



# An optimal 13-point finite difference scheme for a 2D Helmholtz equation with a perfectly matched layer boundary condition

Hatef Dastour<sup>1</sup> · Wenyuan Liao<sup>1</sup>

Received: 23 July 2019 / Accepted: 26 March 2020 / Published online: 2 May 2020

© Springer Science+Business Media, LLC, part of Springer Nature 2020

## Abstract

Efficient and accurate numerical schemes for solving the Helmholtz equation are critical to the success of various wave propagation-related inverse problems, for instance, the full-waveform inversion problem. However, the numerical solution to a multi-dimensional Helmholtz equation is notoriously difficult, especially when a perfectly matched layer (PML) boundary condition is incorporated. In this paper, an optimal 13-point finite difference scheme for the Helmholtz equation with a PML in the two-dimensional domain is presented. An error analysis for the numerical approximation of the exact wavenumber is provided. Based on error analysis, the optimal 13-point finite difference scheme is developed so that the numerical dispersion is minimized. Two practical strategies for selecting optimal parameters are presented. Several numerical examples are solved by the new method to illustrate its accuracy and effectiveness in reducing numerical dispersion.

**Keywords** Helmholtz equation · Perfectly matched layer · Optimal finite difference scheme · Numerical dispersion

## 1 Introduction

One of the popular numerical methods for solving the Helmholtz equation is the finite difference method, and it has been used in numerous practical applications where the analytical solution is not available. The finite differences method can be considered as the classical and the most frequently applied method for the numerical

---

✉ Hatef Dastour  
hatef.dastour@ucalgary.ca

Wenyuan Liao  
wliao@ucalgary.ca

<sup>1</sup> Department of Mathematics & Statistics, University of Calgary, AB, T2N 1N4, Canada

simulation of wave propagation. The finite difference method can be used as an efficient forward-modeling engine of full-waveform inversion (FWI), which is a data-fitting procedure based on full-waveform modeling for extracting quantitative information from seismograms [35]. In particular, in the frequency domain, numerically solving the Helmholtz equation with high wavenumbers is a challenging task in the field of computational mathematics [5, 6, 22, 24, 28, 37]. The main challenge is that as the wavenumber increases, the accuracy of the numerical results usually deteriorates, and the solution of the Helmholtz equation oscillates drastically. This phenomenon is regarded as the pollution effect of high wavenumbers [20, 21, 37]. In particular, the pollution effect of high wavenumbers is almost inevitable in two- and three-dimensional domains [20, 37]. Due to the pollution effect, the wavenumber of the numerical solution is different from the exact wavenumber, which is known as numerical dispersion [7]. Since the numerical dispersion is closely related to the pollution effect, minimizing the numerical dispersion can effectively mitigate the pollution effect [6, 22, 28].

Finite difference frequency domain (FDFD) modeling for the generation of synthetic seismograms and cross-hole tomography has been an active field of research since the 1980s [28]. Pratt and Worthington [27] developed the classical 5-point finite difference scheme, which requires ten grid points per wavelength. However, this can lead to a linear system with a huge and ill-conditioned matrix, especially for large wavenumbers.

Furthermore, due to the limitation of the computing resources, artificial boundary conditions are used to truncate the infinite computing domain into a finite domain. However, boundary conditions on the artificial boundary are not available in general. To specify the artificial boundary conditions, one must ensure that there is no reflection at the artificial boundary. Theoretically speaking, artificial boundary conditions should absorb waves of any wavelength and any frequency without reflection [9]; however, an ideal artificial boundary condition should be computationally stable, not require extensive computational resources, and have an acceptable level of accuracy [9]. Arguably, the most popular method is the perfectly matched layer which was introduced by Bérenger in 1994 [3] and is used to eliminate artificial reflection near the boundary. PML technique introduces an artificial layer with an attenuation parameter around the interior area (the domain of interest). On the other hand, applying PML will modify the original Helmholtz equation and make it even more difficult to solve, as existing numerical methods may fail to solve the modified Helmholtz equation effectively and accurately. Moreover, Medvinsky et al. [23] compared and analyzed a number of other absorbing boundary conditions for solving the Helmholtz equation. For more details about PML, see [3, 14, 23, 30, 34].

In the past decades, many researchers devoted a great deal of efforts in the development of optimal finite difference methods to resolve these issues. In 1996, Jo et al. presented the rotated 9-point finite difference method which consists of linearly combining the two discretizations of the second derivative operator on the classical Cartesian coordinate system and the 45° rotated system [22]. In 1998, Shin and Sohn extended the idea of the rotated 9-point scheme to the 25-point formula, and they obtained a group of optimal parameters by the singular-value decomposition method [28]. Although the 25-point formula reduces the number of grid points per

wavelength to 2, the resulting matrix's bandwidth is much wider than that of the 9-point scheme. The rotated 9-point FDM was followed by another 9-point FDM, which is consistent with PML [6]. The authors also proposed global and refined choice strategies for choosing optimal parameters of the scheme based on minimizing the numerical dispersion. This idea later was extended to the optimal 27-point finite difference scheme for the 3D Helmholtz equation with PML [7]. The rotated 9-point scheme was also extended to a generalized optimal 9-point scheme for the frequency-domain scalar wave equation [5]. Moreover, in 2017, Cheng et al. [8] presented a new dispersion minimizing finite difference method in which the combination weights are determined by minimizing the numerical dispersion with a flexible selection strategy.

There has been a large number of researches on higher-order finite difference schemes for the Helmholtz equation. In particular, various fourth-order methods for time-harmonic wave propagation based on the angle of wave propagation were developed and presented by Harari et al. [19]. Furthermore, Singer et al. [29] constructed and analyzed a fourth-order compact finite difference scheme which depends on uniform grids for the two-dimensional Helmholtz equation with constant wavenumber and is less sensitive to the direction of the propagation. Moreover, an optimal compact finite difference scheme whose parameters are chosen based on minimizing the numerical dispersion was proposed by Wu [37]. Furthermore, Britt et al. [4] also constructed a fourth-order accurate finite difference scheme for the variable coefficient Helmholtz equation that reduces phase error compared with a second order. As for sixth-order finite difference methods, Sutmann [31] derived sixth-order compact finite difference schemes for the 2D and 3D Helmholtz equation with constant coefficients. Interested readers are referred to [33, 38] for more details about sixth-order finite difference schemes for the 2D and 3D Helmholtz equation. Nevertheless, many of these higher-order schemes, in particular, compact finite difference schemes, require the source term to be smooth enough to obtain higher-order accuracy, and this is not always the case in many practical problems. A possible solution is to use non-compact finite difference schemes that do not need this requirement. In 2019, Dastour et al. [10] proposed two non-compact optimal finite difference schemes, optimal 25-point and optimal 17-point finite difference schemes, for the Helmholtz equation with PML. They demonstrated that the 17-point finite difference method is inconsistent with the Helmholtz equation with PML and is impractical when different spatial increments along the  $x$ -axis and  $z$ -axis are used.

Moreover, using finite difference schemes on many practical size problems, especially in two and three dimensional, often leads to huge linear systems that are usually solved using a number of parallel and preconditioned iterative solvers [13, 16, 17]. Some of the popular methods for solving the generated linear systems are CARP-CG [18], parallel sweeping preconditioner (PSP) [15, 26], and an unsymmetric-pattern multifrontal (UMF) method for sparse LU factorization [11, 12].

In this paper, to further reduce the numerical dispersion, we combine the ideas from [6] and [8] to develop an optimal 13-point finite difference method using point-weighting strategy. The rest of this paper is organized as follows. In Section 2, we construct an optimal 13-point finite difference scheme for the Helmholtz equation with PML. We also prove that the 13-point scheme is pointwise consistent with the Helmholtz equation with PML, and the scheme is at least second-order. In Section 3,

we analyze the error between the numerical and the exact wavenumbers and propose refined and optimal choice strategies for choosing optimal parameters to minimize the numerical dispersion. In Section 4, numerical experiments are given to demonstrate the efficiency of the scheme. We show that the new method is accurate and effective in reducing numerical dispersion. Finally, in Section 5, some conclusions of this paper and possible future works are discussed.

## 2 Development of the new optimal 13-point finite difference scheme

In this work, we consider the numerical solution of the 2D Helmholtz equation with PML given by [30, 34]:

$$\frac{\partial}{\partial x} \left( A(x, z) \frac{\partial}{\partial x} p(x, z) \right) + \frac{\partial}{\partial z} \left( B(x, z) \frac{\partial}{\partial z} p(x, z) \right) + C(x, z) k^2(x, z) p(x, z) = \tilde{g}(x, z), \quad (1)$$

where  $k = 2\pi f/v$  is the wavenumber in which  $f$  and  $v$  represent the frequency and the velocity, respectively, and  $p$  is the Fourier component of the wavefield pressure. Moreover,  $A(x, z) = s_z/s_x$ ,  $B(x, z) = s_x/s_z$ , and  $C(x, z) = s_x s_z$  in which  $s_x = 1 - i\sigma_x/\omega$ ,  $s_z = 1 - i\sigma_z/\omega$  with  $\omega = 2\pi f$  denotes the angular frequency, and

$$\tilde{g} = \begin{cases} 0, & \text{inside PML} \\ g, & \text{outside PML} \end{cases}$$

with  $g$  is the Fourier transform of the source function.

Here,  $\sigma_x$  and  $\sigma_z$  are usually chosen as differentiable functions depending on the variables  $x$  and  $z$  only, respectively. For example, one may consider defining them as follows:

$$\sigma_x = \begin{cases} 2\pi a_0 f_M \left( \frac{l_x}{L_{PML}} \right)^2, & \text{inside PML,} \\ 0, & \text{outside PML,} \end{cases} \quad (2)$$

$$\sigma_z = \begin{cases} 2\pi a_0 f_M \left( \frac{l_z}{L_{PML}} \right)^2, & \text{inside PML,} \\ 0, & \text{outside PML,} \end{cases} \quad (3)$$

where  $f_M$  is the peak frequency of the source,  $L_{PML}$  is the thickness of PML, and  $l_x$  and  $l_z$  are the distance from the point  $(x, z)$  inside PML to the interface between the interior region and PML region. Furthermore,  $a_0$  is a constant, and we choose  $a_0 = 1.79$  according to the paper [39]. Equation (1) can be seen as a general form of the Helmholtz equation with its corresponding PML, since in the interior domain  $s_x = 1$  and  $s_z = 1$  lead to  $A = B = C = 1$  and the two-dimensional Helmholtz equation:

$$\Delta p(x, z) + k^2(x, z) p(x, z) = g(x, z), \quad (4)$$

where  $\Delta = \partial^2/\partial x^2 + \partial^2/\partial z^2$  is the Laplacian.

The wavelength is defined by  $\lambda = v/f$ , and the number of wavelengths in a square domain of size  $H$  equals  $H/\lambda$ . The 2D square computational domain is often normalized into  $[0, 1] \times [0, 1]$  for the convenience of analysis. Then, the dimensionless wavenumber is equal to  $2\pi f H/v$  [8, 20]. In the remainder of the paper, the wavenumber refers to a dimensionless wavenumber, which is also denoted by  $k$ .

Consider the network of grid points  $(x_m, z_n) = (x_0 + m \Delta x, z_0 + n \Delta z)$  for  $m, n = 0, 1, 2, \dots$ . Moreover, let  $p_{m,n} = p|_{x=x_m, z=z_n}$  and  $k_{m,n} = k|_{x=x_m, z=z_n}$  represent the pressure of the wavefield and the wavenumber at the location  $(x_m, z_n)$ , respectively. Moreover, the discretizations of  $A(x, z)$ ,  $B(x, z)$  and  $C(x, z)$  at point  $(m, n)$  are denoted by  $A_{m,n}$ ,  $B_{m,n}$ , and  $C_{m,n}$ , respectively. In addition, we have:

$$\begin{cases} A_{m+\frac{j}{2},n+\frac{l}{2}} = A\left(x_m + \frac{j}{2}\Delta x, z_n + \frac{l}{2}\Delta z\right), \\ B_{m+\frac{j}{2},n+\frac{l}{2}} = B\left(x_m + \frac{j}{2}\Delta x, z_n + \frac{l}{2}\Delta z\right), \quad j, l \in \{-3, -1, 0, 1, 3\}. \\ C_{m,n} = C(x_m, z_n), \end{cases} \quad (5)$$

To construct an optimal finite difference scheme for the Helmholtz equation with PML (1), we approximate  $\frac{\partial}{\partial x} \left( A \frac{\partial p}{\partial x} \right)$  and  $\frac{\partial}{\partial z} \left( B \frac{\partial p}{\partial z} \right)$  in different ways. There are two general ways for doing this, derivative-weighting schemes and point-weighting schemes. In [8], Cheng et al. discussed the main differences between a derivative-weighting scheme and a point-weighting scheme. The first difference lies in their constructions, that is, the way for discretizing the Laplacian operator with PML  $\frac{\partial}{\partial x} \left( A \frac{\partial p}{\partial x} \right) + \frac{\partial}{\partial z} \left( B \frac{\partial p}{\partial z} \right)$  [6], while the second and also more important difference lies in their capability of reducing numerical dispersion [6]. For more details about the derivative-weighting scheme, please refer to [6, 8].

In this article, we first construct our 13-point finite difference scheme based on point-weighting. For this end, we need to first approximate the first two terms of the left-hand side of (1) with fourth-order accuracy. It follows from (1) that:

$$\begin{aligned} \frac{\partial}{\partial x} \left( A \frac{\partial p}{\partial x} \right) \Big|_{x=x_m, z=z_n} &= \alpha_1 \left( A \frac{\partial p}{\partial x} \right) \Big|_{x=x_m - \frac{3\Delta x}{2}, z=z_n} + \alpha_2 \left( A \frac{\partial p}{\partial x} \right) \Big|_{x=x_m - \frac{\Delta x}{2}, z=z_n} \\ &\quad + \alpha_3 \left( A \frac{\partial p}{\partial x} \right) \Big|_{x=x_m + \frac{\Delta x}{2}, z=z_n} \\ &\quad + \alpha_4 \left( A \frac{\partial p}{\partial x} \right) \Big|_{x=x_m - \frac{3\Delta x}{2}, z=z_n}, \end{aligned} \quad (6)$$

$$\begin{aligned} \frac{\partial}{\partial z} \left( B \frac{\partial p}{\partial z} \right) \Big|_{x=x_m, z=z_n} &= \beta_1 \left( B \frac{\partial p}{\partial z} \right) \Big|_{x=x_m, z=z_n - \frac{3\Delta z}{2}} + \beta_2 \left( B \frac{\partial p}{\partial z} \right) \Big|_{x=x_m, z=z_n - \frac{\Delta z}{2}} \\ &\quad + \beta_3 \left( B \frac{\partial p}{\partial z} \right) \Big|_{x=x_m, z=z_n + \frac{\Delta z}{2}} \\ &\quad + \beta_4 \left( B \frac{\partial p}{\partial z} \right) \Big|_{x=x_m, z=z_n - \frac{3\Delta z}{2}}. \end{aligned} \quad (7)$$

Here, the coefficients  $\alpha_i$  and  $\beta_i$ ,  $i = 1, \dots, 4$  need to be determined in a way that (6) and (7) can approximate  $\frac{\partial}{\partial x} \left( A \frac{\partial p}{\partial x} \right) \Big|_{x=x_m, z=z_n}$  and  $\frac{\partial}{\partial z} \left( B \frac{\partial p}{\partial z} \right) \Big|_{x=x_m, z=z_n}$  with 4th-order accuracy. Applying the Taylor theorem on the right-hand sides of (6) and (7), and then solving the generated linear systems for  $\alpha_i$  and  $\beta_i$ , we have:

$$\begin{cases} \alpha_1 = \frac{1}{24 \Delta x}, \quad \alpha_2 = -\frac{9}{8 \Delta x}, \quad \alpha_3 = \frac{9}{8 \Delta x}, \quad \text{and} \quad \alpha_4 = -\frac{1}{24 \Delta x}, \\ \beta_1 = \frac{1}{24 \Delta z}, \quad \beta_2 = -\frac{9}{8 \Delta z}, \quad \beta_3 = \frac{9}{8 \Delta z}, \quad \text{and} \quad \beta_4 = -\frac{1}{24 \Delta z}. \end{cases}$$

Next, we need to approximate  $\frac{\partial p}{\partial x}$  and  $\frac{\partial p}{\partial z}$  at points  $(x_m - \frac{3}{2} \Delta x, z_n), \dots, (x_m, z_n + \frac{3}{2} \Delta z)$  with fourth-order accuracy, which can be done in various ways. For example, we can let:

$$\begin{aligned} \frac{\partial p}{\partial x} \Big|_{x=x_m - \frac{3h}{2}, z=z_n} &= w_1 p_{m-2,n} + w_2 p_{m-1,n} + w_3 p_{m,n} + w_4 p_{m+1,n} \\ &\quad + w_5 p_{m+2,n} \end{aligned} \tag{8}$$

with

$$w_1 = -\frac{11}{12 \Delta x}, \quad w_2 = \frac{17}{24 \Delta x}, \quad w_3 = \frac{3}{8 \Delta x}, \quad w_4 = -\frac{5}{24 \Delta x} \quad \text{and} \quad w_5 = \frac{1}{24 \Delta x}. \tag{9}$$

Therefore, the first two terms of the left-hand side of (1) can be approximated as follows:

$$\begin{aligned} \mathcal{L}_x^{(1)} p_{m,n} &= \frac{1}{\Delta x^2} \left[ -\frac{9}{8} A_{m-\frac{1}{2},n} \left( \frac{1}{24} p_{m-2,n} - \frac{9}{8} p_{m-1,n} + \frac{9}{8} p_{m,n} - \frac{1}{24} p_{m+1,n} \right) \right. \\ &\quad + \frac{1}{24} A_{m-\frac{3}{2},n} \left( -\frac{11}{12} p_{m-2,n} + \frac{17}{24} p_{m-1,n} + \frac{3}{8} p_{m,n} \right. \\ &\quad \left. \left. - \frac{5}{24} p_{m+1,n} + \frac{1}{24} p_{m+2,n} \right) \right. \\ &\quad - \frac{1}{24} A_{m+\frac{3}{2},n} \left( -\frac{1}{24} p_{m-2,n} + \frac{5}{24} p_{m-1,n} - \frac{3}{8} p_{m,n} \right. \\ &\quad \left. \left. - \frac{17}{24} p_{m+1,n} + \frac{11}{12} p_{m+2,n} \right) \right. \\ &\quad \left. + \frac{9}{8} A_{m+\frac{1}{2},n} \left( \frac{1}{24} p_{m-1,n} - \frac{9}{8} p_{m,n} + \frac{9}{8} p_{m+1,n} - \frac{1}{24} p_{m+2,n} \right) \right], \end{aligned} \tag{10}$$

$$\begin{aligned}
\mathcal{L}_z^{(1)} p_{m,n} = & \frac{1}{\Delta z^2} \left[ -\frac{9}{8} B_{m,n-\frac{1}{2}} \left( \frac{1}{24} p_{m,n-2} - \frac{9}{8} p_{m,n-1} + \frac{9}{8} p_{m,n} - \frac{1}{24} p_{m,n+1} \right) \right. \\
& + \frac{1}{24} B_{m,n-\frac{3}{2}} \left( -\frac{11}{12} p_{m,n-2} + \frac{17}{24} p_{m,n-1} + \frac{3}{8} p_{m,n} \right. \\
& \quad \left. \left. - \frac{5}{24} p_{m,n+1} + \frac{1}{24} p_{m,n+2} \right) \right. \\
& - \frac{1}{24} B_{m,n+\frac{3}{2}} \left( -\frac{1}{24} p_{m,n-2} + \frac{5}{24} p_{m,n-1} - \frac{3}{8} p_{m,n} \right. \\
& \quad \left. \left. - \frac{17}{24} p_{m,n+1} + \frac{11}{12} p_{m,n+2} \right) \right. \\
& \left. + \frac{9}{8} B_{m,n+\frac{1}{2}} \left( \frac{1}{24} p_{m,n-1} - \frac{9}{8} p_{m,n} + \frac{9}{8} p_{m,n+1} - \frac{1}{24} p_{m,n+2} \right) \right]. \quad (11)
\end{aligned}$$

Moreover, the first two terms of the left-hand side of (1) can be also approximated with second-order accuracy as follows:

$$\mathcal{L}_x^{(2)} p_{m,n} = \frac{1}{\Delta x^2} \left[ A_{m+\frac{1}{2},n} p_{m+1,n} - \left( A_{m+\frac{1}{2},n} + A_{m-\frac{1}{2},n} \right) p_{m,n} + A_{m-\frac{1}{2},n} p_{m-1,n} \right], \quad (12)$$

$$\mathcal{L}_z^{(2)} p_{m,n} = \frac{1}{\Delta z^2} \left[ B_{m,n+\frac{1}{2}} p_{m,n+1} - \left( B_{m,n+\frac{1}{2}} + B_{m,n-\frac{1}{2}} \right) p_{m,n} + B_{m,n-\frac{1}{2}} p_{m,n-1} \right]. \quad (13)$$

Alternatively, we can approximate  $p_{m-1,n}$ ,  $p_{m,n}$ ,  $\dots$ ,  $p_{m,n-1}$  and  $p_{m,n+1}$  using Taylor's theorem with second-order accuracy and then replace them with the new approximations in (12) and (13). For example, it can be seen that  $(p_{m+1,n+1} + p_{m+1,n-1})/2$  approximates  $p_{m+1,n}$  with a second order of accuracy. It follows that:

$$\begin{aligned}
\mathcal{L}_x^{(3)} p_{m,n} = & \frac{1}{2\Delta x^2} \left[ A_{m+\frac{1}{2},n} (p_{m+1,n+1} + p_{m+1,n-1}) - \left( A_{m+\frac{1}{2},n} + A_{m-\frac{1}{2},n} \right) \right. \\
& \times (p_{m,n+1} + p_{m,n-1}) + A_{m-\frac{1}{2},n} (p_{m-1,n+1} + p_{m-1,n-1}) \left. \right], \quad (14)
\end{aligned}$$

$$\begin{aligned}
\mathcal{L}_z^{(3)} p_{m,n} = & \frac{1}{2\Delta z^2} \left[ B_{m,n+\frac{1}{2}} (p_{m+1,n+1} + p_{m-1,n+1}) - \left( B_{m,n+\frac{1}{2}} + B_{m,n-\frac{1}{2}} \right) \right. \\
& \times (p_{m+1,n} + p_{m-1,n}) + B_{m,n-\frac{1}{2}} (p_{m+1,n-1} + p_{m-1,n-1}) \left. \right]. \quad (15)
\end{aligned}$$

As a result, the first two terms of the left-hand side of (1) can be approximated as follows:

$$\frac{\partial}{\partial x} \left( A \frac{\partial p}{\partial x} \right) + \frac{\partial}{\partial z} \left( B \frac{\partial p}{\partial z} \right) \approx \mathcal{L}(p_{m,n}) = \sum_{j=1}^3 b_j \left( \mathcal{L}_x^{(j)} p_{m,n} + \mathcal{L}_z^{(j)} p_{m,n} \right), \quad (16)$$

where  $b_1$ ,  $b_2$ , and  $b_3$  are parameters to be determined subject to the constraint  $\sum_{j=1}^3 b_j = 1$ .

Moreover, consider  $\mathcal{Q}_{m,n} = k_{m,n}^2 C_{m,n} p_{m,n}$ , and let:

$$I^{(1)}(\mathcal{Q}_{m,n}) = \mathcal{Q}_{m,n}, \quad (17)$$

$$\begin{aligned} I^{(2)}(\mathcal{Q}_{m,n}) &= \frac{1}{3} (\mathcal{Q}_{m-1,n} + \mathcal{Q}_{m+1,n} + \mathcal{Q}_{m,n-1} + \mathcal{Q}_{m,n+1}) \\ &\quad - \frac{1}{12} (\mathcal{Q}_{m-2,n} + \mathcal{Q}_{m+2,n} + \mathcal{Q}_{m,n-2} + \mathcal{Q}_{m,n+2}), \end{aligned} \quad (18)$$

$$I^{(3)}(\mathcal{Q}_{m,n}) = \frac{1}{4} (\mathcal{Q}_{m-1,n} + \mathcal{Q}_{m+1,n} + \mathcal{Q}_{m,n+1} + \mathcal{Q}_{m,n-1}), \quad (19)$$

$$I^{(4)}(\mathcal{Q}_{m,n}) = \frac{1}{4} (\mathcal{Q}_{m-1,n-1} + \mathcal{Q}_{m+1,n+1} + \mathcal{Q}_{m-1,n+1} + \mathcal{Q}_{m+1,n-1}). \quad (20)$$

Therefore,

$$I(k_{m,n}^2 C_{m,n} p_{m,n}) = \sum_{j=1}^4 c_j I^{(j)}(k_{m,n}^2 C_{m,n} p_{m,n}), \quad (21)$$

where  $c_j$  are parameters satisfying  $\sum_{j=1}^4 c_j = 1$ .

As a result, an optimal 13-point FDM for the Helmholtz-PML (1) can be obtained as follows:

$$\mathcal{L}(p_{m,n}) + I(k_{m,n}^2 C_{m,n} p_{m,n}) = \tilde{g}_{m,n}. \quad (22)$$

Moreover, let

$$\mathcal{L}^{(1)} = \mathcal{L}_x^{(1)} p_{m,n} + \mathcal{L}_z^{(1)} p_{m,n}, \quad (23)$$

$$\mathcal{L}^{(2)} = \mathcal{L}_x^{(2)} p_{m,n} + \mathcal{L}_z^{(2)} p_{m,n}. \quad (24)$$

We refer:

$$\mathcal{L}^{(1)}(p_{m,n}) + k_{m,n}^2 C_{m,n} p_{m,n} = \tilde{g}_{m,n}, \quad (25)$$

$$\mathcal{L}^{(2)}(p_{m,n}) + k_{m,n}^2 C_{m,n} p_{m,n} = \tilde{g}_{m,n} \quad (26)$$

as the non-compact fourth-order (NC fourth-order) and conventional 5p schemes, respectively. These two schemes will be included for comparison in our final analysis in Section 4.

**Definition 1** Let  $(x_m, z_n) = (x_0 + m\Delta x, z_0 + n\Delta z)$  for  $m, n = 0, 1, 2, \dots$ , and suppose that the partial differential equation under consideration is  $(\Delta + k^2)p = g$ , and the corresponding finite difference approximation is  $\mathcal{L}P_{m,n} = G_{m,n}$  where  $G_{m,n} = g(x_n, y_n)$ . The finite difference scheme  $\mathcal{L}P_{m,n} = G_{m,n}$  is pointwise consistent with the partial differential equation  $(\Delta + k^2)p = g$  at  $(x, z)$ , if for any smooth function  $\phi(x, z)$ :

$$\left\| \left( (\Delta + k^2)\phi - g \right) \Big|_{x=x_m, z=z_n} - [\mathcal{L}\phi(x_m, z_n) - G_{m,n}] \right\| \rightarrow 0 \quad (27)$$

as  $\Delta x, \Delta z \rightarrow 0$ .

In many practical applications, different step sizes  $\Delta x$  and  $\Delta z$  for variables  $x$  and  $z$ , respectively, are used. For simplicity, we set  $\gamma = \Delta z/\Delta x$ , and let  $\Delta x = h$ ,  $\Delta z = \gamma h$  and  $\eta = 1 + 1/\gamma^2$ .

**Proposition 1** If  $\sum_{j=1}^2 b_j = 1$  and  $\sum_{j=1}^4 c_j = 1$ , then the finite difference approximation (22) is pointwise consistent with the Helmholtz-PML (1) and is second-order.

*Proof* Let  $(x, z) \in [x_m, x_{m+1}) \times [z_n, z_{n+1})$ . It follows from Taylor's theorem that:

$$\mathcal{L}(p) = \frac{\partial}{\partial x} \left( A \frac{\partial p}{\partial x} \right) + \frac{\partial}{\partial z} \left( B \frac{\partial p}{\partial z} \right) + \xi_1 h^2 + \xi_2 h^4 + O(h^6), \quad (28)$$

$$I(k^2 C p) = k^2 C p + \xi_3 h^2 + \xi_4 h^4 + O(h^6), \quad (29)$$

where  $\xi_1, \xi_2, \xi_3$ , and  $\xi_4$  are given as follows:

$$\begin{aligned} \xi_1 = & \frac{1}{24} \left[ (b_2 + b_3) \left( 2A \frac{\partial^4}{\partial x^4} p + 2B \gamma^2 \frac{\partial^4}{\partial z^4} p + 4 \frac{\partial^3}{\partial x^3} p \frac{\partial}{\partial x} A + 3 \frac{\partial^2}{\partial x^2} p \frac{\partial^2}{\partial x^2} A \right. \right. \\ & + 3\gamma^2 \frac{\partial^2}{\partial z^2} p \frac{\partial^2}{\partial z^2} B + \frac{\partial}{\partial x} p \frac{\partial^3}{\partial x^3} A + \gamma^2 \frac{\partial}{\partial z} p \frac{\partial^3}{\partial z^3} B + 4\gamma^2 \frac{\partial}{\partial z} B \frac{\partial^3}{\partial z^3} p \left. \left. \right) \right. \\ & \left. + 12b_3 \left( (A \gamma^2 + B) \frac{\partial^2}{\partial z^2} \frac{\partial^2}{\partial x^2} p + \frac{\partial}{\partial z} B \frac{\partial}{\partial z} \frac{\partial^2}{\partial x^2} p + \gamma^2 \frac{\partial}{\partial x} A \frac{\partial^2}{\partial z^2} \frac{\partial}{\partial x} p \right) \right], \end{aligned} \quad (30)$$

$$\begin{aligned}
\xi_2 = & -\frac{1}{5760} \left[ 30(9b_1 - 2b_2 - 2b_3) \left( \frac{\partial^2}{\partial x^2} A \frac{\partial^4}{\partial x^4} p + \gamma^4 \frac{\partial^2}{\partial z^2} B \frac{\partial^4}{\partial z^4} p \right) \right. \\
& + 3(9b_1 - b_2 - b_3) \left( \frac{\partial^5}{\partial x^5} A \frac{\partial}{\partial x} p + \gamma^4 \frac{\partial^5}{\partial z^5} B \frac{\partial}{\partial z} p \right. \\
& \quad \left. \left. + 5 \left( \frac{\partial^4}{\partial x^4} A \frac{\partial^2}{\partial x^2} p + \gamma^4 \frac{\partial^4}{\partial z^4} B \frac{\partial^2}{\partial z^2} p \right) \right) \right] \\
& + 16(4b_1 - b_2 - b_3) \left( A \frac{\partial^6}{\partial x^6} p + B \gamma^4 \frac{\partial^6}{\partial z^6} p \right. \\
& \quad \left. + 3 \left( \frac{\partial}{\partial x} A \frac{\partial^5}{\partial x^5} p + \gamma^4 \frac{\partial}{\partial z} B \frac{\partial^5}{\partial z^5} p \right) \right) \\
& + 10(27b_1 - 4b_2 - 4b_3) \left( \frac{\partial^3}{\partial x^3} A \frac{\partial^3}{\partial x^3} p + \gamma^4 \frac{\partial^3}{\partial z^3} B \frac{\partial^3}{\partial z^3} p \right) \\
& - 120b_3 \left( 3\gamma^2 \left( \frac{\partial^2}{\partial x^2} A + \frac{\partial^2}{\partial z^2} B \right) \frac{\partial^2}{\partial z^2} \frac{\partial^2}{\partial x^2} p \right. \\
& \quad \left. + \gamma^2 \left( \frac{\partial^3}{\partial z^3} B \frac{\partial}{\partial z} \frac{\partial^2}{\partial x^2} p + \frac{\partial^3}{\partial x^3} A \frac{\partial^2}{\partial z^2} \frac{\partial}{\partial x} p \right) \right. \\
& \quad \left. + \left( 4\gamma^2 \frac{\partial^2}{\partial z^2} \frac{\partial^3}{\partial x^3} p + 2\gamma^4 \frac{\partial^4}{\partial z^4} \frac{\partial}{\partial x} p \right) \frac{\partial}{\partial x} A \right. \\
& \quad \left. + \left( 2 \frac{\partial}{\partial z} \frac{\partial^4}{\partial x^4} p + 4\gamma^2 \frac{\partial^3}{\partial z^3} \frac{\partial^2}{\partial x^2} p \right) \frac{\partial}{\partial z} B \right. \\
& \quad \left. \left. + 2 \left( Ag^2 + B \right) \left( \frac{\partial^2}{\partial z^2} \frac{\partial^4}{\partial x^4} p + \gamma^2 \frac{\partial^4}{\partial z^4} \frac{\partial^2}{\partial x^2} p \right) \right) \right], \tag{31}
\end{aligned}$$

$$\xi_3 = \frac{1}{4} (c_3 + 2c_4) \left( \frac{\partial^2}{\partial x^2} + \gamma^2 \frac{\partial^2}{\partial z^2} \right) (k^2 C p), \tag{32}$$

$$\xi_4 = -\frac{1}{48} \left( (4c_2 - c_3 - 2c_4) \left( \frac{\partial^4}{\partial x^4} + \gamma^4 \frac{\partial^4}{\partial x^4} \right) - 12\gamma^2 c_4 \frac{\partial^4}{\partial z^2 \partial x^2} \right) (k^2 C p). \tag{33}$$

Then, the left-hand side of (22) is equivalent to:

$$\begin{aligned}
\mathcal{L}(p) + I(k^2 C p) = & \frac{\partial}{\partial x} \left( A \frac{\partial p}{\partial x} \right) + \frac{\partial}{\partial z} \left( B \frac{\partial p}{\partial z} \right) + C k^2 p \\
& + (\xi_1 + \xi_3) h^2 + (\xi_2 + \xi_4) h^4 + O(h^6). \tag{34}
\end{aligned}$$

The results of this proposition can be concluded from (34) and (1).  $\square$

What can be seen from the above proposition is that the finite difference approximation (11) is second-order for arbitrary constants  $b_i$  and  $c_j$  under the conditions  $\sum_{i=1}^3 b_i = 1$  and  $\sum_{j=1}^4 c_j = 1$ . However, if  $b_2, b_3, c_3$ , and  $c_4$  are chosen from values close to 0, then the finite difference approximation (11) can reach fourth-order.

This means the order of accuracy of finite difference approximation (11) can vary between two and four.

### 3 Numerical dispersion analysis and parameter selection strategy

In this section, a numerical dispersion analysis for the new difference scheme (22) is presented. To do dispersion analysis, consider a homogeneous model with constant velocity  $v$ . Let  $P(x, z) = \exp(-ik(x \cos \theta + z \sin \theta))$ , where  $\theta$  is the propagation angle from the  $z$ -axis, and the wavenumber  $k = 2\pi f/v$  is a positive constant.

In the interior area,  $A = B = C = 1$ ; thus, replacing  $p_{m+i,n+j}$  with  $P_{m+i,n+j}$  ( $i, j \in \mathbb{Z}_3$ ) in the formula (11) gives:

$$\begin{aligned} & \hat{T}_1 P_{m,n-2} + \hat{T}_2 P_{m-1,n-1} + \hat{T}_3 P_{m,n-1} + \hat{T}_2 P_{m+1,n-1} + \hat{T}_4 P_{m-2,n} + \hat{T}_5 P_{m-1,n} \\ & + \hat{T}_6 P_{m,n} + \hat{T}_5 P_{m+1,n} + \hat{T}_4 P_{m+2,n} + \hat{T}_2 P_{m-1,n+1} + \hat{T}_3 P_{m,n+1} + \hat{T}_2 P_{m+1,n+1} \\ & + \hat{T}_1 P_{m,n+2} = 0, \end{aligned} \quad (35)$$

where

$$\begin{cases} \hat{T}_1 = \frac{(1-\eta)b_1}{12h^2} - \frac{c_2}{12}k^2, & \hat{T}_2 = \frac{(1-b_1-b_2)\eta}{2h^2} + \frac{c_4}{4}k^2, \\ \hat{T}_3 = \frac{(4\eta-1)b_1+3(b_2\eta-1)}{3h^2} + \frac{4c_2+3c_3}{12}k^2, & \hat{T}_4 = -\frac{b_1}{12h^2} - \frac{c_2}{12}k^2, \\ \hat{T}_5 = \frac{(3\eta+1)b_1+3(b_2\eta-\eta+1)}{3h^2} + \frac{4c_2+3c_3}{12}k^2, & \hat{T}_6 = -\frac{(5b_1+4b_2)\eta}{2h^2} + (1-c_2-c_3-c_4)k^2. \end{cases}$$

Let  $\lambda = 2\pi v/\omega$  and  $G = \lambda/h$  denote the wavelength and the number of grid points per wavelength, respectively. Moreover, let:

$$\begin{cases} P = \cos(k_x \Delta x) = \cos(kh \cos \theta) = \cos((2\pi/G) \cos \theta), \\ Q = \cos(k_z \Delta z) = \cos(\gamma kh \sin \theta) = \cos((2\gamma\pi/G) \sin \theta). \end{cases} \quad (36)$$

It follows from substituting  $P_{m,n} = \exp(-ik(x \cos \theta + z \sin \theta))$  into (35), and simplifying that:

$$(4Q^2 - 2) \hat{T}_1 + 4PQ\hat{T}_2 + 2Q\hat{T}_3 + (4P^2 - 2) \hat{T}_4 + 2P\hat{T}_5 + \hat{T}_6 = 0. \quad (37)$$

Furthermore, let  $k_N$  represent the numerical wavenumber. It follows from replacing the variable  $k$  in the parameters  $\hat{T}_1, \hat{T}_2, \dots, \hat{T}_6$  with  $k_N$  in (37) that:

$$k_N = \frac{1}{h} \sqrt{\frac{N}{D}}, \quad (38)$$

where

$$\begin{aligned} N &= -2 \left( P^2 + (\eta - 1) Q^2 + 2(Q - P) + \eta(6P(Q - 1) - 8Q + 7) \right) b_1 \\ &\quad - 12\eta(P - 1)(Q - 1)b_2 + 12((\eta P - 1)Q + (1 - \eta)P), \end{aligned} \quad (39)$$

$$D = 2 \left( P^2 + Q^2 - 2(P + Q - 1) \right) c_2 - 3(P + Q - 2)c_3 - 6(PQ - 1)c_4 - 6. \quad (40)$$

The next proposition presents the error between the numerical wavenumber  $k_N$  and the exact wavenumber  $k$  for the finite difference scheme (22).

Moreover, Bayliss et al. [2] demonstrated that in practical applications, the mesh size and the wavenumber are correlated with the accuracy of the desired computation. They showed that the number of points per wavelength is not sufficient to determine the accuracy of a given discretization. In the next proposition, the importance of the pollution effect in numerical computations is highlighted.

**Proposition 2** *For the finite difference scheme (22), there holds:*

$$(k_N)^2 = k^2 \left( 1 + O(k^2 h^2) \right), \quad k h \rightarrow 0. \quad (41)$$

*Proof* Let  $\tau = kh$ ,  $P(\tau) = \cos(\tau \cos \theta)$  and  $Q(\tau) = \cos(\gamma \tau \sin \theta)$ . Then, the (38) can be written as:

$$(k_N)^2 = \frac{1}{h^2} \frac{N(\tau)}{D(\tau)}, \quad (42)$$

where

$$\begin{aligned} N(\tau) = & 2 \left( \gamma^2 \left( P^2(\tau) - 8P(\tau) + 7 \right) + Q^2(\tau) - 8Q(\tau) + 7 \right) b_1 \\ & - 12 \left( \gamma^2 (P(\tau) - 1) Q(\tau) + P(\tau) (Q(\tau) - 1) \right) b_3 \\ & - 12 \left( \gamma^2 (P(\tau) - 1) + Q(\tau) - 1 \right) b_2, \end{aligned}$$

$$\begin{aligned} D(\tau) = & \gamma^2 (6c_1 - 2(P(\tau)(P(\tau) - 2) + Q(\tau)(Q(\tau) - 2) - 1)c_2 \\ & + 3(P(\tau) + Q(\tau))c_3 + 6P(\tau)Q(\tau)c_4). \end{aligned}$$

Applying Taylor theorem on  $N(\tau)$  and  $\frac{1}{D(\tau)}$  at the point  $\tau = 0$ , we have:

$$\begin{aligned} N(\tau) = & 6\gamma^2\tau^2 - \frac{\gamma^2\tau^4}{2} \left[ b_2 + b_3 - 2 \left( b_2 - (3\gamma^2 + 2)b_3 \right) \sin^2(\theta) \right. \\ & \left. + (\gamma^2 + 1)(b_2 - 5b_3) \sin^4(\theta) \right] \\ & - \frac{\gamma^2\tau^6}{60} [4b_1 - b_2 - b_3 - 3(4b_1 - b_2 \\ & + b_3(5\gamma^2 + 4)) \sin^2(\theta) + (12b_1 - 3b_2 + 3b_3(-5\gamma^4 + 5\gamma^2 + 9)) \\ & \times \sin^4(\theta) + (\gamma^4 - 1)(4b_1 - b_2 + 14b_3) \sin^6(\theta)] + O(\tau^8), \quad (43) \end{aligned}$$

$$\begin{aligned} \frac{1}{D(\tau)} = & \frac{1}{6\gamma^2} + \frac{\tau^2}{24\gamma^2} (c_3 + 2c_4) \left( (\gamma^2 - 1) \sin(\theta)^2 + 1 \right) \\ & + \frac{\tau^4}{288\gamma^2} \left[ \gamma^4 \sin^4(\theta) (4c_2 + c_3(3c_3 + 12c_4 - 1) + 2c_4(6c_4 - 1)) \right. \\ & + 6\gamma^2 \cos^2(\theta) \sin^2(\theta) \left( 2c_4(2c_3 + 2c_4 - 1) + c_3^2 \right) \\ & \left. + \cos^4(\theta) (4c_2 + 2c_4(6c_3 + 6c_4 - 1) + c_3(3c_3 - 1)) \right] + O(\tau^6). \quad (44) \end{aligned}$$

It follows from (42), (43), and (44) that:

$$(k_N)^2 = k^2 \left( 1 - \frac{k^2 h^2}{12} \chi_1 - \frac{k^4 h^4}{720} \chi_2 + O(k^6 h^6) \right), \quad kh \rightarrow 0. \quad (45)$$

with

$$\begin{aligned} \chi_1 &= b_2 + b_3 - 3c_3 - 6c_4 + (\gamma^2 + 1)(b_2 - 5b_3) \sin^4(\theta) \\ &\quad + (-2b_2 + (6\gamma^2 + 4)b_3 + 3(1 - \gamma^2)c_3 + 6(1 - \gamma^2)c_4) \sin^2(\theta), \end{aligned} \quad (46)$$

$$\begin{aligned} \chi_2 &= 30((5b_1 + 6b_2 - 5)c_4 + 15(5b_1 + 6b_2 - 5)c_3 - 10(2b_1 + 3b_2) + 28) \\ &\quad \times \sin^6(\theta) v + (-15b_1(6\gamma^4 - 7\gamma^2 - 9)(c_3 + 2c_4) - 30b_2(\gamma^2 - 2) \\ &\quad \times (\gamma^2 + 1)(3c_3 + 6c_4 - 1)v + 15\gamma^2(\gamma^2 - 1)(2b_1 + 7c_3 - 2) \\ &\quad - 6(5(b_1 + 4c_3) - 10c_2(\gamma^4 - 1) - 5c_4(7\gamma^4 - 13\gamma^2 - 8) - 9) \\ &\quad - 45(\gamma^4 - 2\gamma^2 + 1)(c_3 + 2c_4)(c_3 + 2c_4)) \sin^4(\theta) \\ &\quad + (3(40c_2 + 5c_3) - 15b_1(c_3 + 2c_4)(7\gamma^2 + 3) - 30b_2(\gamma^2 + 1) \\ &\quad \times (3c_3 + 6c_4 - 1) - 90(c_3 + 2c_4)^2(\gamma^2 - 1) \\ &\quad + 15(2b_1 + 7c_3 + 26c_4 - 2)\gamma^2 + 3(10c_4 - 8)) \sin^2(\theta) \\ &\quad - 30c_4(b_1 + 6c_3 - 2) - 15c_3(c_1 - 2) - 45(c_3^2 + 4c_4^2) \\ &\quad + 10(b_1 - 6c_2) - 2. \end{aligned} \quad (47)$$

□

The above proposition indicates that  $k_N$  approximates  $k$  with second-order accuracy for arbitrary constants  $b_i$  and  $c_j$  under the conditions  $\sum_{i=1}^3 b_i = 1$  and  $\sum_{j=1}^4 c_j = 1$ . Moreover, the term associated with  $k^4 h^4$  presents the pollution effect, which depends on the wavenumber  $k$ , the parameters of the finite difference formula, and the wave propagation angle  $\theta$  from the  $z$ -axis. It also can be seen from Proposition (2) that  $k_N$  can approximate  $k$  with fourth-order accuracy when  $b_2$ ,  $b_3$ ,  $c_3$ , and  $c_4$  amount to 0.

In the remainder of this section, we incorporate the refined choice strategy (rule 3.8 from [6]) and present an algorithm for parameter selection of the optimal 13-point FDM (22) based on minimizing the numerical dispersion.

Given  $h = \frac{2\pi}{Gk}$ , the relationship of the numerical wavenumber  $k_N$  and the exact wavenumber  $k$  can be presented as follows:

$$\frac{k_N}{k} = \frac{G}{2\pi} \sqrt{\frac{N}{D}}. \quad (48)$$

Furthermore, from the physical point of view, the ratio  $k_N/k$  amounts to the normalized phase velocity (see [8, 22, 25, 28, 32] for more details). The normalized phase velocity is a reliable tool for measuring the numerical dispersion. The

normalized numerical phase velocity and the normalized numerical group velocity can be found as follows, respectively [6, 22, 28, 32]:

$$\frac{V_{ph}^N}{v} = \frac{G}{2\pi} \sqrt{\frac{N}{D}}, \quad (49)$$

$$\frac{V_{gr}^N}{v} = \frac{v}{V_{ph}^N} \left[ \frac{\left( \frac{1}{h} \frac{\partial N}{\partial k} \right) D - N \left( \frac{1}{h} \frac{\partial D}{\partial k} \right)}{D^2} \right]. \quad (50)$$

It can be seen that there would be no numerical dispersion if the normalized numerical phase velocity equals 1. Therefore, to minimize the error between  $k_N$  and  $k$ , one can estimate the parameters of the finite difference scheme (22) in a way that the normalized numerical phase velocity can maintain its value close to 1.

Consider the following function:

$$J(b_1, \dots, c_4; G, \theta) = \frac{G}{2\pi} \sqrt{\frac{N}{D}} - 1, \quad (51)$$

where  $b_1 \in (0, 1]$ , and  $b_2, c_2, c_3, c_4 \in \mathbb{R}$ , and  $(G, \theta) \in I_G \times I_\theta$  with  $I_G$  and  $I_\theta$  are two intervals. In general, one can choose  $I_\theta = [0, \frac{\pi}{2}]$  and  $I_G = [G_{\min}, G_{\max}] \subseteq [2, 400]$ . We remark that the interval  $[0, \frac{\pi}{2}]$  can be replaced by  $[0, \frac{\pi}{4}]$  because of the symmetry, and  $G_{\min} \geq 2$  based on the Nyquist sampling limit (see [28] for more details).

Therefore, minimizing the error between  $k_N$  and  $k$  is equivalent to minimizing the norm  $\|J(b_1, \dots, c_4; \cdot, \cdot)\|_{\infty, I_G \times I_\theta}$ . One way to estimate the optimal parameters is to solve:

$$(b_1, \dots, c_4) = \arg \min \{ \|J(b_1, \dots, c_4; G, \theta)\|_{I_G \times I_\theta} : b_1 \in (0, 1], b_2, c_2, c_3, c_4 \in \mathbb{R} \} \quad (52)$$

using the least-squares method.

Therefore, it follows from  $J(b_1, \dots, c_4; G, \theta) = 0$  that:

$$\frac{G^2}{4\pi^2} \frac{N}{D} = 1, \quad (53)$$

and

$$\begin{aligned} & G^2 ((P - Q)(P + Q - 2) + \eta(Q - 1)(6P + Q - 7)) b_1 + 6\eta G^2 (Q - 1) \\ & (P - 1) b_2 + 4\pi^2 (P^2 + Q^2 + 2(1 - P - Q)) c_2 - 6\pi^2 (P + Q - 2) c_3 \\ & - 12\pi^2 (PQ - 1) c_4 = 6(P(Q - 1)\eta + P - Q) G^2 + 12\pi^2. \end{aligned} \quad (54)$$

Let

$$\left\{ \begin{array}{l} \theta = \theta_m = \frac{(m-1)}{4(l-1)}\pi \in I_\theta = \left[0, \frac{\pi}{2}\right], \quad m = 1, 2, \dots, l, \\ \frac{1}{G} = \frac{1}{G_n} = \frac{1}{G_{\max}} + (n-1) \frac{\frac{1}{G_{\min}} - \frac{1}{G_{\max}}}{r-1} \in \left[\frac{1}{G_{\max}}, \frac{1}{G_{\min}}\right], \quad n = 1, 2, \dots, r. \end{array} \right.$$

Then, (54) leads to the following overdetermined linear system:

$$\begin{bmatrix} S_{1,1}^1 & S_{1,1}^2 & S_{1,1}^3 & S_{1,1}^4 & S_{1,1}^5 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ S_{1,r}^1 & S_{1,r}^2 & S_{1,r}^3 & S_{1,r}^4 & S_{1,r}^5 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ S_{m,n}^1 & S_{m,n}^2 & S_{m,n}^3 & S_{m,n}^4 & S_{m,n}^5 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ S_{l,r}^1 & S_{l,r}^2 & S_{l,r}^3 & S_{l,r}^4 & S_{l,r}^5 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix} = \begin{bmatrix} S_{1,1}^6 \\ \vdots \\ S_{1,r}^6 \\ \vdots \\ S_{m,n}^6 \\ \vdots \\ S_{l,r}^6 \end{bmatrix}, \quad (55)$$

where

$$\begin{cases} S_{m,n}^1 = G_n^2 ((P_{m,n} - Q_{m,n})(P_{m,n} + Q_{m,n} - 2) \\ \quad + \eta(Q_{m,n} - 1)(6P + Q_{m,n} - 7)), \\ S_{m,n}^2 = 6\eta G_n^2 (Q_{m,n} - 1)(P_{m,n} - 1), \\ S_{m,n}^3 = 4\pi^2 (P_{m,n}^2 + Q_{m,n}^2 + 2(1 - P_{m,n} - Q_{m,n})), \\ S_{m,n}^4 = -6\pi^2 (P_{m,n} + Q_{m,n} - 2), \\ S_{m,n}^5 = -12\pi^2 (PQ - 1), \\ S_{m,n}^6 = 6(P_{m,n}(Q_{m,n} - 1)\eta + P_{m,n} - Q_{m,n})G_n^2 + 12\pi^2. \end{cases} \quad (56)$$

with

$$P_{m,n} = \cos((2\pi/G_n)\cos(\theta_m)), \quad Q_{m,n} = \cos((2\pi\gamma/G_n)\sin(\theta_m)). \quad (57)$$

As it was mentioned before, the linear system (55) can be solved using the least-squares method.

We next present refined and optimal algorithms for parameter selection and reducing the numerical dispersion and improving the accuracy of the 13-point finite difference scheme (22). The optimal algorithm is based on the refined choice strategy (rule 3.8 from [6]). According to the rule, first, the interval  $I_G = [G_{\min}, G_{\max}]$  is estimated by using a priori information. For example, for a given step size  $h$ ,  $I_G$  can be considered as follows:

$$I_G = \left[ \frac{v_{\min}}{hf_{\max}}, \frac{v_{\max}}{hf_{\min}} \right], \quad (58)$$

where  $f \in [f_{\min}, f_{\max}]$  and  $v \in [v_{\min}, v_{\max}]$  are the frequency and the velocity, respectively.

Then, the parameters of the finite difference scheme (22) are estimated such that  $(b_1, \dots, c_4) = \arg \min \{ \|J(b_1, \dots, c_4; G, \theta)\|_{I_G \times I_\theta} : b_1 \in (0, 1], b_2, \dots, c_4 \in \mathbb{R}\}$ . (59)

In the remainder of this article, we refer to the 13-point finite difference scheme (22) whose parameters are estimated using the refined choice strategy as the refined 13-point finite difference scheme (refined 13p).

Moreover, according to Proposition (1) and Proposition (2), the order of accuracy of the 13-point finite difference scheme can vary between second and fourth orders. Especially, when  $b_2, b_3, c_3$ , and  $c_4$  amount to 0. We also know that achieving a

certain accuracy requires a minimum number of grid points per wavelength,  $G$ . This number is usually smaller with high-order methods than with low-order methods [36]. Therefore, we propose Algorithm 1 by which the method can obtain a higher order of accuracy whenever  $G_{min}$  is greater than a certain value, which is introduced here by  $G_{mid}$ . Experimentally, we observed that  $G_{mid}$  can take a value greater than 10. Basically, we let all parameters associated with second-order schemes be equal to 0. That is,  $b_2 = b_3 = c_3 = c_4 = 0$ . It follows from (53) and (55) that:

$$\begin{bmatrix} W_{1,1}^1 & W_{1,1}^2 \\ \vdots & \vdots \\ W_{1,r}^1 & W_{1,r}^2 \\ \vdots & \vdots \\ W_{m,n}^1 & W_{m,n}^2 \\ \vdots & \vdots \\ W_{l,r}^1 & W_{l,r}^2 \end{bmatrix} \begin{bmatrix} b_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} W_{1,1}^3 \\ \vdots \\ W_{1,r}^3 \\ \vdots \\ W_{m,n}^3 \\ \vdots \\ W_{l,r}^3 \end{bmatrix}, \quad (60)$$

where

$$\begin{cases} W_{m,n}^1 = (P(8-P) + Q(1-\eta)(Q-8) - 7\eta) G_n^2, \\ W_{m,n}^2 = 4(P(2-P) + Q(2-Q) - 2)\pi^2, \\ W_{m,n}^3 = -12\pi^2. \end{cases} \quad (61)$$

As a result, we propose the optimal 13-point finite difference scheme (optimal 13p), whose parameters are estimated using Algorithm 1.

---

**Algorithm 1** Optimal parameters selection for the scheme (22)

---

**Data:**  $v$ ,  $f$  and  $h$

**Result:**  $b_1, b_2, c_2, c_3$  and  $c_4$

initialization;

**if**  $G_{min} < G_{mid}$  **then**

| Solve the least square problem (55) for  $b_1, b_2, \dots, c_4$ ;

**else**

| Solve the least square problem (60) with  $b_1 = 1$  for  $c_2$ ;

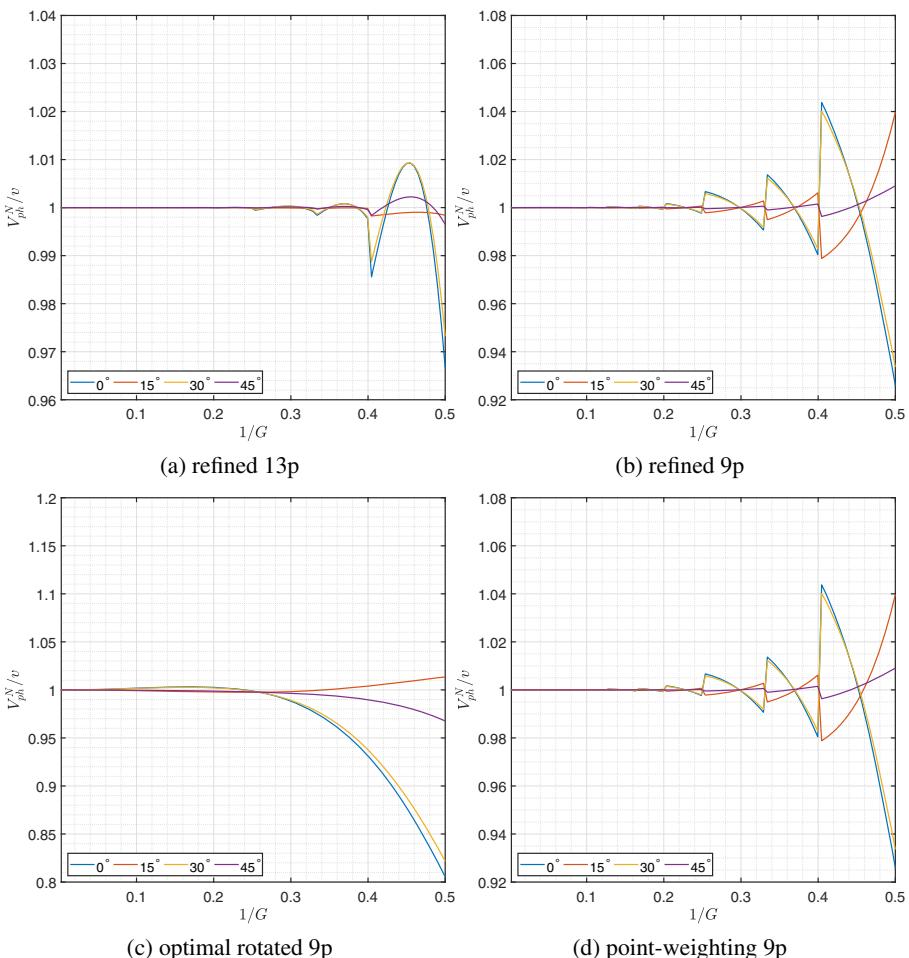
---

From now on, for simplicity, we present all of our following results for the case that  $\Delta x = \Delta z$ . Moreover, let refined 9p represent the optimal 9-point finite difference scheme [6] such that its parameters are estimated using the refined choice strategy (rule 3.8 from [6]). In addition, optimal rotated 9p represents the rotated 9-point FDM [6, 22] with parameters  $a = 0.5461$ ,  $d = 0.3752$ , and  $e = -4 \times 10^{-5}$ . This group of optimal parameters is provided by Jo, Shin, and Suh [6, 22] as optimal parameters. Moreover, let point-weighting 9p represent the point-weighting scheme

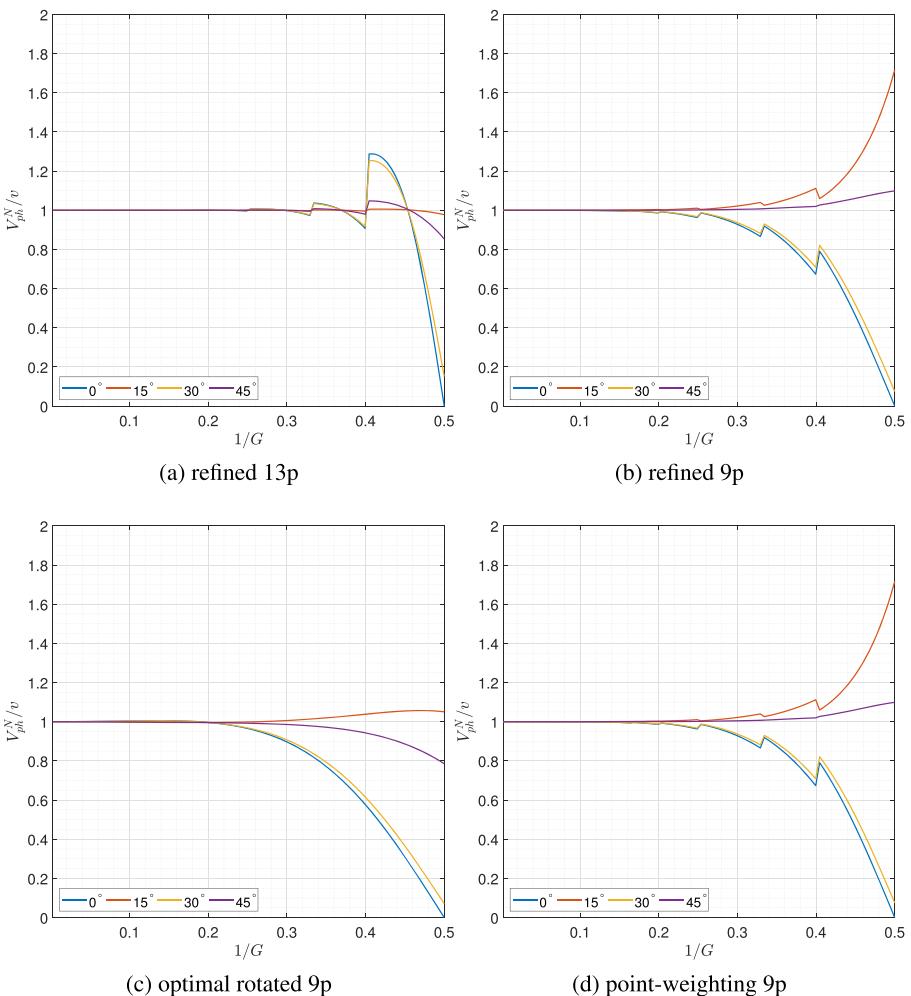
[8] whose parameters are estimated using the flexible strategy for selection of weights (rule 3.3 at p. 2354 of [8]).

The normalized phase and group velocity curves of refined 13p, refined 9p, optimal rotated 9p, and point-weighting 9p are presented in Figs. 1 and 2, respectively. The optimal parameters are estimated on  $I_G$  intervals [2, 2.5], [2.5, 3] [4, 5], [5, 6], [6, 8], [8, 10], and [10, 400]. It can be seen that refined 13p has the least dispersion among all schemes in the study.

Furthermore, since refined 13p and optimal 13p are basically the same schemes (except when  $G_{min} > G_{mid}$ ), we only presented normalized phase and group velocity curves for one of them, refined 13p. With  $G_{mid} = 10$ , the normalized phase and group velocity curves of these two schemes will be identical. However, in terms of computational costs, efficiency, and accuracy, optimal 13p has some advantages over refined 13p. This will be discussed in Example 4.1.



**Fig. 1** Normalized phase velocity curves for various schemes



**Fig. 2** Normalized group velocity curves for various schemes

## 4 Numerical experiments

In this section, we present three numerical examples. Example 4.1 is meant for illustrating the accuracy and efficiency of the new schemes and comparing them with some other optimal finite difference schemes that are widely used for solving the Helmholtz equation with PML. To test the order of accuracy, in Example 4.1, Dirichlet boundary conditions are imposed on the boundary. However, we would like to emphasize that solving the Helmholtz equation with available boundary conditions is not the main focus of this article. There are a large number of efficient and high-order finite difference methods such as [4, 24, 29, 31, 33, 37, 38] that are compact methods. To our best knowledge, currently, there is no compact finite difference method

compatible with the Helmholtz equation with PML. In Example 4.2, we analyze the numerical solutions of the schemes, that we use in 4.1, and compare them with the exact solution in a homogeneous model. Finally, in Example 4.3, a more realistic problem is solved by the new schemes, refined 13p and optimal 13p.

In addition to those that have been introduced before, the new schemes, refined 13p and optimal 13p, are also compared against the refined 25p and the global 25p from [10] that are the optimal 25-point schemes whose parameters are estimated using the refined and global strategies, respectively.

#### 4.1 Example 1

Consider:

$$\Delta p + k^2 p = g(x, z), \quad \text{in } \Omega := (0, 1) \times (0, 1), \quad (62)$$

with

$$(k_x, k_z) = k_0 \left( e^{-k_0(x+z)} + 1 \right) (\cos(\theta), \sin(\theta)), \quad (63)$$

$$g(x, z) = e^{ik_0(x \cos(\theta) + z \sin(\theta))} \left[ \sin(\pi x) \sin(\pi z) \left( k_0^2 e^{-2k_0(x+z)} \left( 2e^{k_0(x+z)} + 1 \right) - 2\pi^2 \right) - 2\pi i k_0 (\cos(\pi x) \sin(\pi z) \cos(\theta) + \cos(\pi z) \sin(\pi x) \sin(\theta)) \right]. \quad (64)$$

Dirichlet boundary conditions are imposed on the boundary, and its analytical solution is available:

$$p(x, z) = \sin(\pi x) \sin(\pi z) e^{-i(xk_x + zk_z)}. \quad (65)$$

In this example, we compare refined 13p and optimal 13p with a number of popular optimal finite difference schemes that are used for the Helmholtz equation with PML. The new schemes, refined 13p and optimal 13p, are compared against refined 25p [10], the global 25p [10], refined 9p [6], optimal rotated 9p [6, 22], point-weighting 9p [8], non-compact fourth-order (NC 4th-order) (25), and conventional 5p (26). Moreover, the interval  $I_G = [G_{\min}, G_{\max}] = \left[ \frac{2\pi}{hk_{\max}}, \frac{2\pi}{hk_{\min}} \right]$  is estimated by using a priori information.

All the experiments in this example are performed with MATLAB 9.5.0.1067069 (R2018b) Update 4 on a Dell laptop equipped with Windows 10 Home Edition (64-bit), Intel(R) Core(TM) i5-4210U CPU, and 8.00 GB physical memory (RAM). We used an unsymmetric-pattern multifrontal (UMF) method for sparse LU factorization [11, 12] for solving linear systems generated by each FDM.

Additionally, the error between the numerical solution and the exact solution is measured in C-norm [6], which is defined for any  $M \times N$  complex matrix  $\mathbf{Z}$  as:

$$\|\mathbf{Z}\|_C = \max_{1 \leq i \leq M, 1 \leq j \leq N} |z_{i,j}|. \quad (66)$$

where  $|z_{i,j}|$  is the complex modulus of  $z_{i,j}$ .

In Table 1, we demonstrate the C-norm for different schemes for different grid-points  $N$  per line when  $\theta = \pi/4$ ,  $k_0 = 10$ , and  $G_{mid} = 10$ . As can be seen, both

**Table 1** The error in the C-norm for  $k_0 = 10$ 

$N$	41	81	161
Refined 13p	9.3344e–03	3.0742e–03	2.5983e–03
Optimal 13p	2.4503e–05	2.8669e–06	7.1886e–07
Refined 25p [10]	7.3191e–05	4.0650e–06	3.7812e–07
Global 25p [10]	9.8230e–05	6.6231e–06	8.8407e–07
NC 4th-order (25)	2.3078e–04	3.2921e–05	8.9589e–06
Refined 9p [6]	9.8639e–03	2.4627e–03	6.1646e–04
Conventional 5p (26)	7.3125e–02	2.1160e–02	5.5046e–03
Optimal rotated 9p [6, 22]	1.1260e–02	2.8265e–03	7.0816e–04
Point-weighting 9p [8]	9.8639e–03	2.4627e–03	6.1646e–04
$N$	321	641	1281
Refined 13p	2.7940e–04	3.6236e–05	8.5970e–06
Optimal 13p	6.0049e–08	3.5808e–09	7.9634e–10
Refined 25p [10]	2.1754e–08	1.0023e–08	7.9202e–10
Global 25p [10]	9.2679e–08	3.2737e–09	1.9928e–10
NC 4th-order (25)	6.1128e–07	1.7979e–08	1.0581e–09
Refined 9p [6]	1.5411e–04	3.8528e–05	9.6322e–06
Conventional 5p (26)	1.3907e–03	3.4858e–04	8.7200e–05
Optimal rotated 9p [6, 22]	1.7707e–04	4.4270e–05	1.1068e–05
Point-weighting 9p [8]	1.5411e–04	3.8528e–05	9.6321e–06

non-optimal methods, NC 4th-order and conventional 5p, perform well. However, based on our numerical experiments, as  $k_0$  increases, the accuracy of the non-optimal methods decreases.

Table 2 shows the error in the C-norm for different schemes for different grid points  $N$  per line when  $\theta = \pi/4$  and  $k_0 = 75$ . It can be seen that optimal 13p schemes provide the best accuracy among all methods when  $k_0 = 75$  and  $G_{mid} = 10$ . This table also shows that refined 13p is second-order, while the order of accuracy of optimal 13p is better than the second-order. In addition, for smaller  $N$ , 13-point schemes have shown a better level of accuracy than the other schemes. While the accuracy of refined 13p is comparable with 9-point schemes for large values of  $N$ , the optimal 13p provides a notable level of accuracy, even better than NC 4th-order. In addition, the refined and global 25p schemes as expected have the best performance among all schemes. However, for large values of  $N$ , the C-norm of optimal 13p is quite comparable with those of the 25-point schemes.

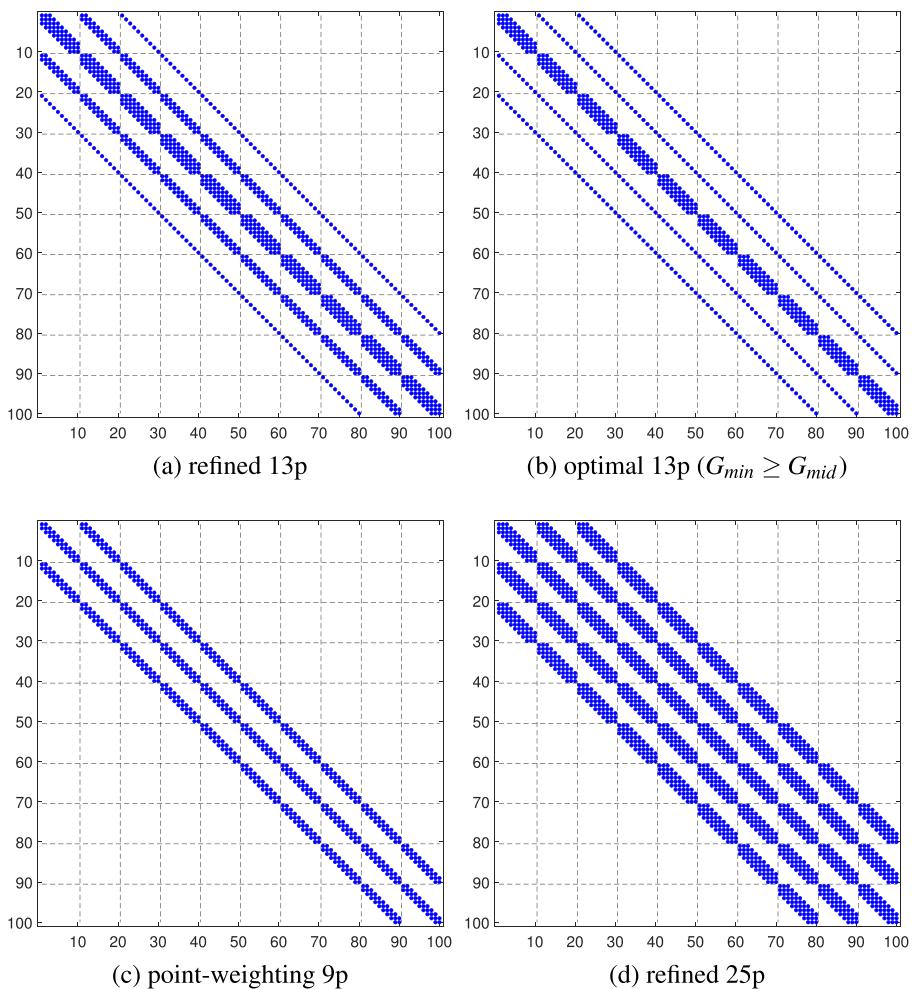
Moreover, in Table 3, we increased  $k_0$  to  $k_0 = 100$  and set  $G_{mid} = 16$ . What stands out from this table is that both refined 13p and optimal 13p have demonstrated the best accuracy among all schemes under study. Again, optimal 13p can take advantage of large values of  $N$ , as its accuracy is better than NC 4th-order even though the associated matrix with optimal 13p and NC 4th-order have similar structures (see Fig. 3). Again, the fourth-order 25-point schemes as expected outperform the rest of schemes under study in terms of having smaller C-norm.

**Table 2** The error in the C-norm for  $k_0 = 75$ 

$N$	41	81	161
Refined 13p	4.7948e−01	8.4428e−02	1.7127e−02
Optimal 13p	4.7948e−01	8.4428e−02	1.7127e−02
Refined 25p [10]	3.2695e−02	4.5151e−03	2.4812e−04
Global 25p [10]	3.8055e−01	2.7117e−02	2.1780e−03
NC 4th-order (25)	1.6949e+00	2.6088e−01	9.9464e−03
Refined 9p [6]	5.1165e−01	9.7380e−02	2.2318e−02
Conventional 5p (26)	9.5996e+00	3.2143e+00	1.6601e+00
Optimal rotated 9p [6, 22]	6.3398e−01	3.1629e−01	8.0844e−02
Point-weighting 9p [8]	5.1165e−01	9.7380e−02	2.2318e−02
$N$	321	641	1281
Refined 13p	5.3125e−03	1.7139e−03	4.8247e−04
Optimal 13p	3.5799e−04	2.2696e−05	1.5766e−06
Refined 25p [10]	2.0716e−05	9.9044e−07	6.4142e−08
Global 25p [10]	1.1650e−04	7.8467e−06	5.8990e−07
NC 4th-order (25)	4.8396e−04	3.1317e−05	2.3352e−06
Refined 9p [6]	5.3031e−03	1.2950e−03	3.2106e−04
Conventional 5p (26)	1.9918e−01	4.6184e−02	1.1442e−02
Optimal rotated 9p [6, 22]	2.0072e−02	5.0364e−03	1.2585e−03
Point-weighting 9p [8]	5.3031e−03	1.2950e−03	3.2106e−04

**Table 3** The error in the C-norm for  $k_0 = 100$ 

$N$	41	81	161
Refined 13p	8.0860e−01	1.5112e−01	2.9006e−02
Optimal 13p	8.0860e−01	1.5112e−01	2.9006e−02
Refined 25p [10]	3.8692e−01	1.0178e−02	7.0577e−04
Global 25p [10]	1.3662e+00	3.5404e−02	8.5880e−03
NC 4th-order (25)	8.8833e+01	2.5897e+00	3.0612e−02
Refined 9p [6]	2.9291e+00	9.0141e−01	2.8377e−01
Conventional 5p (26)	5.7217e+00	1.2101e+01	5.1107e+00
Optimal rotated 9p [6, 22]	1.5824e+00	1.9596e+00	6.0908e−01
Point-weighting 9p [8]	2.9291e+00	9.0141e−01	2.8377e−01
$N$	321	641	1281
Refined 13p	9.8585e−03	3.0541e−03	1.2024e−03
Optimal 13p	9.8585e−03	1.9835e−04	5.3343e−05
Refined 25p [10]	8.7337e−05	3.7472e−06	6.0353e−07
Global 25p [10]	8.0437e−04	6.6983e−05	1.9299e−05
NC 4th-order (25)	3.4756e−03	2.6391e−04	7.4180e−05
Refined 9p [6]	6.5930e−02	1.3619e−02	3.0793e−03
Conventional 5p (26)	4.0883e+00	4.9499e−01	2.8923e−01
Optimal rotated 9p [6, 22]	4.1182e−01	1.7273e−01	5.1743e−02
Point-weighting 9p [8]	6.5930e−02	1.3619e−02	3.0793e−03



**Fig. 3** The matrix in the linear systems with  $N = 10$

Furthermore, the order of accuracy of the new schemes depends on the parameters of the finite difference scheme (22). In general, optimal finite difference schemes are arithmetic-weighted arithmetic averages of several schemes. In particular, the new 13-point schemes are weighted arithmetic averages of second-order and fourth-order schemes,  $\mathcal{L}_1$ ,  $\mathcal{L}_2$ , and  $\mathcal{L}_3$ . For example, for a given  $h$ , the fourth-order scheme  $\mathcal{L}_1$ , from (25), can contribute to this average with weight of  $b_1$ . However, when the size of increment  $h$  is decreased to  $h/2$ ,  $\mathcal{L}_1$  can contribute to the average with a different  $b_1$ . Therefore, the value of  $N$  can affect the order of accuracy of the method.

In addition, Tables 4 and 5 show the error in the C-norm with  $k_0 = 100$ ,  $G_{mid} = 10$  and  $\theta = 0, \pi/16, \dots, \pi/4$ , for  $N = 101$  and  $N = 201$ , respectively. Overall, refined 13p and optimal 13p provide the most accurate results among all schemes in this study. When  $N = 101$  ( $G_{min} < G_{mid}$ ), the errors in the C-norm for refined 13p and

**Table 4** The error in the C-norm for  $k_0 = 100$  and  $N = 101$ 

$\theta$	0	$\pi/16$	$\pi/8$	$3\pi/16$	$\pi/4$
Refined 13p	7.5222e−02	8.9361e−02	8.7764e−02	7.8789e−02	8.3067e−02
Optimal 13p	7.5222e−02	8.9361e−02	8.7764e−02	7.8789e−02	8.3067e−02
Refined 25p	1.0537e−02	1.4637e−02	9.1759e−03	7.2469e−03	3.3765e−02
Global 25p	9.7047e−02	3.6421e−02	1.9046e−02	1.5790e−02	3.3981e−01
NC 4th-order	3.9183e−01	7.2159e−01	5.1360e−01	2.4293e−01	5.7859e−01
Refined 9p	9.4140e−02	1.2232e−01	9.7428e−02	1.1577e−01	6.5001e−01
Conventional 5p	2.8297e+00	3.9821e+00	3.6443e+00	4.4336e+00	4.7288e+01
Optimal rotated 9p	2.4409e−01	1.9197e+00	1.8917e−01	4.5804e−01	4.8616e+00
Point-weighting 9p	9.4140e−02	1.2232e−01	9.7428e−02	1.1577e−01	6.5001e−01

optimal 13p are identical. However, when  $N = 201$ , ( $G_{min} \geq G_{mid}$ ), the optimal 13p provides a more accurate result.

However, all finite difference schemes require solving a linear system  $AX = R$ , where  $A$  is the matrix of coefficients associated with each finite difference scheme,  $X$  is a vector consisting of the unknowns, and  $R$  is the right-hand side. The matrix in the linear system associated with each finite difference scheme, matrix  $A$ , is a sparse matrix with complex values (see Fig. 3). Matrix  $A$  associated with the optimal 9-point schemes (refined 9p, optimal rotated 9p, and point weighting 9p) has the same number of non-zero entries. For a given  $N$ , the numbers of non-zero entries of the matrix  $A$  (NNE) for refined 13p, optimal 13p ( $G_{min} \geq G_{mid}$ ), point-weighting 9p, and refined 25p are  $13 N^2 - 20 N + 4$ ,  $9 N^2 - 12 N$ ,  $9 N^2 - 18 N + 10$  and  $25 N^2 - 60 N + 36$ , respectively). For example, when  $N = 10$ , NNE for refined 13p, optimal 13p ( $G_{min} \geq G_{mid}$ ), point-weighting 9p, and refined 25p are 1104, 780, 784, and 1936, respectively. Therefore, using LU decomposition, theoretically speaking, the computational complexity of the optimal 9-point schemes is comparable when ( $G_{min} \geq G_{mid}$ ) while the computational complexity of refined 13p is

**Table 5** The error in the C-norm for  $k_0 = 100$  and  $N = 201$ 

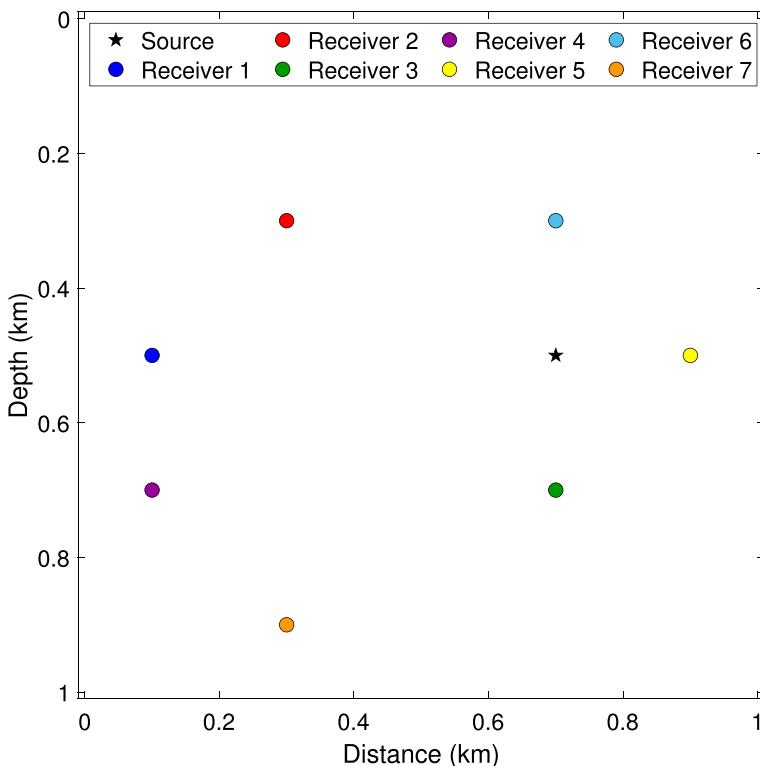
$\theta$	0	$\pi/16$	$\pi/8$	$3\pi/16$	$\pi/4$
Refined 13p	1.9633e−02	2.5184e−02	1.9384e−02	2.1496e−02	1.8735e−02
Optimal 13p	1.9633e−02	2.5184e−02	1.9384e−02	2.1496e−02	1.8735e−02
Refined 25p	7.3531e−04	8.4408e−04	5.2335e−04	4.4990e−04	3.0142e−04
Global 25p	3.0820e−03	8.4830e−03	2.3530e−03	1.7268e−03	2.5818e−03
NC 4th-order	2.2008e−02	6.9477e−02	3.2024e−02	1.8711e−02	1.2017e−02
Refined 9p	2.1567e−02	2.4293e−02	2.2840e−02	2.7603e−02	1.8795e−01
Conventional 5p	1.4974e+00	1.2433e+00	1.3061e+00	2.6444e+00	1.4253e+00
Optimal rotated 9p	1.1361e−01	4.3817e−01	7.2447e−02	6.4544e−02	5.5283e−01
Point-weighting 9p	2.1567e−02	2.4293e−02	2.2840e−02	2.7603e−02	1.8795e−01

slightly higher than the other schemes under study in this example. Moreover, a 25-point scheme (such as refined 25p) also has more computational complexity than the 13-point schemes. Especially, when ( $G_{min} \geq G_{mid}$ ), computational costs using a 25-point scheme can be quite high. In practice, on average, we could not find a significant difference between the CPU time of 9-point schemes and 13-point schemes with the same number of grid points,  $N$ , on our machine using the UMF method for sparse LU factorization.

#### 4.2 Problem 2: a homogeneous model

In this example, a homogeneous velocity model is considered. The velocity of the medium is 2000 m/s, and horizontal and vertical samplings are  $nx = nz = 51$  with sampling intervals  $h = \Delta x = \Delta z = 20$  m, and the time sampling is  $\Delta t = 8$  ms. A point source  $\delta(x - x_s, z - z_s)R(\omega, f_M)$  is located at the point (700 m, 500 m), where  $R(\omega, f_M)$  is the Ricker wavelet, defined in (67), with the peak frequency  $f_M = 15$  Hz.

$$R(t, f_M) = \left(1 - 2\pi^2 f_M^2 t^2\right) / \exp\left(\pi^2 f_M^2 t^2\right). \quad (67)$$



**Fig. 4** Source and receiver locations

For this homogeneous model, the analytical solution is available as follows [1]:

$$p(x, z, t) = i\pi \mathcal{F}^{-1} \left( H_0^{(2)} \left( \frac{\omega}{v} \sqrt{(x - x_s)^2 + (z - z_s)^2} \right) \mathcal{F}(R(t, f_M)) \right) \quad (68)$$

where  $\mathcal{F}$  and  $\mathcal{F}^{-1}$  are Fourier and inverse Fourier transformations with respect to time, respectively, and  $H_0^{(2)}$  is the second Hankel function of order zero.

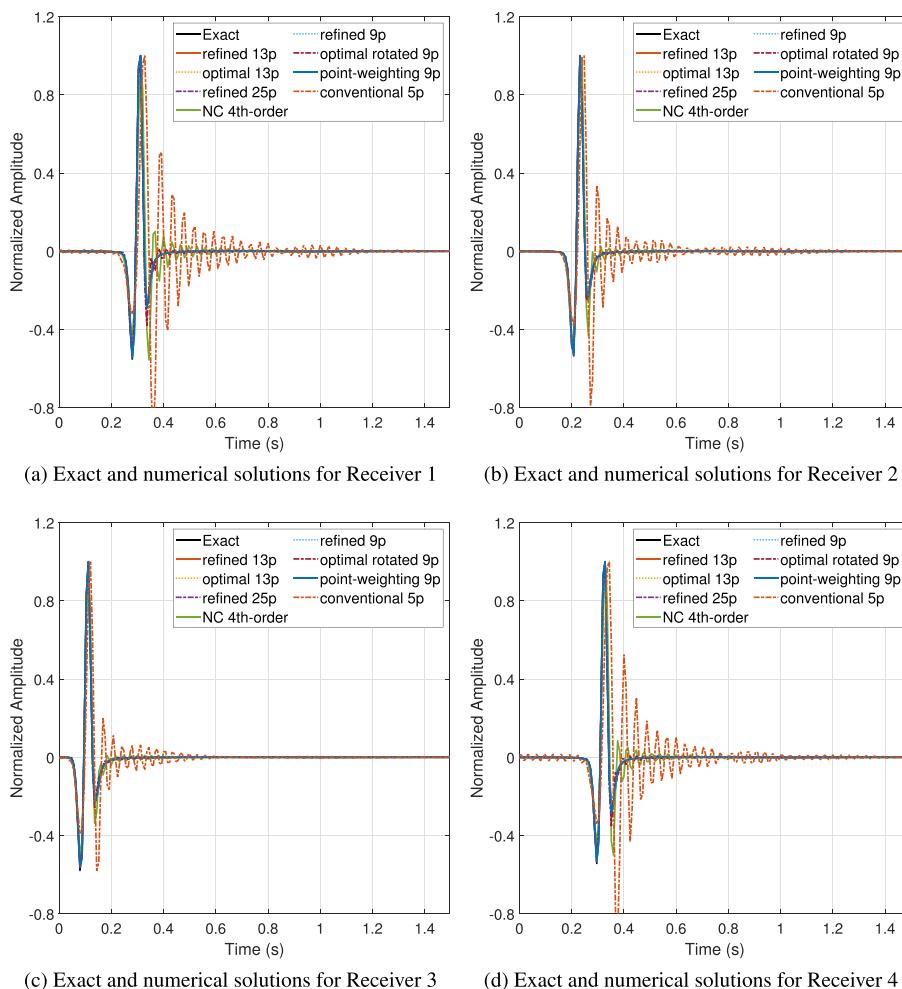
Due to the fact that the numerical dispersion is dependent on the propagation angle, we have placed receivers 1,2,...,7 separately at (100 m, 500 m), (300 m, 300 m), (700, 700 m), (100 m, 700 m), (900 m, 500 m), (700 m, 300 m), and (300 m, 900 m), respectively (please see Fig. 4).

Here, we compare the refined 13p, optimal 13p (with  $G_{mid} = 16$ ), NC 4th-order, refined 9p, optimal rotated 9p, point-weighting 9p, refined 25p, and conventional 5p in terms of accuracy. Table 6 highlights the error in the C-norm for receivers 1,...,7. It can be seen that refined 13p and optimal 13p provide the best accuracy among all schemes. However, optimal 13p provides the best level of accuracy overall. In Figs. 5 and 6, the exact and numerical solutions obtained using refined 13p, optimal 13p (with  $G_{mid} = 16$ ), NC 4th-order, refined 9p, optimal rotated 9p, and point-weighting 9p are compared. It can be seen that the NC 4th-order is the least efficient scheme.

In this example, when  $G_{mid} \geq 34$ , the errors in the C-norm using refined 13p and optimal 13p are identical. Therefore, we ran all of the experiments in this example using  $G_{mid} = 16$  to differentiate the numerical results between these two schemes. As can be seen, in this example, refined 13p is performing better than optimal 13p in terms of accuracy. However, the main reason behind using optimal 13p is the

**Table 6** The error in the C-norm (receivers 1, 2,...,7)

$(x_r, z_r)$	(100, 500)	(300, 300)	(700, 700)	(100, 700)
Refined 13p	4.9071e-02	5.0956e-02	4.7928e-02	4.5169e-02
Optimal 13p	4.9393e-02	5.1436e-02	4.8484e-02	4.5480e-02
Refined 25p	1.9892e-02	1.3034e-02	1.9484e-02	1.6327e-02
NC 4th-order	3.4299e-01	2.2120e-01	1.6034e-01	2.9828e-01
Refined 9p	5.3459e-02	5.8838e-02	5.3181e-02	5.1615e-02
Optimal rotated 9p	1.1694e-01	6.2835e-02	9.3922e-02	9.5977e-02
Point-weighting 9p	5.3459e-02	5.8838e-02	5.3181e-02	5.1615e-02
Conventional 5p	1.0620e+00	8.1231e-01	5.8804e-01	1.0586e+00
$(x_r, z_r)$	(900, 500)	(700, 300)	(300, 900)	
Refined 13p	4.7460e-02	4.7921e-02	5.0916e-02	
Optimal 13p	4.8012e-02	4.8477e-02	5.1311e-02	
Refined 25p	1.9475e-02	1.9484e-02	1.0649e-02	
NC 4th-order	1.6119e-01	1.6039e-01	1.4991e-01	
Refined 9p	5.3168e-02	5.3184e-02	5.9944e-02	
Optimal rotated 9p	9.4310e-02	9.3899e-02	8.3041e-02	
Point-weighting 9p	5.3168e-02	5.3184e-02	5.9944e-02	
Conventional 5p	5.9055e-01	5.8796e-01	7.1227e-01	

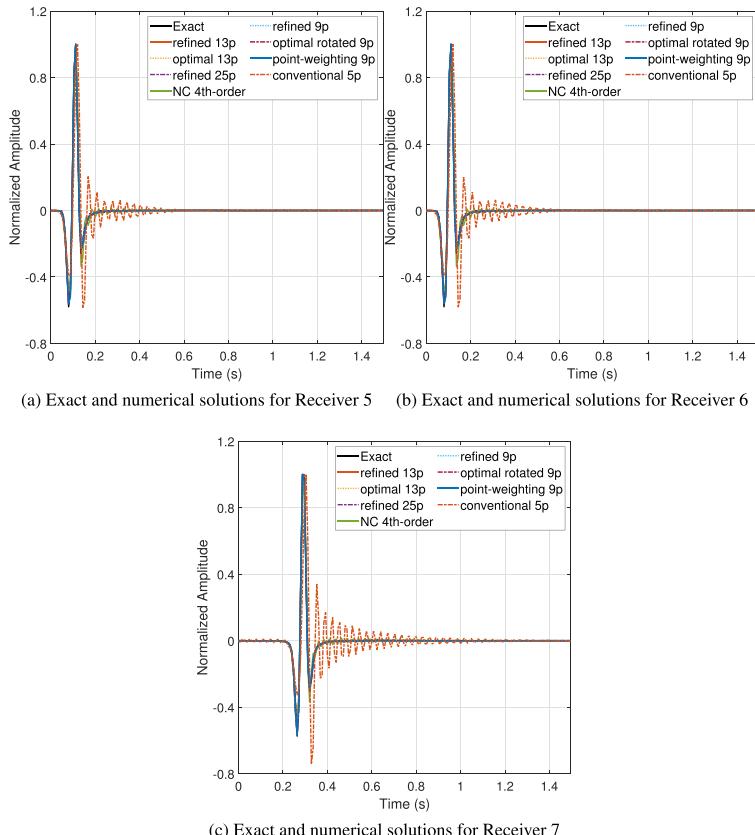


**Fig. 5** Exact and numerical solutions for receivers 1, 2, 3, and 4

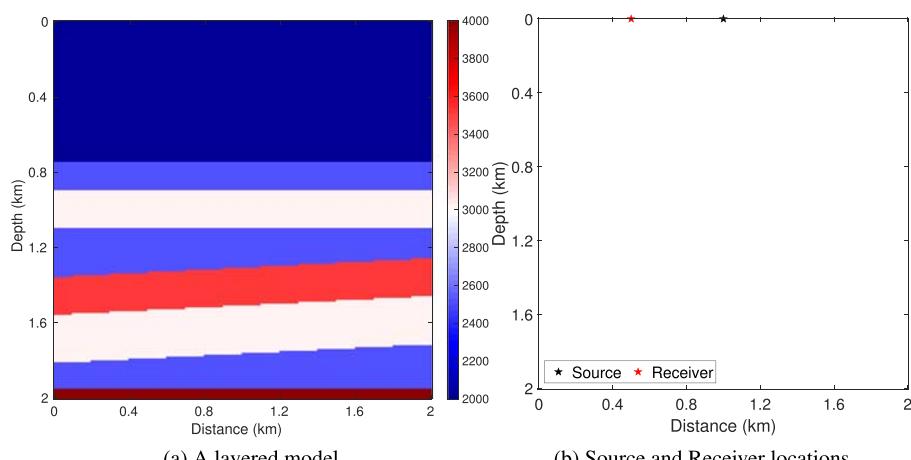
computational cost, as discussed in Example 4.1. Moreover, for a small grid,  $nx = nz = 51$ , there is no significant computational difference among all of the schemes under study (Fig. 7).

Furthermore, the approximated solutions by NC 4th-order and conventional 5p emphasize the necessity of using an optimal method over other methods. Even though the NC 4th-order is a fourth-order scheme, it cannot approximate the solution (68) at all. The main reason responsible for this is the impact of the pollution effect.

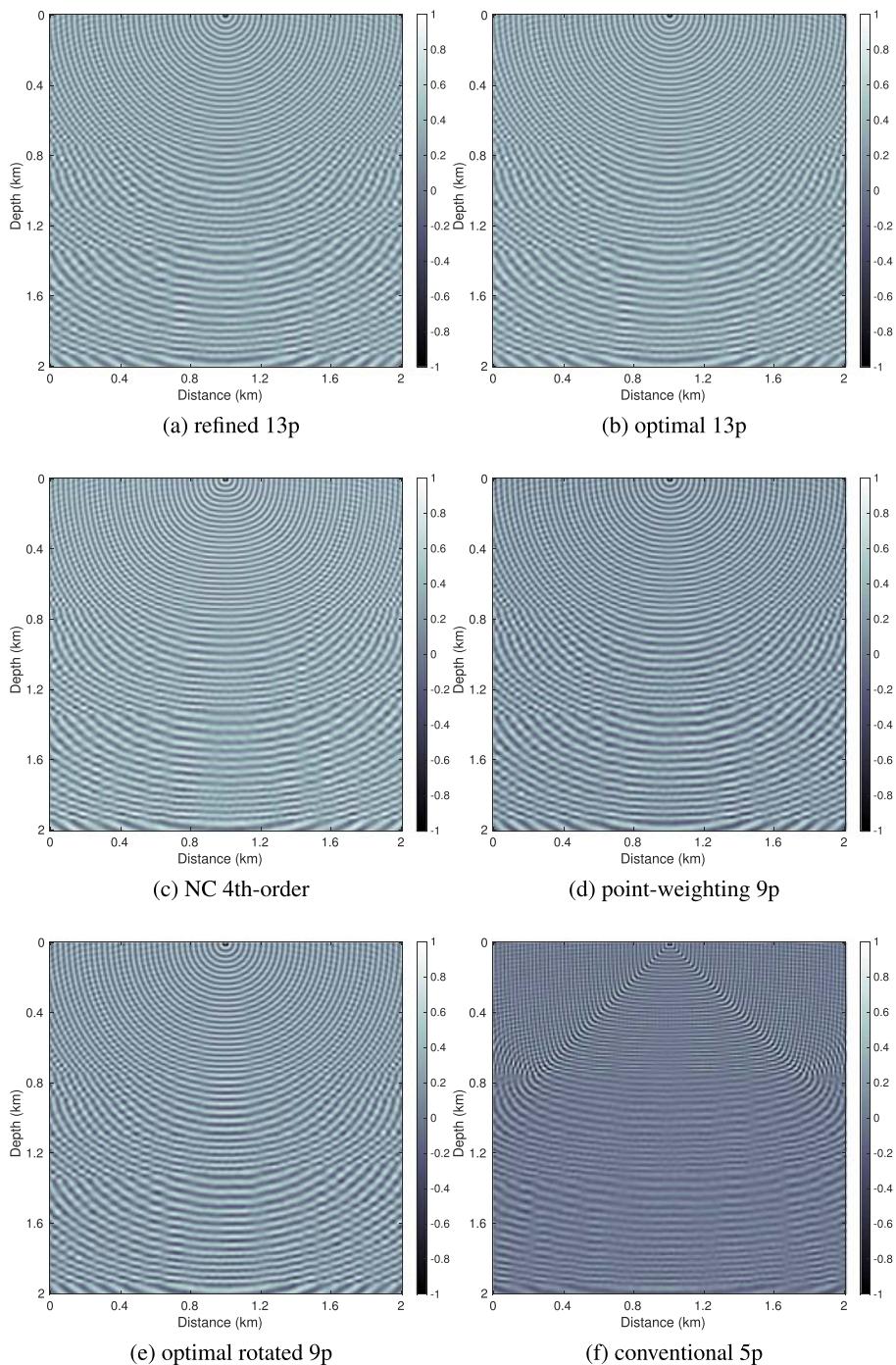
Moreover, the approximation of the solution (68) requires computing and solving the Helmholtz equation with PML for a considerable number of times and then applying inverse Fourier transform. Although the refined 25p provides the most accurate solution among all schemes under study, it requires much higher computational resources than the 13-point schemes.



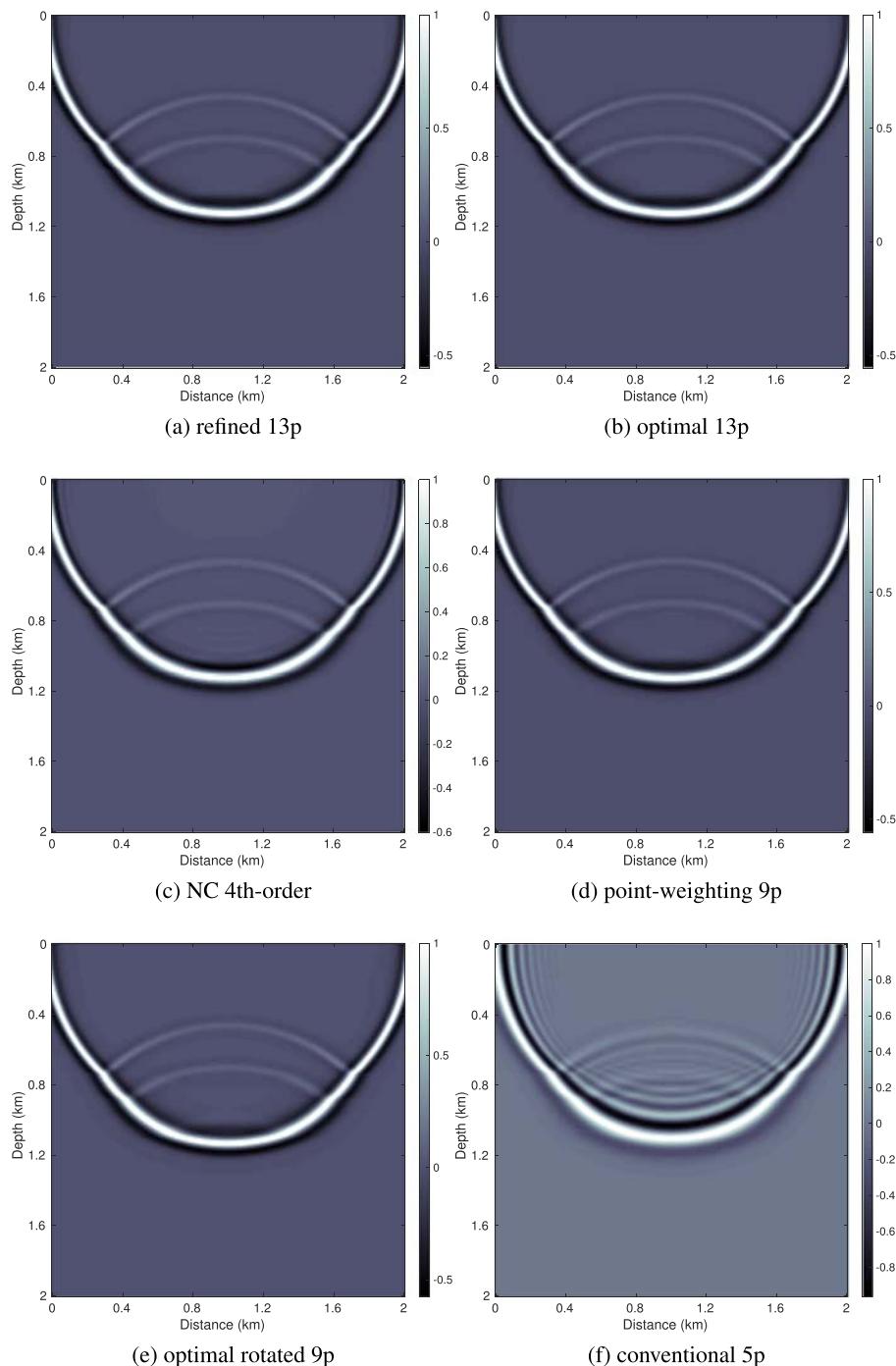
**Fig. 6** Exact and numerical solutions for receivers 5, 6, and 7



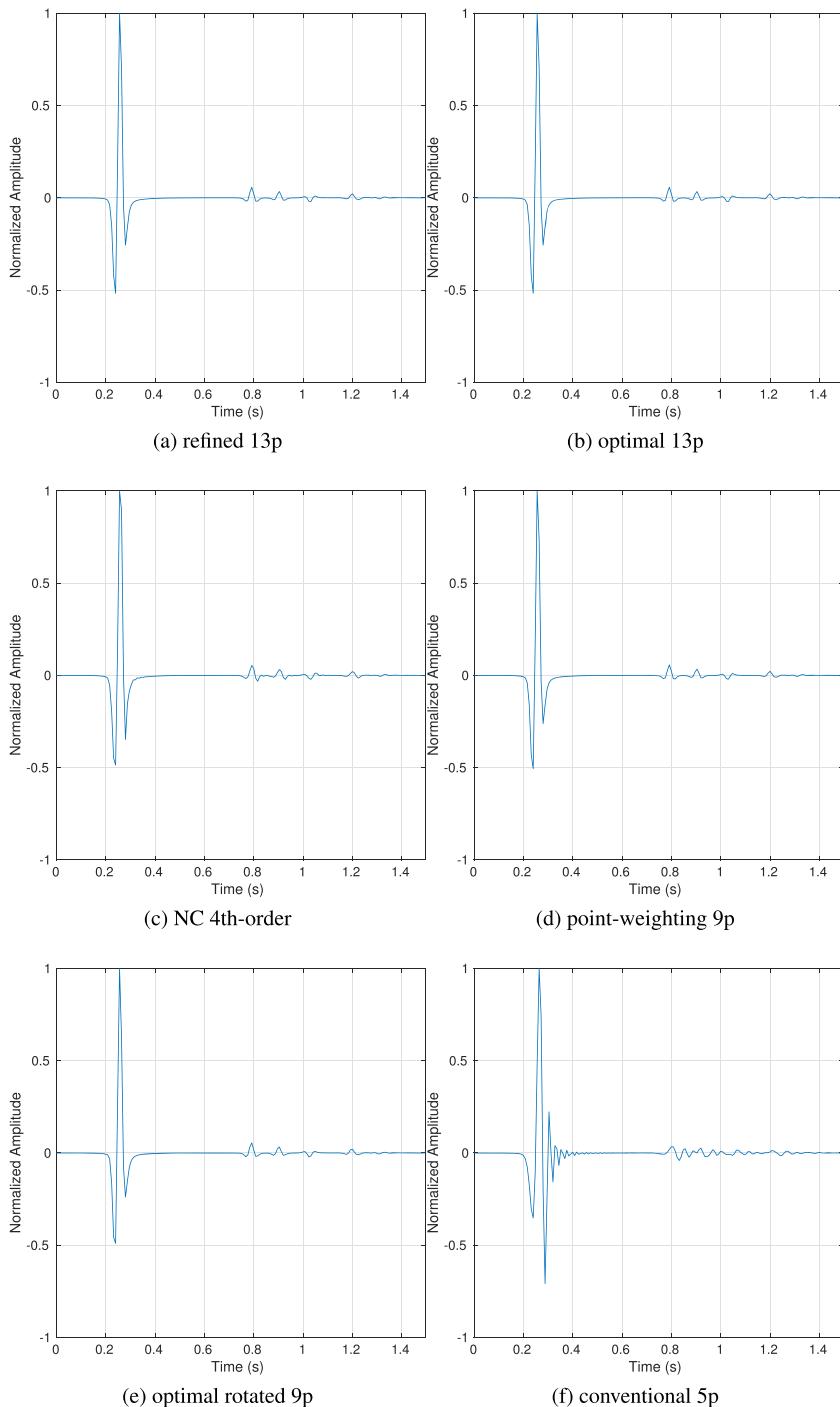
**Fig. 7** A layered model and the locations of the source and the receiver.



**Fig. 8** The monofrequency wavefield (real part) for  $f = 62.5$  Hz



**Fig. 9** Snapshots for  $t = 520$  ms generated by various schemes



**Fig. 10** The numerical solution computed by various schemes

#### 4.3 Problem 3: a layered model

In this example, we consider a layered P-wave velocity model, shown in Fig. 7a, to test the new schemes in a more practical usage. Horizontal and vertical samplings are  $nx = nz = 201$  with sampling intervals  $\Delta x = \Delta z = 10$  m, and the time sampling is  $\Delta t = 8$  ms. A point source  $\delta(x - x_s, z - z_s)R(\omega, f_M)$  is located at the point  $(x_s, z_s) = (1000 \text{ m}, 0 \text{ m})$ , where  $R(\omega, f_M)$  is the Ricker wavelet with the peak frequency  $f_M = 20$  Hz.

In addition, snapshots for  $t = 520$  ms generated by refined 13p, optimal 13p, NC 4th-order, optimal rotated 9p, point-weighting 9p, and conventional 5p are available in Fig. 9. No boundary reflections can be observed, and the upward incident waves, the downward incident waves, and transmissive waves are all clear. It can be seen that the snapshot generated by conventional 5p, shown in Fig. 9d, does not represent the expected solution. The snapshot generated by NC 4th-order has more numerical artifacts than the remaining figures. In particular, close to the source area and above the center, we can observe these numerical artifacts. We also observed some numerical artifacts in Fig. 9e, generated by optimal rotated 9p, in the center area. Moreover, no significant differences have been observed among the snapshots generated by refined 13p, optimal 13p, and point-weighting 9p.

We have also placed a receiver at  $(x_r, z_r) = (500, 0)$  (Fig. 7b) and plotted the corresponding numerical solution in Fig. 10. It can be seen that the numerical solution computed by conventional 5p does not represent the expected result. The numerical solution computed by NC 4th-order also has lots of non-physical oscillations. In the remaining figures, we can see that the first arrival and reflections are all clear, and non-physical oscillations in the synthetic seismogram are negligible.

We have already shown in Example 4.2 that the optimal and refined 13p provide the most reliable results in comparison with other schemes in this study. Overall, monofrequency wavefields, see Fig. 8, are the figures that come directly from solving the Helmholtz equation with PML (1). The more accurate the solution of this equation is estimated, the better results can be found in Figs. 9 and 10.

## 5 Conclusions

In this article, we proposed a 13-point finite difference method (FDM) based on the point-weighting scheme derivation strategy. It was shown that the 13-point FDM is consistent with the Helmholtz equation with PML, and its order of accuracy can vary between two and four. We presented an error analysis for the numerical wavenumber and recommended two strategies, the optimal and refined strategies, for parameter selection of the 13-point FDM. The normalized phase and group velocity curves of the 13-point FDM confirm that it is very effective in numerical dispersion reduction. The new numerical schemes, the refined and optimal 13-point schemes, were compared against a number of existing finite difference methods that are widely used for the Helmholtz equation with PML. Our numerical examples demonstrate that the new schemes are less dispersive than any other optimal 9-point schemes. Moreover, it has been shown that the optimal 13-point has less computational complexity and

provides better accuracy on refined grids than any other schemes under study. In the future, we plan to extend the new schemes for solving the Helmholtz equation with PML in three dimensions.

**Funding information** The work is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) through the individual Discovery Grant (RGPIN-2019-04830). The first author is also supported by the Alberta Innovates Graduate Student Scholarship that he received during his PhD study.

## References

1. Alford, R.M., Kelly, K.R., Boore, D.: Accuracy of finite-difference modeling of the acoustic wave equation. *Geophysics* **39**(6), 834–842 (1974)
2. Bayliss, A., Goldstein, C.I., Turkel, E.: On accuracy conditions for the numerical computation of waves. *J. Comput. Phys.* **59**(3), 396–404 (1985)
3. Berenger, J.P.: A perfectly matched layer for the absorption of electromagnetic waves. *J. Comput. Phys.* **114**(2), 185–200 (1994)
4. Britt, S., Tsynkov, S., Turkel, E.: Numerical simulation of time-harmonic waves in inhomogeneous media using compact high order schemes. *Commun. Comput. Phys.* **9**(3), 520–541 (2011)
5. Chen, J.B.: A generalized optimal 9-point scheme for frequency-domain scalar wave equation. *J. Appl. Geophys.* **92**, 1–7 (2013)
6. Chen, Z., Cheng, D., Feng, W., Wu, T.: An optimal 9-point finite difference scheme for the Helmholtz equation with PML. *International Journal of Numerical Analysis & Modeling.* **10**(2) (2013)
7. Chen, Z., Cheng, D., Wu, T.: A dispersion minimizing finite difference scheme and preconditioned solver for the 3D Helmholtz equation. *J. Comput. Phys.* **231**(24), 8152–8175 (2012)
8. Cheng, D., Tan, X., Zeng, T.: A dispersion minimizing finite difference scheme for the Helmholtz equation based on point-weighting. *Comput. Math. Appl.* **73**(11), 2345–2359 (2017)
9. Collino, F., Monk, P.B.: Optimizing the perfectly matched layer. *Comput. Methods Appl. Mech. Eng.* **164**(1–2), 157–171 (1998)
10. Dastour, H., Liao, W.: A fourth-order optimal finite difference scheme for the Helmholtz equation with PML. *Comput. Math. Appl.* **6**(78), 2147–2165 (2019)
11. Davis, T.A.: Algorithm 832: UMFPACK v4. 3—an unsymmetric-pattern multifrontal method. *ACM Trans. Math. Softw. (TOMS)* **30**(2), 196–199 (2004)
12. Davis, T.A., Duff, I.S.: An unsymmetric-pattern multifrontal method for sparse lu factorization. *SIAM J. Matrix Anal. Appl.* **18**(1), 140–158 (1997)
13. De Zeeuw, P.M.: Matrix-dependent prolongations and restrictions in a blackbox multigrid solver. *J. Comput. Appl. Math.* **33**(1), 1–27 (1990)
14. Engquist, B., Majda, A.: Absorbing boundary conditions for numerical simulation of waves. *Proc. Natl. Acad. Sci.* **74**(5), 1765–1766 (1977)
15. Engquist, B., Ying, L.: Sweeping preconditioner for the Helmholtz equation: moving perfectly matched layers. *Multiscale Model. Simul.* **9**(2), 686–710 (2011)
16. Erlangga, Y.A., Oosterlee, C.W., Vuik, C.: A novel multigrid based preconditioner for heterogeneous Helmholtz problems. *SIAM J. Sci. Comput.* **27**(4), 1471–1492 (2006)
17. van Gijzen, M.B., Erlangga, Y.A., Vuik, C.: Spectral analysis of the discrete Helmholtz operator preconditioned with a shifted Laplacian. *SIAM J. Sci. Comput.* **29**(5), 1942–1958 (2007)
18. Gordon, D., Gordon, R.: Carp-cg: A robust and efficient parallel solver for linear systems, applied to strongly convection dominated pdes. *Parallel Comput.* **36**(9), 495–515 (2010)
19. Harari, I., Turkel, E.: Accurate finite difference methods for time-harmonic wave propagation. *J. Comput. Phys.* **119**(2), 252–270 (1995)
20. Ihlenburg, F., Babuška, I.: Dispersion analysis and error estimation of Galerkin finite element methods for the Helmholtz equation. *Int. J. Numer. Methods Eng.* **38**(22), 3745–3774 (1995)
21. Ihlenburg, F., Babuška, I.: Finite element solution of the Helmholtz equation with high wave number part i: The h-version of the fem. *Comput. Math. Appl.* **30**(9), 9–37 (1995)
22. Jo, C.H., Shin, C., Suh, J.H.: An optimal 9-point, finite-difference, frequency-space, 2-D scalar wave extrapolator. *Geophysics* **61**(2), 529–537 (1996)

23. Medvinsky, M., Turkel, E., Hetmaniuk, U.: Local absorbing boundary conditions for elliptical shaped boundaries. *J. Comput. Phys.* **227**(18), 8254–8267 (2008)
24. Nabavi, M., Siddiqui, M.K., Dargahi, J.: A new 9-point sixth-order accurate compact finite-difference method for the Helmholtz equation. *J. Sound Vib.* **307**(3–5), 972–982 (2007)
25. Operto, S., Virieux, J., Amestoy, P., L'Excellent, J.Y., Giraud, L., Ali, H.B.H.: 3D finite-difference frequency-domain modeling of visco-acoustic wave propagation using a massively parallel direct solver: A feasibility study. *Geophysics* **72**(5), SM195–SM211 (2007)
26. Poulson, J., Engquist, B., Li, S., Ying, L.: A parallel sweeping preconditioner for heterogeneous 3D Helmholtz equations. *SIAM J. Sci. Comput.* **35**(3), C194–C212 (2013)
27. Pratt, R.G., Worthington, M.H.: Inverse theory applied to multi-source cross-hole tomography. Part 1: Acoustic wave-equation method. *Geophys. Prospect.* **38**(3), 287–310 (1990)
28. Shin, C., Sohn, H.: A frequency-space 2-D scalar wave extrapolator using extended 25-point finite-difference operator. *Geophysics* **63**(1), 289–296 (1998)
29. Singer, I., Turkel, E.: High-order finite difference methods for the Helmholtz equation. *Comput. Methods Appl. Mech. Eng.* **163**(1–4), 343–358 (1998)
30. Singer, I., Turkel, E.: A perfectly matched layer for the Helmholtz equation in a semi-infinite strip. *J. Comput. Phys.* **201**(2), 439–465 (2004)
31. Sutmann, G.: Compact finite difference schemes of sixth order for the Helmholtz equation. *J. Comput. Appl. Math.* **203**(1), 15–31 (2007)
32. Trefethen, L.N.: Group velocity in finite difference schemes. *SIAM Rev.* **24**(2), 113–136 (1982)
33. Turkel, E., Gordon, D., Gordon, R., Tsynkov, S.: Compact 2D and 3D sixth order schemes for the Helmholtz equation with variable wave number. *J. Comput. Phys.* **232**(1), 272–287 (2013)
34. Turkel, E., Yefet, A.: Absorbing PML boundary layers for wave-like equations. *Appl. Numer. Math.* **27**(4), 533–557 (1998)
35. Virieux, J., Operto, S.: An overview of full-waveform inversion in exploration geophysics. *Geophysics* **74**(6), WCC1–WCC26 (2009)
36. Wang, S.: An improved high order finite difference method for non-conforming grid interfaces for the wave equation. *J. Sci. Comput.* **77**(2), 775–792 (2018)
37. Wu, T.: A dispersion minimizing compact finite difference scheme for the 2D Helmholtz equation. *J. Comput. Appl. Math.* **311**, 497–512 (2017)
38. Wu, T., Xu, R.: An optimal compact sixth-order finite difference scheme for the Helmholtz equation. *Comput. Math. Appl.* **75**(7), 2520–2537 (2018)
39. Zeng, Y., He, J., Liu, Q.: The application of the perfectly matched layer in numerical modeling of wave propagation in poroelastic media. *Geophysics* **66**(4), 1258–1266 (2001)

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.