ASYMPTOTIC DISPERSION CORRECTION IN GENERAL FINITE DIFFERENCE SCHEMES FOR HELMHOLTZ PROBLEMS

PIERRE-HENRI COCQUET, MARTIN J. GANDER

Abstract. Most numerical approximations of frequency-domain wave propagation problems suffer from the so-called dispersion error, which is the fact that plane waves at the discrete level oscillate at a frequency different from the continuous one. In this paper, we introduce a new technique to reduce the dispersion error in general Finite Difference (FD) schemes for frequency-domain wave propagation using the Helmholtz equation as guiding example. Our method is based on the introduction of a shifted wavenumber in the FD stencil which we use to reduce the numerical dispersion for large enough numbers of grid points per wavelength (or for small enough meshsize), and thus we call the method *asymptotic dispersion correction*. The advantage of this technique is that the asymptotically optimal shift can be determined in closed form by computing the extrema of a function over a compact set. For 1d Helmholtz equations, we prove that the standard 3-point stencil with shifted wavenumber does not have any dispersion error, and that the so-called pollution effect is completely suppressed. For higher dimensional Helmholtz problems, we give easy to use closed form formulas for the asymptotically optimal shift associated to the second order 5-point scheme and a sixth-order 9-point scheme in 2d, and the 7-point scheme in 3d that yield substantially less dispersion error than their standard (unshifted) version. We illustrate this also with numerical experiments.

Key words. Frequency-Domain wave propagation, Finite difference method, Helmholtz equation, Numerical dispersion, Asymptotic dispersion correction.

1. Introduction.

The Helmholtz equation is a model problem for time-harmonic wave propagation. On a bounded domain $\Omega \subset \mathbb{R}^d$, it is given by

$$-\Delta u(x) - k^2 u(x) = f(x), \ x \in \Omega, \tag{1.1}$$

where k is the so-called *wavenumber*, f is a given right hand side, and we will specify the necessary boundary conditions later when needed.

Solving this problem numerically for large k is difficult (see e.g. [14]), mainly because of its elliptic yet non-coercive nature, and that solutions oscillate with period proportional to $1/k$. In addition, at the continuous level, plane waves are given by $\mathrm{e}^{\mathrm{i}k\boldsymbol{x}\cdot\boldsymbol{\theta}}$ for $\boldsymbol{\theta} \in \mathcal{S}^{d-1}$ whereas, at the discrete level, plane waves are given by $\mathrm{e}^{\mathrm{i}k_d\boldsymbol{x}\cdot\boldsymbol{\theta}}$ where k_d is the discrete wavenumber, which depends on $\boldsymbol{\theta}$ and the meshsize h, and we usually have $k_d(\boldsymbol{\theta}, h) \neq k$, which is called the dispersion error. The dispersion error is also responsible for the *pollution effect* [17, 20, 24, 32], which is the fact that keeping kh small is not enough to prevent the relative error to grow with the wavenumber.

For the hp-Finite Element method for Helmholtz problems, it is known that the pollution effect can actually be suppressed (see e.g. [17, 19, 20, 21, 24, 32]) if kh/p is small enough, and $p \geq C\log(k)$ for a large enough constant C. Such results have been obtained for Discontinuous-Galerkin methods as well in [22].

In addition to the previous results, the pollution effect can be suppressed in 1d. We refer for instance to [1], where a stabilized FEM without dispersion error is built, or to [31] where a CIP-FEM is shown to be pollution free if some parameter is suitably chosen. For 2d problems, we refer for example to [12, 16, 33] where several methods have been designed to reduce the dispersion error and pollution effect.

For Finite-Difference (FD) methods, techniques have also been derived to reduce the dispersion error. For the 1d Helmholtz equation, a FD scheme without dispersion error is given in [27, 15], and this suppresses the pollution effect. It is derived using a Taylor series of the solution which permits to define a generalized 3-point stencil. For 2d Helmholtz problems, a dispersion correction using eigenvalues has been designed in [13]. Although the matrix associated to the stencil is modified, numerical results indicate that this method heavily reduces the pollution effect. Another widely used strategy is to consider FD stencils with free parameters that are then optimized to minimize the dispersion error. This technique has been applied to various stencils and we refer for example to [6, 28, 5, 10, 11, 29, 30]. This approach has then been investigated further in [8], where a sixth-order 9-point stencil is considered, with coefficients that are polynomials in (kh) with free parameters. These parameters are then determined numerically by minimizing the averaged truncation error of plane waves. For three-dimensional problems, a dispersion-minimizing scheme based on free parameters that are determined by minimizing the dispersion error can be found in [5]. A similar method is used in [26, 25] where the behavior of a multigrid method is also numerically studied. It is shown that a FD scheme with dispersion correction leads to a convergent multigrid method for some wavenumber/meshsize combinations for which the un-corrected scheme leads to divergent multigrid methods (see also [3, 7, 4]).

Dispersion minimizing schemes that do not rely on numerical optimization to determine the free parameters for 1d can be found in [2, 3], and for 2d in [4] for a 9-point stencil, where a shifted wavenumber is introduced in the stencil. For 1d Helmholtz problems, the shift suppresses the dispersion error. In 2d, the shift is explicitly determined so that the dispersion error is minimized for a large enough number of grid points per wavelength. Numerical simulations then show that this asymptotically optimal shift is close to the numerically best one even for a small number of grid points per wavelength. The major drawback of these approaches is the derivation of the explicit shift itself, which is based on minimizing the distance between the discrete and continuous dispersion relations (see [4, Theorem 4.1]), and can thus not be extended easily to other FD schemes or 3d.

We show here that the dispersion error associated to a general FD scheme can be reduced without relying on numerical optimization. Our method is based on the expansion of the discrete wavenumber k_d as the meshsize goes to zero. A shifted wavenumber is next introduced in the stencil to minimize the leading-order term in the expansion of $(k_d(\boldsymbol{\theta}, h) - k)$. We show that this shifted wavenumber can be determined in closed form by computing the extrema of the remainder which is a trigonometric polynomial in $d-1$ variables defined on a compact set.

Our paper is organized as follows: We first present the new shifted wavenumber idea for the 1d Helmholtz equation and prove that the resulting FD scheme has neither dispersion error nor does it suffer from the pollution effect. We present next the general dispersion minimizing scheme based on a shifted wavenumber, and compute the shift in closed form for the 5-point and 9-point stencils in 2d, and for the 7-point stencil in 3d, so they can easily be used in existing codes. We conclude with numerical experiments to illustrate how much the shift reduces the relative error.

2. Suppressing the dispersion error for the 3-point stencil in 1d.

We consider the one dimensional Helmholtz equation on $\Omega = [0,1]$ with homogeneous Dirichlet boundary conditions,

$$\left\{ \begin{array}{rcll} -u''(x) - k^2 u(x) & = & f(x) & \text{in } (0,1), \\ u(x) & = & 0, & x \in \{0,1\}, \end{array} \right. \tag{2.1}$$

where f is a given source term. Since Problem (2.1) can be singular for some values of k, we assume in what follows that

$$k^2 \notin \pi\mathbb{N},$$

which ensures that k^2 is not an eigenvalue of the Laplace operator with homogeneous Dirichlet boundary conditions at $\{0,1\}$.

We consider a uniform grid $\{x_j\}_{j=1}^n = \{j/(n+1)\}_{j=1}^n$ with $n \in \mathbb{N}^+$ interior points and meshsize $h = 1/(n+1)$. Using a 3-point stencil for the second order derivative, the discrete problem associated to (2.1) reads

$$-\frac{1}{h^2}\left(u_{i-1} - 2u_i + u_{i+1}\right) - \widehat{k}^2 u_i = f(x_i),\ i = 1\cdots, n, \tag{2.2}$$

where $\widehat{k}$ is the shifted wavenumber introduced in [14] given by

$$\widehat{k} = \sqrt{\frac{2}{h^2}(1-\cos(kh))}.$$

Inserting $u_j := \mathrm{e}^{\mathrm{i}k_d x_j}$ into (2.2) with $f = 0$ and neglecting the boundaries, we get that the discrete wavenumber k_d is solution to

$$\cos(k_d h) = 1 - \frac{\widehat{k}^2 h^2}{2}.$$

This yields

$$k_d = k,$$

and thus this scheme does not have dispersion error. We also assume that

$$kh \notin \pi\mathbb{N},$$

since otherwise, we would have $\widehat{k} \in \left\{0, \sqrt{2}/h, 2/h\right\}$ and therefore $\widehat{k}$ no longer converges to k as $h \to 0$, and the stencil would no longer be consistent.

In what follows, we compute the l^∞ error for the 3-point stencil with shifted wavenumber. We begin by computing the local truncation error.

Theorem 2.1. *Assume that $f \in \mathcal{C}^2(0,1)$ and let τ_i be the local truncation error,*

$$\tau_i = -\frac{1}{h^2}\left(u(x_{i-1}) - 2u(x_i) + u(x_{i+1})\right) - \widehat{k}^2 u(x_i) - f(x_i).$$

We then have the estimate

$$\begin{aligned} \|\boldsymbol{\tau}\|_\infty &:= \max_{1\le i\le n} |\tau_i| \le \frac{h^2}{12}\|f''\|_{L^\infty(0,1)} + \frac{k^2h^2}{12}\|f\|_{L^\infty(0,1)} \\ &+ \left(\frac{k^4h^2}{12} + \left|\widehat{k}^2 - k^2\right|\right)\|u\|_{L^\infty(0,1)}. \end{aligned}$$

Proof. Using a Taylor expansion, there exists $\xi_i^- \in (x_{i-1}, x_i)$ and $\xi_i^+ \in (x_i, x_{i+1})$ such that

$$\tau_i = -u''(x_i) - \widehat{k}^2 u(x_i) - f(x_i) - \frac{h^2}{24}\left(u^{(4)}(\xi_i^-) + u^{(4)}(\xi_i^+)\right).$$

Adding and subtracting $k^2u(x_i)$ and using Eq. (2.1), we get

$$\begin{aligned} \tau_i &= -u''(x_i) - k^2u(x_i) - f(x_i) - (\widehat{k}^2 - k^2)u(x_i) - \frac{h^2}{24}\left(u^{(4)}(\xi_i^-) + u^{(4)}(\xi_i^+)\right) \\ &= -(\widehat{k}^2 - k^2)u(x_i) - \frac{h^2}{24}\left(u^{(4)}(\xi_i^-) + u^{(4)}(\xi_i^+)\right). \end{aligned}$$

Noting that

$$u^{(4)}(x) = -f''(x) - k^2(-f(x) - k^2 u(x)) = -f''(x) + k^2 f(x) + k^4 u(x),$$

we obtain the estimate. □

Using a Taylor expansion, we have

$$\left|\widehat{k}^2 - k^2\right| = \frac{2}{h^2}\left|1 - \cos(kh) - \frac{(kh)^2}{2}\right| \leq \frac{2}{h^2}\frac{(kh)^4}{4!} = \frac{k^4 h^2}{12}, \tag{2.3}$$

from which, together with Theorem 2.1, we can see that the shift does not modify the dependence with respect to k, h of the upper bound of the truncation error.

The discrete problem (2.2) can be written as a linear system

$$A_{\widehat{k}}\boldsymbol{u} = \boldsymbol{f},$$

where $A_k := h^{-2}\text{tridiag}(-1, 2 - k^2h^2, -1)$, $\boldsymbol{u} := (u_i)_{i=1}^n$ and $\boldsymbol{f} := (f(x_i))_{i=1}^n$. The eigenvalues of A_k are

$$\lambda_j(k) = \frac{4}{h^2}\sin\left(\frac{j\pi h}{2}\right)^2 - k^2,\ j = 1, \cdots, n.$$

As a result,

$$\lambda_j(\widehat{k}) = 0 \iff k^2 = j\pi,$$

and thus the matrix $A_{\widehat{k}}$ is non-singular as soon as k is not an eigenvalue of the (continuous) Laplace operator acting on $H_0^1(0,1)$.

Let $\boldsymbol{e}_{\widetilde{k}} := (u_i - u(x_i))_{i=1}^n$ be the error, which satisfies

$$A_{\widetilde{k}}\boldsymbol{e}_{\widetilde{k}} = \boldsymbol{\tau},$$

where $\boldsymbol{\tau} = (\tau_i)_{i=1}^n$ is the vector of local truncation errors. We now estimate $\left\|A_{\widetilde{k}}^{-1}\right\|_\infty$ without dispersion correction, $\widetilde{k} = k$, and with dispersion correction, $\widetilde{k} = \widehat{k}$.

Theorem 2.2.

- *No dispersion correction: Assume that* $kh < 2$ *and that* $\lambda_j(k) \neq 0$, *then*

$$\left\|A_k^{-1}\right\|_\infty \leq \frac{h}{|\sin(\theta)|}\frac{1}{|\sin(\theta/h)|},$$

with $\cos(\theta) = 1 - (kh)^2/2$.

- *With dispersion correction: Assume that* $kh \notin \pi\mathbb{N}$ *and* $k \notin \pi\mathbb{N}$, *then*

$$\left\|A_{\widehat{k}}^{-1}\right\|_\infty \leq \frac{kh}{|\sin(kh)|}\frac{1}{k|\sin(k)|}.$$

Proof. We use [9, p. 15, Corollary 4.2] to compute explicitly the elements of the inverse of $A_{\widetilde{k}}$ which yields

$$\left(A_{\widetilde{k}}^{-1}\right)_{i,j} = \begin{cases} (-1)^{i+j}\,\frac{b^{j-i}}{|b|^{j-i+1}}\,\frac{U_{i-1}\left(\frac{a}{2|b|}\right)U_{n-j}\left(\frac{a}{2|b|}\right)}{U_n\left(\frac{a}{2|b|}\right)} & i \leq j, \\ (-1)^{i+j}\,\frac{b^{i-j}}{|b|^{i-j+1}}\,\frac{U_{j-1}\left(\frac{a}{2|b|}\right)U_{n-i}\left(\frac{a}{2|b|}\right)}{U_n\left(\frac{a}{2|b|}\right)} & i > j, \end{cases}$$

where $a := (2 - \widetilde{k}^2 h^2)/h^2$, $b := -1/h^2$ and $U_l(x)$ are the Chebychev polynomials of the second kind that are defined as

$$U_l(x) := \begin{cases} \frac{\sin((l+1)\theta)}{\sin(\theta)} \text{ with } x := \cos(\theta) & \text{if } |x| < 1, \\ \frac{\sinh((l+1)\theta)}{\sinh(\theta)} \text{ with } x := \cosh(\theta) & \text{if } |x| > 1. \end{cases}$$

Note that

$$U_n\left(\frac{a}{2|b|}\right) = \frac{\sin((n+1)\theta)}{\sin(\theta)} = \frac{\sin(\theta/h)}{\sin(\theta)},$$

and that the following bound holds:

$$\left|U_l\left(\frac{a}{2|b|}\right)\right| \leq \left|\frac{\sin((l+1)\theta)}{\sin(\theta)}\right| \leq \frac{1}{|\sin(\theta)|}.$$

The infinity norm of $A_{\widetilde{k}}^{-1}$ can then be estimated as

$$\begin{aligned} \left\|A_{\widetilde{k}}^{-1}\right\|_\infty &= \max_{1\leq i\leq n} \sum_{j=1}^{n} \left|\left(A_{\widehat{k}}^{-1}\right)_{\boldsymbol{i}}\right| \leq \frac{n}{|b|\left|U_n\left(\frac{a}{2|b|}\right)\right|} \frac{1}{|\sin(\theta)|^2} \\ &\leq \frac{h^2 n}{|\sin(\theta)||\sin(\theta/h)|} \leq \frac{h}{|\sin(\theta)||\sin(\theta/h)|}. \end{aligned}$$

If no dispersion correction is used, $\widetilde{k} = k$, the assumptions ensure that $\theta = \arccos(1 - (kh)^2/2)$ is well-defined and the previous estimate gives the result. If dispersion correction is used, $\widetilde{k} = \widehat{k}$, note that

$$\frac{a}{2|b|} = 1 - \frac{\widehat{k}^2 h^2}{2} = 1 - (1 - \cos(kh)) = \cos(kh),$$

which is strictly smaller than 1 since $kh \notin \pi\mathbb{N}$. This gives $\theta = \pm kh$ and the bound on $\left\|A_{\widehat{k}}^{-1}\right\|_\infty$ translates into

$$\left\|A_{\widehat{k}}^{-1}\right\|_\infty \leq \frac{h}{|\sin(kh)||\sin(k)|},$$

which concludes the proof. □

Using Theorem 2.2, we can now get the final error estimate. We are also going to consider a source term f that can appear in some physical applications which may depend on the wavenumber k.

Theorem 2.3. *Let the assumptions of Theorem 2.2 hold, and assume that the source term f satisfies*

$$\|f\|_{L^\infty} \lesssim 1, \;\; \|f''\|_{L^\infty} \lesssim k^2,$$

where the notation $\lesssim$ means that the omitted constants do not depend on k and h.

- *No dispersion correction: If $\widetilde{k} = k$, then the error satisfies*

$$\|\boldsymbol{e}_k\|_\infty \lesssim \frac{k^2 h^3}{|\sin(\theta)||\sin(\theta/h)|}\left(1 + k\left(1 + \frac{1}{|\sin(k)|}\right)\right),$$

where θ is given in Theorem 2.2.

- *With dispersion correction: If $\widetilde{k} = \widehat{k}$, then the error satisfies*

$$\|\boldsymbol{e}_{\widehat{k}}\|_\infty \lesssim \frac{kh}{|\sin(kh)|}\frac{1}{|\sin(k)|}\left(\frac{(kh)^2}{k} + (kh)^2\left(1 + \frac{1}{|\sin(k)|}\right)\right),$$

from which one can see, for any k such that $k, kh \notin \pi\mathbb{N}$, that the error decreases like $O(G^{-2})$ for any wavenumber, where $G := 2\pi/(kh)$ denotes the number of points per wavelength.

Proof. To get the final error estimate, we need some bounds on u satisfying (2.1), which is actually explicitly given by

$$u(x) = \frac{\sin(kx - k)}{k\sin(k)}\int_0^1 \sin(ky)f(y)dy + \frac{1}{k}\int_x^1 \sin(k(y - x))f(y)dy,$$

and thus satisfies the estimate

$$\|u\|_{L^\infty(0,1)} \leq \|f\|_{L^\infty(0,1)}\left(\frac{1}{k|\sin(k)|} + \frac{1}{k}\right). \tag{2.4}$$

Using Theorem 2.1 and the estimate (2.4), we get for the infinity norm of the error the upper bound

$$\begin{aligned}
\|\boldsymbol{e}_{\widetilde{k}}\|_\infty &\leq \left\|A_{\widetilde{k}}^{-1}\right\|_\infty \|\boldsymbol{\tau}\|_\infty \\
&\leq \left\|A_{\widetilde{k}}^{-1}\right\|_\infty \left(\frac{h^2}{12}\|f''\|_{L^\infty(0,1)} + \frac{k^2h^2}{12}\|f\|_{L^\infty(0,1)}\right) \\
&+ \left\|A_{\widetilde{k}}^{-1}\right\|_\infty \left(\frac{k^4h^2}{12} + \left|\widetilde{k}^2 - k^2\right|\right)\|f\|_{L^\infty(0,1)}\left(\frac{1}{k|\sin(k)|} + \frac{1}{k}\right) \\
&\lesssim \left\|A_{\widetilde{k}}^{-1}\right\|_\infty \left((kh)^2 + k(kh)^2\left(1 + \frac{1}{|\sin(k)|}\right)\right),
\end{aligned}$$

where we used (2.3) and the assumptions on f to get the last upper bound. The proof can then be completed by applying Theorem 2.2. □

We now make some comments regarding the results of Theorem 2.3:

- *No dispersion correction:* If no dispersion correction is used, $\widetilde{k} = k$, Theorem 2.3 together with a Taylor expansion gives

$$\frac{1}{\sin(\theta)}\frac{1}{\sin(\theta/h)} = \frac{1}{kh\sin(k)} + \frac{2kh}{\sin(k)}\left(\frac{1}{16} - k\frac{\cos(k)}{48\sin(k)}\right) + O(h^3),$$

from which we see that as $h \to 0$, there is a term of the form $k(kh)^4$ in the expansion of the upper bound of the error $\|\boldsymbol{e}_k\|_\infty$. The presence of the pollution effect suggests that the above bound cannot be improved.

- *Using dispersion correction:* If dispersion correction is used, $\widetilde{k} = \widehat{k}$, then, since $\lim_{kh\to 0}(kh)/\sin(kh) = 1$, this term does not contribute to the convergence rate. The FD scheme with dispersion correction does not suffer from the pollution effect since, for any k such that $k, kh \notin \pi\mathbb{N}$, the error decreases like $O(G^{-2})$ for any wavenumber. Notice however that the convergence rate is deteriorating if k comes close to a continuous or a discrete eigenvalue.

Remark 2.4 (Suppressing dispersion error for general 1d FD schemes). *We consider a uniform grid and the following general stencil associated to the Helmholtz operator:*

$$\left(\mathcal{H}_h u\right)_i := -\left(D_h^2 u\right)_i - k^2\left(M_h u\right)_i,$$

where the subscript i means the approximation is computed at grid point x_i. The (discrete) symbol can then be defined as $\sigma_d(k,\xi,h) = \mathrm{e}^{-\mathrm{i}\xi x_i}\left(\mathcal{H}_h \mathrm{e}^{\mathrm{i}\xi x}\right)_i$ and it can always be written as

$$\sigma_d(k,\xi,h) := \sigma_{-\partial_x^2}(\xi,h) - k^2\sigma_M(\xi,h),$$

where $\sigma_{-\partial_x^2}(\xi,h)$ and $\sigma_M(\xi,h)$ are the discrete symbols associated to the FD discretization of the operator $\varphi \mapsto -\partial_x^2\varphi$ and the constant multiplication operator $\varphi \mapsto 1 \times \varphi$, and $\sigma_{-\partial_x^2}(\xi,h) \to \xi^2$ and $\sigma_M(\xi,h) \to 1$ as $h \to 0$.

We recall the discrete wavenumber is defined as k_d satisfying $\sigma_d(k,k_d,h) = 0$. Now setting $\widehat{k}$ to

$$\widehat{k}^2 = \frac{\sigma_{-\partial_x^2}(k,h)}{\sigma_M(k,h)}, \tag{2.5}$$

we have $\lim_{h\to 0}\widehat{k}^2 = k^2$. In addition, when using this shifted wavenumber, the discrete wavenumber $\widehat{k}_d$ verifies

$$\sigma_d(\widehat{k},\widehat{k}_d,h) = 0 = \sigma_{-\partial_x^2}(\widehat{k}_d,h) - \widehat{k}^2\sigma_M(\hat{k}_d,h).$$

Therefore, $\widehat{k}_d$ satisfies

$$\frac{\sigma_{-\partial_x^2}(\widehat{k}_d,h)}{\sigma_M(\widehat{k}_d,h)} = \widehat{k}^2 = \frac{\sigma_{-\partial_x^2}(k,h)}{\sigma_M(k,h)},$$

from which we see that $\widehat{k}_d = k$ is a solution showing there is no dispersion error.

We now apply the previous derivation to the FD scheme from [18, Eq. (2.4)] whose stencil is (with the notations of the present paper)

$$-\frac{1}{h^2}\left(u_{i+1} - 2u_i + u_{i-1}\right) - k^2\frac{\alpha u_{i+1} + 2(3-\alpha)u_i + \alpha u_{i-1}}{6}.$$

Note that we get the standard 3-point second order stencil for $\alpha = 0$ and a Taylor expansion also shows that, when applied to the homogeneous Helmholtz equation, this stencil is fourth order for $\alpha = 1/2$ (see also [23]). The discrete symbols for this stencil are

$$\sigma_{-\partial_x^2}(\xi,h) = \frac{2}{h^2}\left(1-\cos(\xi h)\right),\ \sigma_M(\xi,h) = \frac{1}{6}\left(2\alpha\cos(\xi h) + 6 - 2\alpha\right).$$

The shifted wavenumber is then defined by (2.5) which gives

$$\widehat{k}^2 = \frac{6}{h^2}\left(\frac{1-\cos(kh)}{\alpha\cos(kh) + 3 - \alpha}\right),$$

and one can check the FD stencil using $\widehat{k}$ instead of k is now free from dispersion error. A Taylor expansion also yields

$$\widehat{k}^2 = k^2 + k^4\frac{h^2}{12}\left(2\alpha - 1\right) + O(h^4)$$

which shows that the FD scheme with shifted wavenumber is again 2nd order for $\alpha = 0$ and 4-th order for $\alpha = 1/2$. We also emphasize that the proof of Theorem 2.2 could be extended to the $\alpha-$scheme above since the matrix is again tri-diagonal.

3. Reducing dispersion error for general FD schemes in higher dimensions.

For Helmholtz problems in dimension $d > 1$, we cannot suppress the dispersion error completely, but we can reduce it. To do so, we introduce the symbol of the continuous Helmholtz operator $\mathcal{H} = -(\Delta + k^2)$ which is

$$\sigma_c(k, \boldsymbol{\xi}) = |\boldsymbol{\xi}|^2 - k^2.$$

We now consider a uniform grid embedded in $\mathbb{R}^d$, with meshsize h, and a general finite difference discretization of $\mathcal{H}$ defined as

$$(\mathcal{H}_h u)_{\boldsymbol{i}} = -(\Delta_h u)_{\boldsymbol{i}} - k^2 (M_h u)_{\boldsymbol{i}},$$

where the subscript $\boldsymbol{i}$ indicates that the approximation is computed at the grid point $\boldsymbol{x}_{\boldsymbol{i}}$. The discrete symbol is then

$$\sigma_d(k, \boldsymbol{\xi}, h) = \left(e^{-\mathrm{i}\boldsymbol{x}\cdot\boldsymbol{\xi}}\right)_{\boldsymbol{i}} \left(\mathcal{H}_h e^{\mathrm{i}\boldsymbol{x}\cdot\boldsymbol{\xi}}\right)_{\boldsymbol{i}}.$$

The discrete wavenumber is, for any $\boldsymbol{\theta} \in \mathcal{S}^{d-1}$, $k_d := k_d(k, \boldsymbol{\theta}, h)$ that satisfies

$$\sigma_d(k, k_d\boldsymbol{\theta}, h) = 0,$$

and we usually have

$$k_d(k, \boldsymbol{\theta}, h) \neq k,$$

which is again the dispersion error. We also introduce the discrete and continuous dispersion relations

$$\begin{aligned} \mathcal{D}_c &:= \left\{\boldsymbol{\xi} \in \mathbb{R}^d \mid |\boldsymbol{\xi}|^2 - k^2 = 0\right\}, \\ \mathcal{D}_d &:= \left\{\boldsymbol{\xi} \in \mathbb{R}^d \mid \sigma_d(k, \boldsymbol{\xi}, h) = 0\right\}. \end{aligned}$$

For a consistent numerical scheme, we have $\lim_{h\to 0} \sigma_d(k, \boldsymbol{\xi}, h) = \sigma_c(k, \boldsymbol{\xi})$ for all $k, \boldsymbol{\xi}$. As a result,

$$\lim_{h\to 0} k_d(k, k\boldsymbol{\theta}, h) = k,$$

for every $k, \boldsymbol{\theta}$. In what follows, we first compute an expansion of k_d as h goes to 0 and next introduce the so-called asymptotic optimal shifted wavenumber which is actually defined up to a free parameter that is next used to minimize the dispersion error for h small enough.

3.1. Expansion of the discrete wavenumber for small meshsize.

From now on, we assume that

(H1) The discrete symbol admits the expansion

$$\sigma_d(k, k\boldsymbol{\theta}, h) = h^p \mathcal{E}(k, k\boldsymbol{\theta}) + O(h^{p+1}),$$

for a smooth function $\mathcal{E}$.

(H2) For a given wavenumber k, the sequence of functions $\left(\nabla_{\boldsymbol{\xi}}\sigma_d(k, \cdot, h)\right)_h$ converges uniformly to $\nabla_{\boldsymbol{\xi}}\sigma_c(k, \cdot)$ on a compact neighborhood of $\boldsymbol{\xi} = k\boldsymbol{\theta}$ for $\boldsymbol{\theta} \in \mathcal{S}^{d-1}$.

(H3) For a given $\boldsymbol{\xi}$, the sequence of functions $\left(\partial_k \sigma_d(\cdot, \boldsymbol{\xi}, h)\right)_h$ converges uniformly to $\partial_k \sigma_c(\cdot, \boldsymbol{\xi})$ on a compact neighborhood of the wavenumber k.

We emphasize that (H1) is satisfied for any FD scheme that is of order p on plane waves. This means that

$$(\mathcal{H}_h u_{\boldsymbol{\theta}})_{\boldsymbol{i}} = h^p \mathcal{E}(k, k\boldsymbol{\theta}) u_{\boldsymbol{\theta}}(\boldsymbol{x}_{\boldsymbol{i}}) + O(h^{p+1}),$$

for any plane wave $u_{\boldsymbol{\theta}}(\boldsymbol{x}) = e^{\mathrm{i}k\boldsymbol{x}\cdot\boldsymbol{\theta}}$ where $\boldsymbol{\theta} \in \mathcal{S}^{d-1}$. As a result, assumption (H1) can be derived by computing directly the Taylor expansion of the discrete symbol at $\boldsymbol{\xi} = k\boldsymbol{\theta}$ and keeping the leading order term. It is also worth noting that, although (H2) and (H3) seem rather technical, they are also easily checked by computing the Taylor expansion of the derivatives of the discrete symbol with respect to $(k, \boldsymbol{\xi})$.

We show below the existence of a discrete wavenumber for small enough meshsize.

Proposition 3.1. *Assume that $(H1) - (H2)$ hold. Then for each $k, \boldsymbol{\theta}$ we have some $h_0 > 0$ such that for all $h \leq h_0$, there exists a discrete wavenumber $k_d(k, \boldsymbol{\theta}, h)$ satisfying $\sigma_d(k, k_d\boldsymbol{\theta}, h) = 0$.*

Proof. Let us fix $k, \boldsymbol{\theta}$ and let $\delta \in \mathbb{R}$ be some given constant. A Taylor expansion gives

$$\sigma_d(k, (k \pm \delta h^p)\boldsymbol{\theta}, h) = \sigma_d(k, k\boldsymbol{\theta}, h) \pm \delta h^p \nabla_\xi \sigma_d(k, k\boldsymbol{\theta}, h) \cdot \boldsymbol{\theta} + O(h^{2p}).$$

From $(H2)$, as $h \to 0$, we have that

$$\nabla_\xi \sigma_d(k, k\boldsymbol{\theta}, h) = \nabla_\xi \sigma(k, k\boldsymbol{\theta}) + o(1) = 2k\boldsymbol{\theta} + o(1),$$

and using then (H1), we get

$$\sigma_d(k, (k \pm \delta h^p)\boldsymbol{\theta}, h) = h^p \left(\mathcal{E}(k, k\boldsymbol{\theta}) \pm 2k\delta\right) + O(h^{p+1}).$$

For some positive constant C such that $2\mathcal{E}(k, k\boldsymbol{\theta}) < C$, we now set

$$\delta = \frac{1}{2k} \left(\mathcal{E}(k, k\boldsymbol{\theta}) - C\right),$$

which gives

$$\begin{aligned} \sigma_d(k, (k + \delta h^p)\boldsymbol{\theta}, h) &= h^p \left(2\mathcal{E}(k, k\boldsymbol{\theta}) - C\right) + O(h^{p+1}), \\ \sigma_d(k, (k - \delta h^p)\boldsymbol{\theta}, h) &= Ch^p + O(h^{p+1}). \end{aligned}$$

From these, we see that there exists h_0 such that for all $h \leq h_0$, we have $\sigma_d(k, (k + \delta h^p)\boldsymbol{\theta}, h) < 0$ and $\sigma_d(k, (k - \delta h^p)\boldsymbol{\theta}, h) > 0$. Since the discrete symbol σ_d is continuous in all its variables, we obtain that there exists $k_d(k, \boldsymbol{\theta}, h) \in (k - \delta h^p, k + \delta h^p)$ such that $\sigma_d(k, k_d\boldsymbol{\theta}, h) = 0$. □

We can now compute the expansion of the discrete wavenumber k_d as $h \to 0$.

Theorem 3.2. *Assume that (H1) and (H2) hold.*

(i) Then the discrete wavenumber has the expansion

$$k_d(k, \boldsymbol{\theta}, h) = k - \frac{h^p}{2k} \mathcal{E}(k, k\boldsymbol{\theta}) + O(h^{p+1}).$$

(ii) If in addition, $\mathcal{D}_d$ has a polar representation of the form

$$\forall \boldsymbol{\theta} \in \mathcal{S}^{d-1} \text{ there is a unique } \boldsymbol{\xi} \in \mathcal{D}_d \text{ such that } \boldsymbol{\xi} = |\boldsymbol{\xi}|\boldsymbol{\theta},$$

then the dispersion error can be defined as below and satisfies the estimate

$$\max_{\boldsymbol{\theta} \in \mathcal{S}^{d-1}} \operatorname{dist}\left(\mathcal{D}_c \cap L(\boldsymbol{\theta}), \mathcal{D}_d \cap L(\boldsymbol{\theta})\right) = \frac{h^p}{2k} \max_{\boldsymbol{\theta}} |\mathcal{E}(k, k\boldsymbol{\theta})| + O(h^{p+1}),$$

where $L(\boldsymbol{\theta})$ is the line passing through the origin with direction vector $\boldsymbol{\theta}$.

Proof. For (i), we start from $(H1)$ which gives

$$\sigma_d(k, k_d\boldsymbol{\theta}, h) - \sigma_d(k, k\boldsymbol{\theta}, h) = 0 - h^p\mathcal{E}(k, k\boldsymbol{\theta}) + O(h^{p+1}).$$

Using a Taylor expansion with integral remainder, this gives

$$(k_d - k)R(\boldsymbol{\theta}, h) = -h^p\mathcal{E}(k, k\boldsymbol{\theta}) + O(h^{p+1}),$$

with

$$R(\boldsymbol{\theta}, h) = \int_0^1 \boldsymbol{\theta}\cdot\nabla_{\boldsymbol{\xi}}\sigma_d(k, \{k + s(k_d - k)\}\boldsymbol{\theta}, h)\, ds.$$

Since $k_d \to k$ as $h \to 0$, we have $\boldsymbol{\xi}_h := \{k+s(k_d-k)\}\boldsymbol{\theta} \to k\boldsymbol{\theta}$ and, for h small enough, the sequence $\boldsymbol{\xi}_h$ remains in a compact neighborhood of $k\boldsymbol{\theta}$. Assumption (H2) then gives that

$$\nabla_{\boldsymbol{\xi}}\sigma_d(k, \{k + s(k_d - k)\}\boldsymbol{\theta}, h) \to \nabla_{\boldsymbol{\xi}}\sigma_c(k, k\boldsymbol{\theta})$$

uniformly, and we can exchange the limit and integral symbols. Since $\nabla_{\boldsymbol{\xi}}\sigma_c = 2\boldsymbol{\xi}$, we obtain

$$\lim_{h\to 0} R(\boldsymbol{\theta}, h) = 2k\int_0^1 \boldsymbol{\theta}\cdot\boldsymbol{\theta}\, ds = 2k.$$

We then finally get

$$(k_d - k) = \frac{1}{2k + o(1)}\left(-h^p\mathcal{E}(k, k\boldsymbol{\theta}) + O(h^{p+1})\right) = -\frac{h^p}{2k}\mathcal{E}(k, k\boldsymbol{\theta}) + O(h^{p+1}),$$

which is the desired estimate.

To prove (ii), note that our assumption and the definition of the discrete wavenumber imply that the discrete dispersion relation can be written as

$$\mathcal{D}_d = \left\{k_d(k, \boldsymbol{\theta}, h)\boldsymbol{\theta} \mid \boldsymbol{\theta} \in \mathcal{S}^{d-1}\right\}.$$

Since the continuous dispersion relation is the sphere centered at 0 with radius k, the dispersion error satisfies

$$\begin{aligned}\max_{\boldsymbol{\theta}\in\mathcal{S}^{d-1}} \operatorname{dist}\left(\mathcal{D}_c \cap L(\boldsymbol{\theta}), \mathcal{D}_d \cap L(\boldsymbol{\theta})\right) &= \max_{\boldsymbol{\theta}\in\mathcal{S}^{d-1}} |k\boldsymbol{\theta} - k_d\boldsymbol{\theta}| \\ &= \max_{\boldsymbol{\theta}\in\mathcal{S}^{d-1}} \left|\frac{h^p}{2k}\mathcal{E}(k, k\boldsymbol{\theta})\right| + O(h^{p+1}),\end{aligned}$$

where we used the result from (i) to get the last estimate. □

3.2. Definition of the shifted wavenumber. We now introduce a real shift $\widehat{k} := \widehat{k}(k, h)$ in the finite difference stencil, which leads to the discrete symbol

$$\sigma_d(k, \boldsymbol{\xi}, h) = \sigma_{-\Delta}(\boldsymbol{\xi}, h) - \widehat{k}^2\sigma_M(\boldsymbol{\xi}, h),$$

where $\sigma_{-\Delta}(\boldsymbol{\xi}, h) \to |\boldsymbol{\xi}|^2$ and $\sigma_M(\boldsymbol{\xi}, h) \to 1$, for all $\boldsymbol{\xi}$, as $h \to 0$. We assume that

$$\widehat{k} = k + k_p h^p,$$

and compute the expansion of the discrete symbol $\sigma_d(\widehat{k}, k\boldsymbol{\theta}, h)$ as h goes to 0. A Taylor formula with integral remainder gives

$$\sigma_d(\widehat{k}, k\boldsymbol{\theta}, h) = \sigma_d(k, k\boldsymbol{\theta}, h) + k_p h^p \int_0^1 \partial_k \sigma_d(k + s k_p h^p, k\boldsymbol{\theta}, h)\, ds.$$

Using (H3), we can invert the limit and integral signs to get

$$\lim_{h\to 0} \int_0^1 \partial_k \sigma_d(k + s k_p h^p, k\boldsymbol{\theta}, h)\, ds = \int_0^1 \partial_k \sigma_c(k, k\boldsymbol{\theta}, h)\, ds = -2k.$$

From (H1), we then obtain

$$\sigma_d(\widehat{k}, k\boldsymbol{\theta}, h) = -2kk_p h^p + h^p \mathcal{E}(k, k\boldsymbol{\theta}) + O(h^{p+1}) = h^p\left(-2kk_p + \mathcal{E}(k, k\boldsymbol{\theta})\right) + O(h^{p+1}),$$

and Theorem 3.2 gives

$$k_d(k, \boldsymbol{\theta}, h) = k - \frac{h^p}{2k}\left(-2kk_p + \mathcal{E}(k, k\boldsymbol{\theta})\right) + O(h^{p+1}).$$

From this, we can minimize the dispersion error for small enough meshsize by taking

$$k_p^{\text{asy}} := \arg\min_{k_p} \left(\max_{\boldsymbol{\theta}\in\mathcal{S}^{d-1}} \left|-2kk_p + \mathcal{E}(k, k\boldsymbol{\theta})\right| \right). \tag{3.1}$$

In what follows, we call k_p^{asy} the *asymptotically optimal shift*, since it minimizes the dispersion error as $h \to 0$.

We now give an explicit formula for the asymptotically optimal shift.

THEOREM 3.3. *Assume that there exists some* $\boldsymbol{\theta}_{\min}$, $\boldsymbol{\theta}_{\max}$ *such that*

$$\mathcal{E}_{\min} := \mathcal{E}(k, k\boldsymbol{\theta}_{\min}) \le \mathcal{E}(k, k\boldsymbol{\theta}) \le \mathcal{E}(k, k\boldsymbol{\theta}_{\max}) =: \mathcal{E}_{\max},$$

where the lower and upper bounds may depend on the wavenumber k. Then the solution of (3.1) *is unique, and is given by*

$$k_p^{\text{asy}} = \frac{1}{4k}\left(\mathcal{E}_{\max} + \mathcal{E}_{\min}\right),$$

and the relative dispersion error satisfies

$$\max_{\boldsymbol{\theta}\in\mathcal{S}^{d-1}} \left|\frac{k_d(\widehat{k}^{\text{asy}}, \boldsymbol{\theta}, h) - k}{k}\right| = \frac{h^p}{2k^2}\left|\frac{\mathcal{E}_{\max} - \mathcal{E}_{\min}}{2}\right| + O(h^{p+1}),$$

where $\widehat{k}^{\text{asy}} = k + h^p k_p^{\text{asy}}$.

Proof. It can be checked by direct computations that the function $k_p \in \mathbb{R} \mapsto \max_{\boldsymbol{\theta}\in\mathcal{S}^{d-1}} \left|-2kk_p + \mathcal{E}(k, k\boldsymbol{\theta})\right|$ is convex. As a result, it has a unique minimum from which the uniqueness of k_p^{asy} follows.

For all $\boldsymbol{\theta} \in \mathcal{S}^{d-1}$, we have

$$-2kk_p + \mathcal{E}_{\min} \le -2kk_p + \mathcal{E}(k, k\boldsymbol{\theta}) \le -2kk_p + \mathcal{E}_{\max}.$$

Since $\mathcal{E}(k, k\boldsymbol{\theta})$ reaches its extrema, this gives

$$\max_{\boldsymbol{\theta}} \left|-2kk_p + \mathcal{E}(k, k\boldsymbol{\theta})\right| = \max\left\{\left|-2kk_p + \mathcal{E}_{\min}\right|, \left|-2kk_p + \mathcal{E}_{\max}\right|\right\} := F(k_p),$$

and we now need to find the argmin of $F(k_p)$ to get k_p^{asy}. We emphasize that both $k_p \mapsto -2kk_p + \mathcal{E}_{\min}$ and $k_p \mapsto -2kk_p + \mathcal{E}_{\max}$ are affine functions with the same slope and thus the minimal value of F is reached for k_p^{asy} such that

$$\left(-2kk_p^{\mathrm{asy}} + \mathcal{E}_{\min}\right) = -\left(-2kk_p^{\mathrm{asy}} + \mathcal{E}_{\max}\right),$$

from which we can derive the announced formula. To get the estimate on the relative dispersion error, we use Theorem 3.2 which yields

$$k_d(\widehat{k}^{\mathrm{asy}}, k\boldsymbol{\theta}, h) = k - \frac{h^p}{2k}\left(-2kk_p^{\mathrm{asy}} + \mathcal{E}(k, k\boldsymbol{\theta})\right) + O(h^{p+1}),$$

and next use that $\max_{\boldsymbol{\theta}} \left|-2kk_p^{\mathrm{asy}} + \mathcal{E}(k, k\boldsymbol{\theta})\right| = \left|-2kk_p^{\mathrm{asy}} + \mathcal{E}_{\max}\right|$. □

The relative dispersion error without shift satisfies

$$\max_{\boldsymbol{\theta}\in\mathcal{S}^{d-1}} \left|\frac{k_d(k, \boldsymbol{\theta}, h) - k}{k}\right| = \frac{h^p}{2k^2} \max\left\{|\mathcal{E}_{\max}|, |\mathcal{E}_{\min}|\right\} + O(h^{p+1}).$$

Using then Theorem 3.3, we obtain that the shift reduces the relative dispersion error by the factor

$$\mathrm{R_f} := \frac{\max_{\boldsymbol{\theta}} |\mathcal{E}(k, k\boldsymbol{\theta})|}{\max_{\boldsymbol{\theta}} |-2kk_p^{\mathrm{asy}} + \mathcal{E}(k, k\boldsymbol{\theta})|} = 2\frac{\max\left\{|\mathcal{E}_{\max}|, |\mathcal{E}_{\min}|\right\}}{|\mathcal{E}_{\max} - \mathcal{E}_{\min}|}. \tag{3.2}$$

We now get some lower bounds for the reduction factor $\mathrm{R_f}\left(\mathcal{E}_{\max}, \mathcal{E}_{\min}\right)$ defined as

$$\mathrm{R_f}(a, b) = 2\frac{\max\left(|a|, |b|\right)}{|a - b|}.$$

Assuming first that $0 < a < b$, we have $-b < a - b < 0$ and thus $|a - b| < |b|$ from which we infer

$$\mathrm{R_f}(a, b) = 2\frac{|b|}{|a - b|} > 2\frac{|b|}{|b|} = 2.$$

Noting that $\mathrm{R_f}(-a, -b) = \mathrm{R_f}(a, b)$ and $\mathrm{R_f}(a, b) = \mathrm{R_f}(b, a)$, we have proved that

$$\forall (a, b) \in \left(\mathbb{R}^+\right)^2 \cup \left(\mathbb{R}^-\right)^2 \text{ with } a \neq b: \ \mathrm{R_f}(a, b) > 2. \tag{3.3}$$

Assuming now that $a < 0 < b$ and $|a| < |b|$, we have $|a - b| < 2|b|$ and thus

$$\mathrm{R_f}(a, b) = 2\frac{|b|}{|a - b|} > 1.$$

Using again the symmetry properties of $\mathrm{R_f}$, we obtain

$$\forall (a, b) \in \left(\mathbb{R}^+ \times \mathbb{R}^-\right) \cup \left(\mathbb{R}^- \times \mathbb{R}^+\right) \text{ with } |a| \neq |b| : \mathrm{R_f}(a, b) > 1.$$

We emphasize that $\mathrm{R_f}(a, b) = 1$ if and only if $a = -b$ but in this case, we would have $\mathcal{E}_{\min} = -\mathcal{E}_{\max}$ and thus the asymptotic shift is $k_p^{\mathrm{asy}} = 0$. It is worth noting that, in the next section, we only end up being in the case (3.3) for each stencil considered. As a result, the reduction factor is in each case greater than 2. More precisely, we prove in Theorem 4.1 that $\mathrm{R_f} = 4$ for the 5-point stencil, Theorem 4.2 gives that $\mathrm{R_f} = 64$ for a sixth-order 9-point stencil and we show in Theorem 4.3 that $\mathrm{R_f} = 3$ for the 7-point stencil in 3d.

4. Asymptotically optimal shift for some standard FD stencils. We now compute the asymptotically optimal shift in 2d for the standard second-order 5-point stencil, a sixth-order 9-point stencil, and the second-order 7-point stencil in 3d. In each case, we first compute the function $\mathcal{E}$ with a Taylor expansion of the discrete symbol, followed by its lower and upper bounds. Applying Theorem 3.3, we can then get the asymptotically optimal shift, as well as the improvement on the relative dispersion error by computing the reduction factor $\mathrm{R_f}$. We also show that the asymptotically optimal shift can be used to reduce the dispersion error even when a relatively small number of grid points per wavelength is used. Note that there is no extra cost when using this asymptotically optimal shift in solving the associated discretized systems, the improvement in the discrete solutions comes for free.

4.1. Application to the 5-point stencil in 2d. The second order 5-point stencil for the Helmholtz operator in 2d is defined as

$$\left(\mathcal{H}_h^{\text{5pt}} u\right)_{i,j} = \frac{-u_{i+1,j} - u_{i-1,j} + 4u_{i,j} - u_{i,j+1} - u_{i,j-1}}{h^2} - k^2 u_{i,j},$$

and the discrete symbol is therefore

$$\sigma_d^{\text{5pt}}(k, \boldsymbol{\xi}, h) = \frac{4 - 2(\cos(h\boldsymbol{\xi}_1) + \cos(h\boldsymbol{\xi}_2))}{h^2} - k^2.$$

A Taylor expansion gives

$$\sigma_d^{\text{5pt}}(k, \boldsymbol{\xi}, h) = \sigma_c(k, \boldsymbol{\xi}) - \frac{h^2}{12}\left(\xi_1^4 + \xi_2^4\right) + O(h^4).$$

Since any $\boldsymbol{\theta} \in \mathcal{S}^1$ can be written as $\boldsymbol{\theta} = (\cos(s), \sin(s))$ for $s \in [0, 2\pi]$, we obtain

$$\mathcal{E}(k, k\boldsymbol{\theta}) = -\frac{k^4}{12}\left(\cos(s)^4 + \sin(s)^4\right) = -\frac{k^4}{12}\left(2\cos(s)^4 - 2\cos(s)^2 + 1\right),$$

and it is easy to verify that the hypotheses (H1),(H2) and (H3) hold. We now introduce the shifted wavenumber as

$$\widehat{k} = k + k_2 h^2,$$

with k_2 defined by (3.1).

THEOREM 4.1. *The asymptotically optimal shift for the standard 5-point difference scheme and the associated reduction factor are*

$$k_2^{\text{asy}} = -\frac{k^3}{32}, \quad \mathrm{R_f} = 4.$$

Proof. To use Theorem 3.3, we have to find the extrema of $\mathcal{E}(k, k\boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \mathcal{S}^1$. Setting $X := \cos(s)^2$, this is equivalent to finding the extrema of

$$f(X) = -\frac{k^4}{12}\left(2X^2 - 2X + 1\right).$$

It is easy to see that $1/2 \le 2X^2 - 2X + 1 \le 1$ for all $X \in [0, 1]$ and thus

$$\mathcal{E}_{\min} = -\frac{k^4}{12}, \; \mathcal{E}_{\max} = -\frac{k^4}{24}.$$

Theorem 3.3 then gives for the asymptotically optimal shift

$$k_2^{\text{asy}} = \frac{\mathcal{E}_{\min} + \mathcal{E}_{\max}}{4k} = -\frac{k^3}{32}.$$

From (3.2), the reduction factor is

$$\mathrm{R_f} = 2\frac{\max\{|\mathcal{E}_{\max}|, |\mathcal{E}_{\min}|\}}{|\mathcal{E}_{\max} - \mathcal{E}_{\min}|} = 4.$$

□

The stencil with shifted wavenumber thus becomes

$$\left(\hat{\mathcal{H}}_h^{5\text{pt}} u\right)_{i,j} = \frac{-u_{i+1,j} - u_{i-1,j} + 4u_{i,j} - u_{i,j+1} - u_{i,j-1}}{h^2} - (k - h^2\frac{k^3}{32})^2 u_{i,j}. \quad (4.1)$$

We now verify the efficiency of the method presented above by computing numerically an optimal shift k_2^{opt} which minimizes the error between the discrete wavenumber k_d and the continuous wavenumber k. The computation of k_d satisfying

$$\sigma_d^{5\text{pt}}(k, k_d\boldsymbol{\theta}, h) = 0,$$

is done numerically since the optimal shift is obtained by computing first $k_d(\boldsymbol{\theta}, k_2)$ satisfying

$$\sigma_d^{5\text{pt}}(k + k_2h^2, k_d(\boldsymbol{\theta}, k_2)\boldsymbol{\theta}, h) = 0,$$

and next by minimizing $k_2 \mapsto \max_{\boldsymbol{\theta}} |k - k_d(k_2)|$. The optimization is done using the Matlab function *fminsearch*.

According to [4, Remark 5.2], the discrete dispersion relation is disconnected for $G < \pi$. When using the asymptotically optimal shift k_2^{asy} (see Theorem 4.1), this requirement translates to

$$G(k_2) = \frac{2\pi}{(k + k_2^{\text{asy}}h^2)h} = \frac{8G^3}{8G^2 - \pi^2} < \pi,$$

where $G = 2\pi/(kh)$ is the number of grid points per wavelength associated to the unshifted discrete Helmholtz equation. Accordingly, the dispersion relation of $\hat{\mathcal{H}}_h^{5\text{pt}}$ becomes disconnected for $G < \pi(1 + \sqrt{5})/4$ and we thus restrict our numerical optimization to $G \geq 2.5$. The relative dispersion error is

$$\mathrm{Err}_{\text{disp}}(\widetilde{k}) = \max_{k\in\mathcal{K}} \max_{\boldsymbol{\theta}} \frac{\left|k_d(\widetilde{k}, \boldsymbol{\theta}) - k\right|}{k},$$

where k_d is the discrete wavenumber and $\widetilde{k}$ is going to be either k, $\widehat{k}^{\text{asy}} = k + h^2k_2^{\text{asy}}$ or $\widehat{k}^{\text{opt}} = k + k_2^{\text{opt}}h^2$. We show in Figure 4.1 the relative error between the asymptotic shift k_2^{asy} and the optimized one k_2^{opt} as well as the relative dispersion error $\mathrm{Err}_{\text{disp}}(\widetilde{k})$. From Figure 4.1, we see that the relative error between k_2^{asy} and k_2^{opt} is smaller than 2% for $G \geq 3$ and thus our asymptotic derivation can be used even for meshsize and wavenumber combinations such that $kh \leq 2\pi/3 \approx 2$. We also note that, for G small, the relative dispersion error is large. This can be explained by the fact that the discrete dispersion relation becomes disconnected for G small and even empty for smaller G (see [4, Theorem 5.1]).

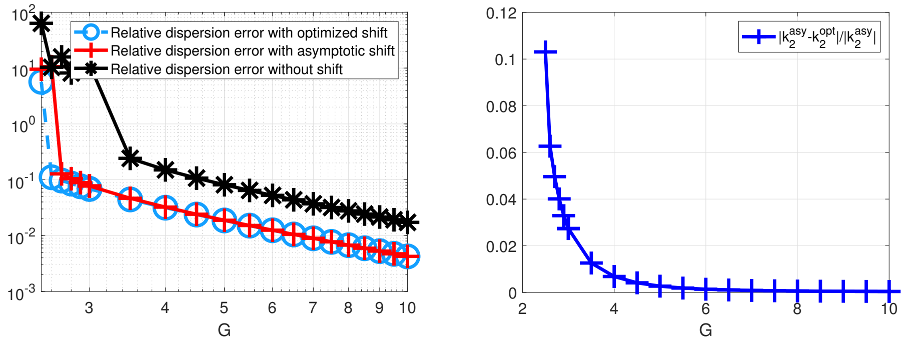

Fig. 4.1. *Left: Log-log plot of the relative dispersion error for the 5-point FD stencil using either $\widetilde{k} = k$ (no dispersion correction), $\widetilde{k} = k + h^2 k_2^{\mathrm{opt}}$ or $\widetilde{k} = k + h^2 k_2^{\mathrm{asy}}$. Right: $\frac{|k_2^{\mathrm{asy}} - k_2^{\mathrm{opt}}|}{|k_2^{\mathrm{asy}}|}$ as a function of G. We used $\mathcal{K} = \{20, 40, 80, 100, 140, 160, 180, 200, 250, 300, 600\}$.*

4.2. Application to a 6th order 9-point stencil. We now derive the asymptotically optimal shift for the 9-point 6-th order FD scheme from [4, Theorem 4.1] whose stencil is

$$\begin{aligned}
&\left(\mathcal{H}_h^{9-\mathrm{pts}} v\right)_{\boldsymbol{i}} := \left(\frac{4a}{h^2} - k_g^2 b\right) v(x_i, y_j) \\
&+ \left(\frac{1-2a}{h^2} - \frac{k_g^2 c}{4}\right) \left(v(x_{i-1}, y_j) + v(x_{i+1}, y_j) + v(x_i, y_{j-1}) + v(x_i, y_{j+1})\right). \qquad (4.2) \\
&- \left(\frac{1-a}{h^2} + k_g^2 \frac{1-b-c}{4}\right) \left(v(x_{i-1}, y_{j-1}) + v(x_{i+1}, y_{j-1}) + v(x_{i-1}, y_{j+1}) + v(x_{i+1}, y_{j+1})\right),
\end{aligned}$$

where a, b, c and k_g are positive constants given by

$$a = \frac{5}{6}, \quad b = \frac{5}{6} - \frac{c}{2}, \quad c = \frac{8}{45} + c_2 G^{-2}, \quad k_g = k - \frac{\pi^4 k}{30} G^{-4},$$

with c_2 being a free-parameter. The discrete symbol associated to $\mathcal{H}_h^{9-\mathrm{pts}}$ satisfies the expansion

$$\begin{aligned}
\sigma_d^{9-\mathrm{pts}}(k, k\boldsymbol{\theta}, h) &= -\frac{k^8 h^6}{6048\pi^2}(2\cos(\theta)^8\pi^2 - 4\cos(\theta)^6\pi^2 + 6\cos(\theta)^4\pi^2 \\
&+ 189 c_2 \cos(\theta)^4 - 4\cos(\theta)^2\pi^2 - 189\cos(\theta)^2 c_2 + \pi^2) + O(h^8) \\
&= h^6 \mathcal{E}\left(k, k\boldsymbol{\theta}, c_2\right) + O(h^8).
\end{aligned}$$

We first determine the constant c_2 by minimizing the asymptotic dispersion error. We thus set

$$c_2^* = \arg\min_{c_2} \left|\mathcal{E}\left(k, k\boldsymbol{\theta}, c_2\right)\right|.$$

Theorem 4.2. *The asymptotically optimal shift for the 9-point stencil and associated reduction factor are*

$$c_2^* = -\frac{\pi^2}{54}, \quad k_6^{\mathrm{asy}} = -\frac{k^7}{12288}, \quad \text{and} \quad \mathrm{R_f} = 64.$$

Proof. We introduce the function F such that

$$\mathcal{E}\left(k, k\boldsymbol{\theta}, c_2\right) = -\frac{k^8}{6048\pi^2}F(\cos(\theta), c_2). \tag{4.3}$$

Setting $X = \cos(\theta) \in [-1, +1]$, we then have to find first the extrema of

$$F(X, c_2) = 2\pi^2X^8 - 4\pi^2X^6 + 6\pi^2X^4 + 189X^4c_2 - 4\pi^2X^2 - 189X^2c_2 + \pi^2.$$

Computing the solution to $F'(X_c, c_2) = 0$, we obtain

$$X_c \in \left\{0,\ \pm\frac{\sqrt{2}}{2},\ \pm\frac{\sqrt{2}}{2\pi}\left(\pi^2 \mp \sqrt{-3\pi^4 - 189\pi^2c_2}\right)\right\}.$$

Assuming that $c_2 < -\pi^2/63$ to avoid complex square roots, we get

$$F(X_c, c_2) \in \left\{\pi^2,\ \frac{\pi^2}{8} - \frac{189}{4}c_2,\ \frac{-1}{8\pi^2}\left(8\pi^4 + 1512\pi^2c_2 + 35721c_2^2\right)\right\}.$$

Studying the variation of these functions (or simply plotting them) for $c_2 \leq -\pi^2/63$, we obtain

$$F_{\min}(c_2) = \frac{-1}{8\pi^2}\left(8\pi^4 + 1512\pi^2c_2 + 35721c_2^2\right) \leq F(X, c_2) \leq F_{\max}(c_2),$$

where

$$F_{\max}(c_2) = \max\left\{\pi^2, \frac{\pi^2}{8} - \frac{189}{4}c_2\right\}.$$

Using (4.3), we then get

$$-\frac{k^8}{6048\pi^2}F_{\max}(c_2) \leq \mathcal{E}\left(k, k\boldsymbol{\theta}, c_2\right) \leq -\frac{k^8}{6048\pi^2}F_{\min}(c_2),$$

from which we finally get

$$\mathcal{E}_{\min}(c_2) = -\frac{k^8}{6048\pi^2}F_{\max}(c_2),\ \mathcal{E}_{\max}(c_2) = -\frac{k^8}{6048\pi^2}F_{\min}(c_2).$$

From these estimates, we also obtain

$$\max_{\boldsymbol{\theta}\in\mathcal{S}^1}\left|\mathcal{E}\left(k, k\boldsymbol{\theta}, c_2\right)\right| = \frac{k^8}{6048\pi^2}\max\left\{\left|F_{\max}(c_2)\right|, \left|F_{\min}(c_2)\right|\right\} := K(c_2),$$

and $c_2^* = \arg\min K(c_2)$. From Figure 4.2, we see that $K(c_2) = |F_{\max}(c_2)|$ and thus

$$c_2^* = [-\frac{\pi^2}{54}, -\frac{\pi^2}{63}].$$

To get a single value for c_2^*, we maximize the reduction factor (see (3.2)) which, for $c_2 \in [-\frac{\pi^2}{54}, -\frac{\pi^2}{63}]$ is given by

$$\mathrm{R_f}(c_2) = 2\frac{\max\left\{\left|\mathcal{E}_{\max}\right|, \left|\mathcal{E}_{\min}\right|\right\}}{\left|\mathcal{E}_{\max} - \mathcal{E}_{\min}\right|} = \frac{16\pi^2}{16\pi^4 + 1512\pi^2c_2 + 35721c_2^2}.$$

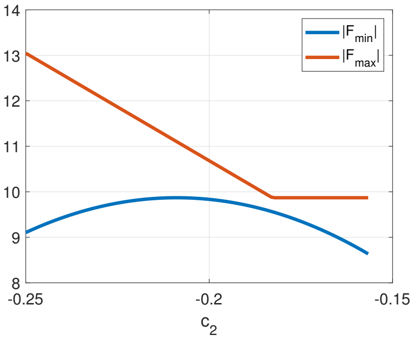

FIG. 4.2. *Graph of* $K(c_2)$.

Since the function $c_2 \mapsto \mathrm{R_f}(c_2)$ is decreasing on $[-\frac{\pi^2}{54}, -\frac{\pi^2}{63}]$, it reaches its maximum at $c_2 = -\frac{\pi^2}{54}$ and we thus set

$$c_2^* := -\frac{\pi^2}{54}.$$

This gives

$$\mathrm{R_f}(c_2^*) = 64$$

as well as

$$\mathcal{E}_{\min} = \mathcal{E}_{\min}(c_2^*) = -\frac{k^8}{6048}, \quad \mathcal{E}_{\max} = \mathcal{E}_{\max}(c_2^*) = -\frac{k^8}{6048}\frac{31}{32}.$$

Theorem 3.3 finally gives that the asymptotically optimal shift is

$$k_6^{\mathrm{asy}} = \frac{1}{4k}\left(\mathcal{E}_{\min} + \mathcal{E}_{\max}\right) = -\frac{k^7}{12288}.$$

□

We show in Figure 4.3 the relative error between the asymptotically optimal shift and the numerically optimized one where the optimization has been performed as in Section 4.1. From these numerical results, we see that the asymptotically optimal shift is close to the numerically optimized one up to $G \geq 5$ and that the relative dispersion error is also reduced even for a small number of grid points per wavelength.

4.3. Application to a 7-point stencil in 3d. The second order 7-point stencil for the Helmholtz operator in 3d is defined as

$$\begin{aligned}\left(\mathcal{H}_h^{7\mathrm{pt}} u\right)_{i,j,k} &= \frac{-u_{i+1,j,k} - u_{i-1,j,k} - u_{i,j,k+1} - u_{i,j,k-1} - u_{i,j+1,k} - u_{i,j-1,k} + 6u_{i,j,k}}{h^2} \\ &- k^2 u_{i,j,k},\end{aligned}$$

and the discrete symbol is thus

$$\sigma_d^{7\mathrm{pt}}(k, \boldsymbol{\xi}, h) = \frac{6 - 2(\cos(h\xi_1) + \cos(h\xi_2) + \cos(h\xi_3))}{h^2} - k^2.$$

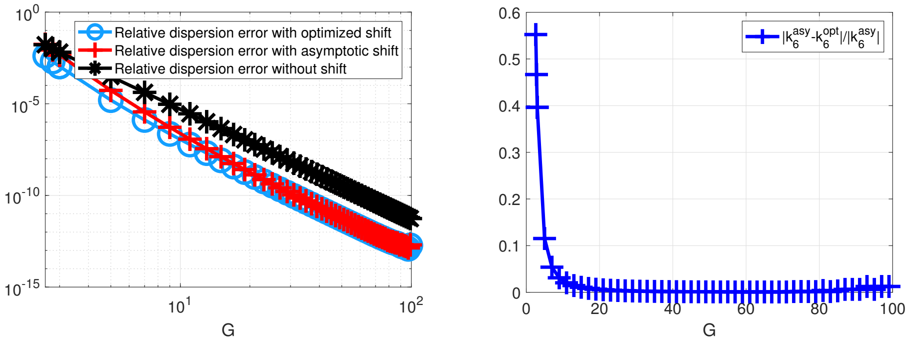

FIG. 4.3. *Left: Log-log plot of the relative dispersion error for the 9-point FD stencil using either $\widetilde{k} = k$ (no dispersion correction), $\widetilde{k} = k + h^6 k_6^{\mathrm{opt}}$ or $\widetilde{k} = k + h^6 k_6^{\mathrm{asy}}$. Right: $\frac{|k_6^{\mathrm{asy}} - k_6^{\mathrm{opt}}|}{|k_6^{\mathrm{asy}}|}$ as a function of G. We used $\mathcal{K} = \{20, 40, 80, 100, 140, 160, 180, 200, 250, 300, 600\}$.*

Since any $\boldsymbol{\theta} \in \mathcal{S}^2$ can be written as $\boldsymbol{\theta} = (\cos(\varphi)\sin(s), \sin(\varphi)\sin(s), \cos(s))$ for $\varphi \in [0, 2\pi]$ and $s \in [0, \pi]$, a Taylor expansion gives

$$\mathcal{E}(k, k\boldsymbol{\theta}) = -\frac{k^4}{12}\left((\cos(\varphi)\sin(s))^4 + (\sin(\varphi)\sin(s))^4 + \cos(s)^4\right),$$

and the asymptotically optimal shift can then be computed.

THEOREM 4.3. *For the 7-point finite difference scheme in 3d, we have the asymptotically optimal shift and related reduction factor*

$$k_2^{\mathrm{asy}} = -\frac{k^3}{36}, \quad \mathrm{R_f} = 3.$$

Proof. Setting $X = \cos(s)^2$ and $Y = \cos(\varphi)^2$, we get

$$\mathcal{E}(k, k\boldsymbol{\theta}) = -\frac{k^4}{12}\left(X^2 + (1-X)^2\left(Y^2 + (1-Y)^2\right)\right) = -\frac{k^4}{12} f(X, Y),$$

and we now have to compute the extrema of f over $[0,1]^2$. A computation gives $\partial_X f(X, Y) = X(2 + g(Y)) - 2g(Y)$ with $g(Y) = Y^2 + (1-Y)^2$. Since $g(Y) \geq 1/2$, $2 + 2g(Y) > 0$, for any fixed Y, the function $X \in [0,1] \mapsto f(X, Y)$ is decreasing for $0 \leq X \leq 2g(Y)/(2 + 2g(Y))$ and increasing otherwise. As a result, we have

$$\forall (X, Y) \in [0,1]^2: \ f\left(\frac{2g(Y)}{2 + 2g(Y)}, Y\right) \leq f(X, Y) \leq \max\left\{f(1, Y), f(0, Y)\right\} \leq 1.$$

Noting then that

$$f\left(\frac{2g(Y)}{2 + 2g(Y)}, Y\right) = \frac{1}{2}\frac{2Y^2 - 2Y + 1}{Y^2 - Y + 1} \geq \frac{1}{2}\frac{1/2}{3/4} = \frac{1}{3},$$

we obtain

$$\frac{1}{3} = f\left(\frac{1}{2}, \frac{1}{3}\right) \leq f(X, Y) \leq f(1, 1) = 1,$$

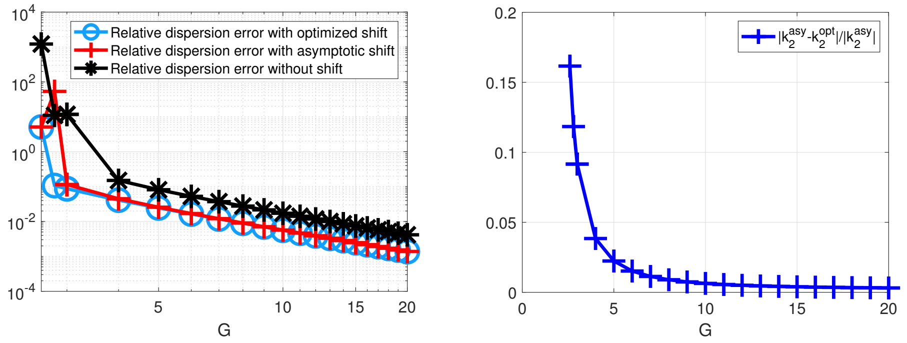

FIG. 4.4. *Left: Log-log plot of the relative dispersion error for the 7-point FD stencil using* $\widetilde{k} = k$ *(no dispersion correction),* $\widetilde{k} = k + h^2 k_2^{\rm opt}$ *or* $\widetilde{k} = k + h^2 k_2^{\rm asy}$*. Right:* $\frac{|k_2^{\rm asy} - k_2^{\rm opt}|}{|k_2^{\rm asy}|}$ *as a function of* G*. We used* $\mathcal{K} = \{20, 40, 80, 100, 140, 160, 180, 200, 250, 300, 600\}$.

and thus

$$\mathcal{E}_{\min} = -\frac{k^4}{12}, \ \mathcal{E}_{\max} = -\frac{k^4}{36}.$$

Using Theorem 3.3, we find $k_2^{\rm asy} = \frac{\mathcal{E}_{\min} + \mathcal{E}_{\max}}{4k} = -\frac{k^3}{36}$.
From (3.2), the reduction factor is

$$\mathrm{R_f} = 2\frac{\max\{|\mathcal{E}_{\max}|, |\mathcal{E}_{\min}|\}}{|\mathcal{E}_{\max} - \mathcal{E}_{\min}|} = 3.$$

□

The 7-point finite difference stencil with shifted wavenumber is therefore

$$\begin{aligned}\left(\hat{\mathcal{H}}_h^{7\rm pt} u\right)_{i,j,k} &= \frac{-u_{i+1,j,k} - u_{i-1,j,k} - u_{i,j,k+1} - u_{i,j,k-1} - u_{i,j+1,k} - u_{i,j-1,k} + 6u_{i,j,k}}{h^2} \\ &- \left(k - \frac{k^3 h^2}{36}\right)^2 u_{i,j,k}.\end{aligned}$$

We show in Figure 4.4 the relative dispersion error $\mathrm{Err}_{\rm disp}(\widetilde{k})$, for $\widetilde{k} \in \{k, \widehat{k}^{\rm asy}, \widehat{k}^{\rm opt}\}$ as well as the relative error between $k_2^{\rm asy}$ and the numerically optimized shift computed as in Section 4.1. This shows that the asymptotically optimal shift is close to the numerically optimized one even for a relatively small number of grid points per wavelength, and that both reduce the relative dispersion error compared to the standard 7-point FD stencil.

5. Numerical experiments. We now test the asymptotically optimal shift numerically to see the effect of dispersion correction when solving some Helmholtz boundary value problems. We start with the 3-point stencil in 1d, followed by the 5-point stencil in 2d, to solve Helmholtz problems with Robin boundary conditions. Then, we test the 9-point stencil with Dirichlet boundary conditions.

5.1. Numerical results for 1d problems. We illustrate now by numerical experiments that, for one dimensional Helmholtz problems, no dispersion error leads

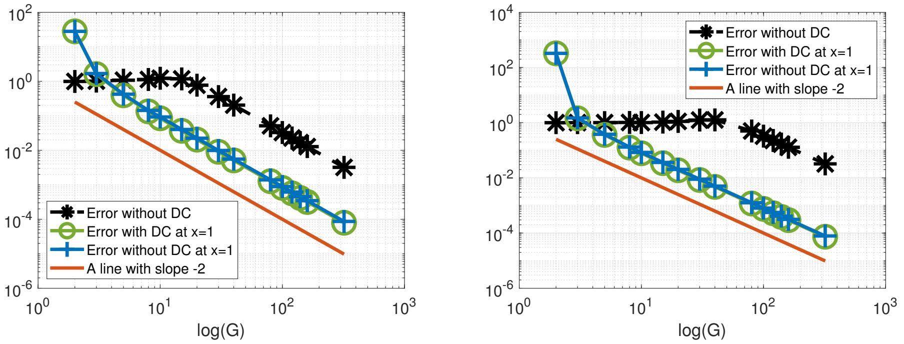

FIG. 5.1. *Log-log plots of the errors for Robin boundary conditions. Left:* $k = 200$. *Right:* $k = 2000$.

to no pollution effect as well. The Helmholtz equation with homogeneous Robin boundary condition at $x = 1$ is

$$\left\{ \begin{array}{rcl} -u''(x) - k^2 u(x) & = & f(x), \quad \text{in }]0,1[, \\ u(0) & = & 0, \\ u'(1) - \mathrm{i}ku(1) & = & 0, \end{array} \right. \tag{5.1}$$

where f is a given source term. We discretize (5.1) with the 3-point stencil (2.2) at n interior grid points, and use a ghost point for the Robin boundary condition. We assume the right-hand-side f to be

$$f = \sin(kx),$$

which gives the closed form solution

$$u_{\mathrm{ex},R}(x) = -\frac{x\cos(kx)}{2k} + \sin(kx)\left(\frac{1 + 2\mathrm{e}^{2\mathrm{i}k} - 2\mathrm{i}k}{4k^2}\right).$$

Denoting by $\boldsymbol{u}$ the discrete solution, we compare the relative errors

$$\mathrm{err} := \frac{\|\boldsymbol{u} - u_{\mathrm{ex},R}(\boldsymbol{x})\|_\infty}{\|u_{\mathrm{ex},R}(\boldsymbol{x})\|_\infty}$$

for the scheme (2.2) with and without the real shift $\widehat{k}$. For the Robin boundary condition, we also compute the error when using the real shift $\widehat{k}$ on the boundary. We compute the error for k fixed and a number of grid points per wavelength given by

$$G := \frac{2\pi}{kh} = 320,\ 160,\ 140,\ 120,\ 100,\ 80,\ 40,\ 30,\ 20,\ 15,\ 10,\ 8,\ 5,\ 3,\ 2.$$

The results are shown in Figure 5.1 and clearly show the pollution effect when no dispersion correction is used which confirms our theoretical results from Section 2 (see Theorem 2.3). It is also worth noting that using the real shift on the Robin condition does not have a major impact on the error.

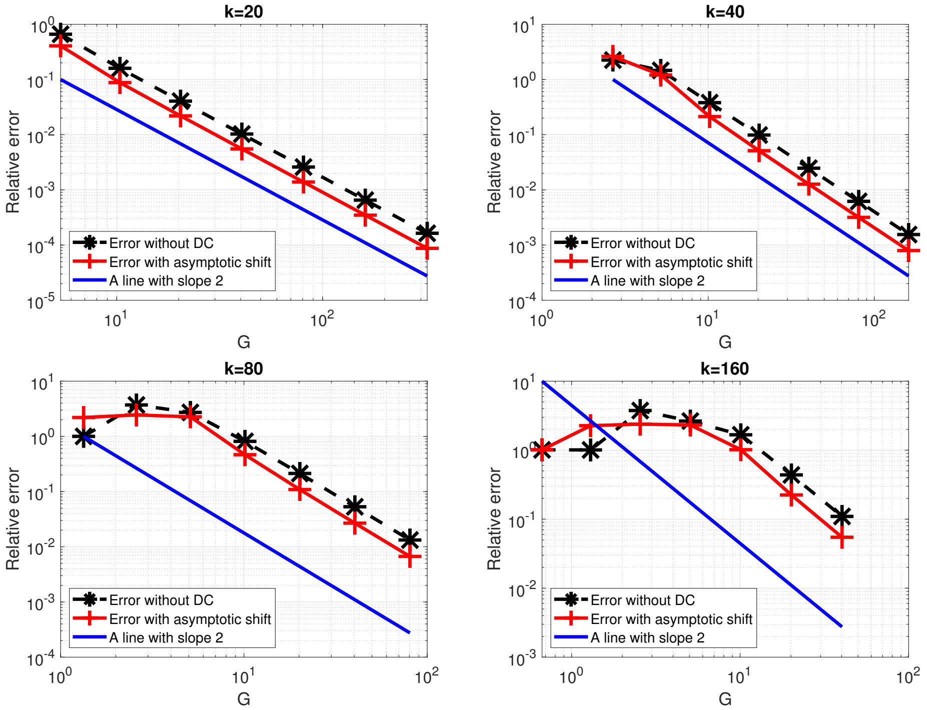

FIG. 5.2. *Relative error with and without shifted wavenumber.*

5.2. Numerical experiments in 2d with the 5-point stencil. We solve the test problem (see also [33, p. 22, Example 2])

$$\left\{\begin{array}{rll} -\Delta u-k^2u &= 0, & \text{in } \Omega=(0,1)^2, \\ u &= f, & \text{on } \{0\}\times[0,1]\cup\{1\}\times[0,1], \\ \partial_n u+\mathrm{i}ku &= g, & \text{on } [0,1]\times\{0\}\cup[0,1]\times\{1\}, \end{array}\right. \tag{5.2}$$

where f, g are defined so that $u(x,y)=\sin(k(x+y)\sqrt{2}/2)$ is the exact solution.

We discretize the Robin boundary condition with a ghost point to achieve second order accuracy and also use the shifted wavenumber in the Robin condition. To compare the efficiency of the 5-point FD scheme with shifted wavenumber, we compute

$$\mathrm{err}_\infty(\widetilde{k})=\frac{\left\|\boldsymbol{u}-u_h(\widetilde{k})\right\|}{\|u\|_\infty},$$

where $u_h(\widetilde{k})$ is the numerical solution without shift (hence $\widetilde{k}=k$) or using the shifted wavenumber (in that case $\widetilde{k}=\widehat{k}=k-h^2k^3/32$).

We compute numerically (see Figure 5.2) the relative error for meshsizes $h=1/(n+1)$ with $n=2^j$ and $j=4,\cdots,10$, and $k=20,40,80,160$. These results show that the shifted wavenumber can not cancel the pollution effect in 2d, but it reduces the error for large enough number of grid points per wavelength. We also compute the reduction factors in Table 5.1. This shows that the shift roughly reduces the relative error by a factor 2 for large enough numbers of grid points per wavelength. Note that it is not beneficial to use the asymptotically optimal shift if too few grid

n	2^4	2^5	2^6	2^7	2^8	2^9	2^{10}
$k = 20$	0.6153	0.5514	0.5406	0.5376	0.5353	0.5342	0.5335
$k = 40$	1.1662	0.8251	0.5594	0.5190	0.5136	0.5126	0.5123
$k = 80$	2.1981	0.6594	0.8327	0.5723	0.5118	0.5052	0.5045
$k = 160$	1.0015	2.2481	0.6354	0.8776	0.6078	0.5119	0.4988

TABLE 5.1

Ratio of the errors $\mathrm{err}_\infty\left(\widehat{k}\right)/\mathrm{err}_\infty(k)$ *for varying meshsize and wavenumber.*

points per wavelength are used. This is expected, since we define k^{asy} by minimizing the dispersion error as the meshsize goes to zero. Nevertheless, since the dispersion error is still reduced when using the asymptotically optimal shift (see Figure 4.1) for $G \geq 5$, we can still expect to reduce the relative error when enough grid points per wavelength are used to get accurate solutions. Also note again that the numerical cost for solving the linear system with or without shift is identical.

5.3. Numerical experiments in 2d with the 9-point stencil. We now solve the test problem

$$\begin{cases} -\Delta u - k^2 u &= 0, \quad \text{in } \Omega = (-1,1)^2, \\ \qquad\quad u &= f, \quad \text{on } \partial\Omega, \end{cases} \tag{5.3}$$

where f is chosen so that $u(x,y) = \sin(k(x+y)\sqrt{2}/2)$ is the exact solution. We solve (5.3) with the 9-point stencil (4.2) where the constants are

$$a = \frac{5}{6}, \quad b = \frac{5}{6} - \frac{c}{2}, \quad c = \frac{8}{45} - \frac{\pi^2}{54}G^{-2}, \quad k_g = k\left(1 - \frac{\pi^4}{30}G^{-4}\right),$$

for the FD scheme without dispersion correction. Since $G = 2\pi/(kh)$, these constants become with dispersion correction

$$\begin{aligned} c &= \frac{8}{45} - \frac{\pi^2}{54}\left(\frac{2\pi}{(k + h^6 k_6^{\text{asy}})h}\right)^{-2}, \\ k_g^{\text{asy}} &= (k + h^6 k_6^{\text{asy}})\left(1 - \frac{\pi^4}{30}\left(\frac{2\pi}{(k + h^6 k_6^{\text{asy}})h}\right)^{-4}\right). \end{aligned} \tag{5.4}$$

The value for k_6^{asy} is defined in Theorem 4.2. We emphasize that the constants from (5.4) satisfy as $h \to 0$ the expansions

$$c^{\text{asy}} = \frac{8}{45} - \frac{\pi^2}{54}G^{-2} + O(h^7), \; k_g^{\text{asy}} = k\left(1 - \frac{\pi^4}{30}G^{-4} - \frac{\pi^6}{192}G^{-6}\right) + O(h^7),$$

where we can use either G or h. Since the 9-point stencil is sixth-order accurate, only the expansion up to order 6 is going to matter for the error and we then neglect the $O(h^7)$ term above to define our FD scheme with dispersion correction. Using these, we now have the same stencil as the one obtained in [4, Theorem 4.1] by minimizing the distance between the discrete and continuous dispersion relations thanks to an asymptotic analysis. Our new approach is however much easier to use, and also to extend to other FD stencils.

Using the same notations as in Subsection 5.2, we show in Figure 5.3 the evolution of the relative error, for a fixed wavenumber, and varying meshsize, and in Table 5.2 the reduction factor. From these results, one can see that, for small enough meshsize, it is highly beneficial to use dispersion correction since it can lower the relative error by a factor up to 100 in some cases.

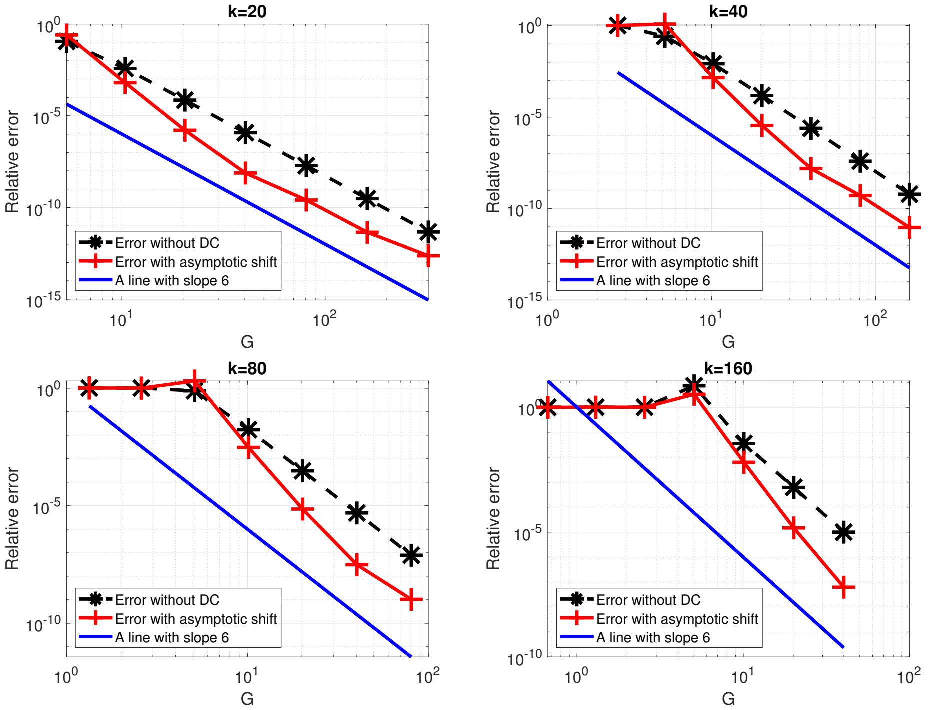

FIG. 5.3. *Relative error with and without shifted wavenumber.*

n	2^4	2^5	2^6	2^7	2^8	2^9	2^{10}
$k = 20$	2.231	0.167	0.022	0.006	0.013	0.014	0.050
$k = 40$	0.994	4.647	0.174	0.023	0.006	0.013	0.015
$k = 80$	0.999	0.997	2.627	0.177	0.023	0.006	0.013
$k = 160$	0.999	0.999	0.998	0.458	0.178	0.023	0.006

TABLE 5.2

Ratio of the errors $\mathrm{err}_\infty\left(\widehat{k}\right)/\mathrm{err}_\infty(k)$ *for varying meshsize and wavenumber.*

6. Conclusions and outlook.

We introduced a new dispersion correction technique for finite difference discretizations of Helmholtz problems. The technique is based on a modified wavenumber, which can be obtained in closed form for arbitrary finite difference discretizations by obtaining the extrema of an associated function defined on a compact set. This function is simply obtained from the Taylor expansion of the discrete symbol of the FD stencil considered. We applied our method to several standard stencils from the literature and our numerical experiments show that, for small enough meshsize, reducing the dispersion error also reduces the relative error in the solution.

A next step is to extend our new technique to finite element discretizations, where dispersion correction is more difficult to achieve. Our technique can also be extended to other time-harmonic wave propagation problems like for instance electromagnetic waves modeled by the Maxwell system, linear elasticity, or even linearized water-wave models (e.g. Serre-Green-Nagdhi or Nwogu equations). It might also be possible to derive asymptotically optimal shifts for finite-difference methods in the time-domain (FDTD).

Acknowledgements: We thanks an anonymous referee who helped us to greatly improve our paper. This research was supported by the Swiss National Science Foundation.

Appendix A. Asymptotically optimal shift as $G \to +\infty$.

We discuss in this appendix the extension of some results from Section 3 as $G \to +\infty$ instead of $h \to 0$ and thus the O appearing below have to be understood as $G \to +\infty$. We are also going to track the dependence with respect to the wavenumber k and thus denote by O_k a O that may depend on k. First of all, we need to replace assumptions $(H1) - (H2) - (H3)$ by the new assumptions

(H1)$'$ The discrete symbol admits the expansion

$$\sigma_d(k, k\boldsymbol{\theta}, G) = k^2 G^{-p} \mathcal{H}(\boldsymbol{\theta}) + k^2 O(G^{-p-1}),$$

for a smooth function $\mathcal{H}$.

(H2)$'$ For a given wavenumber k, the sequence of functions $(\nabla_{\boldsymbol{\xi}} \sigma_d(k, \cdot, G))_G$ converges uniformly to $\nabla_{\boldsymbol{\xi}} \sigma_c(k, \cdot)$ on a compact neighborhood of $\boldsymbol{\xi} = k\boldsymbol{\theta}$ for $\boldsymbol{\theta} \in \mathcal{S}^{d-1}$.

In addition, the derivative of the discrete symbol with respect to ξ verifies

$$\nabla_{\xi} \sigma_d(k, k\boldsymbol{\theta}, G) = \underbrace{\nabla_{\xi} \sigma(k, k\boldsymbol{\theta})}_{=2k\boldsymbol{\theta}} + kO(G^{-p}).$$

(H3)$'$ The derivative of the discrete symbol with respect to its first variable k satisfies

$$\partial_k \sigma_d(k, k\boldsymbol{\theta}, G) = \underbrace{\partial_k \sigma(k, k\boldsymbol{\theta})}_{=-2k} + O(G^{-1}).$$

It is worth noting that $(H1)' - (H2)' - (H3)'$ are satisfied at least for the stencils considered in this paper. Using the above assumptions, we can compute an asymptotic expansion of k_d as $G \to +\infty$.

Theorem A.1. *Assume that $(H1)' - (H2)'$ hold. Then, the discrete wavenumber satisfies as $G \to +\infty$ the asymptotic expansion*

$$k_d(k, \boldsymbol{\theta}, G) = k - G^{-p} \frac{k}{2} \mathcal{H}(\boldsymbol{\theta}) + kO(G^{-p-1}) + kO_k(G^{-2p}).$$

Proof. A Taylor expansion gives

$$\sigma_d(k, k_d\boldsymbol{\theta}, h) - \sigma_d(k, k\boldsymbol{\theta}, h) = (k_d - k) \nabla_{\xi} \sigma_d(k, k\boldsymbol{\theta}, G) + O_k(|k_d - k|^2).$$

Using then (H1)$'$-(H2)$'$ and that $\sigma_d(k, k_d\boldsymbol{\theta}, h) = 0$, we obtain

$$-k^2 G^{-p} \mathcal{H}(\boldsymbol{\theta}) + k^2 O(G^{-p-1}) = (k_d - k) \left(2k + kO(G^{-p})\right) + O_k(|k_d - k|^2), \quad \text{(A.1)}$$

from which we see that $(k_d - k) = kO_k(G^{-p})$. The equality (A.1) can then be recast as

$$\begin{aligned} (k_d - k) &= \frac{-k^2 G^{-p} \mathcal{H}(\boldsymbol{\theta}) + k^2 O(G^{-p-1}) + k^2 O_k(G^{-2p})}{k\left(2 + O(G^{-p})\right)} \\ &= -\frac{k}{2} G^{-p} \mathcal{H}(\boldsymbol{\theta}) + kO(G^{-p-1}) + kO_k(G^{-2p}), \end{aligned}$$

which concludes the proof. □

Below, we always use Theorem A.1 by keeping only terms up to order G^{-p-1} and then we neglect the $O_k(G^{-2p})$ term. We introduce the shifted wavenumber $\widehat{k}$ defined as

$$\widehat{k} := k + G^{-p} k_p.$$

Using $(H3)'$ and Theorem A.1, one can show that the discrete wavenumber associated to the stencil using $\widehat{k}$ instead of k satisfies the expansion

$$k_d(k, \boldsymbol{\theta}, G) = k - G^{-p} \frac{k}{2} \left(-2 \frac{k_p}{k} + \mathcal{H}(\boldsymbol{\theta}) \right) + k O(G^{-p-1}).$$

The asymptotically optimal shift can thus be defined as

$$k_p^{\text{asy}} := \arg\min_{k_p} \left(\max_{\boldsymbol{\theta} \in \mathcal{S}^{d-1}} \left| -2 \frac{k_p}{k} + \mathcal{H}(\boldsymbol{\theta}) \right| \right).$$

Assuming that

$$\forall \boldsymbol{\theta} \in \mathcal{S}^{d-1} : \ \mathcal{H}_{\min} \leq \mathcal{H}(\boldsymbol{\theta}) \leq \mathcal{H}_{\max},$$

we can follow the proof of Theorem 3.3 to obtain an explicit formula for the asymptotically optimal shift, namely

$$k_p^{\text{asy}} = \frac{k}{4} \left(\mathcal{H}_{\min} + \mathcal{H}_{\max} \right).$$

In addition, the relative dispersion error when using the shift satisfies

$$\max_{\boldsymbol{\theta} \in \mathcal{S}^{d-1}} \left| \frac{k_d(\widehat{k}^{\text{asy}}, \boldsymbol{\theta}, G) - k}{k} \right| = G^{-p} \left| \frac{\mathcal{H}_{\max} - \mathcal{H}_{\min}}{2} \right| + O(G^{-p-1}), \qquad \text{(A.2)}$$

where $\widehat{k}^{\text{asy}} = k + k_p^{\text{asy}} G^{-p}$. Since the relative dispersion error without shift verifies

$$\max_{\boldsymbol{\theta} \in \mathcal{S}^{d-1}} \left| \frac{k_d(k, \boldsymbol{\theta}, G) - k}{k} \right| = G^{-p} \max \left\{ |\mathcal{H}_{\max}|, |\mathcal{H}_{\max}| \right\} + O(G^{-p-1}), \qquad \text{(A.3)}$$

we can define the reduction factor $\mathrm{R_f}$ as in Eq. (3.2),

$$\mathrm{R_f} := 2 \frac{\max \left\{ |\mathcal{H}_{\max}|, |\mathcal{H}_{\min}| \right\}}{|\mathcal{H}_{\max} - \mathcal{H}_{\min}|}.$$

To conclude this appendix, it is worth noting that if $(H1)' - (H2)' - (H3)'$ hold then the estimates (A.2) and (A.3) are valid where terms higher than $O(G^{-p-1})$ do not appear and may actually depend on k. Therefore, at least up to order G^{-p-1}, the relative dispersion error behaves like G^{-p} for large G and thus keeping the number of grid point fixed yields a relative dispersion error that is independent of the wavenumber. We emphasize that this claim can be observed numerically.

REFERENCES

[1] Babuska, I. M., & Sauter, S. A. (1997). Is the pollution effect of the FEM avoidable for the Helmholtz equation considering high wavenumbers?. SIAM Journal on numerical analysis, 34(6), 2392-2423.

[2] Cocquet, P. H., Gander, M. J., & Xiang, X. (2018). A finite difference method with optimized dispersion correction for the Helmholtz equation, Domain Decomposition Methods in Science and Engineering XXIV, LNCSE, Springer-Verlag.

[3] Cocquet, P. H., Gander, M. J., & Xiang, X., Dispersion correction for Helmholtz in 1D with piecewise constant wavenumber, accepted for Domain Decomposition Methods in Science and Engineering XXV, LNCSE, Springer-Verlag, 2019.

[4] Cocquet, P. H., Gander, M. J., & Xiang, X. (2021). Closed form dispersion corrections including a real shifted wavenumber for finite difference discretizations of 2d constant coefficient helmholtz problems. SIAM Journal on Scientific Computing, 43(1), A278-A308.

[5] Chen, Z., Cheng, D., & Wu, T. (2012). A dispersion minimizing finite difference scheme and preconditioned solver for the 3D Helmholtz equation. Journal of Computational Physics, 231(24), 8152-8175.

[6] Cheng, D., Tan, X., & Zeng, T. (2017). A dispersion minimizing finite difference scheme for the Helmholtz equation based on point-weighting. Computers & Mathematics with Applications, 73(11), 2345-2359.

[7] Ernst, O. G., & Gander, M. J. (2013). Multigrid methods for Helmholtz problems: A convergent scheme in 1D using standard components. Direct and Inverse Problems in Wave Propagation and Applications, 14.

[8] Feng, Q., Han, B., & Michelle, M. (2021). Sixth Order Compact Finite Difference Method for 2D Helmholtz Equations with Singular Sources and Reduced Pollution Effect. arXiv preprint arXiv:2112.07154.

[9] Da Fonseca, C. M., & Petronilho, J. (2001). Explicit inverses of some tridiagonal matrices. Linear Algebra and its Applications, 325(1-3), 7-21.

[10] Dastour, H., & Liao, W. (2021). An optimal 13-point finite difference scheme for a 2D Helmholtz equation with a perfectly matched layer boundary condition. Numerical Algorithms, 86(3), 1109-1141.

[11] Dastour, H., & Liao, W. (2021). A generalized optimal fourth-order finite difference scheme for a 2D Helmholtz equation with the perfectly matched layer boundary condition. Journal of Computational and Applied Mathematics, 394, 113544.

[12] Deng, Q., & Ern, A. (2021). SoftFEM: revisiting the spectral finite element approximation of second-order elliptic operators. Computers & Mathematics with Applications, 101, 119-133.

[13] Dwarka, V., & Vuik, C. (2021). Pollution and accuracy of solutions of the Helmholtz equation: A novel perspective from the eigenvalues. Journal of Computational and Applied Mathematics, 395, 113549.

[14] Ernst, O. G., & Gander, M. J. (2012). Why it is difficult to solve Helmholtz problems with classical iterative methods. Numerical analysis of multiscale problems, 325-363.

[15] Han, B., Michelle, M., & Wong, Y. S. (2021). Dirac assisted tree method for 1D heterogeneous Helmholtz equations with arbitrary variable wavenumbers. Computers & Mathematics with Applications, 97, 416-438.

[16] Franca, L. P., Farhat, C., Macedo, A. P., & Lesoinne, M. (1997). Residual?free bubbles for the Helmholtz equation. International journal for numerical methods in engineering, 40(21), 4003-4009.

[17] Ihlenburg, F., & Babuska, I. (1995). Finite element solution of the Helmholtz equation with high wavenumber Part I: The h-version of the FEM. Computers & Mathematics with Applications, 30(9), 9-37.

[18] Kaya, A., & Freitag, M. A. (2022). Conditioning analysis for discrete Helmholtz problems. Computers & Mathematics with Applications, 118, 171-182.

[19] Lafontaine, D., Spence, E. A., & Wunsch, J. (2022). Wavenumber-explicit convergence of the hp-FEM for the full-space heterogeneous Helmholtz equation with smooth coefficients. Computers & Mathematics with Applications, 113, 59-69.

[20] Melenk, J. M., & Sauter, S. (2011). Wavenumber explicit convergence analysis for Galerkin discretizations of the Helmholtz equation. SIAM Journal on Numerical Analysis, 49(3), 1210-1243.

[21] Melenk, J., & Sauter, S. (2010). Convergence analysis for finite element discretizations of the Helmholtz equation with Dirichlet-to-Neumann boundary conditions. Mathematics of Computation, 79(272), 1871-1914.

[22] Melenk, J. M., Parsania, A., & Sauter, S. (2013). General DG-methods for highly indefinite Helmholtz problems. Journal of Scientific Computing, 57(3), 536-581.

[23] Singer, I., & Turkel, E. (1998). High-order finite difference methods for the Helmholtz equation.

Computer methods in applied mechanics and engineering, 163(1-4), 343-358.

[24] Spence, E. A. (2022). A simple proof that the *hp*-FEM does not suffer from the pollution effect for the constant-coefficient full-space Helmholtz equation. arXiv preprint arXiv:2202.06939.

[25] Stolk, C. C. (2016). A dispersion minimizing scheme for the 3-D Helmholtz equation based on ray theory. Journal of computational Physics, 314, 618-646.

[26] Stolk, C. C., Ahmed, M., & Bhowmik, S. K. (2014). A multigrid method for the Helmholtz equation with optimized coarse grid corrections. SIAM Journal on Scientific Computing, 36(6), A2819-A2841.

[27] Wang, K., & Wong, Y. S. (2014). Pollution-free finite difference schemes for non-homogeneous Helmholtz equation. International Journal of Numerical Analysis & Modeling, 11(4).

[28] Wu, T., & Chen, Z. (2014). A dispersion minimizing subgridding finite difference scheme for the Helmholtz equation with PML. Journal of Computational and Applied Mathematics, 267, 82-95.

[29] Wu, T. (2017). A dispersion minimizing compact finite difference scheme for the 2D Helmholtz equation. Journal of Computational and Applied Mathematics, 311, 497-512.

[30] Wu, T., & Xu, R. (2018). An optimal compact sixth-order finite difference scheme for the Helmholtz equation. Computers & Mathematics with Applications, 75(7), 2520-2537.

[31] Zhu, L., Burman, E., & Wu, H. (2012). Continuous interior penalty finite element method for Helmholtz equation with high wavenumber: one dimensional analysis. arXiv preprint arXiv:1211.1424.

[32] Zhu, L., & Wu, H. (2013). Preasymptotic error analysis of CIP-FEM and FEM for Helmholtz equation with high wavenumber. Part II: hp version. SIAM Journal on Numerical Analysis, 51(3), 1828-1852.

[33] Zhou, Y., & Wu, H. (2022). Dispersion Analysis of CIP-FEM for Helmholtz Equation. arXiv preprint arXiv:2203.10813.