu^{d_0} \\ \vdots & \ddots & \vdots \\ z_0^{d_{\nu-1}} & \cdots & z_\nu^{d_{\nu-1}} \end{pmatrix} \in \mathbb{C}^{\nu\times(\nu+1)}. \tag{3.11}$$

$$S = \operatorname{diag}(z_0^{\bar{d}},\ldots,z_\nu^{\bar{d}}) \in \mathbb{C}^{(\nu+1)\times(\nu+1)}, \tag{3.12}$$

The coefficients $\boldsymbol{\alpha}$ are thus given by first finding a vector in the null space of Z and then scaling it by the nonsingular diagonal matrix S^{-1}.

From these coefficients α we thus have the full modal solution as

$$\boldsymbol{W}(m,j) = \sum_{r=0}^{\nu} \alpha_r \hat{\mathbf{w}}_r E_r(m,j), \tag{3.13}$$

where

$$E_r(m,j) = e^{ik_x^r mh + ik_y^r jh}, \qquad \hat{\mathbf{w}}_r = \begin{pmatrix} 1 \\ b\tilde{k}_x^r/\tilde{\omega} \\ b\tilde{k}_y^r/\tilde{\omega} \end{pmatrix}.$$

In Theorem 5.1 it is proved that the null space of A is one-dimensional, and that we can always obtain a unique (up to normalization) non-trivial solution $\boldsymbol{\alpha}$ bounded in h.

3.4 Hard boundaries

Next we consider another common type of boundary conditions, which is homogeneous Dirichlet conditions for the normal component of the velocity field, $\hat{\mathbf{n}}\cdot(u,v) = 0$, for $y = \alpha x$, $x \in \mathbb{R}$. In the acoustics community this is sometimes referred to as hard boundaries, and it is equivalent to PEC boundaries for the TE mode in electromagnetics.

In the discretization we use the same boundary cells (2.4) as in the case of soft boundaries. However, the boundary cells are divided into two sets according to

$$\Omega_{\mathrm{cr}} = \{m : j_{m+1} = j_m + 1\}, \qquad \Omega_{\mathrm{hz}} = \{m : j_{m+1} = j_m\}, \tag{3.14}$$

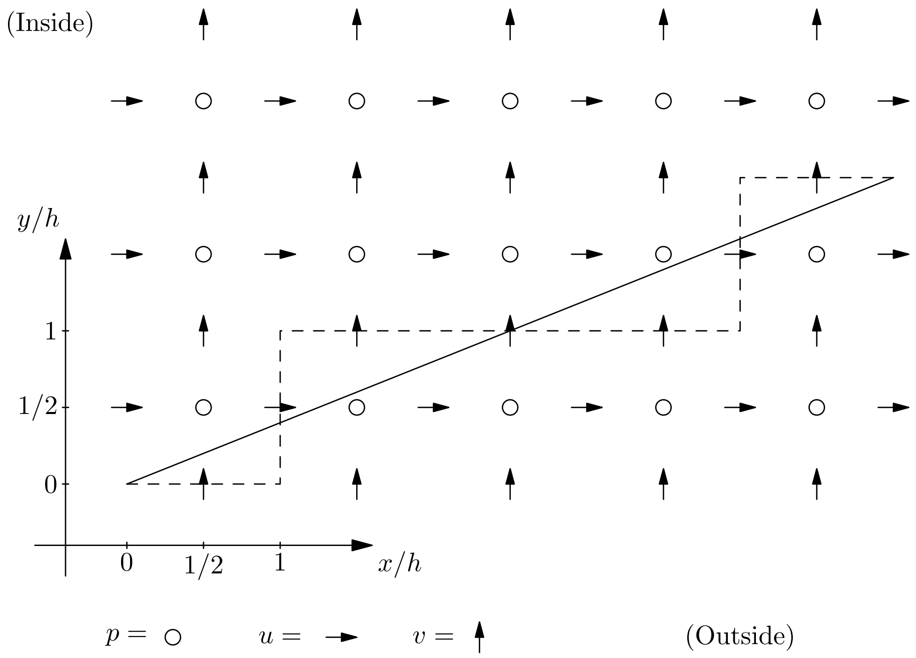

Fig. 3.2 Staircasing of hard boundaries for $\mu = 2$, $\nu = 5$. In the case shown, the boundary cells centered at $x/h = 1/2, 7/2$ are corner cells, in that the boundary occur along both u and v. In the cells centered around $x/h = 3/2, 5/2, 9/2$, the boundary only occurs along v.

where Ω_{cr} refers to the corner cells with two faces adjacent to the boundary, and Ω_{hz} the remaining cells. See Fig. 3.2 for an illustration. Then the staircase boundary condition is enforced by

$$u_{m+1,j_m+\frac{1}{2}} - v_{m+\frac{1}{2},j_m} = 0, \qquad \forall m \in \Omega_{\mathrm{cr}} \tag{3.15}$$

$$v_{m+\frac{1}{2},j_m} = 0, \qquad \forall m \in \Omega_{\mathrm{hz}}. \tag{3.16}$$

Again we look for modal solutions satisfying this for a fixed pair ω and K. Thus, using the wave numbers defined by (3.6)–(3.7) we denote the U and V part of these modes by

$$U_r(m,j) = \frac{b\tilde{k}_x^r}{\omega} e^{ik_x^r mh + ik_y^r jh}, \qquad V_r(m,j) = \frac{b\tilde{k}_y^r}{\omega} e^{ik_x^r mh + ik_y^r jh},$$

for $r = 0, \ldots, \nu$. If we evaluate these expressions on the boundary points, then they reduce to

$$\begin{aligned}
\frac{\omega}{b} U_r(m+1, j_m + \tfrac{1}{2}) &= \tilde{k}_x^r e^{ik_x^r (m+1)h + ik_y^r \left(j_m + \frac{1}{2}\right)h} \\
&= \tilde{k}_x^r e^{ik_x^r (m+1)h + ik_y^r \left(\delta_{m+\frac{1}{2}} + \alpha\left(m + \frac{1}{2}\right)\right)h} \\
&= \tilde{k}_x^r e^{iK\left(m+\frac{1}{2}\right)} e^{ik_x^r h/2 + ik_y^r \delta_{m+\frac{1}{2}} h}
\end{aligned} \tag{3.17}$$

$$\begin{aligned}\frac{\omega}{b}V_r(m+\tfrac{1}{2},j_m) &= \tilde{k}_y^r e^{ik_x^r\left(m+\frac{1}{2}\right)h+ik_y^r j_m h}\\ &= \tilde{k}_y^r e^{ik_x^r\left(m+\frac{1}{2}\right)h+ik_y^r\left(\delta_{m+\frac{1}{2}}-\frac{1}{2}+\alpha\left(m+\frac{1}{2}\right)\right)h}\\ &= \tilde{k}_y^r e^{iK\left(m+\frac{1}{2}\right)}e^{ik_y^r\left(\delta_{m+\frac{1}{2}}-\frac{1}{2}\right)h}. \qquad (3.18)\end{aligned}$$

Recalling that $z_r := \exp\left(ik_y^r h/\nu\right)$ and $z_r^{d_m+\bar{d}} = \exp\left(ik_y^r\delta_{m+\frac{1}{2}}h\right)$, we can now write the linear combination of (3.17) and (3.18) as

$$\begin{aligned}\frac{\omega}{b}U(m+1,j_m+\tfrac{1}{2}) &= e^{iK\left(m+\frac{1}{2}\right)h}\sum_{r=0}^{\nu}\alpha_r\tilde{k}_x^r e^{ik_x^r h/2}z_r^{d_m+\bar{d}},\\ \frac{\omega}{b}V(m+\tfrac{1}{2},j_m) &= e^{iK\left(m+\frac{1}{2}\right)h}\sum_{r=0}^{\nu}\alpha_r\tilde{k}_y^r e^{-ik_y^r h/2}z_r^{d_m+\bar{d}}.\end{aligned}$$

From these expressions we can formulate the discrete boundary conditions (3.15)–(3.16) as

$$\sum_{r=0}^{\nu}\alpha_r\left(\tilde{k}_x^r e^{ik_x^r h/2}-\tilde{k}_y^r e^{-ik_y^r h/2}\right)z_r^{d_m+\bar{d}} = 0, \quad \forall m\in\Omega_{\mathrm{cr}}, \qquad (3.19)$$

$$\sum_{r=0}^{\nu}\alpha_r\tilde{k}_y^r e^{-ik_y^r h/2}z_r^{d_m+\bar{d}} = 0, \qquad \forall m\in\Omega_{\mathrm{hz}}. \qquad (3.20)$$

Thus since these expressions are ν-periodic in m, we only need to satisfy the zero conditions for $m = 0,\dots,\nu-1$. We also point out that there are μ number of equations (3.19), and $\nu-\mu$ number of equations (3.20). We can write (3.19)–(3.20) as a linear system $A\boldsymbol{\alpha} = 0$, with

$$A = \left(QZK(\tilde{k}_x)+ZK(-\tilde{k}_y)\right)S\in\mathbb{C}^{\nu\times(\nu+1)}, \qquad (3.21)$$

and where

$$K(k) = \mathrm{diag}(k^0e^{ik^0h/2},\dots,k^\nu e^{ik^\nu h/2})\in\mathbb{C}^{(\nu+1)\times(\nu+1)}, \qquad k\in\mathbb{C}^{\nu+1},$$

and $Q\in\mathbb{C}^{\nu\times\nu}$ is the diagonal matrix with

$$(Q)_{mm} = \begin{cases}1 & m\in\Omega_{\mathrm{cr}},\\ 0 & m\notin\Omega_{\mathrm{cr}}.\end{cases}$$

In Theorem 5.1 we prove that $(k_x^r,k_y^r)\sim 1/h$, for $r\geq 2$. Therefore we rescale the K matrices as $K' = KD$, where $D = \mathrm{diag}(1,1,h,\dots,h)$. Then by using (3.5) we see that the (diagonal) elements in $K'(k_x)$ and $K'(-k_y)$ for $r\geq 2$ are given by,

$$\begin{aligned}h\tilde{k}_x^r e^{ik_x^r h/2} &= i(1-wz_r^{-\mu}),\\ -h\tilde{k}_y^r e^{-ik_y^r h/2} &= i(1-z_r^{-\nu}),\end{aligned}$$

respectively, where $w := e^{iKh}$, and we use $k_x^r = K-\alpha k_y^r$.

As in the previous case, the full modal solution is then given by

$$W(m,j) = \sum_{r=0}^{\nu} \alpha_r \hat{\mathbf{w}}_r E_r(m,j). \tag{3.22}$$

Remark 3.2 Unlike the situation with soft boundaries, there could potentially be situations where (3.21) is singular. In particular, consider the case when $\nu = \mu = 1$, giving the angle $\alpha = 1$, together with incoming waves along the boundary. For this case then $A \to (0,0)$ when $h \to 0$, which falls outside the theory covered here. This case is, however, considered in [2].

4 Error estimates

In computations involving the Yee scheme, $O(1)$ errors are sometimes observed around boundaries [16]. The aim here is to give precise expressions for the errors in the discrete modal solutions compared to the continuous modal solutions, i.e.

$$\text{error} = \mathbf{W}(m,j) - \bar{\mathbf{W}}(mh, jh),$$

for a given ω and K. The wave vectors for the continuous modes are denoted by $\bar{\mathbf{k}} = (\bar{k}_x, \bar{k}_y)$, $\bar{\mathbf{k}}' = (\bar{k}'_x, \bar{k}'_y)$, and the admissible waves by $\mathbf{k}^r = (k^r_x, k^r_y)$. To normalize the modes in the same way we always let $\alpha_0 = 1$ in the discrete modal expressions, (3.13), (3.22). We can then divide the error as follows

$$\begin{aligned} \text{error} &= \sum_{r=0}^{\nu} \alpha_r \hat{\mathbf{w}}_r e^{ik^r_x mh + ik^r_y jh} - \bar{\mathbf{w}}_{\text{in}} e^{i\bar{k}_x mh + i\bar{k}_y jh} + \bar{\mathbf{w}}_{\text{ref}} e^{i\bar{k}'_x mh + i\bar{k}'_y jh} \\ &= E_{\text{in}} + E_{\text{ref}} + E_{\text{phase}} + E_{\text{eva}}, \end{aligned}$$

where

$$\begin{aligned} E_{\text{in}} &= \hat{\mathbf{w}}_0 e^{ik^0_x mh + ik^0_y jh} - \bar{\mathbf{w}}_{\text{in}} e^{i\bar{k}_x mh + i\bar{k}_y jh}, \\ E_{\text{ref}} &= -\hat{\mathbf{w}}_1 e^{ik^1_x mh + ik^1_y jh} + \bar{\mathbf{w}}_{\text{ref}} e^{i\bar{k}'_x mh + i\bar{k}'_y jh}, \\ E_{\text{phase}} &= (\alpha_1 + 1)\hat{\mathbf{w}}_1 e^{ik^1_x mh + ik^1_y jh}, \\ E_{\text{eva}} &= \sum_{r=2}^{\nu} \alpha_r \hat{\mathbf{w}}_r e^{ik^r_x mh + ik^r_y jh}. \end{aligned}$$

The first two terms are the error in the free space wave propagation. The third term is the error in the reflection coefficient of the approximate boundary, which adds an extra phase factor. The fourth term is the error from the evanescent waves which concentrate on the boundary.

These errors can be explained by using Theorem 5.1 in Section 5. For the free space propagation error we use the result that $\mathbf{k}^0(h) - \bar{\mathbf{k}} = O(h^2)$. Then

$$\begin{aligned} |E_{\text{in}}| &\le |\hat{\mathbf{w}}_0 - \bar{\mathbf{w}}_{\text{in}}| + |\bar{\mathbf{w}}_{\text{in}}| \left| e^{ik^0_x mh + ik^0_y jh} - e^{i\bar{k}_x mh + i\bar{k}_y jh} \right| \\ &= \frac{b}{\omega} \left| \frac{\omega}{\tilde{\omega}} \tilde{\mathbf{k}}^0(h) - \bar{\mathbf{k}} \right| + |\bar{\mathbf{w}}_{\text{in}}| \left| e^{i(\mathbf{k}^0(h) - \bar{\mathbf{k}})\cdot(mh, jh)} - 1 \right| \le Ch^2, \end{aligned}$$

since $\omega/\tilde{\omega} = 1 + O(\Delta t^2)$. Precisely the same argument can be made for E_{ref}. Note that the bars $|\cdot|$ here denote the point wise Euclidean norm in $\mathbb{R}^3$. Hence, the propagation errors in Yee are second order, regardless of boundary conditions, as expected.

For the remaining errors we must consider the particular boundary conditions separately. We find below in Sections 4.1 and 4.2 that for both hard and soft boundaries the phase error E_{phase} is of first order, while the evanescent wave error E_{eva} is of zeroth order on the boundary but decays exponentially with the distance, measured in grid points, away from the boundary. The total error is therefore $O(h)$ away from boundaries, where it is dominated by E_{phase}, and $O(1)$ in the grid points close to the boundary, where it is dominated by E_{eva}. When measured in L^2 norm, this implies an error of size $O(\sqrt{h})$. We recall that the staircase discretization of the boundaries is formally inconsistent and that, therefore, $O(1)$ errors in general will appear. Our conclusion here is that the localization of these errors at the boundaries explains why the staircase discretization still works well for many problems in practice.

4.1 Soft boundaries

As derived in Section 3.3, the system of equations for the modal solutions when we have soft boundaries is $A\boldsymbol{\alpha} = ZS\boldsymbol{\alpha} = 0$, where A, S and Z are given in (3.10), (3.12) and (3.11). We note first that by Theorem 5.1 the null space of Z is one-dimensional so the direction of $\boldsymbol{\alpha}$ is well-defined. To normalize it we fix the first component $\alpha_0 = 1$. Let $\tilde{\boldsymbol{\alpha}} = (\alpha_1, \ldots, \alpha_\nu)^T$ be the remaining part of $\boldsymbol{\alpha}$ and similarly let $\tilde{Z}$, $\tilde{S}$ be the parts of Z and S where the first columns have been removed,

$$\tilde{Z} = \begin{pmatrix} z_1^{d_0} & \cdots & z_\nu^{d_0} \\ \vdots & \ddots & \vdots \\ z_1^{d_{\nu-1}} & \cdots & z_\nu^{d_{\nu-1}} \end{pmatrix}, \qquad \tilde{S} = \mathrm{diag}(z_1^{\bar{d}}, \ldots, z_\nu^{\bar{d}}). \tag{4.1}$$

Then

$$\tilde{Z}(h)\tilde{S}(h)\tilde{\boldsymbol{\alpha}}(h) = -z_0(h)^{\bar{d}} \begin{pmatrix} z_0(h)^{d_0} \\ \vdots \\ z_0(h)^{d_{\nu-1}} \end{pmatrix}.$$

Define $\boldsymbol{\alpha}' = (-1, 0, \ldots, 0)^T \in \mathbb{R}^\nu$. Then

$$\tilde{Z}(h)\tilde{S}(h)(\tilde{\boldsymbol{\alpha}}(h) - \boldsymbol{\alpha}') = \begin{pmatrix} z_1(h)^{d_0+\bar{d}} - z_0(h)^{d_0+\bar{d}} \\ \vdots \\ z_1(h)^{d_{\nu-1}+\bar{d}} - z_0(h)^{d_{\nu-1}+\bar{d}} \end{pmatrix}.$$

By Theorem 5.2 the roots satisfy $z_0(h) = 1 + O(h)$ and $z_1(h) = 1 + O(h)$. Hence, the right hand side is $O(h)$. By Theorem 5.1 the inverse of $\tilde{Z}(h)$ is bounded for small enough h. Therefore,

$$\left\|\tilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}'\right\| \le Ch. \tag{4.2}$$

With this estimate we can now consider the remaining errors. For the phase error we have

$$|E_{\text{phase}}| = \left|(\alpha_1(h)+1)\hat{\mathbf{w}}_1 e^{ik_x^1 mh + ik_y^1 jh}\right| \leq Ch,$$

since (4.2) implies that $|\alpha_1(h)+1| \leq Ch$ and the remaing factors are bounded in h due to $\mathbf{k}^1(h)$ being bounded in h by Theorem 5.1.

The final error from the evanescent waves E_{eva} is a sum of terms of the form

$$\alpha_r \hat{\mathbf{w}}_r e^{ik_x^r mh + ik_y^r jh} = \alpha_r(h) \begin{pmatrix} 1 \\ b\tilde{k}_x^r(h)/\tilde{\omega} \\ b\tilde{k}_y^r(h)/\tilde{\omega} \end{pmatrix} e^{ik_x^r mh + ik_y^r jh}.$$

By Theorem 5.1 there is an $\eta > 0$ such that $\operatorname{Im} k_y^r(h) > \eta/h$. Moreover, since K is real in (3.7) we have $\operatorname{Im} k_x^r(h) = -\alpha \operatorname{Im} k_y^r(h)$. Hence,

$$|e^{ik_x^r mh + ik_y^r jh}| = e^{-\operatorname{Im}(k_x^r mh + k_y^r jh)} = e^{-\operatorname{Im} k_y^r(-\alpha mh + jh)} < e^{-\eta(j-\alpha m)}.$$

For the amplitude we note further that by (4.2) we have $|\alpha_r(h)| \leq Ch$ and by (5.2) in Theorem 5.1 we have $|\tilde{\mathbf{k}}^r| < C/h$. Therefore, we get different errors in the p and the remaining components. More precisely, if $E_{\text{eva}} = (E_{\text{eva}}^p, E_{\text{eva}}^u, E_{\text{eva}}^v)^T$ we get

$$|E_{\text{eva}}^p| = O(he^{-\eta(j-\alpha m)}), \qquad |E_{\text{eva}}^u| = |E_{\text{eva}}^v| = O(e^{-\eta(j-\alpha m)}). \tag{4.3}$$

Note that $j - \alpha m$ is approximately equal to the distance in grid cells from the point (j,m) to the boundary, in the y-direction. The error from the evanescent waves thus die off with the number of grid cells, not with the physical distance. They therefore concentrate more and more at the boundaries at fine resolutions, but never go away.

Thus we see that we have three types of errors: Propagating errors E_{in}, E_{out} of second order $O(h^2)$, phase shift errors E_{phase} of first order $O(h)$, as well as boundary errors E_{eva} of zeroth order $O(e^{-\eta})$ for u,v, and $O(h)$ for p. Close to the boundary then E_{eva} dominates for u,v, elsewhere the phase error E_{phase}. This gives formally the estimated effective L^2 norms

$$\|E^p\|_2 \leq \sqrt{\sum_{m=1}^{N} (Ch)^2 h^2 + \sum_{m=1}^{N^2} (Ch)^2 h^2} = \sqrt{C^2 h^3 + C^2 h^2} = O(h),$$

$$\|E^{u,v}\|_2 \leq \sqrt{\sum_{m=1}^{N} C^2 h^2 + \sum_{m=1}^{N^2} (Ch)^2 h^2} = \sqrt{C^2 h + C^2 h^2} = O(\sqrt{h}),$$

since there are $O(N) = O(1/h)$ boundary cells in a computation on a $N \times N$ grid.

4.2 Hard boundaries

For hard boundaries the system is $A\alpha = 0$, where A is given by (3.21),

$$A = \Big(QZK(\tilde{k}_x) + ZK(-\tilde{k}_y)\Big)S = \Big(QZK'(\tilde{k}_x) + ZK'(-\tilde{k}_y)\Big)D^{-1}S \in \mathbb{C}^{\nu\times(\nu+1)},$$

which is similar to (3.10), besides the extra scalings. Consider the factor $A' = QZK'(\tilde{k}_x) + ZK'(-\tilde{k}_y)$. Summing the columns gives that

$$\mathbf{1}^T A' = (c, -c, 0, \ldots, 0)^T + O(h), \tag{4.4}$$

where $c = \mu k_x^0 - \nu k_y^0$. To see this we first introduce the notation where A_r is the rth column of A. Also, denote $\mathbf{z}_r = Z_r$. Thus for the first two columns, $r = 0, 1$,

$$\begin{aligned}(\mathbf{1}^T QZK'(k_x))_r &= \mathbf{1}^T Q\mathbf{z}_r k_x^r(0) + O(h) = \mu k_x^r(0) + O(h),\\ (\mathbf{1}^T ZK'(-k_y))_r &= \mathbf{1}^T \mathbf{z}_r(-k_y^r(0)) + O(h) = -\nu k_x^r(0) + O(h),\end{aligned}$$

since $\mathbf{z}_r(h) = \mathbf{1} + O(h)$, $r = 0, 1$, and $\mathbf{1}^T Q\mathbf{1} = \mu$. That $\mu k_x^0 - \nu k_y^0 = -\mu k_x^1 + \nu k_y^1$ follows from the boundary condition $\hat{\mathbf{n}} \cdot (u, v) = 0$ for the continuous solution (3.1). For the remaining columns $r = 2, \ldots, \nu - 1$, then

$$\begin{aligned}(\mathbf{1}^T QZK'(k_x))_r &= \left(\sum_{m \in \Omega_{\mathrm{cr}}} z_r^{d_m}\right) i(1 - wz_r^{-\mu})\\ &= \left(\sum_{m=0}^{\mu-1} z_r^m\right) i(1 - wz_r^{-\mu}) = \frac{z_r^\mu - 1}{z_r - 1} i(1 - wz_r^{-\mu}),\end{aligned}$$

$$(\mathbf{1}^T ZK'(-k_y))_r = \left(\sum_{m=0}^{\nu-1} z_r^m\right) i(1 - z_r^{-\nu}) = \frac{z_r^\nu - 1}{z_r - 1} i(1 - z_r^{-\nu}).$$

Here we have used Lemma 5.8. Taking these two together gives, since $w = 1 + O(h)$,

$$(\mathbf{1}^T A')_r = \frac{iz_r^{-\nu}}{z_r - 1} P(z_r) + O(h) = O(h),$$

where P is the polynomial (5.6). Hence (4.4) follows.

Thus we can expand the matrix equation $A\boldsymbol{\alpha} = A'\boldsymbol{\alpha}' = 0$, where $\boldsymbol{\alpha}' = D^{-1}S\boldsymbol{\alpha}$, as $\mathbf{1}^T A'\boldsymbol{\alpha}' = \alpha_0' c - \alpha_1' c + O(h) = 0$, giving $\alpha_1' = 1 + O(h)$, since we normalize the incoming wave to $\alpha_0 = 1$ and $\alpha_0' = 1$. The unprimed $\boldsymbol{\alpha}$ is given by $\boldsymbol{\alpha} = S^{-1}D\boldsymbol{\alpha}'$, and thus $\alpha_r = O(h)$, $r = 2, \ldots, \nu$.

The phase error then becomes

$$|E_{\mathrm{phase}}| = \left|(\alpha_1(h) - 1)\hat{w}_1 e^{ik_x^1 mh + ik_y^1 jh}\right| \le Ch.$$

Since $\alpha_r \le Ch$, $r = 2, \ldots, \nu$ for the evanescent waves, the same arguments as for soft boundaries hold and we get the bounds (4.3). In the end we obtain the same behavior of the different errors as for soft boundaries.

Remark 4.1 The cases $\alpha = 0$ and $\alpha = 1$ can be analyzed in simpler ways than with the techniques used above. Still, our analysis, although restricted to the case $0 < \alpha < 1$, can give some insight also into the limiting cases $\alpha = 0, 1$. The basic second order propagation errors E_{in} and E_{ref} are independent of α and will also be present for $\alpha = 0, 1$. However, the $O(1)$ error $E_{\mathrm{eva}} = 0$ disappears. The reason is that when

$\alpha = 0,1$, then $\nu = 1$, $\mu = \alpha$ and $P(z)$ in (5.6) is just a second order polynomial $(\alpha+1)(z-1)^2$ with two roots. There are therefore only two admissible waves, $\mathbf{k}^0$ and $\mathbf{k}^1$, which correspond to the incoming and reflected waves, and no evanescent waves. When it comes to the $O(h)$ phase error E_{phase}, it may or may not be present. It is easy to check that $d_m = 0$ when $\alpha = 0,1$. In the soft boundary case one then gets $\tilde{Z} = 1$ and $\alpha_1(h) = -(z_0(h)/z_1(h))^{\bar{d}}$. Since $z_0(0) = z_1(0) = 1$, Lemma 5.7 shows that $z_0(h)/z_1(h) = 1 + O(h)$ and the phase error $E_{\text{phase}} = O(|\alpha_1(h)+1|)$ is of size $O(h)$ unless $\bar{d} = 0$, in which case E_{phase} vanishes. But since $d_m = 0$ and $\nu = 1$ we have $\bar{d} = \delta_{m+1/2}$, the offset between the grid cell center and the boundary. Hence, the $O(h)$ phase error disappears when all grid cell centers lie on the boundary. With the convention we have used for boundary cells Γ_C in (2.4) this happens when $\alpha = 1$ (then $\delta_{m+1/2} = 0$) but not when $\alpha = 0$ (then $\delta_{m+1/2} = 1/2$). Upon shifting the grid by half a cell one would remove the phase error also for the case $\alpha = 0$. The argument for hard boundaries is a bit more involved, but renders the same principal result, although in this case our boundary cell convention leads to a $O(h)$ phase error when $\alpha = 1$ and no phase error when $\alpha = 0$, since the quantities involved in the boundary conditions, $\hat{u}, \hat{v}$, are now evaluated on the edges of the boundary cells. In conclusion, when $\alpha = 0,1$ the basic $O(h^2)$ error remains. There is, however, no $O(1)$ error from evanescent waves and also no $O(h)$ phase error if the quantities in the boundary conditions are evaluated precisely on the boundary.

5 Analysis of the admissible wave numbers

We now prove the results we have referred to in the previous sections. The aim is to show the following main theorem.

Theorem 5.1 *Assume $\nu > \mu$ are relatively prime and that $0 < h \le h_0$.*

- *There are 2ν admissible waves $\mathbf{k}^r$ with $r = 0,\dots,2\nu-1$, for all h_0.*
- *For small enough h_0 we can order $\{\mathbf{k}^r(h)\}$ such that they are continuously differentiable in h on $(0,h_0)$.*
- *For small enough h_0 we can choose the r-indexing such that:*
  - *For $r \in \{0,1\}$ the wave vectors $\mathbf{k}^r(h)$ are bounded on $(0,h_0)$ and*

    $$\lim_{h\to 0} \mathbf{k}^0(h) = \bar{\mathbf{k}}, \quad \text{and} \quad \lim_{h\to 0} \mathbf{k}^1(h) = \bar{\mathbf{k}}',$$

    which are the incoming and reflected waves in the continuous case (3.1). *Moreover, if* (3.4) *holds,*

    $$|\mathbf{k}^0(h) - \bar{\mathbf{k}}| \le Ch^2, \qquad |\mathbf{k}^1(h) - \bar{\mathbf{k}}'| \le Ch^2. \tag{5.1}$$

  - *For $2 \le r \le \nu$, the imaginary part of $k_y^r(h)$ is positive and*

    $$\operatorname{Im} k_y^r(h) \ge C/h, \quad |\mathbf{k}^r| \le C/h, \qquad r = 2,\dots,\nu. \tag{5.2}$$

  - *For $\nu+1 \le r \le 2\nu$, the imaginary part of $k_y^r(h)$ is negative.*

– *For small enough h_0 the null space of the matrix Z in* (3.11) *is one-dimensional. The matrix $\tilde{Z}$ in* (4.1) *is non-singular and its inverse is bounded in* $[0,h_0]$.

The first two waves have real-valued wave numbers and correspond to the incoming and reflected waves. The ones with positive imaginary part are evanescent waves that concentrate on the staircase boundary. The remaining waves, with negative imaginary part, are non-physical waves which cannot exist in a bounded solution.

The main steps in the proof are as follows:

1. Rewrite the admissibility conditions (3.6) and (3.7) as a polynomial equation of order 2ν (Section 5.1).
2. Analyze the roots of this polynomial in the limit $h \to 0$ (Theorem 5.2 in Section 5.2).
3. Use perturbation arguments to show that the roots are qualitatively the same also for h small (Section 5.3, Lemma 5.7).
4. Use the properties of the roots and show their implications for the wave vectors $\mathbf{k}^r$ and the matrices Z, $\tilde{Z}$ (Section 5.3.1).

5.1 Preliminaries

The dispersion relation for the Yee scheme, $\tilde{\omega}^2 = c^2(\tilde{k}_x^2 + \tilde{k}_y^2)$, $c^2 = ab$, expands to

$$\sin^2 \frac{\omega\Delta t}{2} = c^2\lambda^2 \left(\sin^2 \frac{hk_x}{2} + \sin^2 \frac{hk_y}{2}\right), \qquad \lambda = \frac{\Delta t}{h}. \tag{5.3}$$

This equation together with the equation

$$k_x + \alpha k_y = K \mod \frac{2\pi}{h}, \tag{5.4}$$

defines the admissible wave numbers for ω and K, as was seen in Section 3.2.

First we note that we can rewrite (5.3) as

$$\cos hk_x + \cos hk_y = \frac{1}{c^2\lambda^2}(\cos\omega\Delta t - 1) + 2.$$

Thus using (5.4) we get an equivalent equation

$$e^{ih(K-\alpha k_y)} + e^{-ih(K-\alpha k_y)} + e^{ihk_y} + e^{-ihk_y} = 4 - \frac{2}{c^2\lambda^2}(1 - \cos\omega\Delta t)$$

to (5.3)–(5.4). This we can simplify by introducing $w = e^{ihK}$, $z = e^{ihk_y/\nu}$, giving the order 2ν polynomial equation

$$z^{2\nu} + \frac{1}{w}z^{\nu+\mu} - Rz^{\nu} + wz^{\nu-\mu} + 1 = 0, \tag{5.5}$$

where $R := 4 - 2(1-\cos\omega\Delta t)/(c^2\lambda^2)$. This gives us 2ν solutions, which we index by r. We finally note that by (5.4), $e^{ik_xh} = e^{iKh}e^{-ihk_y\alpha} = wz^{-\mu}$. Hence, if we pick $\mathrm{Re}\, hk_y$ and $\mathrm{Re}\, hk_x$ as the arguments in $[0,2\pi)$ for z^{ν} and $wz^{-\mu}$ respectively, we have a solution that satisfies (5.3), (5.4) as well as (3.8). Moreover, each z corresponds to precisely one such pair k_x and k_y.

5.2 The limit equation

To study the asymptotics of these waves, we note that $w \to 1$ and $R \to 4$ in the limit $h \to 0$ and $\Delta t \to 0$, and thus the polynomial reduces to

$$P(z) := z^{2\nu} + z^{\nu+\mu} - 4z^{\nu} + z^{\nu-\mu} + 1 = 0. \tag{5.6}$$

For this equation we can show the following theorem.

Theorem 5.2 *The polynomial* (5.6)*, where* $\nu > \mu$ *are relatively prime, has one double root at* $z = 1$*. Its other roots are all distinct. Moreover, half of the remaining roots have magnitude stricly less than one, and the other half have magnitude stricly larger than one.*

The result follows from a sequence of lemmas in which we all make the assumption that $\nu > \mu$ are relatively prime.

Lemma 5.3 *There are no positive real roots other than* $z = 1$*, which is a double root. When* ν *is odd there are no negative real roots.*

Proof. First we observe that the polynomial is symmetric in the sense that if $z_\star \neq 0$ is a root, then so is $1/z_\star$. This follows from the simple relation $z_\star^{2\nu} P(1/z_\star) = P(z_\star)$. The existence of a double root $z = 1$ is established by taking the derivative, i.e.,

$$\begin{aligned} P(1) &= 1 + 1 - 4 + 1 + 1 = 0, \\ P'(1) &= 2\nu + (\nu+\mu) - 4\nu + (\nu-\mu) = 0, \end{aligned}$$

which is not a higher order root since

$$P''(1) = (\nu+\mu)^2 + (\nu-\mu)^2 > 0.$$

To show that this is the only real positive root, assume $s \in \mathbb{R}$, $s > 1$. Then $s^{\mu} + s^{-\mu} > 2$ and

$$P(s) = s^{2\nu} + s^{\nu+\mu} - 4s^{\nu} + s^{\nu-\mu} + 1 > s^{2\nu} - 2s^{\nu} + 1 = (s^{\nu} - 1)^2 > 0.$$

Hence there are no other roots for $s > 1$. Assume now that ν is odd. Then

$$\begin{aligned} P(-s) &= (-s)^{2\nu} + (-s)^{\nu+\mu} - 4(-s)^{\nu} + (-s)^{\nu-\mu} + 1 \\ &= s^{2\nu} + s^{\nu}((-1)^{\nu+\mu} s^{\mu} - 4(-1)^{\nu} + (-1)^{\nu-\mu} s^{-\mu}) + 1 \\ &= s^{2\nu} + s^{\nu}(-(-1)^{\mu} s^{\mu} + 4 - (-1)^{\mu} s^{-\mu}) + 1 \\ &= s^{2\nu} + s^{\nu}(4 - (-1)^{\mu}(s^{\mu} + s^{-\mu})) + 1 \\ &> s^{2\nu} + s^{\nu}(4 - 1 - s^{\mu}) + 1 \\ &> 3s^{\nu} + 1 > 0, \end{aligned}$$

which means that there are no roots for $s < -1$. By symmetry of roots $z_\star$, $1/z_\star$ there are therefore also no roots s with $0 < s < 1$ and, if ν is odd, no roots with $-1 < s < 0$. Finally, when ν is odd,

$$P(-1) = 1 - (-1)^{\mu} + 4 - (-1)^{\mu} + 1 = 6 - 2(-1)^{\mu} \neq 0,$$

which proves the lemma. □

Next we can prove that P has no roots on the lines shooting out from the origin in the complex plane, crossing roots of $z^\nu + 1 = 0$.

Lemma 5.4 *Let $u_j = e^{i\alpha_j}$ and $\alpha_j = 2\pi(j+1/2)/\nu$. Then $P(su_j) \neq 0$ for $s \geq 0$ and $0 \leq j \leq \nu - 1$.*

Proof. We first note that $P(0) = 1 \neq 0$ so we only need to consider $s > 0$. Since $u_j^\nu = -1$ we get

$$\begin{aligned}\operatorname{Im} P(su_j) &= \operatorname{Im}\left(s^{2\nu} - s^\nu (su_j)^\mu - 4s^\nu - s^\nu (su_j)^{-\mu} + 1\right) \\ &= -s^\nu \operatorname{Im}\left((su_j)^\mu + (su_j)^{-\mu}\right).\end{aligned} \tag{5.7}$$

If $\operatorname{Im} u_j^\mu \neq 0$ this can only be zero if $|(su_j)^\mu| = 1$. Since $|(su_j)^\mu| = s^\mu$ we must then have $s = 1$, but

$$\begin{aligned}P(u_j) &= 1 - u_j^\mu + 4 - u_j^{-\mu} + 1 \\ &= 6 - \left(u_j^\mu + u_j^{-\mu}\right) \\ &= 6 - 2\operatorname{Re} u_j^\mu \geq 6 - 2|u_j^\mu|4 > 0.\end{aligned} \tag{5.8}$$

Hence, $P(su_j) \neq 0$ for $s \geq 0$ if $\operatorname{Im} u_j^\mu \neq 0$. On the other hand, if $\operatorname{Im} u_j^\mu = 0$ then $\exp(2\pi i (j+1/2)\mu/\nu) \in \mathbb{R}$, or $(2j+1)\mu/\nu \in \mathbb{Z}$ for some $j = 0, \ldots, \nu - 1$. Since μ and ν are relatively prime, the only possibility is $2j + 1 = \nu$, which corresponds to $u_j = -1$. But then ν is odd and $z = 1$ is the only real root according to Lemma 5.3; $P(su_j) = P(-s)$ is therefore non-zero also for this case, which proves the lemma. □

For the next lemma we construct a pie-shaped region in the complex plane, bounded by the lines

$$\begin{aligned}\gamma_1 &= \left\{z = su_j \mid 0 \leq s \leq R\right\}, \\ \gamma_2 &= \left\{z = Re^{is} \mid \alpha_j \leq s \leq \alpha_{j+1}\right\}, \\ \gamma_3 &= \left\{z = su_{j+1} \mid 0 \leq s \leq R\right\},\end{aligned}$$

where u_j is defined as in Lemma 5.4. We can then show the following result.

Lemma 5.5 *Let γ be the union of the curves γ_1, γ_2 and γ_3 defined above. If R is large enough, there are precisely two roots of* (5.6) *inside γ.*

Proof. We chose R strictly larger than the magnitude of all roots of P. This means that $P(z) \neq 0$ on γ_2. It follows from Lemma 5.4 that $P(z) \neq 0$ also on γ_1 and γ_2, and hence on all of γ. We can then use the argument principle, that for any closed curve $\gamma \subset \mathbb{C}$, and analytic function f, with no zeros on γ, the change in argument in $P(z)$ as z traverses γ is equal to 2π times the number of zeros of $P(z)$ inside γ. First consider γ_1 and noting that $\arg P(0) = \arg 1 = 0$. If $\operatorname{Im} u_j = 0$, then $P(z)$ is purely real and there is no change in the argument. On the other hand, if $\operatorname{Im} u_j \neq 0$ then by (5.7) the only zeros of $\operatorname{Im} P(su_j)$ on γ_1 are when $s = 0$ or $s = 1$. For both these points

$\operatorname{Re} P > 0$ by (5.8) and therefore P never crosses the negative real axis. Setting $v = u_j^{\mu}$ and exploiting the fact that $u_j^{\nu} = -1$, we then get

$$\begin{aligned}
\arg P(Ru_j) &= \arctan \frac{\operatorname{Im} P(Ru_j)}{\operatorname{Re} P(Ru_j)} \\
&= \arctan \frac{-R^{\nu}\operatorname{Im}\left[R^{\mu}v + R^{-\mu}v^{-\mu}\right]}{R^{2\nu} + 4R^{\nu} + 1 - R^{\nu}\operatorname{Re}\left[R^{\mu}v + R^{-\mu}v^{-\mu}\right]} \\
&= \arctan R^{\mu-\nu} \frac{-\operatorname{Im}\left[v + R^{-2\mu}v^{-\mu}\right]}{1 + 4R^{-\nu} + R^{-2\nu} - R^{\mu-\nu}\operatorname{Re}\left[v + R^{-\mu-\nu}v^{-\mu}\right]} \\
&= \arctan R^{\mu-\nu}\left(-1 + O(R^{-\nu})\right) = O(R^{\mu-\nu}).
\end{aligned}$$

The same result also holds for γ_3. Now consider γ_2. For this case we have $dz = iRe^{is}ds$ and

$$\arg P\left(Ru_j\right) - \arg P\left(Ru_{j+1}\right) = \int_{\gamma_2} \frac{P'(z)}{P(z)}dz = i\int_{\alpha_j}^{\alpha_{j+1}} \frac{P'\left(Re^{is}\right)}{P\left(Re^{is}\right)} Re^{is} ds.$$

But the quotient reduces according to

$$\frac{zP'(z)}{P(z)} = \frac{2\nu z^{2\nu} + (\mu+\nu)z^{\mu+\nu} - 4\nu z^{\nu} + (\nu-\mu)z^{\nu-\mu}}{z^{2\nu} + z^{\mu+\nu} - 4z^{\nu} + z^{\nu-\mu} + 1} = 2\nu + O(|z|^{\mu-\nu}).$$

This means that we can simplify according to

$$\int_{\gamma_2} \frac{P'(z)}{P(z)}dz = 2\nu i\int_{\alpha_j}^{\alpha_{j+1}} ds + O(R^{\mu-\nu}) = 2\nu i(\alpha_{j+1} - \alpha_j) + O(R^{\mu-\nu}).$$

The total change in the argument along γ thus becomes $\Delta \arg P(z) = 4\pi i + O(R^{\mu-\nu})$. This is valid for all R large enough, so if we let $R \to \infty$ we can conclude that there are precisely two roots inside γ. □

Finally we consider the unit circle and show the following lemma.

Lemma 5.6 *There are no roots on the unit circle other than* $z = 1$.

Proof. We have

$$\begin{aligned}
P(e^{i\theta}) &= e^{2i\theta\nu} + e^{i\theta(\mu+\nu)} - 4e^{i\theta\nu} + e^{i\theta(\nu-\mu)} + 1 \\
&= e^{2i\theta\nu} + 2e^{i\theta\nu}\left(\cos\theta\mu - 2\right) + 1 \\
&=: \tilde{P}\left(e^{i\theta\nu}, \mu\theta\right),
\end{aligned}$$

where $\tilde{P}(z,\tilde{\theta}) = z^2 + 2z(\cos\tilde{\theta} - 2) + 1$. Hence, if $P\left(e^{i\theta}\right) = 0$, the polynomial $\tilde{P}(z, \theta\mu)$ with fixed $\theta\mu$ has a unit root. This happens precisely when $|\cos\theta\mu - 2| \leq 1$, i.e. only when $\theta\mu = 2\pi n$ for some integer n. But in that case $\cos\theta\mu - 2 = -1$ and the root is one, $1 = e^{i\theta\nu}$, giving $\theta\nu = 2\pi m$ for some integer m. Since μ and ν are relatively prime it follows that $\theta = 2\pi n'$ for some integer n', so the only root with magnitude one is $z = 1$. □

By the symmetry of roots z_* and $1/z_*$ it now follows that the two roots in each pie shaped region considered in Lemma 5.5 are different, except when the root is one. Therefore their magnitudes are also strictly smaller than and larger than one, respectively. This proves Theorem 5.2.

5.3 Results for small h

From the conclusions about the limiting equation (5.5) we can get results also for the full equation (5.6) with small h and Δt by considering it as a perturbation. We let λ, ω and K be fixed. Then the roots of (5.5) only depend on h and we denote them by $z_j(h)$. The following lemma shows that they are all first order perturbations in h.

Lemma 5.7 *Assume $\nu > \mu$ are relatively prime. For small enough h_0 we can order $\{z_j(h)\}$ such that each $z_j(h)$ is continuously differentiable in h on $[0,h_0)$ and $z_j(0)$ are the roots of* (5.6)*. Moreover,*

$$|z_j(h) - z_j(0)| \leq Ch, \quad \forall h \in [0,h_0), \tag{5.9}$$

where C is independent of h and j.

Proof. Let $P(z)$ be the limit polynomial (5.6) and $Q(z,h)$ the full polynomial (5.5) with a fixed λ, ω and K,

$$Q(z,h) = z^{2\nu} + \frac{1}{w(h)} z^{\nu+\mu} - R(h) z^{\nu} + w(h) z^{\nu-\mu} + 1.$$

It is classical perturbation theory that we can order the roots in the way described in the theorem and that (5.9) holds for the distinct roots of $P(z)$. By Theorem 5.2 all roots of $P(z)$ are distinct except a double root at $z = 1$. For this root, classical theory tells us that it can be expanded in a Puiseux series,

$$z(h) = 1 + z_1 h^{\gamma} + o(h^{\gamma}), \tag{5.10}$$

for some exponent $\gamma \leq 1$. Since $w = 1 + iKh + O(h^2)$ and $R = 4 + O(h^2)$ we can write $Q(z,h)$ as

$$Q(z,h) = P(z) + h\tilde{P}(z) + O(h^2), \qquad \tilde{P}(z) = iK(z^{\nu-\mu} - z^{\nu+\mu}).$$

Since $\tilde{P}(1) = 0$, entering (5.10) in this expansion of $Q(z,h)$ we get

$$\begin{aligned} 0 = Q(z(h),h) &= P(1 + z_1 h^{\gamma} + o(h^{\gamma})) + h\tilde{P}(1 + z_1 h^{\gamma} + o(h^{\gamma})) + O(h^2) \\ &= \frac{1}{2} P''(1) z_1^2 h^{2\gamma} + o(h^{2\gamma}) + \tilde{P}'(1) z_1 h^{\gamma+1} + o(h^{\gamma+1}) + O(h^2) \end{aligned}$$

The leading term must vanish and since $P''(1) \neq 1$ this can only happen if $2\gamma = \gamma + 1 = 2$. Hence, $\gamma = 1$ in (5.10) and (5.9) holds also for the double root at $z = 1$. Moreover, $\lim_{h\to 0} z'(h) = z_1$. □

5.3.1 Proof of Theorem 5.1

Since there are 2ν roots to the polynomial (5.5) there are 2ν admissible waves. The ordering is given in the same way as for $z_j(h)$ in Lemma 5.7, and the statement follows since $\mathbf{k}^r(h)$ depends smoothly on $z_r(h)$.

Let $z_0(h)$ and $z_1(h)$ be the roots of $Q(z,h)$ for which $z_0(0) = z_1(0) = 1$. Then again by Lemma 5.7 there is a constant C independent of $h < h_0$, such that

$$|k_y^r(h)| = \left|\frac{\nu}{h}\log(z_r(h))\right| = \left|\frac{\nu}{h}\log(1+[z_r(h)-z_r(0)])\right| \le C, \tag{5.11}$$

for small enough h_0 and $r = 0,1$. Since $k_x^r(h) = K - \alpha k_y^r(h)$, the same holds for $k_x^r(h)$. Moreover, $\tilde{k}_x(h) = k_x + O(k_x^3 h^2)$, $\tilde{k}_y(h) = k_x + O(k_y^3 h^2)$, $\tilde{\omega}(\Delta t) = \omega + O(\omega^3 h^2)$, if $\Delta t = \lambda h$ for a fixed λ. Then,

$$\begin{aligned} &k_x^r(h)^2 + k_y^r(h)^2 = \omega^2 + h^2\Theta(h, k_x^r(h), k_y^r(h)), \\ &k_x^r(h) + \alpha k_y^r(h) = K, \end{aligned} \tag{5.12}$$

where $|\Theta| \le C$ independent of $h \le h_0$ for $r = 0,1$ and small enough h_0. From (5.11) we furthermore see that the limit $\lim_{h\to 0}\mathbf{k}^r(h)$ is well-defined and equal to $z_r'(0)\nu/i$. Defining $\mathbf{k}^r(0)$ thus by continuity we get that

$$k_x^r(0)^2 + k_y^r(0)^2 = \omega^2, \qquad k_x^r(0) + \alpha k_y^r(0) = K, \tag{5.13}$$

which are the conditions for the incoming and reflected wave numbers in the continuous case (3.2)–(3.3). The equation reduces to a second order polynomial equation with two roots. Hence, we can indeed choose our index r such that $\mathbf{k}^0(0) = \bar{\mathbf{k}}$ and $\mathbf{k}^1(0) = \bar{\mathbf{k}}'$. Let us now fix $r \in \{0,1\}$ and define $x(h) = k_x^r(h) - k_x^r(0)$ and $y(h) = k_y^r(h) - k_y^r(0)$. Then, upon subtracting (5.13) from (5.12),

$$\begin{aligned} x(k_x^r(h) + k_x^r(0)) + y(k_y^r(h) + k_y^r(0)) &= h^2\Theta, \\ x + \alpha y &= 0. \end{aligned}$$

Eliminating x with the second equation, gives

$$y = h^2 \frac{\Theta}{k_y^r(h) - \alpha k_x^r(h) + k_y^r(0) - \alpha k_x^r(0)} = h^2 \frac{\Theta}{q(h) + q(0)},$$

where $q(h) := k_y^r(h) - \alpha k_x^r(h)$. From (3.4) and (5.13) we can now deduce that

$$(1+\alpha^2)\omega^2 = (k_y^r(0) - \alpha k_x^r(0))^2 + (k_x^r(0) + \alpha k_y^r(0))^2 = q(0)^2 + K^2 \le q(0)^2 + \eta(1+\alpha^2)\omega^2,$$

which shows that $|q(0)| \ge \delta > 0$. By continuity $|q(0) + q(h)| \ge \tilde{\delta} > 0$ for small enough h and therefore, $|y| \le Ch^2$. Since $|x| = \alpha|y|$ the same holds for x and (5.1) is proved.

For the remaining roots we note first that since $|z_r(h)| = |e^{ik_y^r(h)/\nu}| = e^{-\operatorname{Im} k_y^r h/\nu}$, it follows that the sign of $\operatorname{Im} k_y^r(h)$ is positive when $|z_r(h)| < 1$ and negative if $|z_r(h)| > 1$. By Theorem 5.2, when $h = 0$ there are precisely $\nu - 1$ roots with magnitude strictly smaller than one and the same number with magnitude strictly larger than one. Moreover, by Lemma 5.7 the same thing holds also when $0 \le h \le h_0$ if h_0 is small enough. This proves that we can order the wave numbers such that $r = 2,\dots,\nu$ correspond to those with positive imaginary parts, and $r = \nu+1,\dots,2\nu$ correspond to those with negative imaginary parts. Finally, for $2 \le r \le \nu$ there is a $\delta > 0$ such that for $0 < h \le h_0$, $1-\delta \ge |z_r(h)|$, giving $\operatorname{Im} k_y^r(h) \ge (\nu/h)|\log(1-\delta)|$. Moreover, $|hk_y^r| = \nu|\log z_r(h)| \le C$. Since $k_x^r = K - \alpha k_y^r$ the estimate (5.2) follows.

To investigate the properties of the matrix Z in (3.11) we first note that it is in fact a permutation of a Vandermonde matrix due to the following property of the $\{d_m\}$ indices.

Lemma 5.8 *If $\nu > \mu$ are relatively prime, then $\{d_m\}_{m=0}^{\nu-1}$ as defined in* (3.9) *is a permutation of the integers $\{0,\dots,\nu-1\}$. If m corresponds to the corner cells defined by* (3.14)*, then $\{d_m\}$ is a permutation of the integers $\{0,\dots,\mu-1\}$.*

Proof. Suppose first that $d_{m_1} = d_{m_2}$. Then $0 = d_{m_2} - d_{m_1} = (j_{m_1} - j_{m_2})\nu - (m_1 - m_2)\mu$, and since ν and μ are relatively prime, $j_{m_1} - j_{m_2} = n\mu$, $m_1 - m_2 = n\nu$ for some n. But $0 \le m_1, m_2 \le \nu - 1$ so $\underline{n} = 0$ and $m_1 = m_2$. Therefore all d_m are distinct. Moreover, since $\delta_{m+\frac{1}{2}} \in [0,1)$ and $\bar{d} \in \{0, 1/2\}$,

$$d_m = \nu\delta_{m+\frac{1}{2}} - \bar{d} \in \Big[-\frac{1}{2}, \nu\Big).$$

Hence, by counting, all integers $\{0,\dots,\nu-1\}$ are represented by precisely one d_m.

The second statement follows from the fact that when $j_{m+1} = j_m + 1$,

$$d_{m+1} = j_{m+1}\nu - (m+1)\mu + \left\lfloor \frac{\nu-\mu}{2} \right\rfloor = d_m + \nu - \mu, \qquad m \in \Omega_{\mathrm{cr}},$$

and $d_{m+1} < \nu$. □

It follows from Lemma 5.8 that there is a permutation matrix $W \in \mathbb{R}^{\nu\times\nu}$ such that

$$WZ = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ z_0 & z_1 & \cdots & z_\nu \\ z_0^2 & z_1^2 & \cdots & z_\nu^2 \\ \vdots & \vdots & \ddots & \vdots \\ z_0^{\nu-1} & z_1^{\nu-1} & \cdots & z_\nu^{\nu-1} \end{pmatrix} \in \mathbb{R}^{\nu\times(\nu+1)}.$$

By its size, the column rank of this matrix is at most ν. By Lemma 5.7, for $0 \le h \le h_0$ with h_0 small enough, all z_j are distinct, except possibly z_0 and z_1 which may be equal. Upon removing the first column (the z_0-column) we obtain a Vandermonde matrix with distinct values z_j, which is non-singular. This is $W\tilde{Z}$ in (4.1) showing that $\tilde{Z}$ is non-singular. Moreover, there are hence ν linearly independent columns in WZ, so rank $WZ = \nu$, and therefore the dimension of the null space of WZ, and hence also of WZ, is one. The inverse of $\tilde{Z}$ is a continuous function of h on the compact interval $[0,h_0]$, since it is well-defined on $[0,h_0]$ and each $z_j(h)$ is continuous. It is hence bounded.

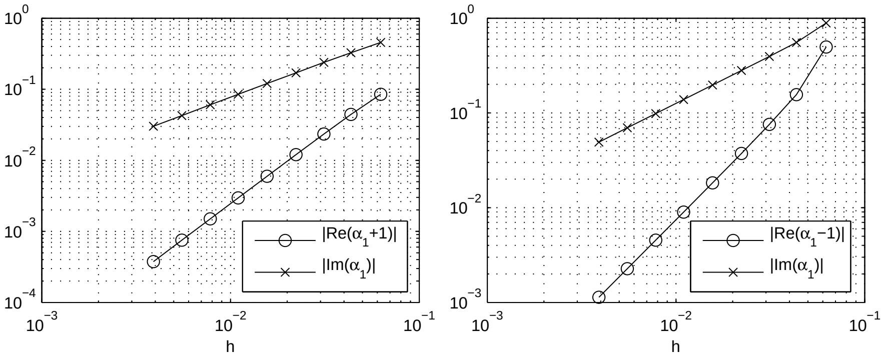

Fig. 6.1 The convergence of α_1, which is the reflection coefficient for the propagating wave. The case considered is $\mu = 3$, $\nu = 2$. *Left:* soft boundary. *Right:* hard boundary.

6 Numerical tests

To verify the analysis, we compute the explicit solution to (3.10) and (3.21). We choose the boundary angle $\alpha = \mu/\nu = 2/3$, which gives 2 propagating modes and $\nu - 1 = 2$ evanescent waves. We compare against the known solution (3.1) of the continuous problem. First we verify the convergence in Fig. 6.1, where we clearly observe first order behavior in the reflection coefficient α_1, for both hard and soft boundaries. The wave vectors for the same solutions are shown in Fig. 6.2, showing the $O(1/h)$ growth in the frequency of the evanescent waves $\mathbf{k}^2, \mathbf{k}^3$, as well as $O(h^2)$ convergence for the reflected wave $\mathbf{k}^1$.

Using the computed α_r and $\mathbf{k}^r$ the full field can be obtained from (3.13). Evaluating this on a $[0,1]^2$ domain for $1/h = 2^6, 2^7, 2^8$, we plot the field along a line $x = 1/2$ in Fig. 6.3, first for a soft boundary, and then a hard boundary. We see the $O(h)$ convergence in the propagating waves and the $O(1)$ spike at the boundary. Zooming in shows the decay as a function of h in Fig. 6.4. While the amplitude of the spike seems to change for the 2^7 resolution, this is only due to the high frequency along the boundary, which we see in Fig. 6.5.

6.1 General boundary shapes

To see that the analysis is applicable to more general domains and boundary shapes, we perform a simulation of harmonic waves scattering against a rigid cylinder. The setup is chosen so that we have an exact continuous solution for the corresponding continuous problem to compare the results with. It can be found in any basic text

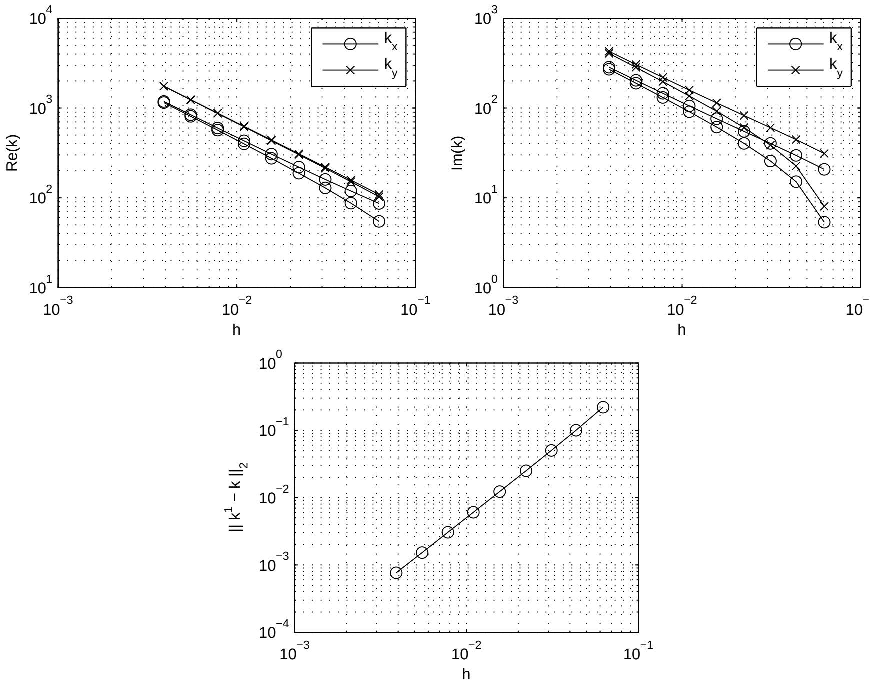

Fig. 6.2 The wave vectors as a function of grid size h. *Top:* since $\nu = 3$, we have 2 evanescent waves. *Bottom:* the wave vector for reflecting wave displaying $O(h^2)$ convergence.

book. In polar coordinates it is given by

$$\varphi^{\text{inc}}(r,\theta) = e^{ikr\cos\theta} = J_0(kr) + \sum_{n=1}^{\infty} 2(i)^n J_n(kr)\cos n\theta,$$

$$\varphi^{\text{ref}}(r,\theta) = \sum_{n=0}^{\infty} M_n H_n^{(1)}(kr)\cos n\theta,$$

together with

$$p(x,y,t) = \frac{1}{b}\text{Re}\,(\partial_t \varphi e^{-i\omega t}), \qquad \mathbf{u}(x,y,t) = \text{Re}\,(\nabla\varphi e^{-i\omega t}). \tag{6.1}$$

The expansion coefficients for the reflected field are determined by the boundary conditions, giving $M_0 = -J_0'(kR)/H_0^{(1)'}(kR)$, $M_n = -2(i)^n J_n'(kR)/H_n^{(1)'}(kR)$. These include Bessel functions J_n as well as Hankel functions of first kind $H_n^{(1)}$.

We use the computational domain $\Omega = \{(x,y) \subset [0,2\pi]\times[0,2\pi] \mid (x-\pi)^2 + (y-\pi)^2 \geq 1\}$, and initialize the field to the exact continuous solution (6.1). We then run the Yee scheme until $t = 0.3$ and compare the result against the exact solution (6.1) at $t = 0.3$. The error is plotted in Fig. 6.6. Here we see that the same characteristic spikes in the error occur along the boundary. These oscillate as $O(1/h)$ and have an amplitude of the same order of magnitude as the incoming field.

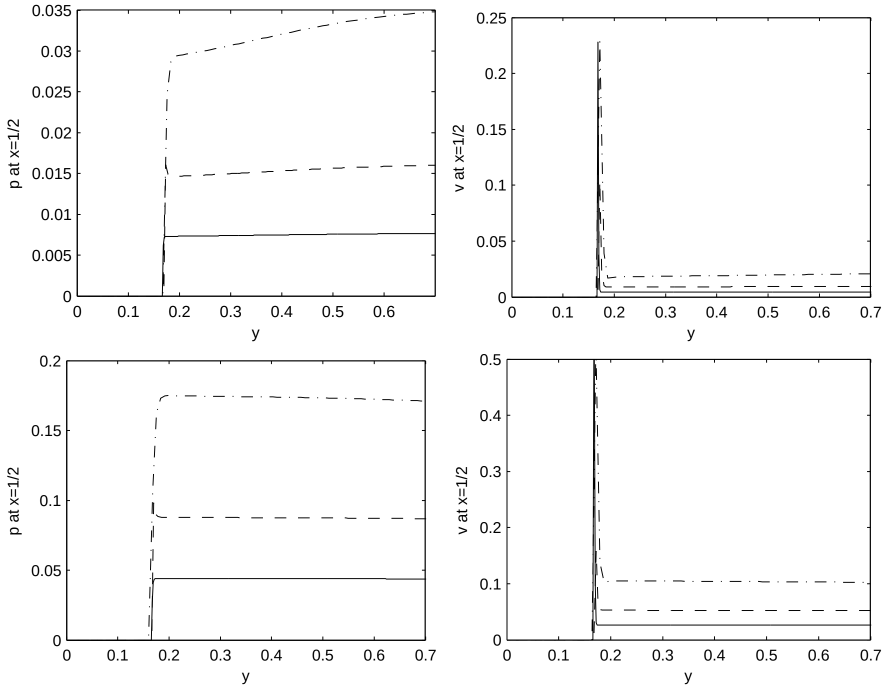

Fig. 6.3 The absolute value of the computed solutions for a soft (*top*) and hard boundary (*bottom*), shown along a line $x = 1/2$, for three different grid resolutions $N = 2^6, 2^7, 2^8$. We see the h convergence of the propagating modes as well as the spike at the boundary in v.

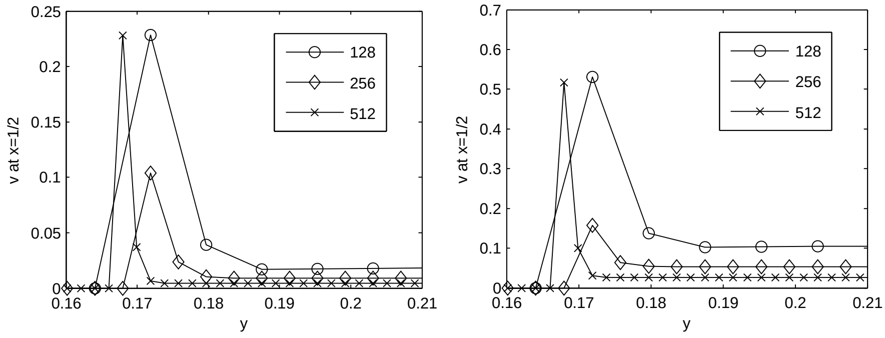

Fig. 6.4 Close-up of the solution in Fig. 6.3 verifies the decay as a function of the number of cells.

7 Conclusion

We have rigorously derived exact solutions to the Yee scheme close to staircase approximated boundaries. This enables a detailed error analysis, showing how the staircasing affects amplitude, phase, frequency and attenuation of waves. In particular, this

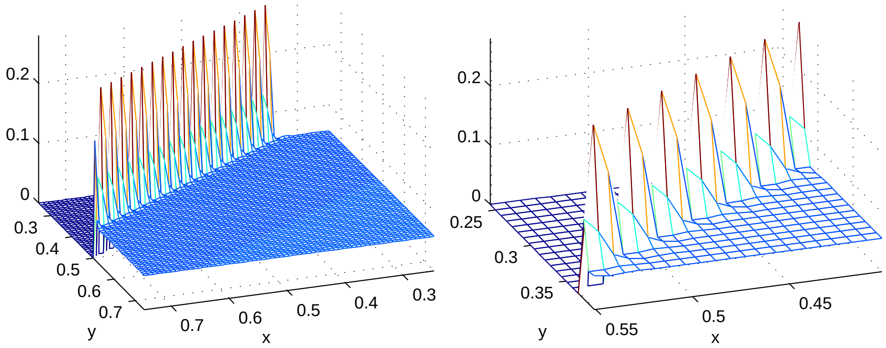

Fig. 6.5 Surface plot of $|v|$ for the soft boundary solution in Fig. 6.3 with $N = 2^7$. The field oscillates along the boundary, with frequency $O(1/h)$.

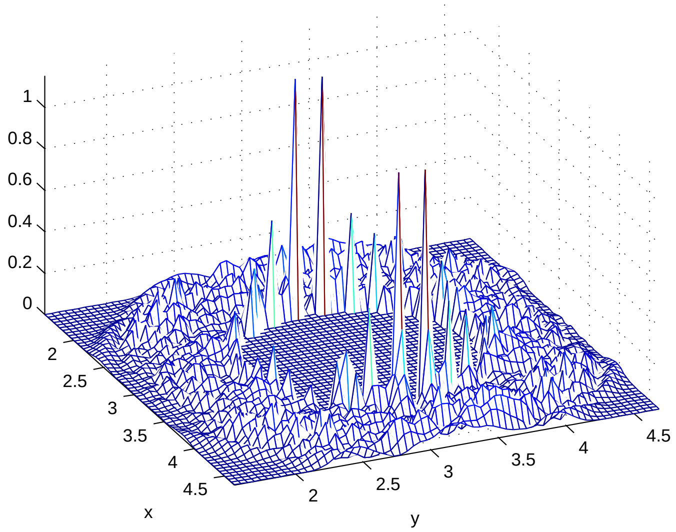

Fig. 6.6 Plot of the error in v for harmonic waves scattering against a hard (rigid) cylinder. The errors are seen to fluctuate with frequency $O(1/h)$ along the cylinder, with an amplitude of the same order of magnitude as the incoming field.

characterizes the $O(1)$ evanescent waves occurring at the boundary which prevents convergence in L^∞ and reduces it in L^2 to $O(\sqrt{h})$. The analysis shows that the errors are local in nature, and explains why they can very often be ignored in applications if the field along the boundary is not the focus of the simulation. The explicit form of the solutions to the Yee scheme should also provide a starting point for deriving more accurate boundary approximations.

References

1. Botteldooren, D.: Finite-difference time-domain simulation of low-frequency room acoustic problems. J. Acoust. Soc. Am. **98**(6), 3302–3308 (1995)
2. Cangellaris, A., Wright, D.: Analysis of the numerical error caused by the stair-stepped approximation of a conducting boundary in FDTD simulations of electromagnetic phenomena. IEEE Trans. Antennas Propag. **39**(10), 1518–1525 (1991)
3. Dey, S., Mittra, R.: A locally conformal finite-difference time-domain (FDTD) algorithm for modeling three-dimensional perfectly conducting objects. IEEE Microw. Guided. W. **7**(9), 273–275 (1997)
4. Ditkowski, A., Dridi, K., Hesthaven, J.S.: Convergent cartesian grid methods for Maxwell's equations in complex geometries. J. Comput. Phys. **170**(1), 39–80 (2001)
5. Engquist, B., Häggblad, J., Runborg, O.: On energy preserving consistent boundary conditions for the Yee scheme in 2D. BIT **52**(3), 615–637 (2012)
6. Haggblad, J., Enquist, B.: Consistent modeling of boundaries in acoustic finite-difference time-domain simulations. J. Acoust. Soc. Am. **132**(3), 1303–1310 (2012)
7. Hao, Y., Railton, C.: Analyzing electromagnetic structures with curved boundaries on Cartesian FDTD meshes. IEEE Trans. Microw. Theory Tech. **46**(1), 82–88 (1998)
8. Holland, R.: Pitfalls of staircase meshing. IEEE Trans. Electromagn. C. **35**(4), 434–439 (1993)
9. Jurgens, T., Taflove, A., Umashankar, K., Moore, T.: Finite-difference time-domain modeling of curved surfaces. IEEE Trans. Antennas Propag. **40**(4), 357–366 (1992)
10. Madsen, N.K.: Divergence preserving discrete surface integral methods for Maxwell's curl equations using non-orthogonal unstructured grids. J. Comput. Phys. **119**(1), 34–45 (1995). DOI http://dx.doi.org/10.1006/jcph.1995.1114
11. Maloney, J.G., Cummings, K.E.: Adaptation of FDTD techniques to acoustic modeling. In: 11th Annual Review of Progress in Applied Computational Electromagnetics, vol. 2, pp. 724–731. Monterey, CA (1995)
12. Railton, C., Craddock, I.: Stabilised CPFDTD algorithm for the analysis of arbitrary 3D PEC structures. IEE Proc. Microwaves Antennas Propag. **143**(5), 367–372 (1996)
13. Railton, C., Schneider, J.: An analytical and numerical analysis of several locally conformal FDTD schemes. IEEE Trans. Microw. Theory. **47**(1), 56–66 (1999)
14. Rickard, Y., Nikolova, N.: Off-grid perfect boundary conditions for the FDTD method. IEEE Trans. Microw. Theory Tech. **53**(7), 2274–2283 (2005)
15. Stephen, R.A.: Modeling sea surface scattering by the time-domain finite-difference method. J. Acoust. Soc. Am. **100**(4), 2070–2078 (1996)
16. Tornberg, A.K., Engquist, B.: Consistent boundary conditions for the Yee scheme. J. Comput. Phys. **227**(14), 6922–6943 (2008)
17. Xiao, T., Liu, Q.H.: A staggered upwind embedded boundary (SUEB) method to eliminate the FDTD staircasing error. IEEE Trans. Antennas Propag. **52**(3), 730–741 (2004)
18. Yee, K.S.: Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media. IEEE Trans. Antennas Propag. **14**, 302–307 (1966)