A study of the Numerical Dispersion for the Continuous Galerkin discretization of the one-dimensional Helmholtz equation

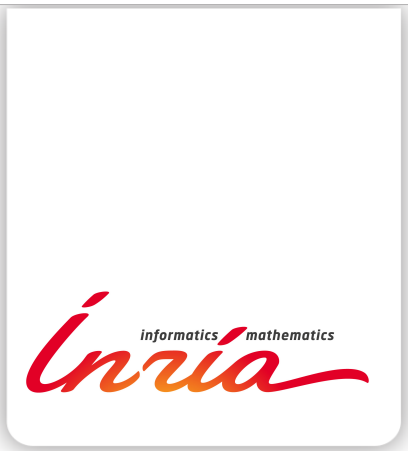
Hélène Barucq, Henri Calandra, Ha Pham, Sébastien Tordeux

A study of the Numerical Dispersion for the Continuous Galerkin discretization of the one-dimensional Helmholtz equation

Hélène Barucq, Henri Calandra, Ha Pham, Sébastien Tordeux

RESEARCH REPORT
N° 9075
June 2017
Project-Team Magique 3D

ISSN 0249-6399 ISRN INRIA/RR--9075--FR+ENG



A study of the Numerical Dispersion for the Continuous Galerkin discretization of the one-dimensional Helmholtz equation

Hélène Barucq*, Henri Calandra†, Ha Pham*, Sébastien Tordeux*

Project-Team Magique 3D

Research Report n° 9075 — June 2017 — 73 pages

Abstract:
Although true solutions of Helmholtz equation are non-dispersive, their discretizations suffer from a phenomenon called numerical dispersion. While the true phase velocity is constant, the numerical one changes with the discretization scheme, order and mesh size. In our work, we study the dispersion associated with classical finite element. For arbitrary order of discretization, without using an ansatz, we construct the numerical solution on the whole $\mathbb{R}$, and obtain an asymptotic expansion for the phase difference between the exact wavenumber and the numerical one. We follow an approach analogous to that employed in the construction of true solutions at positive wavenumbers, which involves z-transform, contour deformation and limiting absorption principle. This perspective allows us to identify the numerical wavenumber with the angle of analytic poles. Such an identification is useful since the latter (analytic poles) can be numerically evaluated by an algorithm based on [5, Theorem 2.4], which then yields the value of numerical wavenumber.

Key-words: z-transform, Numerical dispersion, Helmholtz equation, Finite element, Limiting absorption principle.

* Inria Magique3D
† Total E & P

RESEARCH CENTRE
BORDEAUX – SUD-OUEST

200 avenue de la Vieille Tour
33405 Talence Cedex

Une Étude de la Dispersion Numérique pour la Discrétisation Continue de Galerkine de l'équation de Helmholtz

Résumé :

Alors que les solutions analytiques de l'équation de Helmholtz sont non-dispersives, leurs discrétisations souffrent du phénomème de dispersion numérique. La vitesse de phase exacte est constante; cependant, sa représentation numérique varie selon la méthode de discrétisation, l'ordre et le maillage. Dans notre projet, nous étudions la dispersion associée à la méthode des éléments finis. Pour un ordre de discrétisation quelconque, sans utiliser d'Ansatz, nous construisons directement la solution discrétisée dans $\mathbb{R}$, et nous obtenons une dévéloppement asymptotique de la différence de phase. Nous suivons une approche comparable à celle employée pour la construction de solutions exactes pour un nombre d'onde positif, qui comprend la transformée en Z, la déformation des contours et le principe d'absorption limite. Cette perspective permet d'identifier le nombre d'onde numérique en fonction de l'angle des pôles analytiques. Une telle identification est utile, car ces derniers (pôles analytiques) peuvent être calculés en utilisant un algorithme basé de [5, Theorem 2.4], ce qui donne la valeur du nombre d'onde numérique.

Mots-clés : Transformée en Z, Dispersion numérique, Équation de Helmholtz, Élements Finis, Principe d'absorption limite.

1 Introduction

Dispersion phenomenon The basis for time-harmonic planewaves of frequency $\frac{\omega}{2\pi}$ traveling in a homogeneous medium with constant wavespeed $\mathbf{c}$ are given by $e^{i(\pm\frac{\omega}{\mathbf{c}}\,x-\omega t)}$. The spatial part $e^{\pm i\,\frac{\omega}{\mathbf{c}}\,x}$ are fundamental solutions of the Helmholtz equation at wavenumber κ

$$(-\Delta-\kappa^2)\,\mathbf{u}=0\,,$$

with κ satisfying the dispersion relation $\kappa=\frac{\omega}{\mathbf{c}}$. The phase velocity of the above planewaves is equal to the constant speed $\mathbf{c}$ of the medium, thus depends only on material properties and not ω. In this case, we say that exact solutions to the Helmholtz equation display no dispersion behavior.

On the other hand, the discretized solutions (e.g. by Finite Element or Finite Difference) are of the form $u_h=e^{-i(\kappa_h x-\omega t)}$, where the numerical wavenumber κ_h depends on ω, the discretization step size h and order of discretization r. As a result, the phase velocity of the discretized solution, given by $\frac{\omega}{\kappa_h(\omega,h,r)}$, also depends on these quantities, and is in general different from the constant speed $\mathbf{c}$. In this way, the numerical solution is said to display dispersive behavior. The above discussion describes the numerical dispersion for monochromatic waves (wave of a single frequency). For a polychromatic wave, i.e. one given by a superposition of waves of different frequencies, while the exact solution retains its shape during propagation, the discretized version will separate into a 'train of oscillations' [11], since each monochromatic component travels at a different phase velocity varying with its frequency.

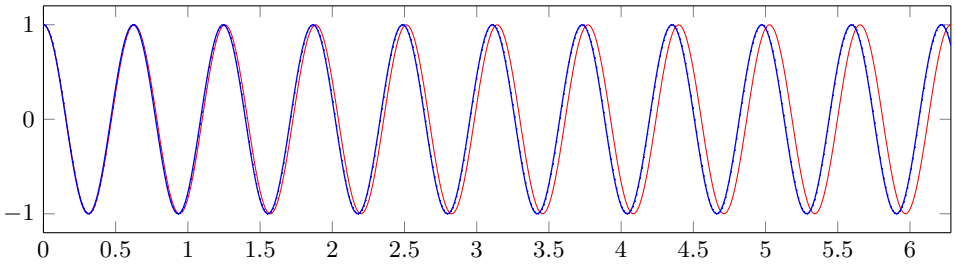


Figure 1: Dispersion associated with Finite Difference discretization of order 2 using planewave analysis with $h=0.05$ and $\frac{\omega}{c}=10$. —— = Re $\mathbf{u}_{\text{exact}}$ =cos(κx), —— Re $u_h=\cos(\kappa_h x)$ where analytic wavenumber $\kappa=10$, and numerical one $\kappa_h\sim 10.107$. The numerical wave with phase velocity $\sim 0.989\mathbf{c}$ lags behind the exact one (with numerical phase velocity $\mathbf{c}$).

For a simple yet informative example, we consider the numerical dispersion associated with Finite Difference of Order 2, c.f. Figure 1. The dispersion relation for the method can be obtained via planewave analysis, c.f. Appendix A, and is given by,

$$\kappa_h \;=\; \tfrac{2}{h}\,\arcsin\left(\tfrac{\kappa\,h}{2}\right) \quad , \quad \kappa:=\tfrac{\omega}{\mathbf{c}}.$$

With $\frac{\omega}{c}=10$, the analytic wavenumber $\kappa=10$. For discretization step size $h=0.05$, the numerical wavenumber has value

$$\kappa_h \;=\; \tfrac{2}{0.05}\,\arcsin\left(\tfrac{10\times 0.05}{2}\right) \;\sim\; 10.10721\,,$$

and the numerical phase velocity

$$\frac{\omega}{\kappa_h} \;=\; \mathbf{c}\frac{\omega/\mathbf{c}}{\kappa_h} \;=\; \mathbf{c}\frac{\kappa}{\kappa_h} \;=\; \mathbf{c}\frac{\kappa}{\frac{2}{h}\arcsin(\frac{\kappa\,h}{2})} \;=\; \mathbf{c}\,\frac{\frac{\kappa\,h}{2}}{\arcsin(\frac{\kappa h}{2})} \;\sim\; 0.98939\,\mathbf{c}.$$

This means that the numerical wave (with phase velocity $\sim 0.989\ \mathbf{c}$) lags behind the true solution, c.f. Figure 1. While this out-of-phaseness results in large error, c.f. Figure 2, the amplitudes of the two solutions agree and their qualitative behaviors resemble. This means that an analysis based only on error bound can give a misleading impression that the numerical solution is low-quality. In more complicated cases, numerical dispersion can cause the apparition of parasitic waves, which greatly affects the quality of the discretized solution, c.f. [3]. In short, in addition to the error analysis, an understand of the dispersion phenomenon, i.e. the difference between the analytical wavenumber and the numerical one, is essential. This understanding also serves as a guideline in the choice of the mesh size and order of approximation, c.f. [1].

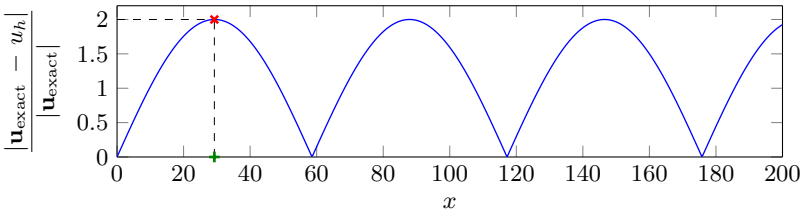


Figure 2: Error caused by dispersion associated with discretization by Finite Different of order 2. The graph shows the relative error between $\mathbf{u}_{\text{exact}} = e^{i\,\kappa\,x}$ and $u_h = e^{i\,\kappa_h\,x}$. The first maximum is indicated in red (×) and corresponds to 200% error.

Literature For a discussion of dispersion analysis, we refer to the introduction of [1] and the book of [3] and the references therein. Analytic results at low order discretizations can be obtained by planewave analysis or discrete Fourier analysis, c.f. [6, 10, 7]. For general orders, an upper bound of the phase difference was given by [8, Theorem 3.2], which states that, for discretization order $r \geq 1$, and if $\frac{h\,\kappa}{r} < 1$, on uniform mesh, the difference between the continuous wave number and the numerical one of the FEM solution is bounded above by,

$$\left|\kappa_h - \kappa\right| \quad \leq \quad \kappa\, C\, \left(\tfrac{e}{4}\right)^{2r} \tfrac{(\pi r)^{-1/2}}{4} \left(\tfrac{\kappa h}{2r}\right)^{2r} \quad . \tag{1}$$

Here, C is a constant not depending on κ, h and r. This upper bound is improved by [1, Theorem 3.2] which gives an asymptotic expansion of κ_h in terms of $\kappa\, h$ and r; in particular,

$$\cos(\kappa_h\, h) - \cos(\kappa\, h) = \tfrac{1}{2}\left[\tfrac{p!}{(2p)!}\right]^2 \tfrac{(\kappa\, h)^{2r+2}}{2p+1} + \mathsf{O}((\kappa h)^{2r+4})\,, \tag{2}$$

and for κh sufficiently small,

$$\kappa_h\, h - \kappa\, h = -\tfrac{1}{2}\left[\tfrac{p!}{(2p)!}\right]^2 \tfrac{(\kappa\, h)^{2r+1}}{2p+1} + \mathsf{O}((\kappa h)^{2r+3}. \tag{3}$$

This implies that

$$\frac{\kappa}{\kappa_h} - 1 = \mathsf{O}((\kappa h)^{2r}), \tag{4}$$

and that the bound (1) by [8, Theorem 3.2] is sharp.

Summary of results In our work, we also carry out an analytic study of the phase difference for Continuous Galerkin FEM for any order, however using a perspective different from that of

[1] and [8]. The results in [1, Theorem 3.2] were obtained by using a Bloch-wave Ansatz for the numerical solution. There are two folds to the novelty of our work: methodology and application for numerical evaluation.

- *Methodology* Instead of starting from an ansatz as in [1], we use an approach that is analogous to and inspired by the limiting absorbing principle used to construct exact solutions at positive wavenumber. The latter uses first Fourier transform to obtain solution at complex wavenumber, then contour deformation to perform analytic continuation, c.f. the discussion in Section 2. To construct the discretized solution (on the whole $\mathbb{R}$), we first use 'blocking' and Z-transform to transform the system of two-sided infinite recurrence relations generated by the discretization by FEM (on the whole $\mathbb{R}$) into one matrix-vector equation,
$$\mathcal{A}(\kappa^2 h^2, z)\, W(\kappa, z) \;=\; h\, z\, H(z).$$
For convenience, this equation is called the characteristic equation and its coefficient matrix the transfer matrix. The numerical solution at complex wavenumber can be obtained by inverse Z-transform and 'de-blocking'. Furthermore, we show that it can be written in the form of a complex integral, which undergoes contour deformation and limiting absorption principle to give rise to a numerical solution at positive wavenumber.

- *Analytic results* With contour deformation, we obtain explicit formula for the discrete Green function and the numerical solution, which allows identifying the numerical wavenumber κ_h with the angle of the (analytic) poles of the inverse of the transfer matrix. In particular, we show that all of the nonzero poles of $[\mathcal{A}(\kappa^2 h^2, z)]^{-1}$ are a pair of conjugate roots lying on the unit circle, denoted by $e^{i\,\pm\,\phi(\kappa h\,,\,r)}$, and that the numerical wave is given in terms of these poles by
$$\kappa_h = \frac{\phi(\kappa\, h\,,\, r)}{h}.$$
Here the phase ϕ is a function of κh and r, hence the notation $\phi(\kappa\, h\,,\, r)$, c.f. Section 6. For any discretization order r, we obtain an analytic expansion for $\cos\phi$, c.f. Proposition 6
$$\cos\phi \;=\; \cos(\kappa h) \;+\; (\kappa h)^2\; \mathsf{O}((\kappa h)^{2r}).$$
This implies dispersion relation 4 for small enough κh, and the bound
$$\frac{\kappa_h - \kappa}{\kappa} \;\leq\; C(r)\,(\kappa\, h)^{2r}.$$
This means that we reobtain the asymptotic expansion (4) and the bound (1). However, our result is weaker than [1, Theorem 3.2], in the sense that the coefficient of the highest order term is not specified, and we do not have knowledge of its sign. Our results only concern $\kappa h < \pi$, and not the high regime which was also studied in [1].

- *Application* Although we arrive at the same result regarding the order of the phase difference, the method we used offers a different perspective to characterize the numerical wavenumber. Identifying the numerical wavenumber κ_h in terms of the (analytic) poles of the inverse of the transfer matrix allows using an algorithm based on [5, Theorem 2.4] to calculate the poles, and hence κ_h, and obtain dispersion curves, for arbitrary order of discretization. For convenience, we will call this Guillaume's algorithm.

The remainder of the report is organized as follows. In Section 2, we review the limiting absorption principle used to obtain true solution at positive wavenumber; as mentioned, this

inspires our approach for the discretized problem. In Section 3, we introduce the discretized problem and its 'blocked' version. In Section 4, after performing Z-transform on the linear system, we study the structure of poles of the resulting characteristic equation. Also in the same section, we obtain the asymptotic expansion for $\cos\phi$. These results rely on technical details in Appendix C. In Section 5, contour deformation and limiting absorption principle are carried out to obtain the explicit form of the discretized solution, which allows the identification of the numerical wavenumber in terms of the phase ϕ. Calculation of numerical wavenumber using Guillaume's algorithm and resulting dispersion curve are in Section 6.

2 Technical motivation - limiting absorption principle for the continuous problem

The Laplacian Δ is an unbounded operator on $L^2(\mathbb{R}$ with domain $H^2(\mathbb{R})$. The spectrum of $-\Delta$ is purely continuous with

$$\boldsymbol{\sigma}(-\Delta) = \boldsymbol{\sigma}_{\text{continuous}}(-\Delta) = \mathbb{R}^+.$$

When $\sigma \notin \boldsymbol{\sigma}(-\Delta)$, the inverse of $-\Delta - \sigma$ exists and called a resolvent $\mathcal{R}(\sigma)$ of Δ.

$$(-\Delta - \sigma)^{-1} : L^2(\mathbb{R}) \to L^2(\mathbb{R}) \text{ bounded}\,,\ \sigma \in \mathbb{C} \setminus \mathbb{R}^+\,.$$

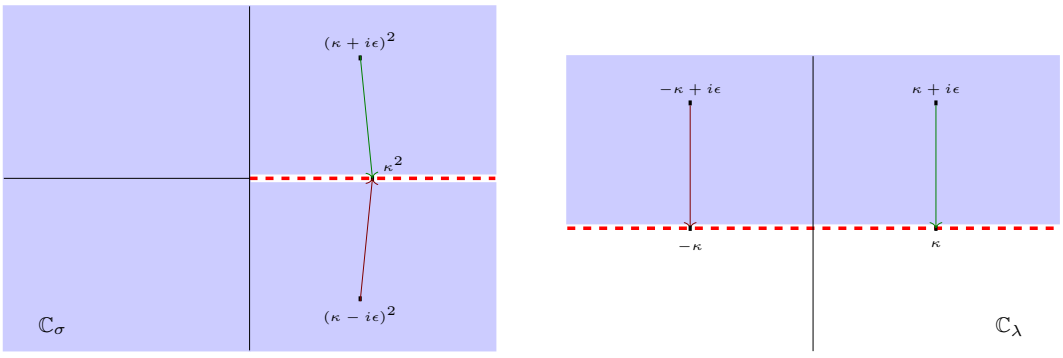
In this case, the variational form is coercive in $L^2(\mathbb{R})$, and the unique solution in $L^2(\mathbb{R})$ can be obtained by Fourier transform

$$(-\Delta - \sigma)\mathsf{u} = \mathsf{f} \text{ in } L^2(\mathbb{R}) \ \Rightarrow\ \mathsf{u} = \mathcal{F}^{-1}\frac{1}{|\xi|^2 - \sigma}\,\mathcal{F}\mathsf{f}\,.$$

When $\sigma = \kappa^2, \kappa \in \mathbb{R}^+$, we can construct a sequence w_n of 'almost eigenfunction' (called the Weyl seq.)

$$\|(-\Delta - \lambda)w_n\|_{L^2(\mathbb{R})} \to 0\ ;\ \|w_n\|_{L^2(\mathbb{R})} = 1.$$

This prevents the existence of a bounded inverse $(-\Delta - \sigma)^{-1}$ in $L^2(\mathbb{R})$ for $\sigma \in [0, +\infty)$.



(a) (b) Complex plane $\lambda = \sqrt{\sigma}$

Figure 3: Highlighted region $\mathbb{C} \setminus \mathbb{R}^+$ is the resolvent set of $-\Delta$. On this region, there exists unique solution to the Helmholtz equation in $L^2(\mathbb{R})$, obtained by Fourier transform.

There are two approaches to construct a solution for positive wavenumbers. Both start with the unique solution at complex wavenumber slightly off the spectrum, and then justify in different way the limit as the wavenumber approaches the positive axis.

Point-wise limit approach By separation of variables, we obtain

$$G_\epsilon(x) := \frac{e^{i\sqrt{\kappa^2+i\epsilon}\,|x|}}{2\,i\,\sqrt{\kappa^2+i\epsilon}}$$

as the fundamental solution to

$$(-\Delta - (\kappa^2 + i\epsilon))\, G_\epsilon \;=\; \delta(y)\;,\;\; \kappa, \epsilon \in \mathbb{R}^+.$$

The unique solution in $L^2(\mathbb{R})$ is given by

$$\mathsf{u}_\epsilon = \int_{-\infty}^{\infty} G_\epsilon(x-y)\; f(y)\, dy.$$

For each x ,

$$G_\epsilon(x) \;\xrightarrow[\epsilon\to 0]{\text{point wise}}\; G_{\text{outgoing}}(x) := \tfrac{e^{i\,\kappa\,|x|}}{2\,i\,\kappa}.$$

As a result of this,

$$\mathsf{u}_\epsilon \;\xrightarrow[\epsilon\to 0]{}\; \mathsf{u}_{\text{outgoing}} := G_{\text{outgoing}} \star \mathsf{f} \text{ in } H^2_{\text{loc}}(\mathbb{R})$$

with $\mathsf{u}_{\text{outgoing}}$ a solution to

$$(-\Delta - \kappa^2)\mathsf{u} = \mathsf{f}\,,\, \mathsf{f} \in L^2_c(\mathbb{R}).$$

This solution satisfies the Sommerfeld radiation condition, and thus called 'outgoing',

$$\lim_{x\to\infty} |\tfrac{1}{i}u'(x) - \kappa u(x)| = 0 \quad , \quad \lim_{x\to-\infty} |\tfrac{1}{i}u'(x) + \kappa u(x)| = 0.$$

See e.g. the lecture note by [9] for further details.

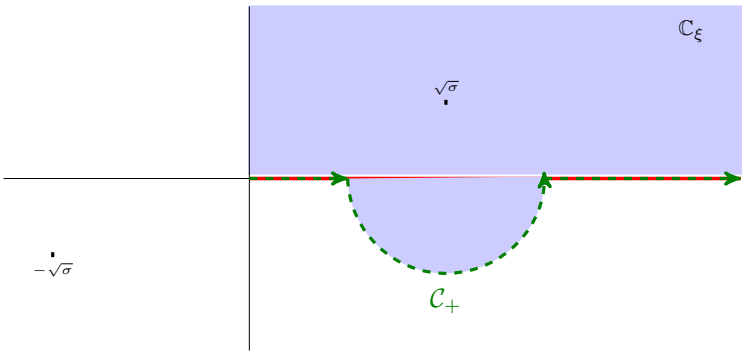


Figure 4: Outgoing solution defined by limiting absorption principle with deforming the positive real line to contour $\mathcal{C}_+$.

Analytic Continuation approach We discuss here the general idea of this approach. Consider $\mathsf{f} \in \mathcal{C}^\infty_c(\mathbb{R})$, then its Fourier transform $\widehat{\mathsf{f}}$ has an analytic extension to $\mathbb{C}$. For $\sigma \in \mathbb{C}$ with

$\mathrm{Re}\,\sigma > 0$, $\mathrm{Im}\,\sigma > 0$, the unique solution in $L^2(\mathbb{R})$ to $(-\Delta - \sigma)\mathsf{u} = \mathsf{f}$ is

$$\begin{aligned}\mathsf{u}(\sigma, x) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{ix\cdot\xi}\frac{\widehat{\mathsf{f}}(\xi)}{|\xi|^2-\sigma}\,d\xi \\ &= \frac{1}{2\pi}\int_{-\infty}^{0} e^{ix\cdot\xi}\frac{\widehat{\mathsf{f}}(\xi)}{|\xi|^2-\sigma}\,d\xi \;+\; \frac{1}{2\pi}\int_{0}^{\infty} e^{ix\cdot\xi}\frac{\widehat{\mathsf{f}}(\xi)}{|\xi|^2-\sigma}\,d\xi \\ &= \frac{1}{2\pi}\int_{0}^{\infty}\left[e^{ix\cdot\xi}\,\widehat{\mathsf{f}}(\xi) + e^{-ix\cdot\xi}\,\widehat{\mathsf{f}}(-\xi)\right]\frac{1}{|\xi|^2-\sigma}\,d\xi \quad , \quad \mathrm{Re}\;\sigma > 0\;,\;\mathrm{Im}\;\sigma > 0.\end{aligned}$$

In the last integral, we can deform $[0,\infty)$ to $\mathcal{C}_+$, as in Figure 4 to obtain an analytic continuation of $\mathsf{u}(x)$ to $\{\sigma : \mathrm{Re}\,\sigma > 0\}$. Denote this function by $\mathsf{u}_{\text{outgoing}}$

$$\mathsf{u}_{\text{outgoing}}(\sigma, x) := \frac{1}{2\pi}\int_{\mathcal{C}_+}\frac{e^{ix\cdot\xi}\,\widehat{\mathsf{f}}(\xi) + e^{-ix\cdot\xi}\,\widehat{\mathsf{f}}(-\xi)}{\xi^2-\sigma}\,d\xi \quad , \quad \mathrm{Re}\;\sigma > 0.$$

The above expression remains valid for $\sigma \in \mathbb{R}^+$ and defines a solution to the Helmholtz equation. In this way, we have 'crossed' the spectrum $\mathbb{R}^+$, see also Figure 3. This solution can be shown to satisfy the outgoing Sommerfeld condition. By uniqueness of such a solution, we reobtain the outgoing solution constructed in the previous approach.

3 The discrete problem

3.1 Discretization

To avoid any parasite reflexion due to the boundary conditions, we will define a fictitious Finite Element method on the complete real line $\mathbb{R}$. Geometrically, $\mathbb{R}$ is partitioned into intervals of length h

$$I_J = [y_J, y_{J+1}] \quad , \quad \text{with } J \in \mathbb{Z},$$

with the geometrical nodes given by

$$y_J := Jh, \quad \text{with } J \in \mathbb{Z}.$$

For a Finite Element method of order r, the global interpolation nodes are

$$x_{J,k} := (J + \tfrac{k}{r})h \quad , \quad \text{with } J \in \mathbb{Z}\,,\, 0 \le k < r.$$

For $J \in \mathbb{Z}$, we make use of the map $F_J : \hat{x} \mapsto h\hat{x} + y_J$ which is an isomorphism between the reference element $[0,1]$ and the interval I_J, to construct a basis on $\mathbb{R}$ using the local Lagrangian polynomials. We recall the definition of the local Lagrangian polynomials associated with the defining set $S = \{\hat{x}_i = i/r;\, 0 \le i \le r\}$,

$$\hat{\phi}_i(\hat{x}) := \prod_{\substack{0\le j\le r,\\ j\neq i}} \frac{(\hat{x}-\hat{x}_j)}{(\hat{x}_i-\hat{x}_j)}.$$

A global basis for the set of piece-wise polynomial functions defined on $\mathbb{R}$,

$$\mathbb{P}_r := \{p \in C^0(\mathbb{R}) \;\Big|\; p|_{I_J} \in P_r(I_J)\,,\, \forall J \in \mathbb{Z}\}$$

is given by the family $\phi_{J,k}$-s defined on all of $\mathbb{R}$,

$$k=0: \quad \phi_{J,0}(x) := \begin{cases} \hat{\phi}_0(F_J^{-1}x) & , \quad x \in I_J \\ \hat{\phi}_r(F_{J-1}^{-1}x) & , \quad x \in I_{J-1} \\ 0 & , \quad \text{otherwise} \end{cases};$$

$$0<k<r: \quad \phi_{J,k}(x) := \begin{cases} \hat{\phi}_k(F_J^{-1}x) & , \quad x \in I_J \\ 0 & , \quad \text{otherwise} \end{cases}.$$

Define the bilinear linear form a,

$$a(v,w) := \int_{-\infty}^{\infty} v'(x)\, w'(x)\, dx - \kappa^2 \int_{-\infty}^{\infty} v(x)\, w(x)\, dx.$$

Denote by $f = \{f_{J,k}\}$ the sequence defined by

$$f_{J,k} := \int_{-\infty}^{\infty} \mathsf{f}(x)\, \phi_{J,k}(x); \quad J \in \mathbb{Z},\, 0 \le k < r. \tag{5}$$

The numerical function $u_h \in \mathbb{P}_r$ approximating exact outgoing solution $\mathsf{u}_{\text{outgoing}}$ to the Helmholtz equation, $-\Delta \mathsf{u} - \kappa^2 \mathsf{u} = \mathsf{f}$, in the $\phi_{J,k}$-s basis is given by

$$u_h = \sum u_{J,k}\, \phi_{J,k}, \tag{6}$$

and satisfies the following relation,

$$a(u_h, \phi_{J,k}) = \int_{-\infty}^{\infty} \mathsf{f}(x)\, \phi_{J,k}(x)\, dx, \quad J \in \mathbb{Z},\, 0 \le k < r. \tag{7}$$

As a result, the coefficients $u_{J,k}$-s in (6) have to satisfy the following system of r recurrence relations at square wave number κ^2. Those at levels Jr, $J \in \mathbb{Z}$ come from applying (7) to $\phi_{J,0}$ at geometrical nodes,

$$\sum_{l=0}^{r-1} \mathcal{M}_{rl}\, u_{J-1,l} \;+\; 2\mathcal{M}_{00}\, u_{J,0} \;+\; \sum_{l=1}^{r-1} \mathcal{M}_{0l}\, u_{J,l} \;+\; \mathcal{M}_{0r} u_{J+1,0} \;=\; h\, f_{J,0}\,. \tag{8}$$

The remaining types at levels $Jr+k$, with $0<k<r$, are obtained from applying (7) to $\phi_{J,k}$ at interpolation nodes,

$$\sum_{l=0}^{r-1} \mathcal{M}_{kl}\, u_{J,l} + \mathcal{M}_{kr} u_{J+1,0} = h\, f_{J,k} \quad , \quad 0<k<r. \tag{9}$$

The coefficients of the above equations are components of the matrix $\mathcal{M}$ defined as

$$\mathcal{M} = \mathcal{M}(\kappa^2 h^2) := \hat{S} - \kappa^2 h^2 \hat{M},$$

from the local mass and local stiff matrix $\hat{M}$ and $\hat{S}$, both of size $(r+1)\times(r+1)$,

$$\hat{M}_{ij} = \mathsf{a}_{\mathsf{M}}(\hat{\phi}_i, \hat{\phi}_j) \quad ; \quad \hat{S}_{ij} = \mathsf{a}_{\mathsf{S}}(\hat{\phi}_i, \hat{\phi}_j) \quad , \quad 0 \le i,j \le r. \tag{10}$$

These matrices describe the interactions between the local Lagrangian polynomials via the (local) bilinear forms a_{M} and a_{S},

$$\begin{aligned} \mathsf{a}_{\mathsf{S}}(f,g) &= \int_0^1 f'(\hat{x})\, g'(\hat{x})\, d\hat{x} \quad , \quad f,g \in H^1(0,1)\,; \\ \mathsf{a}_{\mathsf{M}}(f,g) &= \int_0^1 f(\hat{x})\, g(\hat{x})\, d\hat{x} \quad , \quad f,g \in L^2(0,1). \end{aligned} \tag{11}$$

Remark 1. *Equations* (8) *and* (9) *do not define uniquely the vector* $(u_{J,k})_{J\in\mathbb{Z},0\leq k<r}$. *As for the continuous problem it is necessary to resort to a limiting absorption principle to select the outgoing solution. This procedure consists in adding a small imaginary part* $\epsilon > 0$ *to the wave number* k

$$\kappa_\epsilon = \kappa\,(1 + \mathrm{i}\epsilon)\,. \tag{12}$$

This new term is classically interpreted as a damping term. It can be shown that the associated formulation is coercive in $l^2(\mathbb{Z})$, *we can define the unique vector of* $l^2(\mathbb{Z}\times\mathbb{Z}_r,\mathbb{C})$

$$\sum_{l=0}^{r-1}\mathcal{M}^\epsilon_{rl}\,u^\epsilon_{J-1,l} + 2\mathcal{M}^\epsilon_{00}\,u^\epsilon_{J,0} + \sum_{l=1}^{r-1}\mathcal{M}^\epsilon_{0l}\,u^\epsilon_{J,l} + \mathcal{M}^\epsilon_{0r}u^\epsilon_{J+1,0} \;=\; h\,f_{J,0}; \tag{13}$$

$$\sum_{l=0}^{r-1}\mathcal{M}^\epsilon_{kl}\,u^\epsilon_{J,l} + \mathcal{M}^\epsilon_{kr}u^\epsilon_{J+1,0} \;=\; h\,f_{J,k} \quad , \quad 0<k<r. \tag{14}$$

Here, $\mathcal{M}^\epsilon = \mathcal{M}(\kappa_\epsilon^2 h^2) = \hat{S} - \kappa_\epsilon^2 h^2 \hat{M}$. *The outgoing solution of* (8) *and* (9) *is finally defined by taking the limit of* u_ϵ *for* ϵ *going to* 0.

$$u_{J,k} = \lim_{\epsilon\to 0^+} u^\epsilon_{J,k}. \tag{15}$$

3.2 The 'blocked' discrete problem

We will discuss the procedure of sequential 'blocking', which replaces the current system (8) -(9) with a single equivalent recurrence relation of second order. We will need the following definitions.

Blocking and unblocking sequences Because of their isomorphism, one can use $\mathbb{Z}$ or $\mathbb{Z}\times\mathbb{Z}_r$ to index the elements of a sequence; with $\mathbb{Z}_r$ denoting the set of integers k satisfying $0\leq k\leq r-1$. We denote by $\mathfrak{M}(\mathbb{Z}\,,S)$ the set of sequences taking value in S and indexed by $\mathbb{Z}$; its elements are denoted by $u=\{u_J\}_{J\in\mathbb{Z}}$. We denote by $\mathfrak{M}(\mathbb{Z}\times\mathbb{Z}_r\,,S)$ the set of sequences taking value in S and indexed by $\mathbb{Z}\times\mathbb{Z}_r$; its elements are denoted by $u=\{u_{J,k}\}_{J\in\mathbb{Z},0\leq k<r}$.

The sets of sequences $\mathfrak{M}(\mathbb{Z}\times\mathbb{Z}_r,\mathbb{C})$ and $\mathfrak{M}(\mathbb{Z},\mathbb{C}^r)$ are isomorphic via the 'blocking' mapping $\mathfrak{b}_r$ and its inverse 'unblocking' $\mathfrak{b}_r^{-1}$ as follows. A scalar sequence $u\in\mathfrak{M}(\mathbb{Z}\times\mathbb{Z}_r,\mathbb{C})$ can be 'blocked' to obtain sequence $U\in\mathfrak{M}(\mathbb{Z},\mathbb{C}^r)$, whose elements are given by

$$U=\mathfrak{b}_r u \quad ; \quad U_J=(u_{J,k})_{0\leq k<r}. \tag{16}$$

Conversely, we can unblock a sequence $U\in\mathfrak{M}(\mathbb{Z}\,,\mathbb{C}^r)$ to obtain a scalar-valued sequence $u\in\mathfrak{M}(\mathbb{Z}\times\mathbb{Z}_r,\mathbb{C})$, whose elements are given by

$$u=\mathfrak{b}_r^{-1}U \quad ; \quad u_{J,k}=(U_J)_k, \quad J\in\mathbb{Z}\,,\,0\leq k<r. \tag{17}$$

In the above definition, $(U_J)_k$ is the k-th component of vector U_J. We say that the (vector-valued) sequence $U=\mathfrak{b}_r u$ is the 'blocked' version of the (scalar) sequence u, and $u=\mathfrak{b}_r^{-1}U$ the 'unblocked' version of U.

The interior local matrices We denote by $\hat{S}_{\rm int}$, $\hat{M}_{\rm int}$ the matrices of size $r-1$ obtained from the restriction of the local matrices $\hat{S}$ and $\hat{M}$ to the interior degrees of freedom,

$$(\hat{S}_{\rm int})_{ij} = \hat{S}_{ij} \quad , (\hat{M}_{\rm int})_{ij} = \hat{M}_{ij} \quad , 1 \leq i,j \leq r-1. \tag{18}$$

We next define matrix $\mathcal{M}_{\rm int}(w)$ of size $r-1$ as

$$\mathcal{M}_{\rm int}(w) = \hat{S}_{\rm int} - w\hat{M}_{\rm int}. \tag{19}$$

As a result, $\mathcal{M}_{\rm int}$ is symmetric and $\hat{S}_{\rm int} = \mathcal{M}_{\rm int}\big|_{w=0}$ and $\hat{M}_{\rm int} = -\frac{d}{dw}\mathcal{M}_{\rm int}\big|_{w=0}$.

Interaction between interior and ends The following quantities describe the interaction between the interior Lagrangian polynomials $\{\hat{\phi}_j\}_{1\leq j\leq r-1}$ and the left and right ones $\hat{\phi}_0$ and $\hat{\phi}_r$. We define scalars α_0 and α_1

$$\alpha_0 = \hat{S}_{0r} \quad , \quad \alpha_1 = \hat{M}_{0r}. \tag{20}$$

We define $\mathfrak{a}_0, \mathfrak{a}_1$, $\mathfrak{b}_0$ and $\mathfrak{b}_1$ scalar vectors of size $r-1$, whose components are given by, for $1 \leq j \leq r-1$,

$$\begin{cases} (\mathfrak{a}_0)_j = \hat{S}_{jr} \\ (\mathfrak{a}_1)_j = \hat{M}_{jr}. \end{cases} \quad ; \quad \begin{cases} (\mathfrak{b}_0)_j = \hat{S}_{j0} \\ (\mathfrak{b}_1)_j = \hat{M}_{j0} \end{cases} . \tag{21}$$

We then introduce $\alpha(w)$, $\mathfrak{a}(w)$ and $\mathfrak{b}(w)$

$$\alpha(w) = \alpha_0(w) - w\alpha_1(w) \quad , \quad \mathfrak{a}(w) = \mathfrak{a}_0 - w\mathfrak{a}_1 \quad , \quad \mathfrak{b}(w) = \mathfrak{b}_0 - w\mathfrak{b}_1. \tag{22}$$

The first coefficient matrix We define the first coefficient matrix $B(w)$ of size $r \times r$,

$$B(w) = \begin{pmatrix} 2\mathcal{M}_{00}(w) & \mathfrak{b}^t(w) \\ \mathfrak{b}(w) & \mathcal{M}_{\rm int}(w) \end{pmatrix}. \tag{23}$$

The vector $\mathfrak{b}$ defined in (21) is thus the truncated first row of B. As a first order polynomial in w, we write B

$$B(w) = B_0 - wB_1, \tag{24}$$

where B_0 and B_1 are scalar matrices defined as,

$$\begin{array}{ll} (B_0)_{00} = 2\hat{S}_{00} & (B_1)_{00} = 2\hat{M}_{00}\,; \\ (B_0)_{ij} = \hat{S}_{ij} & (B_1)_{ij} = \hat{M}_{ij} \qquad 0 \leq i,j \leq r-1\,,\, i \neq 0 \text{ or } j \neq 0. \end{array} \tag{25}$$

Note that, $B_0 = B\big|_{w=0}$ and $B_1 = -\frac{d}{dw}B\big|_{w=0}$.

The second coefficient matrix We next introduce the second coefficient matrix $A(w)$ of size $r \times r$,

$$A(w) = \begin{pmatrix} \alpha(w) & \mathfrak{a}^t(w) \\ 0_{(r-1)\times 1} & 0_{(r-1)\times(r-1)} \end{pmatrix}. \tag{26}$$

See also Figure 5 which summarizes the relation between $B, \mathcal{M}_{\rm int}, \alpha, \mathfrak{a}$ and $\mathfrak{b}$ and the local matrix $\mathcal{M} = \hat{S} - w\hat{M}$.

We introduce the equivalent form of system (8)-(9) as a second-order recurrence relation.

Proposition 1 ('Blocked' Recurrence relation)**.** *With sequences u and U, f and F related to each other by the blocking relation, c.f.* (16)-(17)*, i.e*

$$u = \mathfrak{b}_r^{-1}U \quad , \quad f = \mathfrak{b}_r^{-1}F\,,$$

the following statements are equivalent:

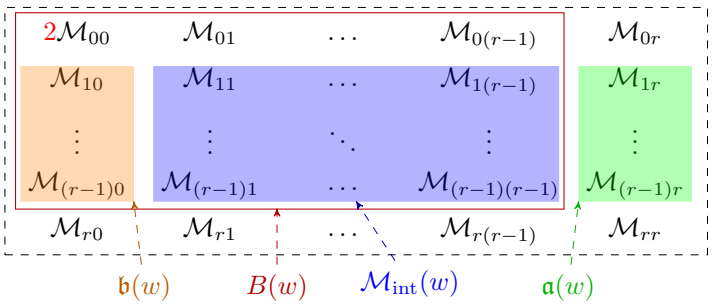


Figure 5: Relation between the interior matrix $\mathcal{M}_{\text{int}}$ defined by (19), coefficient matrices B defined by (24) and A defined by (26), the local matrices $\hat{S}$ and $\hat{M}$ defined by (10).

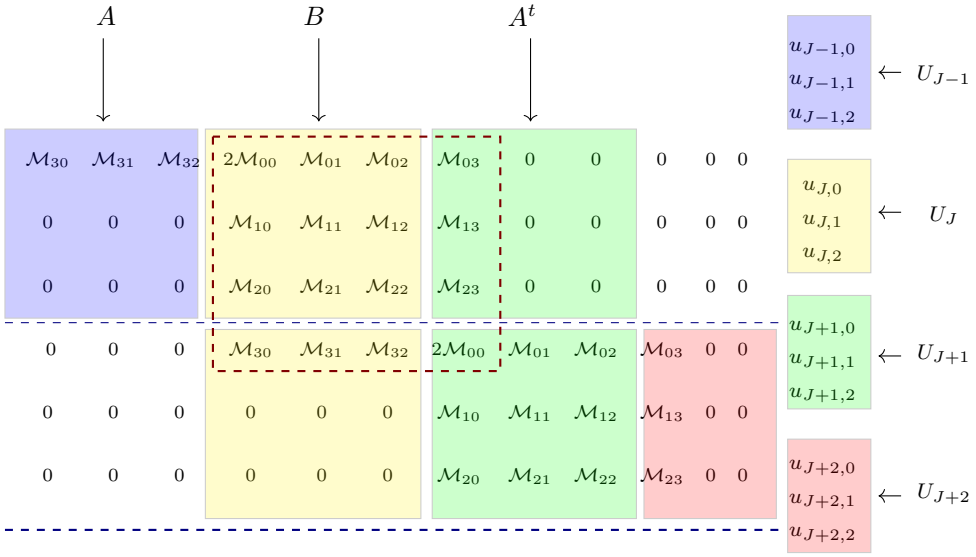


Figure 6: The recurrence pattern of the infinite linear system (8)-(9) and definition of coefficient matrices for recurrence relation (27).

1. *u solves the system of recurrence relations given in* (8)-(9) *at κ with right hand side f.*

2. *With matrices A and B defined in* (26) *and* (24)*, U solves*

$$A(w)\,U_{J-1} + B(w)\,U_J + A^t(w)\,U_{J+1} = h\,F_J \quad , \quad w = \kappa^2 h^2, \tag{27}$$

called the 'blocked' recurrence relation at κ with right hand side F. See also Figure 6.

4 Analytic results (Part 1) : Root structure of the characteristic polynomial and technical lemmas

4.1 Motivation

In Section 3, we have written system (8)-(9) as a second-order constant coefficient recurrence relation (27). In terms of shift operator τ acting on sequences, c.f (131),

$$(\tau_K U)_J := U_{J+K},$$

recurrence relation (27) can be written as

$$A(w)\,\tau_{-1}U + B(w)\,U + A^t(w)\,\tau_1 U = h\,F \quad ; \quad w = \mathsf{k}^2 h^2. \tag{28}$$

On can solve constant coefficient recurrence relations by a process analogous to the resolution of constant coefficient ordinary differential equations by Laplace transform. This process is via the discrete Z-transform, c.f. Appendix E, which first transforms a recurrence relation into an algebraic equation, a solution of the original problem is then obtained by taking Z^{-1}-transform. The above processes however need justification, c.f. Section 5. We give here a heuristic discussion to give a glimpse of the technical difficulties that will entail.

We list our working definition of Z-transform and refer the readers to Appendix E for more details.

Definition 1. *For a scalar sequence $u = \{u_n\}_{n\in\mathbb{Z}}$,*

$$[Zu](z) := \sum_{n=-\infty}^{\infty} u_n\, z^n, \tag{29}$$

if the RHS converges. We will also work with the following version

$$[Zu](\theta) := \sum_{n=-\infty}^{\infty} u_n\, e^{2\pi i n\theta}. \tag{30}$$

∇

Since the relation in question is vector-valued, we will be using the 'blocked' discrete Z-transform $Z_{\mathfrak{B}}$. Definition 5 extends readily to vector-valued sequences, upon noting that for a vector-valued sequence $U \in \mathfrak{M}(\mathbb{Z}, \mathbb{C}^r)$, with π_i denoting the projection onto the i-th component, $\pi_i\, U$ is a scalar sequence.

Definition 2. *For a vector-valued sequence $U \in \mathfrak{M}(\mathbb{Z}, \mathbb{C}^r)$, $[Z_{\mathfrak{B}} U](z)$ is a vector of dimension r with components given by*

$$\pi_i\, Z_{\mathfrak{B}}\, U = Z\, \pi_i\, U \quad , 1 \le i \le r, \tag{31}$$

if the RHS makes sense. Here there are also two versions corresponding to whichever definition of scalar Z used.

In the case where U is obtained from by a scalar sequence u by the 'blocking' map $\mathfrak{b}_r$, c.f. (16)-(17), i.e $U = \mathfrak{b}_r u$,

$$[\pi_k Z_{\mathfrak{B}}\, U](z) = \sum_{J=-\infty}^{\infty} u_{Jr+k}\, z^J\,, \, 1 \le k \le r.$$

This operator transforms the the shift operator $\tau_{\pm K}$ into a multiplication operator, see Proposition 38. In particular, for a vector-valued sequence sequence U of the form $U = \mathfrak{b}_r u$ for some scalar sequence u, we have

$$[Z_{\mathfrak{B}}\, \tau_{\pm K} U](z) = z^{\mp K}\, [Z_{\mathfrak{B}} U](z) \quad , \quad K \in \mathbb{Z}^+.$$

Taking formal $Z_{\mathfrak{B}}$ transform of (28) and write $W(w,z) = Z_{\mathfrak{B}}\, U(w)$ and $H(z) = Z_{\mathfrak{B}}\, F$, we obtain

$$\Big[A(w)\, z + B(w) + A^t(w)\, z^{-1}\Big] W(w,z) = h\, H(z) \quad , \quad w = \kappa^2 h^2.$$

For $z \neq 0$, the above algebraic equation is equivalent to

$$\underbrace{\big[A(w)\, z^2 + B(w)\, z + A^t(w)\big]}_{\mathcal{A}(w,z)} W(w,z) = h\, zH(z) \quad , \quad w = \kappa^2 h^2.$$

We define the transfer matrix $\mathcal{A}(w,z) = (\mathcal{A}(w,z))_{0 \leq i,j < r}$ of size r,

$$\mathcal{A}(w,z) := z^2 A(w) + zB(w) + A^t(w). \tag{32}$$

Substitute the definition of A and B, c.f (26) and (23) into the RHS, the entries of the transfer matrix are given by,

$$\mathcal{A} = \begin{pmatrix} z^2\, \alpha + 2z\, \mathcal{M}_{00} + \alpha & z^2\, \mathfrak{a}^t + z\, \mathfrak{b}^t \\ z\, \mathfrak{b} + \mathfrak{a} & z\mathcal{M}_{\text{int}} \end{pmatrix}. \tag{33}$$

By their definitions, we recall that $A(w)$ and $B(w)$ are matrix-valued polynomial of first order in w, c.f (26) and (23); hence the entries of $\mathcal{A}(w,z)$ are polynomial of first order in w and second order in z. We first make a general remark regarding the inverse of such a matrix.

Remark 2. *For a square matrix D of size r , we recall the definition of its adjugate* $\operatorname{Adj} D$*,*

$$\operatorname{Adj} D(z) = C^t(z) \quad , \quad C_{ij}(z) = (-1)^{i+j} M_{ij}(z) \quad , \quad 1 \leq i,j \leq r;$$

where $M_{ij}(z)$ is the (i,j) minor of $D(z)$. By Cramer's rule, we have

$$\big(\operatorname{Adj} D(z)\big)\; D(z) \quad = \quad D(z)\; \big(\operatorname{Adj} D(z)\big) \quad = \quad \det D(z)\; \operatorname{Id}.$$

The inverse of D is given by

$$D^{-1}(z) := \frac{\operatorname{Adj} D(z)}{\det D(z)}$$

outside the spectrum of D

$$\sigma_D := \{z \in \mathbb{C} \,\big|\, \det D(z) = 0\}.$$

For the rest of the discussion, we suppose that the components of D are polynomials in z. Under this hypothesis, $\det D(z)$ *and the components of* $\operatorname{Adj} D$ *are also polynomials in z, and the function $D^{-1}(z)$, wherever it is defined, is rational in z, with scalar matrix coefficients. In addition, the set σ_D is finite and discrete in $\mathbb{C}$, since $\sigma_D \subset \{z \in \mathbb{C} \,\big|\, \det D(z) = 0\}$, and the latter is the set of zeros of a polynomial hence is finite and discrete itself. In short, the inverse $D^{-1}(z)$ fails to be defined on σ_D, but exists and is analytic on $\mathbb{C} \setminus \sigma_D$. The elements of σ_D is discrete and finite in $\mathbb{C}$, with σ_D finite and discrete and called the poles of D^{-1}, or also as generalized eigenvalues of D.* △

Following discussion and notation of Remark 2, we can define $W(w,z)$ as

$$W(w,z) := h\, z\, \mathcal{A}(w,z)^{-1}\, H(z) \quad , \quad z \notin \sigma_{\mathcal{A}(w,\cdot)}, \tag{34}$$

where

$$\mathcal{A}^{-1}(w,z) := \frac{\operatorname{Adj} \mathcal{A}(w,z)}{\det \mathcal{A}(w,z)} \quad , \quad z \notin \sigma_{\mathcal{A}(w,\cdot)}.$$

The justification of Z-transform of (28) and of Z^{-1} transform of W necessitates a clear understanding of $\sigma_{\mathcal{A}}$ both for real and small enough w and well as its perturbation to the upper half complex plane. This will be the goal of the current section.

The main results of the next sections are that

$$\det \mathcal{A}(w,z) = \delta(w)\, z^{r-1}\, (z^2 - 2\rho(\omega) + 1)$$

and

$$\rho(w) \;=\; \cos w^{1/2} + w\mathsf{O}(w^r).$$

As a result, we will show that there the only nonzero poles are a pair of conjugate complex number having the properties depicted in Figure 7.

4.2 Properties of the inverse of the transfer matrix $\mathcal{A}^{-1}$

In order to state the results, we need the following notations. Define functions $\beta(w)$ and $\delta(w)$ as

$$\beta(w) := \det \begin{pmatrix} 0 & \mathfrak{a}^t(w) \\ \mathfrak{a}(w) & \mathcal{M}_{\text{int}}(w) \end{pmatrix} \quad ; \quad \delta(w) := \det \begin{pmatrix} \alpha(w) & \mathfrak{b}^t(w) \\ \mathfrak{a}(w) & \mathcal{M}_{\text{int}}(w) \end{pmatrix}. \tag{35}$$

We note that β and δ are determinants of matrices whose components are polynomial of first order in w, and thus are polynomials in w, c.f Remark 8.

Lemma 2. *1. The adjugate of the transfer matrix $\mathcal{A}(w,z)$ has the following factorization,*

$$\operatorname{Adj} \mathcal{A}(w,z) = z^{r-2} Q(w,z), \tag{36}$$

where the entries of $Q(w,z)$ are polynomials of second order in z and first order in w.

2. The determinant of $\mathcal{A}(w,z)$ has the following factorization,

$$\det \mathcal{A}(w,z) = \delta(w)\, z^{r-1}\, \mathbf{q}(w,z), \tag{37}$$

with the characteristic quadratic $\mathbf{q}(w,\cdot)$ defined by

$$\mathbf{q}(w,z) = z^2 - 2\rho(w)z + 1 \tag{38}$$

and the rational function ρ

$$\rho(w) := \frac{\beta(w) + \det B(w)}{-2\delta(w)}. \tag{39}$$

Proof. **Property 1 :** We recall the more explicit form of the transfer matrix $\mathcal{A}$, c.f. (33),

$$\mathcal{A} = \begin{pmatrix} z^2\,\alpha + 2z\,\mathcal{M}_{00} + \alpha & z^2\,\mathfrak{a}^t + z\,\mathfrak{b}^t \\ z\,\mathfrak{b} + \mathfrak{a} & z\mathcal{M}_{\text{int}} \end{pmatrix}.$$

By direct calculation that, we obtain that, for some constant c_0, constant vectors $v_1, v_0, \tilde{v}_1, \tilde{v}_0$ of dimension $r-1$ and constant matrices E_0, E_1 and E_2 of dimension $(r-1)\times(r-1)$, $\mathrm{Adj}\,\mathcal{A}$ can be written as,

$$\mathrm{Adj}\,\mathcal{A} = \begin{pmatrix} z^{r-1}c_0 & z^{r-2}(z\,v_1^t + v_0^t) \\ z^{r-1}(z\,\tilde{v}_1 + \tilde{v}_0) & z^{r-2}(z^2E_2 + zE_1 + E_0) \end{pmatrix}^t.$$

This immediately gives (36).

Property 2 : Using the fact the det is a multi-linear form, we develop $\det\mathcal{A}$ along the first row and then along the first column, and obtain

$$\det\mathcal{A} \;=\; \det\begin{pmatrix} 2z\mathcal{M}_{00} & z\,\mathfrak{b}^t \\ z\,\mathfrak{b}+\mathfrak{a} & z\,\mathcal{M}_{\mathrm{int}} \end{pmatrix} + \det\begin{pmatrix} z^2\,\alpha+\alpha & z^2\,\mathfrak{a}^t \\ z\,\mathfrak{b}+\mathfrak{a} & z\,\mathcal{M}_{\mathrm{int}} \end{pmatrix}.$$

Expanding both terms with respect to z,

$$\begin{aligned}
\det\begin{pmatrix} 2z\mathcal{M}_{00} & z\,\mathfrak{b}^t \\ z\,\mathfrak{b}+\mathfrak{a} & z\,\mathcal{M}_{\mathrm{int}} \end{pmatrix} &= \det\begin{pmatrix} 0 & z\,\mathfrak{b}^t \\ \mathfrak{a} & z\,\mathcal{M}_{\mathrm{int}} \end{pmatrix} + \det\begin{pmatrix} 2z\mathcal{M}_{00} & z\,\mathfrak{b}^t \\ z\,\mathfrak{b} & z\,\mathcal{M}_{\mathrm{int}} \end{pmatrix} \\
&= z^{r-1}\det\begin{pmatrix} 0 & \mathfrak{b}^t \\ \mathfrak{a} & \mathcal{M}_{\mathrm{int}} \end{pmatrix} + z^r\det B\,; \\
\det\begin{pmatrix} z^2\,\alpha+\alpha & z^2\,\mathfrak{a}^t \\ z\,\mathfrak{b}+\mathfrak{a} & z\,\mathcal{M}_{\mathrm{int}} \end{pmatrix} &= \det\begin{pmatrix} \alpha & z^2\,\mathfrak{a}^t \\ \mathfrak{a} & z\,\mathcal{M}_{\mathrm{int}} \end{pmatrix} + \det\begin{pmatrix} z^2\,\alpha & z^2\,\mathfrak{a}^t \\ z\,\mathfrak{b} & z\,\mathcal{M}_{\mathrm{int}} \end{pmatrix} \\
&= z^{r-1}\det\begin{pmatrix} \alpha & z\,\mathfrak{a}^t \\ \mathfrak{a} & \mathcal{M}_{\mathrm{int}} \end{pmatrix} + z^{r+1}\begin{pmatrix} \alpha & \mathfrak{a}^t \\ \mathfrak{b} & \mathcal{M}_{\mathrm{int}} \end{pmatrix} \\
&= z^r\det\begin{pmatrix} 0 & \mathfrak{a}^t \\ \mathfrak{a} & \mathcal{M}_{\mathrm{int}} \end{pmatrix} + z^{r-1}\;\det\begin{pmatrix} \alpha & 0 \\ 0 & \mathcal{M}_{\mathrm{int}} \end{pmatrix} + z^{r+1}\delta(w) \\
&= z^{r-1}\;\left[\delta(w) - \det\begin{pmatrix} 0 & \mathfrak{b}^t \\ \mathfrak{a} & \mathcal{M}_{\mathrm{int}} \end{pmatrix}\right] \;+\; z^r\;\beta(w) \;+\; z^{r+1}\;\delta(w).
\end{aligned}$$

Grouping the two last equations we get

$$\det\mathcal{A}(w,z) = z^{r-1}\delta(w) + z^r(\det B(w) + \beta(w)) + z^{r+1}\delta(w) \quad .$$

□

4.3 Coefficients of the characteristic quadratic q

In subsection 4.2, we have obtained an initial decomposition of the determinant of the transfer matrix $\mathcal{A}$,

$$\det\mathcal{A}(w,z) = \delta(w)\,z^{r-1}\left(z^2 - 2\rho(w)z + 1\right) = \delta(w)\,z^{r-1}\,\mathbf{q}(w,z).$$

where the expression ρ in the first order term of the characteristic quadratic $\mathbf{q}$ is given by

$$\rho(w) = \frac{\beta(w) + \det B(w)}{-2\delta(w)}.$$

See (35) and (23) for their definitions.

For small enough w, we will show that

$$\rho(w) = \cos w^{1/2} + w\mathsf{O}(w^r) \qquad \text{and} \qquad \delta(w) \neq 0\,.$$

Step 1 : In Lemma 3, we show

$$\frac{\beta(w) + \det B(w)}{\delta(w)} = -2 - w\frac{\det \mathcal{M}_{\rm int}(w)}{\delta(w)} \ \left(1 + w\ \kappa \cdot \mathcal{M}_{\rm int}^{-1}(w)\,\kappa\right)$$

Step 2 : In Proposition 31 in Appendix C, we show

$$w\,\kappa \cdot \mathcal{M}_{\rm int}^{-1}\kappa = -1 + \tfrac{2-2\cos w^{1/2}}{w^{1/2}\sin w^{1/2}} + \mathsf{O}(w^{2[r/2]+1}).$$

Step 3: In Lemma 4, we show $\delta(w) \neq 0$ and furthermore satisfies

$$-\frac{\delta(w)}{\det \mathcal{M}_{\rm int}(w)} = \tfrac{w^{1/2}}{\sin w^{1/2}} + w^{2[\frac{r-1}{2}]+2}\,\mathsf{e}(\mathfrak{a},\mathfrak{b}).$$

Step 4: After obtained the necessary expansion for the coefficients of the quadratic $\mathbf{q}$, we can calculate discriminant. This is the result of Proposition 5.

We now begin the proofs for Step 1.

Lemma 3. *We have the following identity*

$$\beta(w) + \det B(w) + 2\delta(w) = -w \det \mathcal{M}_{int}\left[1 + w\kappa \cdot \mathcal{M}_{int}^{-1}\kappa\right],$$

where κ is defined by, c.f (77), $\kappa = (\kappa_1, \ldots, \kappa_{r-1})^t$, $\kappa_j = \int_0^1 \hat{\phi}_j(\hat{x})$.

Remark 3. *We next obtain a relation between $\beta(w) + \det B(w)$ and $\delta(w)$. We first remark that the terms appearing in the definitions of ρ are all polynomials in w, being determinants of matrices whose components are polynomial of first order in w, c.f Remark 8; in particular, $\beta(w)$ and $\delta(w)$ are polynomials of degree $r-1$, and $\det B(w)$ of order r. Similarly, since $\mathcal{M}_{int}$ is a matrix valued of size $r-1$ first order polynomial in w, $\det \mathcal{M}_{int}(w)$ is, for the same reason, a polynomial of degree $r-1$. As a result, the ratio of these quantities will be rational in w. In particular, we will obtain expansion for the following functions*

$$\frac{\delta(w)}{\det \mathcal{M}_{int}(w)} \quad \textit{and} \quad \frac{\beta(w) + \det B(w) + 2\delta(w)}{\det \mathcal{M}_{int}(w)}.$$

By Proposition 26, we have that for $|w| < \pi^2$, $\mathcal{M}_{int}(w)$ is invertible, in fact at $w = 0$, $\det \mathcal{M}_{int} = \det \hat{S}_{int}$, and the latter is invertible by Proposition 24. As a result, the above rational functions are analytic in $|w| < \pi^2$ and are equal to their converging Taylor expansion in that neighborhood.

Proof. We first rewrite β, $\det B$ and δ as follows

$$\beta = \det\begin{pmatrix} 0 & \mathfrak{a}^t \\ \mathfrak{a} & \mathcal{M}_{\rm int}\end{pmatrix} = \det\begin{pmatrix} \mathcal{M}_{00} & \mathfrak{a}^t \\ \mathfrak{a} & \mathcal{M}_{\rm int}\end{pmatrix} + \det\begin{pmatrix} -\mathcal{M}_{00} & \mathfrak{a}^t \\ 0 & \mathcal{M}_{\rm int}\end{pmatrix}$$

$$\det B = \det\begin{pmatrix} 2\mathcal{M}_{00} & \mathfrak{b}^t \\ \mathfrak{b} & \mathcal{M}_{\rm int}\end{pmatrix} = \det\begin{pmatrix} \mathcal{M}_{00} & \mathfrak{b}^t \\ \mathfrak{b} & \mathcal{M}_{\rm int}\end{pmatrix} + \det\begin{pmatrix} \mathcal{M}_{00} & \mathfrak{b}^t \\ 0 & \mathcal{M}_{\rm int}\end{pmatrix}$$

and

$$2\delta = 2\det\begin{pmatrix} \alpha & \mathfrak{b}^t \\ \mathfrak{a} & \mathcal{M}_{\rm int}\end{pmatrix} = \det\begin{pmatrix} \alpha & \mathfrak{b}^t \\ \mathfrak{a} & \mathcal{M}_{\rm int}\end{pmatrix} + \det\begin{pmatrix} \alpha & \mathfrak{a}^t \\ \mathfrak{b} & \mathcal{M}_{\rm int}\end{pmatrix}.$$

The second terms in the expression of β and $\det B$ cancel out. We group the first term of β and the second term of 2δ, and the first terms of $\det B$ and 2δ, using the fact that $\mathcal{M}_{\text{int}}$ is symmetric,

$$\beta + \det B + 2\delta = \det \begin{pmatrix} \mathcal{M}_{00} + \alpha & \mathfrak{a}^t \\ \mathfrak{a} + \mathfrak{b} & \mathcal{M}_{\text{int}} \end{pmatrix} + \det \begin{pmatrix} \mathcal{M}_{00} + \alpha & \mathfrak{b}^t \\ \mathfrak{b} + \mathfrak{a} & \mathcal{M}_{\text{int}} \end{pmatrix}. \tag{40}$$

We recall the results from Proposition 21 which gives

$$\mathfrak{b} + \text{Col}(\mathcal{M}_{\text{int}}, 1) + \ldots + \text{Col}(\mathcal{M}_{\text{int}}, r-1) + \mathfrak{a} = -w\,\kappa;$$

$$\mathcal{M}_{00} + \pi_1 \mathfrak{b} + \ldots + \pi_{r-1}\mathfrak{b} + \alpha = -w\,\kappa_0.$$

$$\mathcal{M}_{00} + \pi_1 \mathfrak{a} + \ldots + \pi_{r-1}\mathfrak{a} + \alpha = -w\,\kappa_0.$$

These above identities give the value of the sum of the columns of matrices in (40). Hence, for each of the matrix in (40), by multilinearity of det, we replace their first column by the sum of all column, and obtain

$$\begin{aligned} \beta + \det B + 2\delta &= \det \begin{pmatrix} -w\kappa_0 & \mathfrak{a}^t \\ -w\kappa & \mathcal{M}_{\text{int}} \end{pmatrix} + \det \begin{pmatrix} -w\kappa_0 & \mathfrak{b}^t \\ -w\kappa & \mathcal{M}_{\text{int}} \end{pmatrix} \\ &= \det \begin{pmatrix} -w\kappa_0 & \mathfrak{a}^t + \mathfrak{b}^t \\ -w\kappa & \mathcal{M}_{\text{int}} \end{pmatrix} \quad = \det \begin{pmatrix} -2w\kappa_0 & -w\kappa^t \\ \mathfrak{a} + \mathfrak{b} & \mathcal{M}_{\text{int}} \end{pmatrix} \end{aligned} \tag{41}$$

Note that we can compute the sum of the first row in the last matrix in (41), by using Proposition 19 and Corollary 18 which give

$$\kappa_0 + \pi_1\kappa + \ldots + \pi_{r-1}\kappa + \kappa_r = 1 \quad ; \quad \kappa_r = \kappa_0$$

$$\Rightarrow \quad 2\kappa_0 + \pi_1\kappa + \ldots + \pi_{r-1}\kappa = 1.$$

As a result, for the last matrix in the expression (41), using the multi-linearity of det, we replace its first column by the sum of the columns, and obtain

$$\beta + \det B + 2\delta = \det \begin{pmatrix} -w & -\omega\kappa^t \\ -w\kappa & \mathcal{M}_{\text{int}} \end{pmatrix} = -w \det \mathcal{M}_{\text{int}} + w^2 \det \begin{pmatrix} 0 & \kappa^t \\ \kappa & \mathcal{M}_{\text{int}} \end{pmatrix}.$$

The proof is finished by using Lemma 39 which allows rewriting the determinant term, and we obtain

$$\beta + \det B + 2\delta = -w \det \mathcal{M}_{\text{int}} - w^2\,\kappa \cdot \mathcal{M}_{\text{int}}\,\kappa^{-1}.$$

□

Next we show the proofs for Step 3.

Lemma 4. *For $|w| < \pi^2$, the rational function $\frac{\delta}{\det \mathcal{M}_{int}}$ is analytic in w and is given by its converging Taylor series*

$$-\frac{\delta(w)}{\det \mathcal{M}_{int}(w)} = \frac{w^{1/2}}{\sin w^{1/2}} + w^{2[\frac{r-1}{2}]+2}\,\mathfrak{e}(\mathfrak{a}, \mathfrak{b}).$$

With the analytic error term $\mathfrak{e}(\mathfrak{a}, \mathfrak{b})$ bounded by

$$|\mathfrak{e}(\mathfrak{a}, \mathfrak{b})| \leq 3\,\pi^{4(1-[\frac{r-1}{2}])}\,\frac{(\pi^2 + |w|)^2}{(\pi^2 - |w|)^3}.$$

With the current assumption on w, matrix $\mathcal{M}_{int}$ is invertible, c.f. Proposition 26; as a result, $\delta(w) \neq 0$ for small enough w.

Proof. We recall the definition of $\delta(w)$,

$$\delta(w) = \det \begin{pmatrix} \hat{S}_{r0} - w\hat{M}_{r0} & \mathfrak{b}^t(w) \\ \mathfrak{a}(w) & \mathcal{M}_{\text{int}}(w) \end{pmatrix}.$$

We next use Lemma 39 to rewrite the RHS as

$$\begin{aligned}\delta(w) &= \left(\hat{S}_{r0} - w\hat{M}_{r0}\right) \det \mathcal{M}_{\text{int}}(w) - \det \mathcal{M}_{\text{int}}(w) \times \mathfrak{a} \cdot \mathcal{M}_{\text{int}}^{-1}\mathfrak{b} \\ &= \det \mathcal{M}_{\text{int}}(w) \left[\hat{S}_{r0} - w\hat{M}_{r0} - \mathfrak{a} \cdot \mathcal{M}_{\text{int}}^{-1}\mathfrak{b}\right].\end{aligned}$$

The proof is finished by using the expansion of $\mathfrak{a} \cdot \mathcal{M}_{\text{int}}^{-1}\mathfrak{b}$ given by Proposition 32,

$$\mathfrak{a}(w) \cdot \mathcal{M}_{\text{int}}^{-1}(w)\, \mathfrak{b}(w) = \hat{S}_{r0} - w\hat{M}_{r0} + \frac{w^{1/2}}{\sin w^{1/2}} + w^{2[\frac{r-1}{2}]+2}\, \mathsf{e}(\mathfrak{a}, \mathfrak{b}).$$

□

Main results of subsection After obtained the necessary tools from Step 1- 3, we now obtain the results on the coefficients and discriminant of the characteristic polynomial $\mathbf{q}$. From its definition, $\rho = \frac{\beta + \det B}{-2\delta}$ is a ration of polynomials in w, see also the comments at the beginning of previous Subsection, hence itself is a rational function in w, and thus the discriminant $\Delta(w) = 4(\rho^2 - 1)$ is also rational in w. At $w = 0$, $\delta(0) = -\det \mathcal{M}_{\text{int}}(0) = -\det \hat{S}_{\text{int}}$, and the latter is positive, c.f Proposition 24. As a result, $\rho(w)$ and hence $\Delta(w)$ are analytic in w in a neigbohood of $w = 0$.

Proposition 5. *1. ρ as defined in (39) is a rational function with real coefficients in w and has expansion,*

$$\rho(w) = \cos w^{1/2} + w\mathsf{O}(w^r). \tag{42}$$

2. The discriminant $\Delta(w)$ is a polynomial of degree $\leq 2r$ with real coefficients in w and has expansion,

$$\Delta(w) = -4\sin^2(w^{1/2})\big(1 + \mathsf{O}(w^r)\big). \tag{43}$$

Proof. **Step 1** : We list the relation between $\beta + \det B$ and δ obtained in Lemma 3,

$$\begin{aligned}\frac{\beta(w) + \det B(w)}{-2\delta(w)} &= 1 + \frac{w \det \mathcal{M}_{\text{int}}}{2\delta(w)}\big(1 + w\kappa \cdot \mathcal{M}_{\text{int}}^{-1}\kappa\big) = 1 + w\, \frac{1 + w\kappa \cdot \mathcal{M}_{\text{int}}^{-1}\kappa}{\frac{2\delta(w)}{\det \mathcal{M}_{\text{int}}}} \\ &\overset{(\star)}{=} 1 - w\, \frac{\frac{2 - 2\cos w^{1/2}}{w^{1/2}\sin w^{1/2}} + \mathsf{O}(w^{2[r/2]+1})}{\frac{2\, w^{1/2}}{\sin w^{1/2}} + \mathsf{O}(w^{2[\frac{r-1}{2}]+2})} \\ &= 1 - \frac{(1 - \cos w^{1/2})\frac{w^{1/2}}{\sin w^{1/2}} + \mathsf{O}(w^{2[r/2]+1})}{\frac{w^{1/2}}{\sin w^{1/2}} + \mathsf{O}(w^{2[\frac{r-1}{2}]+2})} \\ &= 1 - \left[(1 - \cos w^{1/2}) + w\, \mathsf{O}(w^{2[r/2]+1})\frac{\sin w^{1/2}}{w^{1/2}}\right]\left[1 - \mathsf{O}(w^{2[\frac{r-1}{2}]+2})\frac{\sin w^{1/2}}{w^{1/2}}\right] \\ &= \cos w^{1/2} + \mathsf{O}(w^{2[\frac{r-1}{2}]+3}) + \mathsf{O}(w^{2[r/2]+2}) + \text{lower order terms}\,.\end{aligned} \tag{44}$$

Equality ($\star$) was obtained by substituting the calculation of $\kappa \cdot \mathcal{M}_{\text{int}}^{-1}\kappa$ given by Proposition 31,

$$1 + w\,\kappa \cdot \mathcal{M}_{\text{int}}^{-1}\kappa = \frac{2 - 2\cos w^{1/2}}{w^{1/2}\sin w^{1/2}} + \mathsf{O}(w^{2[r/2]+1}),$$

and that of $\frac{-\delta(w)}{\det \mathcal{M}_{\text{int}}(w)}$ given by Proposition 4,

$$-\frac{\delta(w)}{\det \mathcal{M}_{\text{int}}(w)} = \frac{w^{1/2}}{\sin w^{1/2}} + \mathsf{O}(w^{2[\frac{r-1}{2}]+2})\,.$$

We next simplify the error in (44),

$$\mathsf{O}(w^{2[\frac{r-1}{2}]+3}) + \mathsf{O}(w^{2[r/2]+2})$$

can be simplied to $w\mathsf{O}(w^r)$ by

$$2\left[\frac{r-1}{2}\right] + 3 = \begin{cases} r+1 & , \quad r \text{ is even} \\ r+2 & , \quad r \text{ is odd} \end{cases} \quad ; \quad 2 + 2\left[\frac{r}{2}\right] = \begin{cases} r+2 & , \quad r \text{ is even} \\ r+1 & , \quad r \text{ is odd} \end{cases}.$$

As a result, we have obtained,

$$\rho(w) = \frac{\beta + \det B}{-2\delta} = \cos w^{1/2} + w\mathsf{O}(w^r).$$

Step 2 : We next calculate the discriminant $\Delta(w)$. In terms of ρ, the discriminant of $\mathbf{q}(w,\cdot)$ is

$$\begin{aligned}
\Delta &= \frac{(\beta(w) + \det B(w))^2 - 4\delta^2(w)}{\delta^2(w)} = 4(\rho^2(w) - 1) \\
&= -4\Big[\sin^2(w^{1/2}) - 2w\,\mathsf{O}(w^r)\cos w^{1/2} - w^2\,\mathsf{O}(w^r)\mathsf{O}(w^r)\Big] \\
&= -4\left[\sin^2(w^{1/2}) - w\mathsf{O}(w^r)\Big(2\cos w^{1/2} - w\mathsf{O}(w^r)\Big)\right] \\
&= -4\sin^2(w^{1/2})\big(1 + \mathsf{O}(w^r)\big).
\end{aligned}$$

In the last equality, we have used $\sin(w^{1/2}) = w^{1/2}(1 + \mathsf{O}(w))$.

□

4.4 Roots of the characteristic quadratic q

At real wavenumbers We would like to determine the requirements on $w \in \mathbb{R}$ so that the characteristic quadratic $\mathbf{q}(w,z) = z^2 - 2\rho(w)\,z + 1$, introduced in (38) and (37), has complex roots. We first obtain more explicit expansions of the discriminant $\Delta(w)$ and the first order coefficient, $-2\rho(w)$, c.f. (39). As a result of Lemma 5, in particular due to the form of the discriminant $\Delta(w)$ given in (43), for $w > 0$ and small enough, the characteristic quadratic $z \mapsto \mathbf{q}(w,z)$ have complex conjugate roots of norm 1. We denote these roots by $\mathsf{z}_{\pm,0} = e^{\pm i\phi_0(w)}$ with

$$\mathsf{z}_{\pm,0} = e^{\pm i\phi_0(w)} := \rho\big(\mathsf{w}\big) \pm i\tfrac{1}{2}\sqrt{\big|\Delta\big(\mathsf{w}\big)\big|}. \tag{45}$$

In addition, due to the property of ρ given in (42), the argument ϕ_0 satisfies,

$$\cos\phi_0(w) = \rho(w) = \cos w^{1/2} + w\,\mathsf{O}(w^r).$$

From this relation, Proposition (40) in F.2 gives an expansion for the difference between $\phi_0(w)$ and $w^{1/2}$.

We summarize this discussion in the following proposition. This result will be crucial in the dispersion analysis in Subsection 6.2.

Proposition 6. *1. For $w > 0$ and small enough, the roots of characteristic quadratic $\mathbf{q}(w,\cdot)$, introduced in* (38) *and* (37)*, are complex conjugates of magnitude 1. Denote these roots by $z_{\pm,0} = e^{\pm i\phi_0(w)}$, c.f* (45)*.*

2. The principal argument $\phi_0(w) \in (0,\pi)$ is a function analytic in w and depends on r and has the expansion,

$$\cos\phi_0 \;=\; \cos w^{1/2} \;+\; w\mathsf{O}(w^r). \tag{46}$$

Thus for small enough w, implies

$$\phi_0(w) = w^{1/2}\left(1 + \underset{w\to 0}{\mathsf{O}}(w^r)\right). \tag{47}$$

At complex wavenumbers We would like to study how the roots of $\mathbf{q}(w,\cdot)$ behave under small perturbations of w to the complex plane, in particular for w of the form

$$w_\epsilon = (1+i\epsilon)^2\mathsf{w} \quad , \quad \mathsf{w} > 0\,,\, \epsilon \in> 0.$$

We will show that they have the properties depicted in Figure 7. For such w, denote by $\mathsf{z}_\pm(w)$ the corresponding roots of $\mathbf{q}(w,\cdot)$ with

$$\mathsf{z}_{\pm,\epsilon} := \rho(w_\epsilon) \pm \tfrac{1}{2}\sqrt{\Delta(w_\epsilon)}. \tag{48}$$

Here $\sqrt{\cdot}$ is the complex square root branch defined on $\mathbb{C}\setminus\mathbb{R}^+$ with

$$\begin{gathered}\sqrt{z} := |z|^{1/2}\, e^{i\frac{1}{2}\operatorname{Arg} z},\ \text{where } z \in \mathbb{C}\setminus\mathbb{R}^+,\\ \operatorname{Arg} : \mathbb{C} \longrightarrow [0,2\pi).\end{gathered} \tag{49}$$

From Proposition 6, at $\epsilon = 0$ for real and small enough w, the roots are complex conjugates and lie on the unit circle. When w is not real, these roots cannot remain on the unit circle; for if z is a root of magnitude 1, we multiply both sides of $\mathbf{q}(w,z) = 0$ by $\overline{z}^2$

$$z^2 - 2\rho(w)z + 1 = 0 \quad \Longrightarrow \quad 1 - 2\rho(w)\overline{z} + \overline{z}^2 = 0.$$

This means $\overline{z}$ is also a root. By uniqueness of factorization of polynomials, this implies that $\rho(w)$ has to be real. However, this is not generally the case when w is not real, for $\rho(w) = \frac{\beta + \det B}{\delta}$ is a rational function with real coefficients in w since $\beta, \det B$ and δ are all polynomials with real coefficients in w, c.f Remark 8. On the other hand, it remains true that $\mathsf{z}_+\mathsf{z}_- = 1$, one root will lie inside the unit circle, and the other one outside.

To determine the relative position of $\mathsf{z}_{\pm,\epsilon}$ to the unit circle, we first obtain a complex version of Lemma 5.

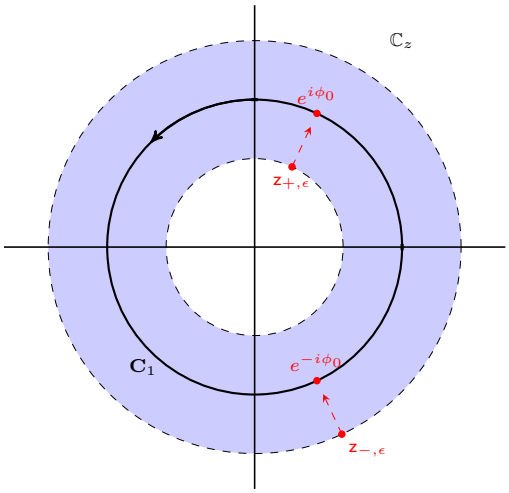


Figure 7: The structure of roots of the characteristic quadratic $\mathbf{q}(w,\cdot)$ for $w = (1+i\epsilon)\mathsf{w}$ with $\epsilon > 0$ and $\mathsf{w} > 0$, c.f Proposition 8.

Lemma 7. *For $w_\epsilon = (1+i\epsilon)\mathsf{w}$ with $\mathsf{w} > 0$ and $\epsilon \in \mathbb{R}^+$, we have*

$$\rho(w_\epsilon) = 1 - \big(\tfrac{1}{2} - \tilde{e}_1 + \epsilon^2\,\tilde{e}_2\big)\mathsf{w} - i\,\mathsf{w}\,\epsilon\big(\tfrac{1}{2} - \tilde{e}_2 - \tilde{e}_1\big).$$
$$\textit{with}\quad \tilde{e}_i \in \mathbb{R}\quad,\quad \tilde{e}_i = \mathsf{O}(\mathsf{w})\,,\ \textit{for } i = 1,2.$$

and

$$\sqrt{\Delta(w_\epsilon)} = 2\,\mathsf{w}^{1/2}\left[i\,(1+e_1) - \epsilon\,\big(\tfrac{1}{2} + e_2\big)\right]$$
$$\textit{with}\quad e_i \in \mathbb{R}\quad,\quad e_i = \mathsf{O}(\mathsf{w})\,,\ \textit{for } i = 1,2.$$

Proof. **Property 1** : Lemma 5 remains true for complex w with small enough norm. As a result, we obtain

$$\rho(w) = 1 - \tfrac{1}{2}w + w\,\mathsf{O}(w)\quad,\quad \text{for } w \in \mathbb{C}\,, |w| \text{ small enough.}$$

Apply (54) to ρ, we have

$$\begin{aligned}\rho(w_\epsilon) &= 1 - \tfrac{1}{2}(1+i\epsilon)\,\mathsf{w} + (1+i\epsilon)\,\mathsf{w}\,(\tilde{e}_1 + i\,\epsilon\,\tilde{e}_2)\\ &= 1 - \big(\tfrac{1}{2} - \tilde{e}_1 + \epsilon^2\,\tilde{e}_2\big)\mathsf{w} - i\,\mathsf{w}\,\epsilon\big(\tfrac{1}{2} - \tilde{e}_2 - \tilde{e}_1\big).\end{aligned} \tag{50}$$

with $\tilde{e}_i \in \mathbb{R}$, $\tilde{e}_i = \mathsf{O}(\mathsf{w})$, for $i = 1,2$.

Property 2 : Lemma 5 remains true for complex w with small enough norm. As a result, we obtain

$$\Delta(w) = -4w\big[1 + \mathsf{O}(w)\big]\quad,\quad \text{for } w \in \mathbb{C}\,, |w| \text{ small enough.}$$

Apply (54) to $\Delta(w)$, we obtain

$$\Delta(w_\epsilon) = 4\,\mathsf{w}\left[-(1+i\epsilon) - 2\,\hat{e}_1 - i\,\epsilon\,2\,\hat{e}_2\right];$$
$$\hat{e}_i \in \mathbb{R}\quad,\quad \hat{e}_i = \mathsf{O}(\mathsf{w})\,,\ \text{for } i = 1,2.$$

Now using the Taylor expansion for $\sqrt{-1-z}$, $z \in \mathbb{C}$ and $|z| < 1$, given by (52), we have,

$$\begin{aligned}\sqrt{\Delta(w_\epsilon)} &= 2\,i\,\mathsf{w}^{1/2}\,\big(1 + \tfrac{1}{2}\,i\,\epsilon + e_1 + i\,e_2\big)\\ &= 2\,\mathsf{w}^{1/2}\left[i\,(1+e_1) - \epsilon\,\big(\tfrac{1}{2} + e_2\big)\right];\end{aligned} \tag{51}$$

with $e_i \in \mathbb{R}$ and $e_i = \mathsf{O}(\mathsf{w})$ for $i = 1, 2$. □

Proposition 8. *The roots of characteristic quadratic $z \mapsto \mathbf{q}(w_\epsilon, z)$ at $w_\epsilon = (1 + i\epsilon)\mathsf{w}$, for $\mathsf{w} > 0$ and small enough, which were denoted by $\mathsf{z}_{\pm,\epsilon}$, c.f (48), have the properties depicted in Figure 7, i.e.*

1. *as $\epsilon \to 0$, $\mathsf{z}_{\pm,\epsilon} \to e^{\pm i\phi_0}$;*
2. *for $\epsilon > 0$ and small enough, $|\mathsf{z}_{+,\epsilon}| < 1 < |\mathsf{z}_{-,\epsilon}|$.*

Proof. **Property 1** : For $\mathsf{w} > 0$, we consider the limits of $\mathsf{z}_{\pm,\epsilon}$ as $\epsilon \to 0$. Recall that $w \mapsto \Delta(w, z)$ is a continuous function of w. Hence as $\epsilon \to 0$, $\Delta\big(w_\epsilon\big) \to \Delta\big(\mathsf{w}\big)$. The latter $\Delta\big(\mathsf{w}\big)$ is strictly negative, for $\mathsf{w} > 0$ and small enough, c.f. Lemma 5. On the other hand, the chosen branch $\sqrt{\cdot}$ is analytic across $\{z \in \mathbb{C} \,|\, \mathrm{Re}\, z < 0\}$. Thus we have

$$\sqrt{\Delta\big(w_\epsilon\big)} \longrightarrow i\sqrt{\big|\Delta\big(\mathsf{w}\big)\big|} \quad , \quad \epsilon \to 0.$$

As a result, for $\mathsf{w} > 0$ and small enough,

$$\mathsf{z}_{\pm,\epsilon} \longrightarrow \mathsf{z}_{\pm,0} = e^{\pm i\phi_0} \quad , \quad \epsilon \to 0.$$

Property 2 : Insert the value of ρ and Δ at w_ϵ given (51) and (50) into the definition of $\mathsf{z}_{\pm,\epsilon}$, we obtain

$$\begin{aligned}
\mathsf{z}_{+,\epsilon} &= \rho(w_\epsilon) + \tfrac{1}{2}\sqrt{\Delta(w_\epsilon)} \\
&= 1 - \big(\tfrac{1}{2} - \tilde{e}_1 + \epsilon^2\,\tilde{e}_2\big)\mathsf{w} - \mathsf{w}^{1/2}\,\epsilon\,\big(\tfrac{1}{2} + e_2\big) \quad + i\,\mathsf{w}^{1/2}\Big[1 + e_1 - \mathsf{w}^{1/2}\,\epsilon\,\big(\tfrac{1}{2} - \tilde{e}_2 - \tilde{e}_1\big)\Big]; \\
\mathsf{z}_{-,\epsilon} &= \rho(w_\epsilon) - \tfrac{1}{2}\sqrt{\Delta(w_\epsilon)} \\
&= 1 - \big(\tfrac{1}{2} - \tilde{e}_1 + \epsilon^2\tilde{e}_2\big)\mathsf{w} + \mathsf{w}^{1/2}\epsilon\,\big(\tfrac{1}{2} + e_2\big) \quad - i\,\mathsf{w}^{1/2}\Big[1 + e_1 + \mathsf{w}^{1/2}\,\epsilon\,\big(\tfrac{1}{2} - \tilde{e}_2 - \tilde{e}_1\big)\Big].
\end{aligned}$$

Since $e_i = \mathsf{O}(\mathsf{w})$ and $\tilde{e}_i = \mathsf{O}(\mathsf{w})$, and for small $\mathsf{w} > 0$ and small $\epsilon > 0$, we have the following comparisions,

$$\begin{array}{rcccl}
0 & < & \mathsf{w}^{1/2}\,\epsilon\,\big(\frac{1}{2} + e_2\big) & < & 1 - \big(\frac{1}{2} - \tilde{e}_1 + \epsilon^2\tilde{e}_2\big)\mathsf{w}\,; \\
0 & < & \mathsf{w}^{1/2}\,\epsilon\,\big(\frac{1}{2} - \tilde{e}_2 - \tilde{e}_1\big) & < & 1 + e_1.
\end{array}$$

As a result, it follows that $|\mathsf{z}_{+,\epsilon}| < |\mathsf{z}_{-,\epsilon}|$. Since $\mathsf{z}_{+,\epsilon}\mathsf{z}_{-,\epsilon} = 1$, we obtain $|\mathsf{z}_{+,\epsilon}| < 1 < |\mathsf{z}_{-,\epsilon}|$.

□

Remark 4. *Since $\{w \in \mathbb{C}\,,\, |w + 1| < 1\}$ is disjoint from the branch cut $\mathbb{R}^+$ of $\sqrt{\cdot}$, the function $\{|z| < 1\} \ni z \mapsto \sqrt{-1 - z}$ is analytic. On the other hand, when restricted to $\{\mathrm{Im}\, z = 0\,,\, |z| < 1\}$, it coincides with $(-1, 1) \ni x \mapsto \sqrt{-1 - x} = i\sqrt{1 + x}$, which is real-analytic and whose Taylor expansion is,*

$$x \in \mathbb{R}\,,\, |x| \leq 1, \quad \sqrt{1 + x} = 1 + \tfrac{1}{2}x + \sum_{n=2}^{\infty} c_n x^n \quad , \quad c_n \in \mathbb{R}.$$

As a result, being the analytic continuation of $i\sqrt{1 + x}$ from $\{\mathrm{Im}\, z = 0\,,\, |z| < 1\}$ to $\{|z| < 1\}$, $\sqrt{-1 - z}$ has the following Taylor expansion,

$$|z| < 1, \quad \sqrt{-1 - z} = i\Big(1 + \tfrac{1}{2}z + \sum_{n=2}^{\infty} c_n z^n\Big). \tag{52}$$

Remark 5. *1. For $\epsilon \ll 1$, we have*

$$(1+i\,\epsilon)^N = e_{1,N} + i\,\epsilon\, e_{2,N} \quad ; \quad e_{i,N} \in \mathbb{R} \text{ and } e_{i,N} = \mathsf{O}(1) \text{ for } i = 1,2. \tag{53}$$

This is seen from direct calculation.

$$(1+i\epsilon)^N = 1 - \sum_{k=1}^{[N/2]} (-1)^k \binom{N}{2k} \epsilon^{2k} + i\epsilon \sum_{k=0}^{N^*} (-1)^k \binom{N}{2k+1} \epsilon^{2k} = e_{1,N} + i\,\epsilon\, e_{2,N}.$$

with $e_i \in \mathbb{R}$ and $e_i = \mathsf{O}(1)$ for $i = 1,2$. Here, $N^* = [\frac{N}{2}] - 1$ for N even, and is equal to $[\frac{N}{2}]$ for N odd.

2. Suppose g is analytic in a neighborhood of zero with real Taylor coefficients, i.e,

$$g(w) = g(0) + \sum_{n=1}^{\infty} g_n\, w^n \quad , \quad g_n \in \mathbb{R}.$$

For $w_\epsilon = (1+i\epsilon)\mathsf{w}$ where $\mathsf{w} \in \mathbb{R}^+$, by (53), we have,

$$g\big(w_\epsilon\big) = \sum_{n=1}^{N} g_n (1+i\epsilon)^n \mathsf{w}^n + E_1 + i\epsilon E_2. \tag{54}$$

Here E_1 and E_2 are real-valued analytic functions in w defined as

$$E_1 = \sum_{n=N+1}^{\infty} g_n\, e_{1,n}\, \mathsf{w}^n = \mathsf{O}(\mathsf{w}^{N+1}) \quad ; \quad E_2 = \sum_{n=N+1}^{\infty} g_n\, e_{2,n}\, \mathsf{w}^n = \mathsf{O}(\mathsf{w}^{N+1}).$$

5 Analytic results (Part 2) : Discrete Limiting Principle and the resolution of the blocked recurrence relation

5.1 Construction of the l^2 solution for complex wave number

We recall that, via Finite element method of order r with step size h, the discretization of the Helmholtz equation $-\mathsf{u}'' - \kappa^2 \mathsf{u} = \mathsf{f}$ gives the system of recurrence relations (8)-(9). Also recall that the inhomogeneous f is defined from f by (5), c.f. Section 3. The latter is equivalent the 'blocked' recurrence relation (27), c.f. Proposition 1. The heuristic discussion at the beginning of Section 4 show how the 'blocked' recurerence relation (27) is 'formally' transformed into an algebraic equation by using $Z_{\mathfrak{B}}$ transform, c.f. E for definition. For $\kappa^2 \in \mathbb{C}\backslash\mathbb{R}^+$ and $f \in l^2_{\text{comp}}(\mathbb{Z})$, this process is readily justified as follows.

It can shown that the system of recurrence relations (8)-(9) is coercive in $l^2(\mathbb{Z})$ at complex wavenumber, which assures the existence and uniqueness of solution $\in l^2(\mathbb{Z})$ in these cases. Here, we proceed to construct directly this solution in $l^2(\mathbb{Z})$. For $U = \mathfrak{b}_r\, u$, $U \in (l^2(\mathbb{Z}))^r$ if and only if $u \in l^2(\mathbb{Z})$. This means that, under the same assumption for κ^2 and with inhomogeneous term $F \in (l^2_{\text{comp}}(\mathbb{Z}))^r$, the same proposition also gives existence and uniqueness of solution in $(l^2(\mathbb{Z}))^r$ to the 'blocked' recurrence relation (27). As a result, under these assumptions on κ^2 and F, the $Z_{\mathfrak{B}}$ of this unique solution exists, together with the applicability of the inversion formula (130); hence, we are allowed to take $Z_{\mathfrak{B}}$ transform on both sides of (27) and use the inversion formula (130) to construct this unique solution. We will do this for the case where κ^2 is of the form

$$\kappa_\epsilon^2 = \kappa^2(1+i\epsilon) \quad \text{, with } \kappa \in \mathbb{R} \setminus \{0\}\,, \epsilon > 0. \tag{55}$$

To lighten the notation, we drop h from the variables, and denote by

$$W(\kappa^2, z) = W(\kappa^2, h, z) = Z_{\mathfrak{B}}\, U(\kappa^2, h) \quad \text{and} \quad H(z) = H(h, z) = Z_{\mathfrak{B}}\, F(h).$$

After taking $Z_{\mathfrak{B}}$ transform on both sides of (27), c.f. the discussion at the beginning of Section 4, we obtain,

$$\Big[z\, A(\kappa_\epsilon^2 h^2) + B(\kappa_\epsilon^2 h^2) + z^{-1}\, A^t(\kappa_\epsilon^2 h^2) \Big] W(\kappa_\epsilon^2, z) = h\, H(z) \quad .$$

For $z \neq 0$, the above equation is equivalent to

$$\mathcal{A}(\kappa_\epsilon^2 h^2, z)\, W(\kappa_\epsilon^2, z) = h\, z\, H(z). \tag{56}$$

where we recall that we have denoted by $\mathcal{A}(w, z)$ the transfer matrix

$$\mathcal{A}(w, z) = A(w)\, z^2 + B(w)\, z + A^t(w).$$

By Cramer's rule, we solve for W and the solution to the 'blocked' equation is then obtained by the inversion formula (130).

Proposition 9. *Consider sequences $F \in (l^2_{comp}(\mathbb{Z}))^r$ and $f \in l^2_{comp}(\mathbb{Z})$ related to each other by $F = \mathfrak{b}_r f$. For $\kappa_\epsilon^2 = (1 + i\epsilon)\kappa^2$ with $\kappa \in \mathbb{R}^\star$, $\epsilon > 0$, we have*

1. *The unique solution in $(l^2(\mathbb{Z}))^r$ to the 'blocked' recurrence relation(27) at such κ^2 with r.h.s F is given by*

$$\big[U(\kappa_\epsilon^2)\big]_J = \frac{1}{2\pi i} \oint_{\mathbf{C}_1} W\big(\kappa_\epsilon^2, z\big)\, \frac{dz}{z^{J+1}} \quad ; \quad W(\kappa_\epsilon^2, z) = h\, \frac{\operatorname{Adj} \mathcal{A}(\kappa_\epsilon^2 h^2, z)}{\det \mathcal{A}(\kappa_\epsilon^2 h^2, z)}\, z\, H(z). \tag{57}$$

2. *The unique solution in ${}^2(\mathbb{Z})$ to system (8)-(9) at such κ^2 with r.hs f is given by*

$$u\big(\kappa_\epsilon^2\big) = \mathfrak{b}_r^{-1} U\big(\kappa_\epsilon^2\big).$$

Remark 6. *The validity of the integral in the RHS of (57) is guaranteed by the fact that the solution we are constructing is in $(l^2(\mathbb{Z}))^r$. In the current case, this fact can be seen from another perspective, since we also know the poles structure of W. From Lemma 37, we have obtained*

$$\operatorname{Adj} \mathcal{A}(w, z) = z^{n-2} Q(w, z) \quad , \quad \textit{with } Q \textit{ polynomial - matrix in } z\,;$$

and

$$\det \mathcal{A}(w, z) = \delta(w)\, z^{n-1} \mathbf{q}(w, z) \quad \textit{with } \mathbf{q} \textit{ a quadratic polynomial in } z.$$

As a result,

$$W(\kappa^2, z) = h\, \frac{\operatorname{Adj} \mathcal{A}(\kappa^2 h^2, z)}{\det \mathcal{A}(\kappa^2 h^2, z)}\, z\, H(z) = h\, \frac{Q(\kappa^2 h^2, z)}{\delta(\kappa^2 h^2)\, \mathbf{q}(\kappa^2 h^2, z)}\, H(z).$$

Since both Q and H are polynomial in z, the poles of $W(\kappa^2, z)$ come only from the roots of $\mathbf{q}$. By Proposition 8, the roots $\mathsf{z}_{\pm,\epsilon}$ of quadratic $z \mapsto \mathfrak{q}(w_\epsilon, z)$ satisfy $|\mathsf{z}_{+,\epsilon}| < 1 < |\mathsf{z}_{-,\epsilon}|$. As a result, the poles of W do not lie on the integrating contour $\mathbf{C}_1$, c.f. Figure 7. Here, $\mathbf{C}_r$ denotes the curve whose image is the circle centered at zero of radius r. ▽

5.2 Construction of the outgoing solutions at real wave number. Limiting absorption principle.

Having constructed solutions $U\big(\kappa^2(1+i\epsilon)\big)$ to the 'blocked' recurrence relation (27) for $\epsilon > 0$, c.f Proposition 9, we next construct a solution at real square wave number κ^2, as the limit as $\epsilon \to 0^+$.

We first recall and simplify the notations.

$$W_\epsilon(z) = W\big((1+i\epsilon)\,\kappa^2 h^2\,,\,z\big) \quad , \quad Q_\epsilon(z) = Q\big((1+i\epsilon)\,\kappa^2 h^2\,,\,z\big)$$
$$\mathbf{q}_\epsilon(z) = \mathbf{q}\big((1+i\epsilon)\,\kappa^2 h^2\,,\,z\big) \quad , \quad \delta_\epsilon = \delta((1+i\epsilon)\kappa^2 h^2).$$

We will denote by $W_0(z)$, $Q_0(z)$, $\mathbf{q}_0$ and δ_0, their corresponding limits as $\epsilon \to 0^+$. We recall the form of the currrent form of $U\big(\kappa^2(1+i\epsilon)\big)$, which is given by a contour integral along $\mathbf{C}_1$, c.f Proposition 9 and Figure 7,

$$[U_\epsilon]_J = \frac{1}{2\pi i}\oint_{\mathbf{C}_1} W_\epsilon(z)\,\frac{dz}{z^{J+1}} \quad ; \quad W_\epsilon(z) = h\,\frac{Q_\epsilon(z)}{\delta_\epsilon\,\mathbf{q}_\epsilon(z)}\,H(z).$$

Here $\mathbf{C}_r$ denotes the curve whose image is the circle centered at zero of radius r.

That the limit of U_ϵ exists as $\epsilon \to 0^+$ is not immediate. For $\epsilon > 0$, $\mathbf{C}_1$ is free of poles of W_ϵ, c.f Remark 6. However, this ceases to be the case at $\epsilon = 0$, for the limit integrand W_0 has poles at $e^{\pm i\phi_0}$, c.f. Proposition 8 and Figure 7. On the other hand, W_ϵ has an analytic extension to $\mathbb{C} \setminus \{\mathsf{z}_{+,\epsilon}\,,\,\mathsf{z}_{-,\epsilon}\}$. As a result, $\mathbf{C}_1$ can be replaced by a homotopic curve, which can be carefully chosen to be pole-free, even from those of W_0. We denote such a curve Γ_{outgoing} and proceed to construct it.

We list the simple closed loops that will be need for the definition of Γ_{outgoing}. For $\delta > 0$ chosen so that $\{|z - e^{i\phi_0}| < \delta\} \cap \{|z - e^{-i\phi_0}| < \delta\} = \emptyset$, we define

$$\Gamma_{\text{left}} : \theta \in [\phi_0, -\phi_0 + 2\pi] \longmapsto e^{i\theta} \quad ; \quad \gamma_{\pm,>} = \{|z - e^{\pm i\phi_0}| = \delta\} \cap \{|z| > 1\};$$
$$\Gamma_{\text{right}} : \theta \in [-\phi_0, \phi_0] \longmapsto e^{i\theta} \quad ; \quad \gamma_{\pm,<} = \{|z - e^{\pm i\phi_0}| = \delta\} \cap \{|z| < 1\}.$$

$\mathbf{C}_r$ denotes the curve whose image is the circle centered at zero of radius r. For the above curves, the orientation is anti-clockwise, and we use the same notation for the curve and its image, c.f Figure 8 as well as 9 and 10. We next define,

$$\begin{array}{lllllllll} \Gamma_{\text{outgoing}} & := \Gamma_{\text{right}} & \cup & \gamma_{+,>} & \cup & \Gamma_{\text{left}} & \cup & \gamma_{-,<} \\ \Gamma_{+} & := \Gamma_{\text{right}} & \cup & \gamma_{+,<} & \cup & \Gamma_{\text{left}} & \cup & \gamma_{-,<} \\ \Gamma_{-} & := \Gamma_{\text{right}} & \cup & \gamma_{+,>} & \cup & \Gamma_{\text{left}} & \cup & \gamma_{-,>}. \end{array} \tag{58}$$

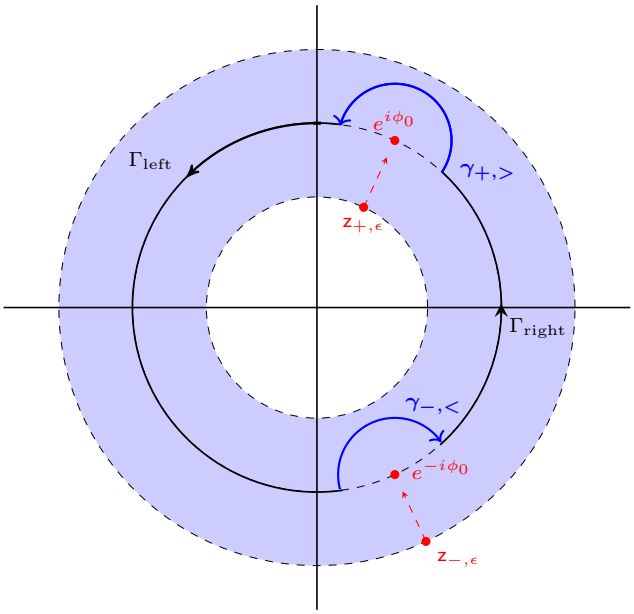


Figure 8: $\Gamma_{\text{outgoing}} = \Gamma_{\text{right}} \cup \gamma_{+,>} \cup \Gamma_{\text{left}} \cup \gamma_{-,<}$ is homotopic to $\mathbf{C}_1$ and are free poles of W_ϵ for $0 \leq \epsilon < \epsilon_0$. Hence, the contour integral along this curve gives an equivalent definition for the solutions given in Proposition 9.

Using the relative position of $\mathsf{z}_{\pm,\epsilon}$ with respect to $\mathbf{C}_1$, given Proposition 8, there exists $\epsilon_0 > 0$ such that $\mathsf{z}_{\pm,\epsilon} \notin \Gamma_{\text{outgoing}}$ for $0 \leq \epsilon < \epsilon_0$. Since Γ_{outgoing} is homotopic to $\mathbf{C}_1$, we have a second formula for U_ϵ

$$\left(U_\epsilon\right)_J = \frac{1}{2\pi i} \oint_{\Gamma_{\text{outgoing}}} W_\epsilon(z) \frac{dz}{z^{J+1}}.$$

In fact, this second formula stays valid at $\epsilon = 0$, since the following integral is valid

$$\oint_{\Gamma_{\text{outgoing}}} W_0(z) \frac{dz}{z^{J+1}},$$

and the fact that we can interchange the order, between taking limit and the contour integration, as shown in the following lemma.

Lemma 10. *We have for all $J \in \mathbb{Z}$,*

$$\lim_{\epsilon \to 0^+} \oint_{\Gamma_{outgoing}} W_\epsilon(z) \frac{dz}{z^{J+1}} = \oint_{\Gamma_{outgoing}} W_0(z) \frac{dz}{z^{J+1}}. \tag{59}$$

Proof. We recall the expression of $W_\epsilon(z)$ and $W_0(z)$,

$$W_\epsilon(z) = \frac{h}{\delta_\epsilon} \frac{Q_\epsilon(z)}{(z - \mathsf{z}_{+,\epsilon})(z - \mathsf{z}_{-,\epsilon})} H(z) \quad , \quad W_0(z) = \frac{h}{\delta_0} \frac{Q_0(z)}{(z - e^{i\phi_0})(z - e^{-i\phi_0})} H(z).$$

By its definition, the origin and $e^{\pm i\phi_0}$ do not lie on Γ_{outgoing}. Since $\mathsf{z}_{\pm,\epsilon} \to e^{\pm i\phi_0}$, there exists $\epsilon_0 > 0$ so that the family $\{\mathsf{z}_{\pm,\epsilon}\}_{0<\epsilon<\epsilon_0}$ stay away from Γ_{outgoing} and the origin. In short, there exists $C(J) > 0$ with

$$\sup_{z \in \Gamma_{\text{outgoing}}} \left| \frac{z^{-J-1}}{(z - \mathsf{z}_{+,\epsilon})(z - \mathsf{z}_{-,\epsilon})} \right| < C(J) \quad ; \quad \sup_{z \in \Gamma_{\text{outgoing}}} \left| \frac{z^{-J-1}}{(z - e^{i\phi_0})(z - e^{-i\phi_0})} \right| < C(J).$$

Similarly, $Q_\epsilon H(z)$ and $Q_0 H(z)$ are finite sum of integer powers of z and polynomials in ϵ; hence, there exists $\tilde{C}(J)$ with

$$\sup_{z\in\Gamma_{\text{outgoing}}} Q_\epsilon H < \tilde{C}(J) \quad ; \quad \sup_{z\in\Gamma_{\text{outgoing}}} Q_0 H < \tilde{C}(J).$$

As a result, we have

$$|W_\epsilon z^{-(J+1)}| \leq C(J)\tilde{C}(J) \quad , \quad |W_0 z^{-(J+1)}| \leq C(J)\tilde{C}(J) \quad ; \quad z \in \Gamma_{\text{outgoing}}.$$

To obtain (59), we next apply dominated convergence theorem to the family $W_\epsilon\, z^{-(J+1)}$, whose limit is $W_0\, z^{-(J+1)}$. □

We now can take the limit as $\epsilon \to 0^+$ on both sides of the 'blocked' Recurrence equation (27) at square wave number $\kappa_\epsilon^2 = (1+i\epsilon)\kappa^2$ and obtain a solution at square wave number κ^2 for $\kappa \in \mathbb{R}^\star$, called the outgoing solution.

Definition 3 (Outgoing solution)**.** *For $\kappa \in \mathbb{R}^\star$, and $f \in l^2_{comp}(\mathbb{Z})$ and $F \in (l^2_{comp}(\mathbb{Z}))r^r)$ with $F = \mathfrak{b}_r f$, we have*

1. *$U_{outgoing}(\kappa^2) := \lim_{\epsilon\to 0^+} U\left(\kappa_\epsilon^2\right)$ is called the outgoing solution to the 'blocked' recurrencre relation (27) at square wave number κ^2 with right hand side (r.h.s) F*
2. *$u_{outgoing}(\kappa^2) = \mathfrak{b}_r^{-1} U_{outgoing}(\kappa^2)$ is called the outgoing solution to the system (8)-(9) at square wave number κ^2 with r.h.s f.*

In addition, as a result of Lemma 10, the above solution can be written in the form of a contour integral.

Proposition 11. *For $\kappa \in \mathbb{R}^\star$, the outgoing solutions defined in Definition 3 are given by,*

$$\left(U_{outgoing}(\kappa^2)\right)_J = \frac{1}{2\pi i} \oint_{\Gamma_{outgoing}} W_0(z)\, \frac{dz}{z^{J+1}}\, dz\,;$$

$$u_{outgoing}(\kappa^2) = \frac{1}{2\pi i} \mathfrak{b}_r^{-1} \oint_{\Gamma_{outgoing}} W_0(z)\, \frac{dz}{z^{J+1}}\, dz\,.$$

5.3 Formula of the 'blocked' outgoing solutions

We can obtain an explicit calculation of U_{outgoing} by using the definition in Proposition 11 and Cauchy residue theorem.

Lemma 12. *Consider a sequence F of compact support with* $\operatorname{Supp} F \subset [N_{\min}, N_{\max}]$. *Define $H(z) := Z_{\mathfrak{B}} F$, i.e. $H(z) = \sum_{k=N_{\min}}^{N_{\max}} z^k\, F_k$. The outgoing solution $U_{outgoing}$ to (27) at κ^2 with $\kappa \in \mathbb{R}^\star$,*

$$A(\kappa^2 h^2)\, U_{J-1} + B(\kappa^2 h^2)\, U_J + A^t(\kappa^2 h^2)\, U_J = h\, F_J.$$

is given by

1. *For $J < N_{\min}$,*

$$(U_{outgoing})_J = \frac{h}{\delta_0} \frac{Q_0(e^{i\phi_0})\, H(e^{i\phi_0})}{2\, i \sin\phi_0}\, e^{-i\,\phi_0\,(J+1)}.$$

2. For $J > N_{\max}$,

$$(U_{outgoing})_J = \frac{h}{\delta_0}\,\frac{Q_0(e^{-i\phi_0})\,H(e^{-i\phi_0})}{2\,i\sin\phi_0}\,e^{i\,\phi_0\,(J+1)}.$$

3. For $N_{\min} < J < N_{\max}$

$$\begin{aligned}(U_{outgoing})_J = {} & \frac{h}{\delta_0}\,\frac{Q_0(e^{-i\phi_0})\,e^{i\phi_0}}{2\,i\sin\phi_0}\sum_{k=N_{\min}}^{J-1} e^{i(J-k)\,\phi_0}F_k \\ & + \frac{h}{2\,\delta_0}\left[\frac{Q_0(e^{-i\phi_0})\,e^{i\phi_0}+Q_0(e^{i\phi_0})\,e^{-i\phi_0}}{2\,i\sin\phi_0} + \ Q_0(0)+Q_0^t(0)\right]F_J \\ & + \frac{h}{\delta_0}\,\frac{Q_0(e^{i\phi_0})\,e^{-i\phi_0}}{2\,i\sin\phi_0}\sum_{k=J+1}^{N_{\max}} e^{-i(J-k)\,\phi_0}F_k.\end{aligned}$$

4. For $J = N_{\min}$ or $J = N_{\max}$,

$$\begin{aligned}(U_{outgoing})_{N_{\min}} = {} & \frac{h}{2\,\delta_0}\left[\frac{Q_0(e^{-i\phi_0})\,e^{i\phi_0}+Q_0(e^{i\phi_0})\,e^{-i\phi_0}}{2\,i\ \sin\phi_0} + Q_0(0)+Q_0^t(0)\right]F_{N_{\min}} \\ & + \frac{h}{\delta_0}\,\frac{Q_0(e^{i\phi_0})}{2\,i\ \sin\phi_0}\,e^{-i\phi_0}\sum_{k=N_{\min}+1}^{N_{\max}} e^{-i(J-k)\,\phi_0}F_k.\end{aligned}$$

$$\begin{aligned}(U_{outgoing})_{N_{\max}} = {} & \frac{h}{\delta_0}\,\frac{Q_0(e^{-i\phi_0})}{2\,i\ \sin\phi_0}\,e^{i\phi_0}\sum_{k=N_{\min}}^{N_{\max}-1} e^{i(J-k)\,\phi_0}F_{N_{\max}} \\ & + \frac{h}{2\delta_0}\left[\frac{Q_0(e^{-i\phi_0})\,e^{i\phi_0}+Q_0(e^{i\phi_0})\,e^{-i\phi_0}}{2\,i\ \sin\phi_0} + Q_0(0)+Q_0^t(0)\right]F_0.\end{aligned}$$

Proof. **Case 1 with $J < N_{\min}$:** Since $N_{\min} - J \geq 1$, for all $j \in \mathbb{Z}$ with $N_{\min} \leq j \leq N_{\max}$,

$$j - J - 1 \geq N_{\min} - J - 1 \geq 0.$$

In addition, Q_0, $\mathbf{q}_0$ and H is analytic at $z = 0$. As a result, for such J,

$$z^{-J-1}W_0\,dz = h\,\frac{Q_0(z)}{\delta_0\,\mathbf{q}(z)}z^{-J-1}\left(\sum_{j=N_{\min}}^{N_{\max}} z^j\,(f_{j,k})_{0\leq k\leq r-1}\right)dz\,.$$

is also analytic at $z = 0$. Because of of this, we deform Γ_{outgoing} into Γ_+, by deforming $\gamma_{+,>}$ to the opposite side of the unit disk into $\gamma_{+,<}$, crossing the pole at $e^{i\phi_0}$; see the following figure.

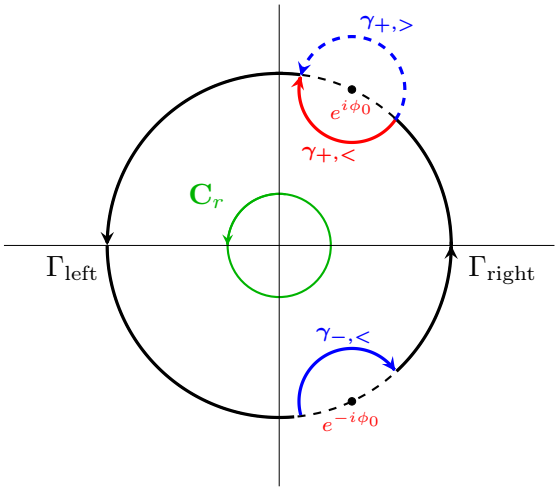


Figure 9: Deform Γ_{outgoing} to Γ_+. $\Gamma_+ := \Gamma_{\text{right}} \cup \boldsymbol{\gamma}_{+,<} \cup \Gamma_{\text{left}} \cup \boldsymbol{\gamma}_{-,<}$ Γ_+ is homotopic to $\mathbf{C}_r$, $r < 1$.

$$
\begin{aligned}
(U_{\text{outgoing}})_J &= \frac{1}{2\pi i} \oint_{\Gamma_{\text{outgoing}}} W_0(z)\, z^{-(J+1)}\, dz \\
&= \frac{1}{2\pi i} \oint_{\Gamma_+} W_0(z)\, z^{-(J+1)}\, dz \quad + \quad \frac{1}{2\pi i} \oint_{\gamma_{+,>} \cup -\gamma_{+,<}} W_0(z)\, z^{-(J+1)}\, dz \\
&= \frac{1}{2\pi i} \oint_{\mathbf{C}_r} W_0(z)\, z^{-(J+1)}\, dz \quad + \quad \text{Res}\left(W_0 z^{-(J+1)}, e^{i\phi_0}\right) \quad , \quad r < 1 \\
&= \text{Res}\left(W_0\, z^{-(J+1)}, e^{i\phi_0}\right) \quad = \quad \frac{h}{\delta_0} \frac{Q_0(e^{i\phi_0})\, H(e^{i\phi_0})}{e^{i\phi_0} - e^{-i\phi_0}} e^{-i\phi_0(J+1)} \\
&= \frac{h}{\delta_0} \frac{Q_0(e^{i\phi_0})\, H(e^{i\phi_0})}{2\, i \sin\phi_0} e^{-i\phi_0(J+1)}.
\end{aligned}
$$

Case 2 with $J > N_{\max}$: With $\tilde{z} := z^{-1}$, by Lemma 13, we have

$$
z^{-J} W_0(z) \frac{dz}{z} = \frac{h}{\delta_0} \frac{Q_0^t(\tilde{z})}{\mathbf{q}_0(\tilde{z})} \tilde{z}^J \sum_{j=0}^{M} \tilde{z}^{-j} \left(f_{j,k}\right)_{0 \le k \le r-1} \frac{d\tilde{z}}{\tilde{z}}.
$$

If $J > N_{\max}$ then $J - N_{\max} \ge 1$. Thus for all $j \in \mathbb{Z}$ with $N_{\min} \le j \le N_{\max}$,

$$
J - j - 1 \ge J - N_{\max} - 1 \ge 0.
$$

As a result, the factor of $d\tilde{z}$ in the above expression is analytic at $\tilde{z} = 0$. Because of of this, we deform Γ_{outgoing} into Γ_-, by deforming $\gamma_{-,<}$ to the opposite side of unit disk into $\gamma_{-,>}$, crossing the simple pole at $e^{-i\gamma_0}$; see the following figure.

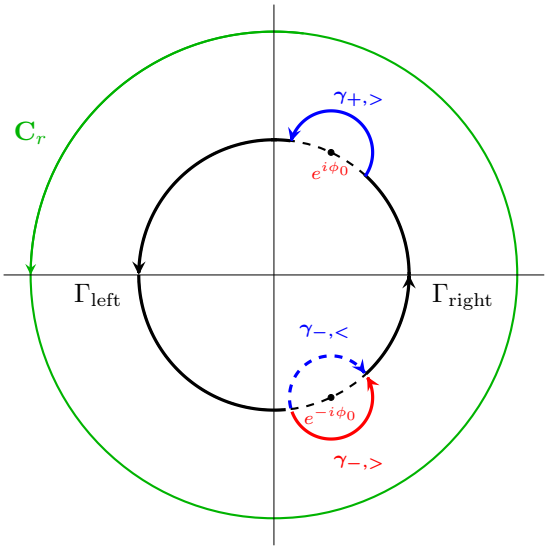


Figure 10: Deform Γ_{outgoing} to Γ_- ; $\Gamma_- := \Gamma_{\text{right}} \cup \boldsymbol{\gamma}_{+,>} \cup \Gamma_{\text{left}} \cup \boldsymbol{\gamma}_{-,>}$ Γ_- is homotopic to $\mathbf{C}_r$, $r > 1$.

$$
\begin{aligned}
(U_{\text{outgoing}})_J &= \frac{1}{2\pi i} \oint_{\Gamma_{\text{outgoing}}} W_0(z)\, z^{-(J+1)}\, dz \\
&= \frac{1}{2\pi i} \oint_{\Gamma_{\text{outgoing}}} W_0(z)\, z^{-(J+1)}\, dz \quad + \quad \frac{1}{2\pi i} \oint_{\gamma_{-,<} \cup -\gamma_{-,>}} W_0(z) z^{-(J+1)} dz \\
&= \frac{1}{2\pi i} \oint_{\mathbf{C}_r} W_0(z)\, z^{-(J+1)}\, dz \quad - \quad \text{Res}\left(W_0\, z^{-(J+1)}, e^{-i\phi_0}\right) \quad , \quad 1 < r \\
&= \frac{1}{2\pi i} \oint_{\mathbf{C}_{r^{-1}}} \frac{h}{\delta_0}\, \tilde{z}^J\, \frac{Q_0^t(\tilde{z})}{\mathbf{q}_0(\tilde{z})} H(\tilde{z}^{-1})\, \frac{d\tilde{z}}{\tilde{z}} - \quad \text{Res}\left(W_0\, z^{-(J+1)}, e^{-i\phi_0}\right) \quad , \quad 1 < r \\
&= -\,\text{Res}\left(W_0\, z^{-(J+1)}, e^{-i\phi_0}\right) \quad = \quad \frac{h}{\delta_0}\, \frac{Q_0(e^{-i\phi_0})\, H(e^{-i\phi_0})}{2\, i \sin\phi_0}\, e^{i\,\phi_0\,(J+1)}.
\end{aligned}
$$

Note that the orientation of $\mathbf{C}_x$ is chosen to be anti-clockwise.

Case 3a with $N_{\min} < J < N_{\max}$: We break up $z^{-J} W_0 \frac{dz}{z}$ into three terms as follows,

$$
\begin{aligned}
& z^{-J} W_0 \frac{dz}{z} \\
&= \frac{h}{\delta_0} \left(\frac{Q_0(z)}{\mathbf{q}_0(z)} z^{-J} \sum_{j=N_{\min}}^{J-1} z^j F_j \frac{dz}{z} \quad + \frac{Q_0(z)}{\mathbf{q}_0(z)} H_J \frac{dz}{z} \quad + \frac{Q_0(z)}{\mathbf{q}_0(z)} z^{-J} \sum_{j=J+1}^{N_{\max}} z^j F_j \frac{dz}{z} \right).
\end{aligned} \tag{60}
$$

Since $\dfrac{Q_0(z)}{\mathbf{q}(z)} = \dfrac{Q_0^t(\tilde{z})}{\mathbf{q}(\tilde{z})}$ by Lemma 13, and $z^{-1}\, dz = \tilde{z}^{-1} d\tilde{z}$, the first term of the RHS in the above expression is analytic at $\tilde{z} = 0$, thus can be calculated by using the choice of contour

deformation in Case 2.

$$\oint_{\Gamma_{\text{outoing}}} \frac{Q_0(z)}{\mathbf{q}(z)} z^{-J-1} \sum_{j=N_{\min}}^{J-1} z^j F_j \, dz = -2\pi i \text{ Res} \left(\frac{Q_0(z)}{z\,\mathbf{q}_0(z)} z^{-J} \sum_{j=N_{\min}}^{J-1} z^j F_j \, , \, e^{-i\phi_0} \right)$$
$$= \pi \, \frac{Q_0(e^{-i\phi_0})}{\sin\phi_0} \, e^{i(J+1)\phi_0} \sum_{j=N_{\min}}^{J-1} e^{-i\,j\,\phi_0} F_j.$$

The third term in the RHS of (60) is analytic at $z = 0$, and thus is calculated by using the choice of contour deformation in Case 1.

$$\oint_{\Gamma_{\text{outoing}}} \frac{Q_0(z)}{\mathbf{q}_0(z)} z^{-J-1} \sum_{j=J+1}^{N_{\max}} z^j F_j \, dz = 2\pi i \text{ Res} \left(\frac{Q_0(z)}{\mathbf{q}_0(z)} z^{-J-1} \sum_{j=J+1}^{N_{\max}} z^j F_j \, , e^{i\phi_0} \right)$$
$$= \pi \, \frac{Q_0(e^{i\phi_0})}{\sin\phi_0} e^{-i(J+1)\phi_0} \sum_{j=J+1}^{N_{\max}} e^{i\,j\,\phi_0} F_j.$$

It remains to calculate the second term in (60). This term can be calculated in two ways, either by reading off its residue in z at $\{0\,,\, e^{i\phi_0}\}$ or at $\{\infty\,,\, e^{-i\phi_0}\}$. The first choice gives,

$$\begin{aligned} \oint_{\Gamma_{\text{outoing}}} \frac{Q_0(z)}{\mathbf{q}_0(z)} \, H_J \, \frac{dz}{z} &= 2\pi i \text{ Res} \left(\frac{Q_0(z)}{z\,\mathbf{q}_0(z)} \, H_J \, , \, e^{i\phi_0} \right) + 2\pi i \text{ Res} \left(\frac{Q_0(z)}{z\,\mathbf{q}_0(z)} \, H_J \, , \, 0 \right) \\ &= \pi \, \frac{Q_0(e^{i\phi_0})}{\sin\phi_0} e^{-i\phi_0} \, H_J + 2\pi \, i \, Q_0(0) H_J. \end{aligned} \tag{61}$$

By the change of variable given in Lemma 13, we obtain the second equivalent version,

$$\begin{aligned} \oint_{\Gamma_{\text{outoing}}} \frac{Q_0(z)}{\mathbf{q}_0(z)} \, H_J \, \frac{dz}{z} &= \int_{\mathbf{C}_r} \frac{Q_0(z)}{\mathbf{q}_0(z)} \, H_J \, \frac{dz}{z} - 2\pi i \, \text{Res} \left(\frac{Q_0(z)}{\mathbf{q}_0(z)} \, H_J \, \frac{1}{z} \, , \, e^{-i\phi_0} \right) \\ &= \int_{\mathbf{C}_{r^{-1}}} \frac{Q_0^t(\tilde{z})}{\mathbf{q}_0(\tilde{z})} \, H_J \, \frac{d\tilde{z}}{\tilde{z}} - 2\pi \, i \text{ Res} \left(\frac{Q_0(z)}{\mathbf{q}_0(z)} \, H_J \, \frac{1}{z} \, , \, e^{-i\phi_0} \right) \\ &= 2\pi i \, Q_0^t(0) \, H_J + \pi \, \frac{Q_0(e^{-i\phi_0})}{\sin\phi_0} e^{i\phi_0} H_J. \end{aligned} \tag{62}$$

We combine the two versions to obtain,

$$\oint_{\Gamma_{\text{outoing}}} \frac{Q_0(z)}{\mathbf{q}_0(z)} \, H_J \, \frac{dz}{z} = \pi \, \left(\frac{Q_0(e^{i\phi_0})e^{-i\phi_0} + Q_0(e^{-i\phi_0})e^{i\phi_0}}{2\sin\phi_0} + iQ_0(0) + iQ^t(0) \right) H_J. \tag{63}$$

Case 3b with $J = N_{\min}$ or $J = N_{\max}$: The calculation for these two cases can be read off from that for Case 3a, with a difference being that expression (60), for these cases, will contain only 2 terms; explicitly

$$W_0 z^{-N_{\min}} \frac{dz}{z} = \frac{h}{\delta_0} \left[\frac{Q_0(z)}{\mathbf{q}_0(z)} \, H_{N_{\min}} \, \frac{dz}{z} + \frac{Q_0(z)}{\mathbf{q}_0(z)} z^{-J} \sum_{j=N_{\min}+1}^{N_{\max}} z^j F_j \frac{dz}{z} \right].$$

and

$$W_0 z^{-N_{\max}} \frac{dz}{z} = \frac{h}{\delta_0} \left[\frac{Q_0(z)}{\mathbf{q}_0(z)} z^{-J} \sum_{j=N_{\min}}^{N_{\max}-1} z^j F_j \frac{dz}{z} \quad + \frac{Q_0(z)}{\mathbf{q}_0(z)} \, H_{N_{\max}} \, \frac{dz}{z} \right].$$

□

The following calculation applies for both $\epsilon > 0$ and $\epsilon = 0$, thus we will drop this subscript.

Lemma 13. *With $\tilde{z} = z^{-1}$, we have the following identities.*

$$\frac{Q(z)}{\mathbf{q}(z)} = \frac{Q^t(\tilde{z})}{\mathbf{q}(\tilde{z})} \quad ; \quad z^{-J} W(z) \frac{dz}{z} = h\, \tilde{z}^J\, \frac{Q^t(\tilde{z})}{\mathbf{q}(\tilde{z})}\, H(\tilde{z}^{-1}).$$

Proof. Recall that $\operatorname{Adj} \mathcal{A}(z) = z^{r-2} Q(z)$ and $\det(\tilde{z}^2 A + \tilde{z} B + A^t) = z^{r-1} \mathbf{q}(z)$, c.f. Lemma 37, thus

$$\frac{\operatorname{Adj} \mathcal{A}(z)}{\det \mathcal{A}(z)} = \frac{Q(z)}{\tilde{z}\, \mathbf{q}(z)}. \tag{64}$$

We next write $\frac{\operatorname{Adj} \mathcal{A}(z)}{\det \mathcal{A}(z)}$ in terms of $\tilde{z} = z^{-1}$,

$$\begin{aligned} \frac{\operatorname{Adj} \mathcal{A}(z)}{\det \mathcal{A}(z)} &= \frac{\operatorname{Adj}(\tilde{z}^{-2} A + \tilde{z}^{-1} B + A^t)}{\det(\tilde{z}^{-2} A + \tilde{z} B + A^t)} \\ &= \frac{\tilde{z}^{-2(r-1)} \operatorname{Adj}(\tilde{z} A^t + \tilde{z} B + A)}{\tilde{z}^{-2r} \det(\tilde{z}^2 A^t + \tilde{z} B + A)} \quad = \tilde{z}^2 \frac{\operatorname{Adj}(\tilde{z} A^t + \tilde{z} B + A)}{\det(\tilde{z}^2 A^t + \tilde{z} B + A)} \\ &= \tilde{z}^2 \left[\frac{\operatorname{Adj}(\tilde{z} A + \tilde{z} B + A^t)}{\det(\tilde{z}^2 A + \tilde{z} B + A^t)} \right]^t \quad \overset{(64)}{=} \tilde{z} \frac{Q^t(\tilde{z})}{\mathbf{q}(\tilde{z})}. \end{aligned}$$

We deduce immediately the first identity. As for the second one, we can write $z^{-J} W(z)$ in terms of $\tilde{z}$ as

$$z^{-J} W(z) = h\, z^{-J}\, \frac{\operatorname{Adj}(z^2 A + z B + A^t)}{\det(z^2 A + z B + A^t)}\, z H(z) = h\, \tilde{z}^J\, \tilde{z} \frac{Q^t(\tilde{z})}{\mathbf{q}(\tilde{z})} \tilde{z}^{-1} H(\tilde{z}^{-1}).$$

And thus

$$z^{-J} W(z) \frac{dz}{z} = h\, \tilde{z}^J\, \frac{Q^t(\tilde{z})}{\mathbf{q}(\tilde{z})} H(\tilde{z}^{-1})\, \frac{d\tilde{z}}{\tilde{z}}.$$

□

5.4 The outgoing discrete 'blocked' Green's function

We have denoted by $\mathfrak{M}(\mathbb{Z}, S)$ the set of sequences taking value in S and indexed by $\mathbb{Z}$, c.f. SubSection (3.2). Denote by $\mathbf{M}_{m\times n}(\mathbb{F})$ the set of matrices of size $m \times n$ with $\mathbb{F}$-valued entries.

For a sequence $B \in \mathfrak{M}\big(\mathbb{Z}, \mathbf{M}_{m\times n}(\mathbb{F})\big)$ and sequence $V \in \mathfrak{M}\big(\mathbb{Z}, \mathbf{M}_{n\times k}(\mathbb{F})\big)$, the discrete convolution $B \underset{\text{discrete}}{\star} V$ is defined to be the sequence in $\mathfrak{M}\big(\mathbb{Z}, \mathbf{M}_{m\times k}(\mathbb{F})\big)$ with

$$\big(B \underset{\text{discrete}}{\star} V\big)_J := \sum_{K\in\mathbb{Z}} B_{J-K}\, V_K \tag{65}$$

under suitable hypothesis on the support of B and V.

Definition 4 (Outgoing Discrete 'blocked' Green's function)**.** *For $\kappa \in \mathbb{R}^\star$, the outgoing discrete 'blocked' Green's function for the 'blocked' recurrence relation* (27) *at κ^2, denoted by $G_{outgoing}(\kappa^2)$, is a sequence in $\mathfrak{M}(\mathbb{Z}, \mathbf{M}_{r\times r}(\mathbb{C}))$ with the property*

$$G_{outgoing}(\kappa^2) \underset{discrete}{\star} F = U_{outgoing}(\kappa^2),$$

where $U_{outgoing}$ is the outgoing solution to (27) *with F as the right hand side, defined in Definition 3.* ▽

Given the explicit calculation of outgoing solution U_{outgoing} by Lemma 12, we now obtain an explicit formula for G_{outgoing}.

Proposition 14.

$$(G_{outgoing})_J = \frac{h}{2\,i\,\delta_0\,\sin\phi_0} e^{i\,|J|\,\phi_0} \left(\tilde{G}_{outgoing}\right)_J ; \tag{66}$$

where

$$\left(\tilde{G}_{outgoing}\right)_J = \begin{cases} Q_0(e^{i\phi_0})\,e^{-i\,\phi_0} & , J<0 \\ \frac{1}{2}\big[Q_0(e^{-i\phi_0})\,e^{i\phi_0} + Q_0(e^{i\phi_0})\,e^{-i\phi_0}\big] + i\big[Q_0(0) + Q_0^t(0)\big]\,\sin\phi_0 & , J=0 \\ Q_0(e^{-i\phi_0})\,e^{i\,\phi_0} & , J>0 \end{cases} .$$

Proof. Denote by G the matrix-valued sequence defined by the RHS of (66). Consider a sequence F of compact support with $\operatorname{Supp} F \subset [N_{\min}, N_{\max}]$. Denote by $H(z) := Z_{\mathfrak{B}}\,F$, i.e. $H(z)$ is given by,

$$H(z) = \sum_{J=N_{\min}}^{N_{\max}} z^J F_J \quad , \quad F_J := (f_{J,k})_{0\le k\le r-1} .$$

We will show that sequence $U := G \underset{\text{discrete}}{\star} F$ is equal to the outgoing solution U_{outgoing} (corresponding to the same F), by comparing with the explicit form of latter given in Lemma 12.

By definition,

$$U_J := \sum_{J\in\mathbb{Z}} G_{J-K}\,F_K = \sum_{K=N_{\min}}^{N_{\max}} G_{J-K}\,F_K .$$

Case 1 : If $J < N_{\min}$ then $J-K<0$ for all $k\in\mathbb{Z}$ with $N_{\min}\le K\le N_{\max}$. As a result,

$$\begin{aligned} U_J &= \sum_{K=N_{\min}}^{N_{\max}} \frac{h}{2\,i\,\delta_0\sin\phi_0}\,Q_0(e^{i\phi_0})\,e^{-i(J-K+1)\phi_0}F_K \\ &= \frac{h}{\delta_0}\,\frac{Q_0(e^{i\phi_0})}{2\,i\sin\phi_0} e^{-i(J+1)\phi_0} \sum_{K=N_{\min}}^{N_{\max}} e^{i\,K\,\phi_0}F_K \\ &= \frac{h}{\delta_0}\,\frac{Q_0(e^{i\phi_0})\,H(e^{i\phi_0})}{2\,i\sin\phi_0}\,e^{-i\,\phi_0\,(J+1)} \quad = \quad (U_{\text{outgoing}})_J . \end{aligned}$$

Case 2 : If $J > N_{\max}$ then $J-K>0$ for all $K\in\mathbb{Z}$ with $N_{\min}\le K\le N_{\max}$. As a result,

$$\begin{aligned} U_J &= \sum_{K=N_{\min}}^{N_{\max}} \frac{h}{2\,i\,\delta_0\,\sin\phi_0}\,Q_0(e^{-i\phi_0})\,e^{i(J-K+1)\phi_0}F_K \\ &= \frac{h}{\delta_0}\,\frac{Q_0(e^{-i\phi_0})}{2\,i\sin\phi_0} e^{i(J+1)\phi_0} \sum_{K=N_{\min}}^{N_{\max}} e^{-i\,K\,\phi_0}F_K \\ &= \frac{h}{\delta_0}\,\frac{Q_0(e^{-i\phi_0})\,H(e^{-i\phi_0})}{2\,i\sin\phi_0}\,e^{i\,\phi_0(J+1)} \quad = \quad (U_{\text{outgoing}})_J . \end{aligned}$$

Case 3 : If $N_{\min}\le J\le N_{\max}$, we split the expression defining U into

$$U_{N_{\min}} = G_0\,F_{N_{\min}} + \sum_{K=N_{\min}+1}^{N_{\max}} G_{N_{\min}-K}\,F_K .$$

$$U_{N_{\max}} = \sum_{K=N_{\min}}^{N_{\max}-1} G_{N_{\max}-K}\, F_K + G_0\, F_{N_{\max}}.$$

And for $N_{\min} < J < N_{\max}$,

$$U_J = \sum_{K=N_{\min}}^{J-1} G_{J-K}\, F_K \quad + \quad G_0\, F_J \quad + \quad \sum_{K=J+1}^{N_{\max}} G_{J-K}\, F_K.$$

We will carry out the calculation for the more general case $N_{\min} < J < N_{\max}$. From this, the result for $J = N_{\min}$ and $J = N_{\max}$ can be readily deduced.

For J with $N_{\min} < J < N_{\max}$, we have

$$\begin{aligned} U_J = & \frac{h}{\delta_0}\frac{Q_0(e^{i\phi_0})}{2\,i\sin\phi_0} e^{-i(J+1)\phi_0} \sum_{K=N_{\min}}^{J-1} e^{i\,K\,\phi_0} F_k \\ & + \frac{h}{2\,\delta_0}\left(\frac{Q_0(e^{-i\phi_0})\,e^{i\phi_0} + Q_0(e^{i\phi_0})\,e^{-i\phi_0}}{2\,i\sin\phi_0} + Q_0(0) + Q_0^t(0)\right) F_J \\ & + \frac{h}{\delta_0}\frac{Q_0(e^{-i\phi_0})}{2\,i\sin\phi_0} e^{i(J+1)\phi_0} \sum_{K=J+1}^{N_{\max}} e^{-i\,K\,\phi_0} F_k \quad = \quad (U_{\text{outgoing}})_J. \end{aligned}$$

This finishes the proof. □

5.5 Formula for the scalar outgoing discrete solution

In previous sections, for $\kappa \in \mathbb{R}^\star$ and sequence $F \in (l^2_{\text{comp}}(\mathbb{Z}))^r$, we have written the outgoing 'blocked' solution $U_{\text{outgoing}}(\kappa^2)$, c.f Definition 3, to the 'blocked' recurrence relation(27) with right hand side (r.h.s) F,

$$A(w)\,U_{J-1} + B(w)\,U_J + A^t(w)\,U_{J+1} = h\,F_J \quad , \quad w = \kappa^2 h^2\,,$$

in terms of the outgoing discrete 'blocked' Green's function $G_{\text{outgoing}}(\kappa^2)$ computed in (66),

$$U_{\text{outgoing}}(\kappa^2) = G_{\text{outgoing}}(\kappa^2) \underset{\text{discrete}}{\star} F.$$

Using the equivalence in Proposition 1, we 'unblock' $U_{\text{outgoing}}(\kappa^2)$ to obtain a solution of the original system of the recurrence relation (8)-(9) with r.h.s given by $f = \mathfrak{b}_r^{-1} F \in \mathfrak{M}(\mathbb{Z} \times \mathbb{Z}_r, \mathbb{C})$, i.e

$$F_J = (f_{J,l})_{0 \le l \le r-1}\,,\ J \in \mathbb{Z}.$$

Under the current hypothesis on F, $f \in l^2_{\text{comp}}(\mathbb{Z})$. We have denoted this solution by $u_{\text{outgoing}}(\kappa^2)$, c.f Definition 3 and Proposition 11. In terms of G_{outgoing}, the outgoing solution $u_{\text{outgoing}}(\kappa^2)$ can be written as,

$$\begin{aligned} u_{\text{outgoing}}(\kappa^2)_{J,k} &= \pi_{k+1}\, U_{\text{outgoing}}(\kappa^2)_J \\ &= \pi_{k+1}\left(G_{\text{outgoing}}(\kappa^2) \underset{\text{discrete}}{\star} F\right)_J \\ &\overset{\text{c.f (65)}}{=} e_{k+1} \cdot \sum_{I \in \mathbb{Z}} G_{\text{outgoing}}(\kappa^2)_{J-I}\, F_I \\ &= \sum_{I \in \mathbb{Z}} \sum_{0 \le l \le r-1} \left(G_{\text{outgoing}}(\kappa^2)^t_{J-I}\right)_{(l+1)(k+1)}\, f_{I,l}. \end{aligned} \tag{67}$$

We recall that $G_{\text{outgoing}}(\kappa^2)^t_{J-I}$ is a matrix of size $r\times r$, c.f Proposition 14, and $\big(G_{\text{outgoing}}(\kappa^2)^t_{J-I}\big)_{kl}$ is the (k,l)-th entry of matrix.

As a result, we have

Proposition 15. *For $\kappa\in\mathbb{R}^\star$ and $f\in l^2_{comp}(\mathbb{Z})$, the outgoing solution $u_{outgoing}(\kappa^2)$ to the system of the recurrence relation (8)-(9) with r.h.s f is given by,*

$$u_{outgoing}(\kappa^2)_{J,k}=\sum_{(\tilde J,\tilde k)\in\mathbb{Z}\times\mathbb{Z}_r} g_{outgoing}(\kappa^2)_{(J,k),(\tilde J,\tilde k)}\, f_{\tilde J,\tilde k}\,, \tag{68}$$

where scalar sequence $g_{outgoing}(\kappa^2)$ indexed by $\mathbb{Z}\times\mathbb{Z}_r\times\mathbb{Z}\times\mathbb{Z}_r$ is defined in terms of $G_{outgoing}$, c.f (68),

$$g_{outgoing}(\kappa^2)_{(J_1,k_1),(J_2,k_2)}:=\big(G_{outgoing}(\kappa^2)^t_{J_1-J_2}\big)_{(k_2+1)(k_1+1)}\,. \tag{69}$$

Suppose $\operatorname{Supp} F\subset[N_{\min},N_{\max}]$. We consider the nodal points at $x=Jh$ with $J>N_{\max}$ and $J<N_{\min}$. In this case, we will use the corresponding formula of G_{outgoing}, c.f Proposition 14,

$$\begin{aligned}
G_{\text{outgoing}}(\kappa^2)_J&=\frac{h}{2\,i\,\delta_0\,\sin\phi_0}e^{i\,|J|\,\phi_0}Q_0(e^{i\phi_0})\,e^{-i\,\phi_0}\quad,\quad J<0;\\
G_{\text{outgoing}}(\kappa^2)_J&=\frac{h}{2\,i\,\delta_0\,\sin\phi_0}e^{i\,|J|\,\phi_0}Q_0(e^{-i\phi_0})\,e^{i\,\phi_0}\quad,\quad J>0.
\end{aligned}$$

in the formula for the scalar solution given by (68), we obtain

$$\begin{aligned}
u_{\text{outgoing}}(\kappa^2)_{J,0}&=\frac{h\,e^{iJ\phi_0}}{2\,i\,\delta_0\,\sin\phi_0}\sum_{\substack{0\le l\le r-1\\ N_{\min}\le\tilde J\le N_{\max}}} e^{i\,(1-\tilde J)\,\phi_0}\,f_{\tilde J,l}\,\big[Q_0^t(e^{-i\,\phi_0})\big]_{(l+1)1}\quad,J>N_{\max};\\
u_{\text{outgoing}}(\kappa^2)_{J,0}&=\frac{h\,e^{-iJ\phi_0}}{2\,i\,\delta_0\,\sin\phi_0}\sum_{\substack{0\le l\le r-1\\ N_{\min}\le\tilde J\le N_{\max}}} e^{i\,(\tilde J-1)\,\phi_0}\,f_{\tilde J,l}\,\big[Q_0^t(e^{i\,\phi_0})\big]_{(l+1)1}\quad,J<N_{\min}.
\end{aligned} \tag{70}$$

6 Dispersion Analysis

6.1 Numerical Wavenumber

After having obtain the approximating solution in Subsection 5.5, we now compare its form with that the form of the analytic (/exact) outgoing solution at large x, to determine the role played by the wave number k in their oscillatory behavior.

Exact solution For $\kappa>0$ and $\mathsf{f}\in L^2_c(\mathbb{R})$, from the discussion in Section 2, we recall analytic problem,

$$\text{find}\ \ \mathsf{u}\in H^2_{\text{loc}}(\mathbb{R})\ \text{satisfying}\begin{cases}-\mathsf{u}''-\kappa^2\mathsf{u}=\mathsf{f}\ \text{ in }\ \mathcal{D}'(\mathbb{R});\\ \lim_{x\to\pm\infty}|\frac{1}{i}\mathsf{u}'(x)\mp\kappa\,\mathsf{u}(x)|=0\end{cases}.$$

with unique solution, denoted by $\mathsf{u}_{\text{outgoing}}$ given by the convolution of the outgoing Green's function $\mathsf{G}_{\text{outgoing}}$ with f,

$$\mathsf{u}_{\text{outgoing}}(x)=\int_{-\infty}^{\infty}\mathsf{G}_{\text{outgoing}}(x-y)\,\mathsf{f}(y)\,dy\quad,\quad\text{where }\mathsf{G}_{\text{outgoing}}(x)=\frac{i\,e^{i\,\kappa\,|x|}}{2\kappa}.$$

Suppose $\mathrm{Supp}\,\mathsf{f} = [a,b]$. When $x > b$, $x \notin \mathrm{Supp}\,\mathsf{f}$, then

$$\begin{aligned}\mathsf{u}_{\text{outgoing}}(x) &= \int_{-\infty}^{x} \mathsf{G}_{\text{outgoing}}(x-y)\,\mathsf{f}(y)\,dy + \int_{x}^{\infty} \underbrace{\mathsf{G}_{\text{outgoing}}(x-y)\,\mathsf{f}(y)}_{0}\,dy \\ &= \int_{-\infty}^{x} \frac{i}{2\kappa} e^{i\,\kappa\,(x-y)}\,\mathsf{f}(y)\,dy \quad = \frac{i}{2\kappa} e^{i\,\kappa\,x} \int_{-\infty}^{\infty} e^{-i\,\kappa\,y}\,\mathsf{f}(y)\,dy.\end{aligned}$$

We carry out the same computation for $x < a$, to obtain that

$$\mathsf{u}_{\text{outgoing}}(x) = \begin{cases} \frac{i}{2\kappa} e^{i\,\kappa\,x}\,\hat{\mathsf{f}}(\kappa) & ,x > b \\ \frac{i}{2\kappa} e^{-i\,\kappa\,x}\,\hat{\mathsf{f}}(-\kappa) & ,x < a \end{cases} \quad ; \quad \hat{f}(\xi) := \int_{-\infty}^{\infty} e^{-ix\xi} f(x)\,dx.$$

In particular, for $x_J = Jh$ with $|J|$ large, we have

$$\mathsf{u}_{\text{outgoing}}(x_J) = \frac{i}{2\kappa} e^{i\,\mathrm{sgn}(J)\,J\,h\,\kappa}\;\hat{\mathsf{f}}(\mathrm{sgn}(J)\kappa). \tag{71}$$

Here sgn is the signed function.

Discretized solution With discretization step size h, using the technique of Garlekin Finite Element of order r, the Helmholtz equation $-\mathsf{u}'' - \kappa^2\mathsf{u} = \mathsf{f}$ is discretized to give the system of recurrence relations (8)-(9) at square wave number κ^2, with the r.h.s being the approximating sequence f created from f by (5). A solution to the latter is given by the sequence $u_{\text{outgoing}}(\kappa^2)$, with $u_{\text{outgoing}}(\kappa^2)_{J,k}$ approximating the analytic solution $\mathsf{u}_{\text{outgoing}}$ at $x = Jh + k\frac{r}{h}$, $J \in \mathbb{Z}$ and $0 \le k \le r-1$.

In particular, at $x_J = Jh$, the value of the analytic solution is approximated by, c.f (70)

$$u_{\text{outgoing}}(\kappa^2)_{J,0} = \frac{h\,e^{iJ\phi_0}}{2\,i\,\delta_0\,\sin\phi_0} \sum_{\substack{0\le l\le r-1 \\ N_{\min}\le \tilde{J}\le N_{\max}}} e^{i\,(1-\tilde{J})\,\phi_0}\,f_{\tilde{J},l}\left[Q_0^t(e^{-i\,\phi_0})\right]_{(l+1)1} \quad ,J > N_{\max};$$

$$u_{\text{outgoing}}(\kappa^2)_{J,0} = \frac{h\,e^{-iJ\phi_0}}{2\,i\,\delta_0\,\sin\phi_0} \sum_{\substack{0\le l\le r-1 \\ N_{\min}\le \tilde{J}\le N_{\max}}} e^{i\,(\tilde{J}-1)\,\phi_0}\,f_{\tilde{J},l}\left[Q_0^t(e^{i\,\phi_0})\right]_{(l+1)1} \quad ,J < N_{\min}.$$

Here $N_{\min}$ and $N_{\max}$ are defined as $N_{\min} = \left[\frac{\alpha}{h}\right]$ and $N_{\max} = \left[\frac{\beta}{h}\right] + 1$.

Upon comparing, we see that, for the numerical solution u_{outgoing}, the role played by

$$\kappa_h := \frac{\phi_0}{h} \tag{72}$$

is equivalent to that of the analytic wave number κ for the analytic solution $\mathsf{u}_{\text{outgoing}}$.

6.2 Dispersion Analysis

We recall that by Proposition 6, for real $w > 0$ and small enough, the characteristic quadratic $\mathbf{q}(w,\cdot)$ has conjugate complex roots on the unit circle, denoted by $e^{\pm i\phi_0(w)}$. We write $\phi_0 = \phi_0(w)$. With $\phi_0 \in (0,\pi)$ being the principal argument of $e^{\pm i\phi_0}$ the roots of characteristic quadratic $\mathbf{q}$. For w real with $w > 0$ and small enough, we have ϕ_0 is analytic in w and has the following expansion, c.f (47) in Proposition 6,

$$\phi_0 = w^{1/2}(1 + \mathsf{O}(w^r)).$$

On the other hand, at the end of Section 5.4, we have introduced the numerical wave number $k_h := \frac{\phi_0}{h}$, c.f. (72). As a result, we obtain readily the dispersion relation between the numerical wavenumber k_h and the analytic one k from the expansion of ϕ_0. The following proposition is just the restatement of Proposition 6 with $(\kappa\, h)^2$ replacing w.

Proposition 16 (Phase difference)**.** *For small enough $\kappa\, h$, $\kappa_h h$ is an analytic function of $\kappa\, h$ and is given by*

$$\cos(\kappa_h h) \;=\; \cos(\kappa\, h) \;+\; \kappa\, h\; \mathsf{O}((\kappa\, h)^{2r}).$$

This implies for small $\kappa\, h$ that

$$\kappa_h = \kappa + \kappa\, \mathsf{O}\big((\kappa\, h)^{2r}\big).$$

The approximation error is $\mathsf{O}((\kappa h)^{2r})$ and is in fact is bounded by $C(r)(\kappa\, h)^{2r}$ where the constant C depends only on r, the order of the method.

6.3 Pole location algorithm

Suppose M is a matrix of size $r \times r$, and whose components are polynomials in z. Denote by $\mathcal{R}_M(z) := \frac{\mathrm{Adj}\, M(z)}{\det M(z)}$, wherever the RHS is defined. Define the set of generalized eigenvalues of M,

$$\sigma_M := \{z \in \mathbb{C} \,\big|\, M(z) \text{ is not invertible}\}.$$

These are also called the poles of $\mathcal{R}_M$, since $\mathcal{R}_M(\cdot)$ fails to be defined on σ_M, but exists and is analytic on $\mathbb{C} \setminus \sigma_M$. For $z_0 \notin \sigma_M$, [5, Theorem 2.4] provides an algorithm to locate the closest pole of $\mathcal{R}_M$ to z_0. First, to find all poles of $\mathcal{R}_M$, we look at the poles of $x(\cdot) = \mathcal{R}_M(\cdot)\, b$ for arbitrary scalar vector $b \in \mathbb{C}^r$. Write

$$x(z) = \sum_{k=0}^{\infty} x_k (z - z_0)^k.$$

Denote by π_l the projection onto the l-th component. If λ_0 is one of the closest poles to z_0, then for $1 \le l \le r$, the ratio of the $\pi_l\, x_k$ and $\pi_l\, x_{k+1}$ converges to $\lambda_0 - z_0$, c.f., [5, Theorem 2.4].

In the results of previous sections, we have shown that the numerical wave number k_h is related to the argument of the poles of W which solves

$$\mathcal{A}_w(z)\, W_w(\cdot) \;=\; h\, z\, H(z)$$

with $H(z) = Z_{\mathfrak{B}} f$ and $\mathcal{A} = z^2 A_w + z B_w + A_w^t$, $w = \kappa^2 h^2$ and $\kappa \in \mathbb{R}^\star$. By Guillaume's algorithm, to look for these poles, we approximate the poles of $x(z)$, which solves $\mathcal{A}(z)\, x(z) = h\, z\, \mathbf{b}$, for arbitrary scalar vector $\mathbf{b} \in \mathbb{C}^r$. Specifically, for $z_0 \notin \sigma_{\mathcal{A}}$, we first expand b and x about an *analytic point* z_0

$$z\, h\, \mathbf{b} = h\, z_0\, \mathbf{b} + h\, (z - z_0)\, \mathbf{b} \quad ; \quad x(z) = \sum_{k=0}^{\infty} x_k(z_0)\, (z - z_0)^k.$$

Then we expand $\mathcal{A}(z)$ about z_0

$$\mathcal{A}(z) \;=\; M_0 + M_1 (z - z_0) + M_2 (z - z_0)^2,$$

where the coefficients are given by

$$M_0(z_0) = z_0 B + A^t + z_0^2 A \quad ; \quad M_1(z_0) = 2 z_0 A + B \quad ; \quad M_2(z_0) = A\,.$$

The coefficient x_k of the expansion of x at z_0 solves the following algebraic equation,

$$\begin{aligned} M_0(z_0)\, x_0 &= h\, z_0\, \mathbf{b} \quad ; \\ M_0(z_0)\, x_1 &= -M_1(z_0)\, x_0 + h\, \mathbf{b} \quad ; \\ M_0(z_0)\, x_k &= -M_1(z_0)\, x_{k-1} - M_2(z_0)\, x_{k-2} \quad , \quad k \geq 2\,. \end{aligned}$$

Since the vector $\mathbf{b}$ will be chosen arbitrarily, which means x_0 can be chose arbitrarily, x_1 is then obtained by,

$$M_0(z_0)\, x_1 = -M_1(z_0) x_0 + z_0^{-1} M_0(z_0)\, x_0. \tag{73}$$

If λ_0 is the unique closest pole to z_0, we have

$$\frac{\pi_l\, x_k}{\pi_l\, x_{k+1}} \longrightarrow \lambda_0 - z_0 \ , \ k \to \infty.$$

Here π_l the projection on the l-th component of a vector.

Since $\sigma_{\mathcal{A}} = \{e^{i\phi_0}\, , e^{-i\phi_0}\}$, c.f. Section 4.2, we can restrict ourselves to looking for poles in region $\Omega = [-2, 2] \times [-2, 2] \subset \mathbb{C}$. We partition Ω into smaller squares of width $\delta_z > 0$,

$$\Omega_{k,l} = \big[\, -2 + k\, \delta_z\, ,\, -2 + (k+1)\, \delta_z \big] \ \times \ [-2 + l\, \delta_z\, ,\, -2 + (l+1)\, \delta_z \big].$$

Our choice of initial analytic point z_0 will be an arbitrary non-zero point in $\Omega_{i,j}$ as long as $M_0(\cdot)$ is invertible there. We then use the following procedure to approach poles closest to this initial guess. We compare condition number $\operatorname{cond} M_0(z_0)$ to ϵ_{cond}, as a criteria for the invertibility of M_0 at z_0. Note that, when $\operatorname{cond} M_0(z_0) < \epsilon_{\text{cond}}$, this means z_0 is close enough to a pole, and thus z_0 can be considered one numerically.

Algorithm We summarize from the above discussion the operations carried out for each square $\Omega_{i,j}$. In addition to the condition number of M_0, the second stop criteria is the number of iterations N_{iter}. Denote by n_{der} the number of derivative we will use to approximate the distance between z_0 and the closest pole.

1. **Start** : Choose initial data $z_0 \neq 0$ arbitrarily in $\Omega_{i,j}$.
   - If $\operatorname{cond} M_0(z_0) > \epsilon_{\text{cond}}$, then z_0 is an analytic point, and continue to step 2;
   - If this is not the case, z_0 is a pole numerically, and move onto the next square.
2. Choose x_0 arbitrarily. Calculate $x_1(z_0), \ldots, x_{n_{\text{der}}+1}(z_0)$, using

   $$\begin{aligned} M_0(z_0)\, x_1 &= -M_1(z_0)\, x_0 + z_0^{-1} M_0(z_0)\, x_0 \\ M_0(z_0)\, x_k &= -M_1(z_0)\, x_{k-1} - M_2(z_0)\, x_{k-2} \ , \ k \geq 2. \end{aligned}$$

   The ratio $\mathfrak{r} = \frac{\pi_1\, x_{n_{\text{der}}+1}}{\pi_1\, x_{n_{\text{der}}}}$ gives an approximation of $\lambda_0 - z_0$, where λ_0 is the closest pole to initial data z_0, hence an approximation of the direction to get from z_0 to λ_0.
3. **Restart** : For the next iteration, we update the initial data as $z_0 \longmapsto z_0 + \mathfrak{r}$, and reiterate the process, as long as the stop criteria have not been met.
4. *Stop criteria* are

   $$\text{the condition number of } M_0 \quad \text{and} \quad \text{the number of iterations } N_{\text{iter}}.$$

Numerical Results Denote by $\mathbf{z}_{ij}$ the result given when apply the above algorithm using initial guess located in square Ω_{ij}. As discussed, $\mathbf{z}_{ij}$ is an approximation/candidate for the poles that we are looking for (i.e. of $\mathbf{A}^{-1}$). Denote by $\kappa_{h,ij}$ the numerical wavenumber extracted from theses poles $\mathbf{z}_{ij}$-s via the relation

$$\mathbf{z}_{h,ij} \;=\; e^{i\,\kappa_{h,ij}\,h}\,.$$

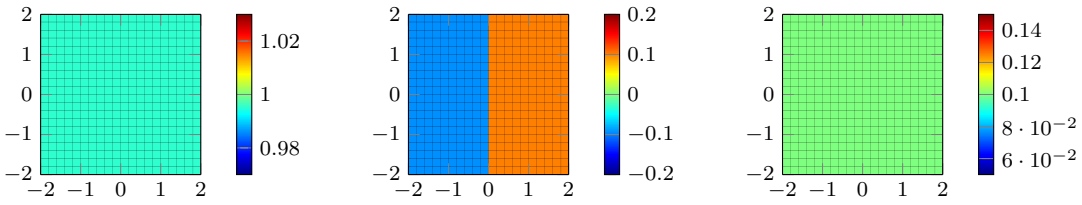
Figure 11,12 show the results for discretization order 4 and Figure 13 for order 9. The parameters for these results are:

$$\kappa = 1\ ,\ h = 0.1\ ,\ \text{dimension of } \Omega_{ij} : \delta_z = 0.2.$$

Stop criteria parameters are

$$N_{\text{iter}} = 5\ ;\ \epsilon_{\text{cond}} = 10^{-13}\ ;\ \text{Nb of derivatives for approx } n_{\text{der}} = 20.$$

The numerical results confirm what is predicted by the theory: the nonzero poles consist of only two conjugate complex numbers (which are expected to represent $e^{i\,\phi_0}$ and $e^{-i\phi_0}$ by the theory). For order 4, subfigure (a) in Figure 11 says that values of the real part of these $\mathbf{z}_{ij}$ take on approximately one value (same color for each of the square), while subfigure (b) and (c) show that set of the imaginary part of $\mathbf{z}_{ij}$ take on two values of equal absolute value and opposite sign. For order 9, Figure 13 show that variation in the values $\{\text{Re}\,\mathbf{z}_{ij}\}$ and $\{|\text{Im}\,\mathbf{z}|\}$ are of order 10^{-8} and 10^{-7}, hence each set corresponds to only one number. In short, the value of $\{\mathbf{z}_{ij}\}$ for each case, consists of exactly a pair of conjugate numbers. As a result, each of these square predicts the same numerical wavenumber κ_h. The final representative numerical wavenumber for the current parameters (in particular discretization order r, analytic wavenumber κ and discretization size h), can be obtained by taking the average of this family $\{\kappa_{h,ij}\}$.



(a) Re $\mathbf{z}_{ij}$ (b) Im $\mathbf{z}_{ij}$ (c) |Im $\mathbf{z}_{ij}$|

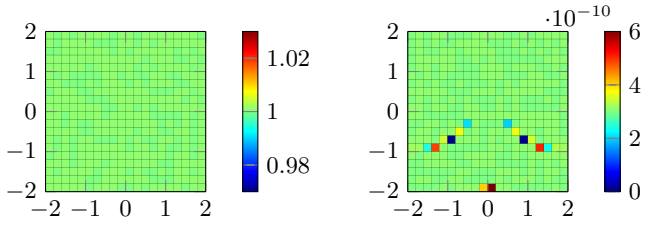
Figure 11: Application of Guillaume's algorithm for Order 4. $\mathbf{z}_{ij}$ - the numerical poles calculated with Guillaume's algorithm, using initial guess from square located in Ω_{ij}. The value of $\{\mathbf{z}_{ij}\}$ consists of exactly a pair of conjugate numbers.

In the previous paragraph, for a discretization order r, analytic wavenumber κ and discretization size h, we have discussed how to obtain a representative numerical wavenumber κ_h. In order to obtain the dispersion curve which represents the relative difference in wavenumber $\frac{\kappa_h - \kappa}{\kappa}$ in terms of $\kappa\,h$, for each discretization order r, and analytic wavenumber fixed at $\kappa = 1$, and discretization stepsize h in the range

$$\texttt{grid in}\ \kappa\,h \;=\; 0.01 : 0.01 : 3,$$

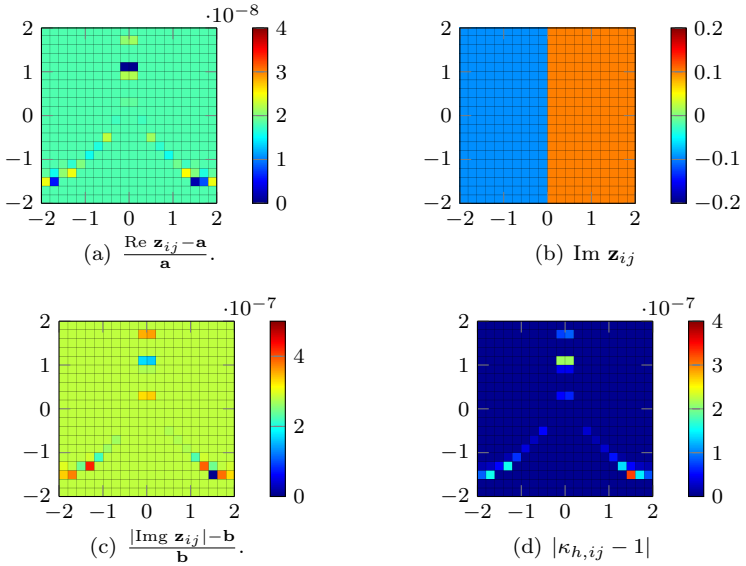
we apply this algorithm to calculate the corresponding numerical wavenumber.

Remark 7. *In the Figure 14 and 15, we show the curve an an interval away from the region of numerical instability. As the order grows, this region starts with large κh.*



(a) Re $\kappa_{h,ij}$ (b) Im $\kappa_{h,ij}$

Figure 12: Application of Guillaume's algorithm for Order 4. Here $\kappa_{h,ij}$ is obtained from $\mathbf{z}_{ij}$ via the relation $\mathbf{z}_{ij} = e^{i\kappa_{h,ij}\,h}$. Note that the imaginary part is of very small order, which is expected since $\kappa_{h,ij}$-s approximate the analytic wavenumber $\kappa = 1$.



(a) $\frac{\text{Re } \mathbf{z}_{ij} - \mathbf{a}}{\mathbf{a}}$. (b) Im $\mathbf{z}_{ij}$

(c) $\frac{|\text{Img } \mathbf{z}_{ij}| - \mathbf{b}}{\mathbf{b}}$. (d) $|\kappa_{h,ij} - 1|$

Figure 13: Application of Guillaume's algorithm for Order 9. Here $\mathbf{z}_{ij}$ = the numerical poles calculated with Guillaume's algorithm, using initial guess located in square Ω_{ij}.
Fig. (a) shows the variation of the real part of $\mathbf{z}_{ij}$ with $\mathbf{a} := \min_{1\le i,j\le 21} \text{Re } \mathbf{z}_{ij}$. Since this is order 10^{-8}, they are approximately equal to one another.
Fig. (b) shows the variation of the absolute value of the imaginary part of $\mathbf{z}_{ij}$ $\mathbf{b} = \min_{1\le i,j\le 21} |\text{Im } \mathbf{z}_{ij}|$. Since this is order 10^{-7}, they are approximately equal to one another.
Fig. (c) shows that imaginary part of $\mathbf{z}_{ij}$ consists of exactly two numbers of equal magnitude and opposite sign.
Fig. (d) showing the variation of $\kappa_{h,ij}$ compared with the true $\kappa = 1$, with $\kappa_{h,ij}$ obtained from $\mathbf{z}_{ij}$ by $\mathbf{z}_{ij} = e^{i\kappa_{h,ij}\,h}$.
As a result, the value of $\{\mathbf{z}_{ij}\}$ consists of exactly a pair of conjugate numbers.

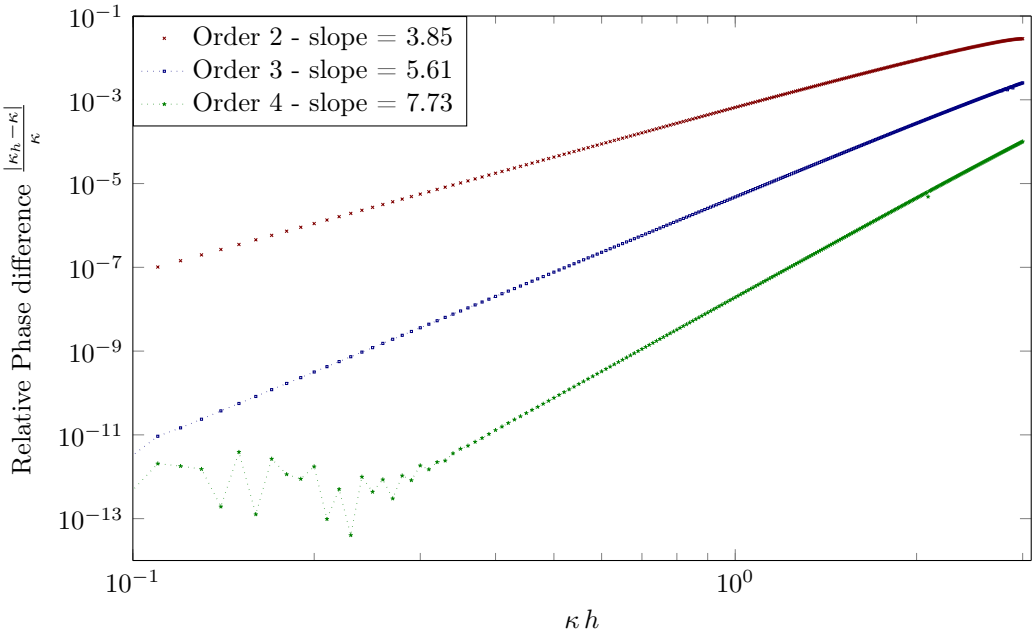


Figure 14: Dispersion curve for κh taking values in $0.1 : 0.01 : 3$. The slopes of each curve are approximately twice the discretization order, as predicted by the asymptotics (4).

7 Conclusion

In this project, we have used a method different from that in [1] and [8] to characterize and quantize the numerical dispersion associated with continuous Galerkin finite element. Without starting from an ansatz, we construct directly the numerical solution (on the whole real line) via the limiting absorption principle. We also obtain an asymptotic expansion for the difference between the analytic and numerical wavenumber, which agrees with the results of [1] and [8]. With technique in complex analysis like contour deformation, we arrive at an explicit formula for the numerical solution, which allows the identification of the numerical wavenumber in terms of analytic poles of an algebraic equation. This identification is useful since we can evaluate concretely the numerical wave number from analytic poles, and the latter are calculated via Guillaume's algorithm.

There are few points of ameliorations and questions for the current work.

1. Although we arrive an asymptotic expansion of the same order, our coefficient for the highest order term is a generic constant, and is not explicit, as in [1]. The question is whether with the same method, we can do more careful analysis and calculate this term, or at least its sign. Furthermore, can we make statements in higher regime, i.e. when $\kappa\, h \gg 1$?

2. The algorithm uses local quantities described in terms of the interactions of the local Lagrangian polynomials with one another via the local pairing a_w on $[0, 1]$. In the current codes, these quantities are calculated explicitly in terms of the coefficient of the Lagrangian polynomials in the basis $1, \hat{x}, \ldots, \hat{x}^r$. For higher orders, this approach suffers from numerical

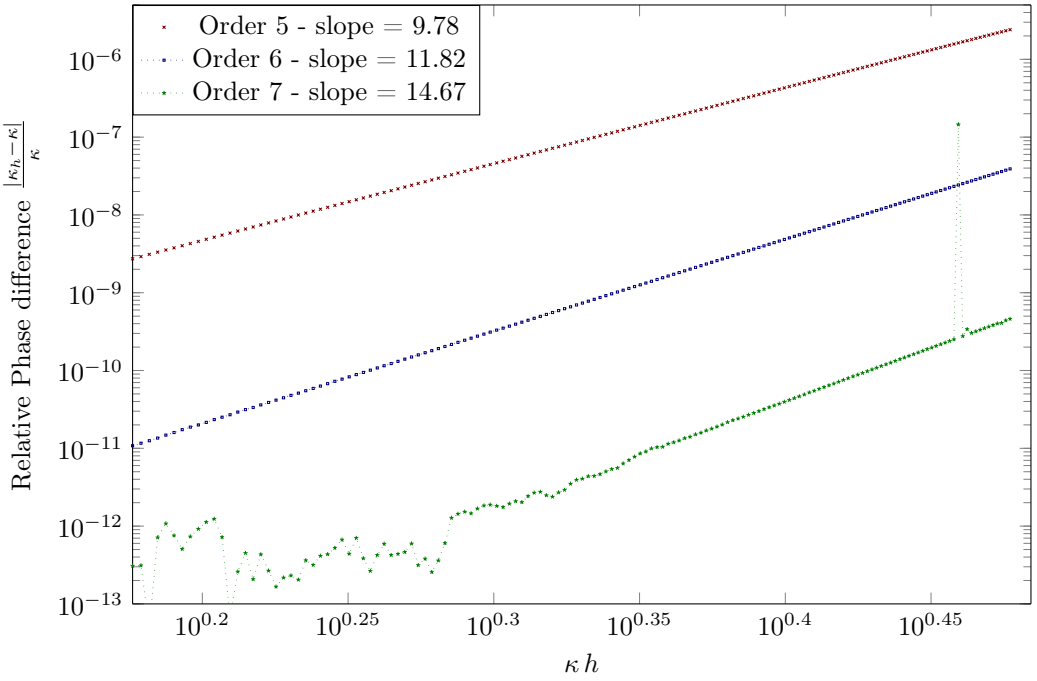


Figure 15: Dispersion curve for κh taking values in $1.5 : 0.01 : 3$. The slopes of each curve are approximately twice the discretization order, as predicted by the asymptotics (4).

instability. In order to have a more stable numerical evaluation, perhaps it is better to describe the Lagrangian polynomials in terms of Legendre polynomials, and make use of the orthogonality of the second family. We also note that [1] used Legendre polynomials.

3. Can we extend this idea, in particular the identification of numerical wavenumber with analytic poles and thus the pole evaluation algorithm, to other discretization schemes like discontious galerkin, and in higher dimension, e.g. a waveguide, and inhomogeneous media with periodic heterogeneity?

Acknowledgement This Project has received funding from the European Union's Horizon 2020, research and innovation program under the Marie Sklodowska-Curie grant agreement No. 644202.

A Toy example with Finite Difference order 2

Uniform discretization of $\mathbb{R}$ with step size h by nodes $x_n = n\,h$ with $n \in \mathbb{Z}$. The recurrence relation given by second order Finite Difference

$$-u_{n-1} + (2 - \kappa^2 h^2)\, u_n - u_{n+1} = 0 \quad , \quad n \in \mathbb{Z}.$$

The characteristic polynomial of the recurrence relation,

$$z^2 - (2 - \kappa^2 h^2)\, z + 1 = 0 \quad ,$$

for $0 \leq \kappa h < 2$, has conjugate complex roots of norm 1: $e^{\pm i\,\gamma_h}$. The solution of the recurrence relation

$$u_n \;=\; a_+\, e^{i\,\gamma_h\, n} \;+\; a_- e^{-i\,\gamma_h\, n} \quad .$$

(Analytic) wavenumber κ controls the oscillatory behavior of u_{exact},

$$\mathbf{u}_{\text{exact}}(x) \;=\; a_+\, e^{i\,\kappa\, x} \;+\; a_- e^{-i\,\kappa}\, x \quad .$$

The numerical wave number $\kappa_h := \frac{\gamma_h}{h}$ controls the oscillatory behavior of numerical solution. At $x = nh$, the value of the numerical solution

$$\begin{aligned} u_n = u_{\text{dist}}(nh) = \; & a_+ e^{i\,\frac{\gamma_h}{h}\,(nh)} \;+\; a_- e^{-i\,\frac{\gamma_h}{h}\,(nh)} \\ = \; & a_+\, e^{i\,\kappa_h\, x} \;+\; a_-\, e^{-i\,\kappa_h\, x} \quad . \end{aligned}$$

Since $e^{i\gamma_h}$ solves $z^2 - (2-\kappa^2 h^2)z + 1 = 0$ thus $e^{i\gamma_h}$ satisfies $z - (2 - \kappa^2 h^2) + z^{-1} = 0$, and we have

$$\underbrace{e^{i\gamma_h} + e^{-i\gamma_h}}_{2\cos(\gamma_h)} = 2 - \kappa^2 h^2 \quad \Rightarrow \quad 2\,(1 - \cos(\gamma_h)) = \kappa^2 h^2\,.$$

As a result,

$$4\sin^2(\tfrac{1}{2}\gamma_h) = \kappa^2 h^2.$$

For $\kappa > 0$ and h is chosen small enough so that $\sin(\frac{1}{2}\gamma_h) > 0$, we have

$$\sin(\tfrac{1}{2}\gamma_h) \;\; \tfrac{1}{2}\kappa\, h \quad \Rightarrow \quad \gamma_h \;=\; 2\arcsin(\tfrac{1}{2}\kappa\, h)$$

Thus we have obtain the dispersion relation, for $|\kappa h| < 2$,

$$\kappa_h = \frac{2}{h}\arcsin\left(\frac{1}{2}\kappa\, h\right) = \frac{2}{h}\sum_{n=0}^{\infty} \frac{(2n)!}{4^n (n!)^2 (2n+1)} \frac{1}{2\cdot 4^n}\,(\kappa\, h)^{2n+1}.$$

The phase difference is given by

$$\delta_h \;=\; \kappa_h - \kappa \;=\; \frac{1}{24}\kappa^3\, h^2 \;+\; \mathcal{E}_h(\kappa\, h) \quad , \quad |\kappa\, h| < 2$$

with $\mathcal{E}_h$ analytic,

$$\mathcal{E}_h(\kappa\, h) \;=\; \kappa\, h \sum_{n=2}^{\infty} \frac{(2n)!}{4^{2n}(n!)^2(2n+1)} (\kappa\, h)^{2n} \;=\; \kappa\, h\; \mathsf{O}\big((\kappa\, h)^4\big) \quad ; \quad |\kappa\, h| < 2 \quad .$$

B Properties of the local quantities

B.1 Reference Lagrangian polynomials

Denote by $P_r = P_r(I)$ the set of polynomials of degree less than or equal to r defined on interval I. A basis of P_r on the reference interval $[0,1]$ are given by the Lagrangian polynomials $\hat{\phi}_i$, $0 \leq i \leq r$, for the defining set $S = \{\hat{x}_i = i/r;\ 0 \leq i \leq r\}$. In particular,

$$\hat{\phi}_i(\hat{x}) := \prod_{\substack{0 \leq j \leq r, \\ j \neq i}} \frac{(\hat{x} - \hat{x}_j)}{(\hat{x}_i - \hat{x}_j)}.$$

It follows from their definitions that $\hat{\phi}_i(\hat{x}_j) = \delta_{ij}$ for $0 \leq i, j \leq r$.

We denote by $I_f \in P_r$ the interpolated version of a function $f : [0,1] \mapsto \mathbb{C}$,

$$I_f = \sum_{j=0}^{r} f(\hat{x}_j)\hat{\phi}_j. \tag{74}$$

Since the $\hat{\phi}_j$-s form a basis for P_r, any polynomial $p \in P_r$ is equal to its interpolated version I_p by $\hat{\phi}_j$-s.

$$p = I_p. \tag{75}$$

B.2 Symmetries of the local quantities

Proposition 17. *The Lagrangian polynomials $\{\hat{\phi}_j\}_{0 \leq j \leq r}$ have the following symmetries,*

$$\hat{\phi}_j(1-\hat{x}) = \hat{\phi}_{r-j}(\hat{x}) \quad , \quad -\hat{\phi}'_j(1-\hat{x}) = \hat{\phi}'_{r-j}(\hat{x}) \quad , \quad 0 \leq j \leq r. \tag{76}$$

Proof. For j with $0 \leq j \leq r$, the value of function $f : x \mapsto 1 - x$ at interpolation node $x_j = j/r$ is

$$f(x_j) = 1 - x_j = 1 - j/r = (r-j)/r = x_{r-j}.$$

Hence, j, k with $0 \leq j, k \leq r$, we have

$$f^*\hat{\phi}_j(x_k) = \hat{\phi}_j(f(x_k)) = \hat{\phi}_j(x_{r-k}) = \delta_{j(r-k)} = \delta_{(r-j)k} = \hat{\phi}_{r-j}(x_k).$$

The second to last equality follows from the fact $j = r - k$ is equivalent to $k = r - j$. This means $f^*\hat{\phi}$ agrees with $\hat{\phi}_{r-j}$ on the defining set S which is of cardinality $r+1$. On the other hand, both $f^*\hat{\phi}_j$ and $\hat{\phi}_j$ are polynomials of degree r. As a result, $f^*\hat{\phi}_j \equiv \hat{\phi}_{r-j}$ on $[0,1]$.

□

Define κ_j and vector $\kappa \in \mathbb{R}^{r-1}$ as follows

$$\kappa_j = \int_0^1 \hat{\phi}_j \, dx \,, \, 0 \leq j \leq r \quad ; \quad \kappa = (\kappa_j)_{1 \leq j \leq r-1}. \tag{77}$$

The above symmetry of $\hat{\phi}_j$-s gives the following properties for κ, $\mathfrak{a}$, $\mathfrak{b}$, c.f (21), and the local matrices.

Corollary 18.

$$\kappa_j = \kappa_{r-j} \quad , \quad 0 \leq j \leq r.$$

$$\hat{M}_{ij} = \hat{M}_{(r-i)(r-j)} \quad , \quad \hat{S}_{ij} = \hat{S}_{(r-i)(r-j)} \quad , \quad 0 \leq i,j \leq r.$$

$$\mathfrak{b}_j = \mathfrak{a}_{r-j} \quad , \quad 1 \leq j \leq r-1.$$

Proof. The identity of κ follows symmetry (76) of $\hat{\phi}_j$-s and a change of integration variables. With the same idea, we write out completely the identities with the local matrices. For $0 \leq i,j \leq r$, we have

$$\int_0^1 \hat{\phi}_i \, \hat{\phi}_j \, d\hat{x} = \int_0^1 (f^* \hat{\phi}_{r-i}) \, (f^* \hat{\phi}_{r-j}) \, d\hat{x} = \int_0^1 \hat{\phi}_{r-i} \, \hat{\phi}_{r-j} \, d\hat{x} \quad ;$$

$$\int_0^1 \hat{\phi}'_i \, \hat{\phi}'_j \, d\hat{x} = \int_0^1 [f^* \hat{\phi}_{r-i}]' \, [\hat{\phi}_{r-j}]' \, d\hat{x} = \int_0^1 \hat{\phi}'_{r-i} \, \hat{\phi}'_{r-j} \, d\hat{x}.$$

This gives $\hat{M}_{ij} = \hat{M}_{(r-i)(r-j)}$ and $\hat{S}_{ij} = \hat{S}_{(r-i)(r-j)}$. As a result of this, for $1 \leq j \leq r-1$, we have

$$\hat{M}_{0j} = \hat{M}_{r(r-j)} \qquad \Rightarrow \quad \pi_j \mathfrak{b}_0 = \pi_{r-j} \mathfrak{a}_0.$$

$$\hat{S}_{0j} = \hat{S}_{r(r-j)} \qquad \Rightarrow \quad \pi_j \mathfrak{b}_1 = \pi_{r-j} \mathfrak{a}_1.$$

The proof is finished by using the definition of $\mathfrak{a} = \mathfrak{a}_0 - w\mathfrak{a}_1$ and $\mathfrak{b} = \mathfrak{b}_0 - w\mathfrak{b}_1$.

□

B.3 Property in terms sum of row (column)

We obtain properties of local mass matrix $\hat{M}$ and local stiff matrix $\hat{S}$, c.f (10), regarding sum of row/column.

Proposition 19. *1. The sum of any row / column of $\hat{S}$ is zero.*

2. For $0 \leq j \leq r$, the sum of the j-th row / column of $\hat{M}$ is equal to κ_j.

3. The sum of all components of $\hat{M}$ is equal to 1, which in terms of κ_j, is $\sum_{j=0}^{r} \kappa_j = 1$.

Proof. Since $\hat{M}$ and $\hat{S}$ are symmetric, it suffices to prove the statement regarding row summation. Using (75), we can write the constant function 1 is equal to its interpolated version by $\hat{\phi}_j$-s, c.f (74),

$$1 = I_1 = \sum_{j=0}^{r} \hat{\phi}_j.$$

Using the above identity and the definition of $\hat{M}$ and $\hat{S}$ in terms of $\hat{\phi}_j$, c.f (10), we obtain

$$\kappa_j = \int_0^1 \hat{\phi}_j(\hat{x}).1 \, d\hat{x} = \int_0^1 \hat{\phi}_j(\hat{x}) \left(\sum_{i=0}^{r} \hat{\phi}_i(\hat{x}) \right) d\hat{x} = \sum_{j=0}^{r} \hat{M}_{ij} \quad , \quad 0 \leq j \leq r;$$

$$1 = \int_0^1 1.1 \, d\hat{x} = \int_0^1 \left(\sum_{j=0}^{r} \hat{\phi}_j(\hat{x}) \right) \left(\sum_{i=0}^{r} \hat{\phi}_i(\hat{x}) \right) d\hat{x} = \sum_{i,j=0}^{r} \hat{M}_{ij} \quad , \quad 0 \leq i,j \leq r.$$

$$0 = \int_0^1 \hat{\phi}_i'.1'\, d\hat{x} = \int_0^1 \hat{\phi}_j' \left(\sum_{i=0}^{r} \hat{\phi}_i' \right) d\hat{x} = \sum_{j=0}^{r} \hat{S}_{ij} \quad , \quad 0 \le i,j \le r.$$

□

We state the above results in terms of the components of matrix $\mathcal{M}$ which has been defined as $\mathcal{M}(w) = \hat{S} - w\hat{M}$.

Corollary 20. *$\mathcal{M}$ is symmetric and has the following properties,*

1. *$\mathcal{M}_{ij} = \mathcal{M}_{(r-i)(r-j)}$, $0 \le i,j \le r$; in particular, $\mathcal{M}_{00} = \mathcal{M}_{rr}$.*
2. *The sum of the j-th row of $\mathcal{M}$, $0 \le j \le r$, satisfies*

$$\mathcal{M}_{j0} + \ldots + \mathcal{M}_{jr} = -w\kappa_j \quad , \quad 0 \le j \le r.$$

Since $\mathcal{M}$ is symmetric, we also have the same statements stated above in terms of columns .

Proposition 21. *With* $\mathrm{Row}(C,k)$ *and* $\mathrm{Col}(C,k)$ *denoting respectively the k-th row and column vector of matrix C, and π_j the projection onto the j-th component of a vector, we have*

$$\begin{aligned} \mathfrak{b}^t + \mathrm{Row}(\mathcal{M}_{int},1) + \ldots + \mathrm{Row}(\mathcal{M}_{int},r-1) + \mathfrak{a}^t &= -w\,\kappa^t; \\ \mathfrak{b} + \mathrm{Col}(\mathcal{M}_{int},1) + \ldots + \mathrm{Col}(\mathcal{M}_{int},r-1) + \mathfrak{a} &= -w\,\kappa \\ \mathcal{M}_{00} + \pi_1\mathfrak{b} + \ldots + \pi_{r-1}\mathfrak{b} + \alpha &= -w\,\kappa_0; \\ \mathcal{M}_{00} + \pi_1\mathfrak{a} + \ldots + \pi_{r-1}\mathfrak{a} + \alpha &= -w\,\kappa_0. \end{aligned}$$

Proof. By Corollary 20, the sum of entries in a row of $\mathcal{M}$ is equal to

$$\mathcal{M}_{0j} + \mathcal{M}_{1j} + \ldots + \mathcal{M}_{(r-1)j} + \mathcal{M}_{rj} = -w\,\kappa_j, \quad 0 \le j \le r. \tag{78}$$

We recall the definitions of $\mathfrak{a}, \mathfrak{b}, \alpha$ and $\mathcal{M}_{\mathrm{int}}$ in terms of $\mathcal{M}$, c.f (21), (20), and (19),

$$(\mathcal{M}_{\mathrm{int}})_{ii} = \mathcal{M}_{ij} \quad , 1 \le i,j \le r-1 \quad ; \quad \alpha = \mathcal{M}_{r0} = \mathcal{M}_{0r};$$

$$\mathfrak{b}_j^t = \mathcal{M}_{0j} = \mathcal{M}_{j0} \quad ; \quad \mathfrak{a}_j^t = \mathcal{M}_{rj} = \mathcal{M}_{jr} \quad , \quad 1 \le j \le r-1.$$

In terms these quantities, (78) can be rewritten as,

$$\pi_j\mathfrak{b}^t + \pi_j\,\mathrm{Row}(\mathcal{M}_{\mathrm{int}},1) + \ldots + \pi_j\,\mathrm{Row}(\mathcal{M}_{\mathrm{int}},r-1) + \pi_j\mathfrak{a}^t = -w\,\kappa_j, \quad 1 \le j \le r-1.$$

In the form of summation of row vectors, the above expression can be written as,

$$\mathfrak{b}^t + \mathrm{Row}(\mathcal{M}_{\mathrm{int}},1) + \ldots \mathrm{Row}(\mathcal{M}_{\mathrm{int}},r-1) + \mathfrak{a}^t = -w\,\kappa^t.$$

For $j = 0$, (78) can be written as,

$$\mathcal{M}_{00} + \pi_1\mathfrak{b} + \ldots + \pi_{r-1}\mathfrak{b} + \alpha = -w\,\kappa_0.$$

The relation between $\mathfrak{a}$ and $\mathfrak{b}$, $\mathfrak{b}_j = \mathfrak{a}_{r-j}$, $1 \le j \le r-1$, c.f Corollary 18, then gives the last identity in the statement.

□

B.4 Relation between local matrices and the bilinear forms

We have introduced $P_r(I)$ as the set of polynomials of degree $\leq r$ defined on interval I. We will also work with its subsets $P_{r,D}$ defined as

$$P_{r,D}(I) := \{g \in P_r(I)\,;\, g(0) = g(1) = 0\} \quad , \text{ with } I = [0,1]. \tag{79}$$

For a continuous function f defined on $[0,1]$, introduce the (column) vectors $[f] \in \mathbb{R}^{r+1}$ and $[f]_{\rm int} \in \mathbb{R}^{r-1}$ whose components are created from the values of f on the interpolation nodes $\{k/r\}_{0\leq k\leq r}$ as follows,

$$\begin{aligned} [f] \in \mathbb{R}^{r+1} \text{ with } \quad & [f]_j = f(j/r)\,,\, 0 \leq j \leq r; \\ [f]_{\rm int} \in \mathbb{R}^{r-1} \text{ with } \quad & ([f]_{\rm int})_j = f(j/r)\,,\, 1 \leq j \leq r-1. \end{aligned} \tag{80}$$

Note that for f with $f(0) = f(1) = 0$ then $[f] = (0\,,\, [f]_{\rm int}\,,\, 0)^t$. In the current notation, that every elements of P_r is equal to its interpolation version, c.f (75) and (74), can be written as

$$\begin{aligned} g \in P_r \quad &: \quad g = I_g = \sum_{j=0}^{r} [g]_j\, \hat{\phi}_j; \\ g \in P_{r,D} \quad &: \quad g = I_g = \sum_{j=0}^{r} [g]_j\, \hat{\phi}_j = \sum_{j=1}^{r-1} ([g]_{\rm int})_j\, \hat{\phi}_j. \end{aligned} \tag{81}$$

We have introduced the bilinear forms $\mathsf{a_M}$ and $\mathsf{a_S}$, c.f (11),

$$\begin{aligned} \mathsf{a_S}(f,g) &= \int_0^1 f'(\hat{x})\, g'(\hat{x})\, d\hat{x} \quad , && f, g \in H^1(0,1)\,; \\ \mathsf{a_M}(f,g) &= \int_0^1 f(\hat{x})\, g(\hat{x})\, d\hat{x} \quad , && f, g \in L^2(0,1). \end{aligned}$$

The components of the local stiff matrix $\hat{S}$ and local mass matrix $\hat{M}$ which give the interaction among the Lagrangian polynomials $\hat{\phi}_j$ via the bilinear forms $\mathsf{a_S}$ and $\mathsf{a_M}$, c.f. (18), are given by,

$$\hat{S}_{ij} = \mathsf{a_S}\left(\hat{\phi}_i, \hat{\phi}_j\right) \quad , \quad \hat{M}_{ij} = \mathsf{a_M}\left(\hat{\phi}_i, \hat{\phi}_j\right) \quad , \quad 0 \leq i, j \leq r.$$

The interior matrices $\hat{S}_{\rm int}$, $\hat{M}_{\rm int}$ introduced in Subsection 3.2,

$$(\hat{S}_{\rm int})_{ij} = \hat{S}_{ij} \quad , \quad (\hat{M}_{\rm int})_{ij} = \hat{M}_{ij} \quad , \quad 1 \leq i, j \leq r-1.$$

We introduce $\hat{S}_{{\rm int},\star}$ and $\hat{M}_{{\rm int},\star}$ of size $(r-1)\times(r+1)$ as

$$(\hat{S}_{{\rm int},\star})_{ij} = \hat{S}_{ij} \quad , \quad (\hat{M}_{{\rm int},\star})_{ij} = \hat{M}_{ij} \quad , \quad 1 \leq i \leq r-1\,,\, 0 \leq j \leq r. \tag{82}$$

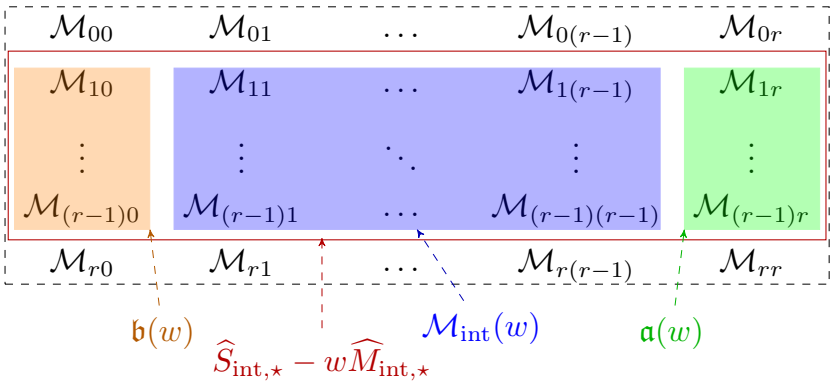


Figure 16: Local matrices $\hat{S}_{\mathrm{int},\star}$, $\hat{M}_{\mathrm{int},\star}$, interior matrix $\mathcal{M}_{\mathrm{int}}$.

The following proposition describes the action of the matrices listed above in terms of the bilinear forms $\mathsf{a_M}$ and $\mathsf{a_S}$.

Proposition 22. *1. For $g \in P_r$*

$$\begin{aligned} \left(\hat{S}\,[g]\right)_j &= \mathsf{a_S}\left(\hat{\phi}_j, g\right) \quad ; \quad \left(\hat{M}\,[g]\right)_j = \mathsf{a_M}\left(\hat{\phi}_j, g\right); \quad 0 \le j \le r; \\ \left(\hat{S}_{int,\star}\,[g]\right)_j &= \mathsf{a_S}\left(\hat{\phi}_j, g\right) \quad ; \quad \left(\hat{M}_{int,\star}\,[g]\right)_j = \mathsf{a_M}\left(\hat{\phi}_j, g\right); \quad 1 \le j \le r-1. \end{aligned} \tag{83}$$

2. For $g \in P_{r,D}$ and $1 \le j \le r-1$,

$$\begin{aligned} \left(\hat{S}_{int,\star}\,[g]\right)_j &= \left(\hat{S}_{int}\,[g]_{int}\right)_j = \mathsf{a_S}\left(\hat{\phi}_j, g\right) = \mathsf{a_M}(\hat{\phi}_j, -g'') \quad ; \\ \left(\hat{M}_{int,\star}\,[g]\right)_j &= \left(\hat{M}_{int}\,[g]_{int}\right)_j = \mathsf{a_M}\left(\hat{\phi}_j, g\right). \end{aligned} \tag{84}$$

Proof. We will write out the proof associated to $\mathsf{a_S}$, that for $\mathsf{a_M}$ is completely analogous.

Property 1 : For $g \in P_r$, we have

$$\left(\hat{S}\,[g]\right)_i = \sum_{k=0}^{r} \hat{S}_{ik}\,[g]_k = \sum_{k=0}^{r} \mathsf{a_S}(\hat{\phi}_i\,, \hat{\phi}_k)\,[g]_k = \mathsf{a_S}\left(\hat{\phi}_i\,, \sum_{k=0}^{r} [g]_k\,\hat{\phi}_k\right) \overset{(81)}{=} \mathsf{a_S}(\hat{\phi}_i\,, g).$$

Property 2 : For $g \in P_{r,D}$, since $g(0) = g(1) = 0$, we have $[g] = (0, [g]_{\mathrm{int}}, 0)$. This fact together with the definition of $\hat{M}_{\mathrm{int},\star}$, we obtain the first equality,

$$\hat{S}_{\mathrm{int}}[g]_{\mathrm{int}} = \hat{S}_{\mathrm{int},\star}\,[g].$$

We next compute the components of $\hat{S}_{\mathrm{int}}[g]_{\mathrm{int}}$. For $1 \le i \le r-1$, we have

$$\begin{aligned} \left(\hat{S}_{\mathrm{int}}\,[g]_{\mathrm{int}}\right)_i &= \sum_{k=1}^{r-1} (\hat{S}_{\mathrm{int}})_{ik}\,([g]_{\mathrm{int}})_k \quad = \sum_{k=1}^{r-1} \mathsf{a_S}(\hat{\phi}_i\,, \hat{\phi}_k)\,([g]_{\mathrm{int}})_k \\ &= \mathsf{a_S}\left(\hat{\phi}_i\,, \sum_{k=1}^{r-1} ([g]_{\mathrm{int}})_k\,\hat{\phi}_k\right) \overset{(81)}{=} \mathsf{a_S}(\hat{\phi}_i\,, g) \quad = \mathsf{a_M}(\phi_i\,, -g''). \end{aligned}$$

The last equality is obtained by integrating by parts, using that $\hat{\phi}_j(0) = \hat{\phi}_j(1) = 0$ for $1 \le j \le r-1$. □

Corollary 23.

$$\begin{aligned} f,g \in P_r \ : \quad & \mathsf{a_S}(f,g) = [f]\cdot \hat{S}\,[g] \qquad ; \qquad \mathsf{a_M}(f,g) = [f]\cdot \hat{M}\,[g]; \\ f,g \in P_{r,D} \ : \quad & [f]\cdot \hat{S}\,[g] = [f]_{int}\cdot \hat{S}_{int}\,[g]_{int} \ ; \ [f]\cdot \hat{M}\,[g] = [f]_{int}\cdot \hat{M}_{int}\,[g]_{int}; \\ f \in P_{r,D}\ , \ g \in P_r \ : \quad & [f]_{int}\cdot \hat{S}_{int,\star}\,[g] = [f]\cdot \hat{S}\,[g] \ ; \ [f]_{int}\cdot \hat{M}_{int,\star}\,[g] = [f]\cdot \hat{M}\,[g]. \end{aligned}$$

B.5 Invertibility of the interior matrices

Proposition 24. *The interior matrices $\hat{S}_{int}$, $\hat{M}_{int}$, c.f* (18) *are symmetric and definite positive, and thus invertible.*

Proof. For $v \in \mathbb{R}^{r-1}$, define $g \in P_{r,D}(0,1)$ associated to v by

$$g(\hat{x}) := \sum_{j=1}^{r-1} v_j \hat{\phi}_j(\hat{x}) \quad ; \quad v = [g]_{\rm int}.$$

By (23), we have

$$v\cdot \hat{S}_{\rm int} v = \mathsf{a_S}\,(g\,,g) = \int_0^1 g'(\hat{x})\,g'(\hat{x})\,d\hat{x} \quad \geq 0\,.$$

If $v\cdot \hat{S}_{\rm int} v = 0$, this means $g' = 0$ in $L^2(0,1)$, which implies that g is constant on $(0,1)$. Since, g vanishes at 0 and 1. As a result, g is identically zero on $(0,1)$, which gives that $v = 0$. This means $\hat{S}_{\rm int}$ is positive definite, and since it is symmetric, it is invertible.

The proof is similar for $\hat{M}_{\rm int}$, which is also symmetric. By (23), we have

$$v\cdot \hat{M}_{\rm int} v = \mathsf{a_M}\,(g\,,g) = \int_0^1 g^2(\hat{x})\,d\hat{x} \quad \geq 0.$$

If $v \in \mathbb{R}^{r-1}$ with $v\cdot \hat{M}_{\rm int} v = 0$, the above identity implies $g = 0$ in $L^2(0,1)$. Since g is a polynomial, g is zero everywhere, which means v has to be the zero vector. Thus we obtain the definite-positivity of $\hat{M}_{\rm int}$.

□

The matrix $\hat{M}_{\rm int}$ is definite positive and symmetric, thus the following bilinear form is a scalar product and induces the norm denoted by $|\cdot|_{\hat{M}_{\rm int}}$,

$$v,w \in \mathbb{R}^{r-1} \quad : \quad (v,w)_{\hat{M}_{\rm int}} := v\cdot \hat{M}_{\rm int} w \quad ; \quad |v|^2_{\hat{M}_{\rm int}} = v\cdot \hat{M}_{\rm int} v. \tag{85}$$

We have the Cauchy-Schwartz inequality

$$v\cdot \hat{M}_{\rm int} w \leq \left(v\cdot \hat{M}_{\rm int} v\right)^{1/2} \left(w\cdot \hat{M}_{\rm int} w\right)^{1/2}.$$

$$v \in \mathbb{R}^{r-1} \quad : |v|^2_{\hat{M}_{\rm int}} = v\cdot \hat{M}_{\rm int} v.$$

Lemma 25. *With $\rho(C)$ denoting the spectral radius of matrix C, we have*

$$\rho\left(\hat{M}_{int}\,\hat{S}^{-1}_{int}\right) \leq \pi^2.$$

Proof. **Step 1 :** We recall the fact that for a square matrix C and any matrix norm,

$$\rho(C) = \lim_{k\to\infty} \|C^k\|^{1/k} \leq \|C\|. \tag{86}$$

We will work with the vector norm $|\cdot|_{\hat{M}_{\rm int}}$ introduced in (85).

For $v \in \mathbb{R}^{r-1}$, denote by w

$$w = \hat{S}_{\rm int}^{-1}\hat{M}_{\rm int}v \quad \Rightarrow \quad \hat{S}_{\rm int}w = \hat{M}_{\rm int}v.$$

Scalar product on both sides with w, and by Cauchy-Schwartz to obtain the upper bound, we have

$$w\cdot\hat{S}_{\rm int}w = w\cdot\hat{M}_{\rm int}v = (w,v)_{\hat{M}_{\rm int}} \leq |w|_{\hat{M}_{\rm int}}\,|v|_{\hat{M}_{\rm int}}.$$

For the lower bound we use (116) which gives

$$|w|^2_{\hat{M}_{\rm int}} = w\cdot\hat{M}_{\rm int}w \leq \pi^{-2}\, w\cdot\hat{S}_{\rm int}w.$$

As a result,

$$\forall\, v\in\mathbb{R}^{r-1},\ |w|^2_{\hat{M}_{\rm int}} \leq \pi^{-2}\,|w|_{\hat{M}_{\rm int}}\,|v|_{\hat{M}_{\rm int}} \quad \Rightarrow \quad |w|_{\hat{M}_{\rm int}} \leq \pi^{-2}\,|v|_{\hat{M}_{\rm int}}.$$

This means that

$$\|\hat{S}_{\rm int}^{-1}\hat{M}_{\rm int}\|_{\hat{M}_{\rm int}} \leq \pi^{-2} \qquad \underset{\Longrightarrow}{(86)} \qquad \rho\left(\hat{S}_{\rm int}^{-1}\hat{M}_{\rm int}\right) \leq \pi^{-2}.$$

□

We next discuss the invertibility of $\mathcal{M}_{\rm int} = \hat{S}_{\rm int} - w\hat{M}_{\rm int}$. For a polynomial p of order n we define its reciprocal polynomial $p_{\rm reciprocal}$ by

$$p_{\rm recip}(x) = x^n p(x^{-1}).$$

Denote by χ_C the characteristic polynomial of matrix C of size $n\times n$ i.e $\chi_C(w) = \det(w\,{\rm Id} - C)$. In the above notations, $(\chi_C)_{\rm reciprocal}$ is the reciprocal polynomial of the characteristic polynomial of matrix C.

Proposition 26. *1.* $\det\mathcal{M}_{int}$ *is a polynomial in* w *of order* $r-1$ *and can be written as*

$$\det\mathcal{M}_{int} = \det\hat{S}_{int}\ (\chi_{\hat{M}_{int}\hat{S}_{int}^{-1}})_{recip}(w)\,. \tag{87}$$

The last polynomial can be written as,

$$(\chi_{\hat{M}_{int}\hat{S}_{int}^{-1}})_{recip}(w) = 1 - w\,{\rm tr}\left(\hat{M}_{int}\hat{S}_{int}^{-1}\right) + \sum_{k=2}^{r-2}\tilde{\mathbf{b}}_k\,w^k + (-1)^{r-1}\det\left(\hat{M}_{int}\hat{S}_{int}^{-1}\right)w^{r-1}.$$

2. When w *satisfies* $|w|<\pi^2$, $\mathcal{M}_{int}(w)$ *is invertible with analytic inverse in* w *with expansion*

$$\mathcal{M}_{int}^{-1} = \hat{S}_{int}^{-1} + \sum_{k=1}^{\infty} w^k\left(\hat{S}_{int}^{-1}\hat{M}_{int}\right)^k\hat{S}_{int}^{-1}\,. \tag{88}$$

Proof. **Property 1 - Determinant :** We recall the definition of $\mathcal{M}_{\text{int}}$, c.f (19),

$$\mathcal{M}_{\text{int}}(w) = \hat{S}_{\text{int}} - w\hat{M}_{\text{int}}.$$

Since $\hat{S}_{\text{int}}$ is invertible, c.f Proposition 24, we can factor it out in the definition of $\mathcal{M}_{\text{int}}$,

$$\mathcal{M}_{\text{int}}(w) = \hat{S}_{\text{int}} \left(\text{Id} - w\hat{S}_{\text{int}}^{-1}\, \hat{M}_{\text{int}}\right).$$

As a result,

$$\begin{aligned} \det \mathcal{M}_{\text{int}}(w) &= \det \hat{S}_{\text{int}} \times \det\left(\text{Id} - w\hat{S}_{\text{int}}^{-1}\, \hat{M}_{\text{int}}\right) \\ &= \det \hat{S}_{\text{int}} \times w^{r-1} \det\left(w^{-1}\,\text{Id} - \hat{S}_{\text{int}}^{-1}\, \hat{M}_{\text{int}}\right). \end{aligned} \tag{89}$$

Since the characteristic polynomial of $\hat{S}_{\text{int}}^{-1}\, \hat{M}_{\text{int}}$ is of degree $r-1$ and has the form $\chi_{\hat{S}_{\text{int}}^{-1}\, \hat{M}_{\text{int}}}(w) = w\,\text{Id} - \hat{S}_{\text{int}}^{-1}\, \hat{M}_{\text{int}}$, the second factor in the RHS of (89) is the reciprocal polynomial of $\chi_{\hat{S}_{\text{int}}^{-1}\, \hat{M}_{\text{int}}}$.

It follows immediately that, the coefficient of the zero-th order term is 1, while the coefficient of the first order term, which corresponds to that of the $(r-1)$-th order term of $\chi_{\hat{S}_{\text{int}}^{-1}\, \hat{M}_{\text{int}}}$, is $-\operatorname{tr}(\hat{S}_{\text{int}}^{-1}\, \hat{M}_{\text{int}})$.

Property 2a - Invertibility: We determine w so that the matrix is definite positive. By Proposition 23, for $v \in \mathbb{R}^{r-1}$, we have

$$\begin{aligned} v \cdot \mathcal{M}_{\text{int}} v &= \int_0^1 g'^2\, d\hat{x} - w \int_0^1 g^2\, d\hat{x} \quad , \quad g := \sum_{n=1}^{r-1} v_j \hat{\phi}_j \\ &\geq (\pi^2 - w)\|g\|_{L^2}^2. \end{aligned} \tag{90}$$

The last inequality comes from the Poincaré inequality for all $f \in H_0^1(0,1)$. As a result, if $w < pi^2$, $\mathcal{M}_{\text{int}}$ is definite positive and hence thus is invertible, since $\mathcal{M}_{\text{int}}(w) = \hat{S}_{\text{int}} - w\hat{M}_{\text{int}}$ is symmetric.

Property 2b - Analyticity : For w with $|w|\,\rho(\hat{S}_{\text{int}}^{-1}\hat{M}_{\text{int}}) < 1$, the Neumann series of $w\hat{S}_{\text{int}}^{-1}\hat{M}_{\text{int}}$ converges and gives the inverse of $\text{Id} - w\hat{M}_{\text{int}}\hat{S}_{\text{int}}^{-1}$ which is then analytic in w, i.e

$$\left(\text{Id} - w\hat{S}_{\text{int}}^{-1}\, \hat{M}_{\text{int}}\right)^{-1} = \sum_{k=0}^{\infty} w^k \left(\hat{S}_{\text{int}}^{-1}\, \hat{M}_{\text{int}}\right)^k.$$

The proof is finished by using Lemma 25 which gives a bound on the spectral radius $\rho(\hat{S}_{\text{int}}^{-1}\hat{M}_{\text{int}}) < \pi^2$.

□

Remark 8. *There is another way to see that* $\det \mathcal{M}_{int}$ *is a polynomial in* w*. The* k*-th order derivative of the determinant of a matrix is given by, c.f [2],*

$$\frac{d^k}{dw^k} \det \begin{pmatrix} \text{Row}(C,1) \\ \text{Row}(C,2) \\ \vdots \\ \text{Row}(C,r) \end{pmatrix} = \sum_{\substack{k_i=0 \\ k_1+k_2+\ldots+k_r=k}}^{k} \frac{k!}{k_1!k_2!\ldots k_p!} \det \begin{pmatrix} \frac{d^{k_1}}{dw^{k_1}} \text{Row}(C,1) \\ \frac{d^{k_2}}{dw^{k_2}} \text{Row}(C,2) \\ \vdots \\ \frac{d^{k_r}}{dw^{k_r}} \text{Row}(C,r) \end{pmatrix}.$$

Since the components of $\mathcal{M}_{int}$ *are polynomials of first order in* w*,* $\det \mathcal{M}_{int}$ *is polynomial of degree* $\leq r-1$ *in* w*. In fact, the degree is exactly* $r-1$ *since coefficient of the* $(r-1)$*-th order term is* $(-1)^{r-1} \det \hat{M}_{int}$ *and is not zero by invertibility of* $\hat{M}_{int}$*, c.f Proposition 24.* ▽

C Relation between the Discrete Problems, variational problems and continuous ones

C.1 Variational Formulation

Define for $f, h \in H^1(0,1)$ the bilinear form

$$\mathsf{a}_w(f,h) := \mathsf{a}_\mathsf{S}(f,h) - w\,\mathsf{a}_\mathsf{M}(f,h) \quad ; \quad L_\mathsf{g}(h) := \mathsf{a}_\mathsf{M}(\mathsf{g},h).$$

For $g \in L^2(0,1)$, define the variational problem (VP) for a suitable Hilbert space $\mathbb{H} \subset H^1(0,1)$:

$$\text{Find } f \in \mathbb{H} : \mathsf{a}(f,\mathsf{h}) = L_\mathsf{g}(\mathsf{h}) \quad ; \quad \forall\, \mathsf{h} \in \mathbb{H}. \tag{91}$$

In $H_0^1(0,1)$, by Poincaré Lemma, the H^1 semi-norm $\|\cdot\|_1$ defined by $\|f\|_1 := \|f'\|_{L^2}$ is a norm equivalent to the usual H^1-norm. Using the optimal Poincaré constant $1/\pi$, c.f Subsection D.1, we obtain the boundedness (for all w) and coercitivity (for $0 \le w \le \pi^2$) of a with respect to $\big(H_0^1, \|\cdot\|_1\big)$,

$$|\mathsf{a}_w(f,\mathsf{h})| \le (1 + |w|\,\pi^{-2})\|f'\|_{L^2}\|\mathsf{h}'\|_{L^2} \quad , \quad \forall\, f, \mathsf{h} \in H_0^1; \tag{92}$$

$$\mathsf{a}_w(f,f) = \|f'\|_{L^2}^2 - w\,\|f\|_{L^2}^2 \ge (1 - w\,\pi^{-2})\|f'\|_{L^2}^2 \quad , \quad \forall\, f \in H_0^1. \tag{93}$$

Hence, by Lax-Milgram's theorem, the variational problems in $(H_0^1, \|\cdot\|_1)$ and $(P_{r,D}, \|\cdot\|_1)$ have existence and uniqueness of solution.

For $\mathsf{g} \in L^2$ and $0 \le w \le \pi^2$, we denote by f the solution of the VP for H_0^1, and by f_h that for $P_{r,D}$, i.e.

$$\begin{aligned} \mathsf{f} \in H_0^1 \quad &: \quad \mathsf{a}_w(\mathsf{f},\mathsf{h}) = L_\mathsf{g}\mathsf{h} \quad , \forall\, \mathsf{h} \in H_0^1; \\ f_h \in P_{r,D} \quad &: \quad \mathsf{a}_w(f_h,\mathsf{h}) = L_\mathsf{g}\mathsf{h} \quad , \forall\, \mathsf{h} \in P_{r,D}. \end{aligned}$$

Follows directly from this is the a -orthogonality of $\mathsf{f} - f_h$ to $P_{r,D}$.

$$\mathsf{a}_w\,(\mathsf{f} - f_h\,,\,\mathsf{h}) = 0 \quad , \quad \forall\, h \in P_{r,D}. \tag{94}$$

We also recall Lemma of Céa for symmetric bilinear form gives

$$\big\|(\mathsf{f} - f_h)'\big\|_{L^2} \le \left(\frac{1 + |w|\pi^{-2}}{1 - |w|\pi^{-2}}\right)^{1/2} \|(\mathsf{f} - \mathsf{h})'\|_{L^2} \quad , \quad \forall\, \mathsf{h} \in P_{r,D}. \tag{95}$$

C.2 Main results

With $0 \le w < \pi^2$ and $\mathsf{g} \in P_r(0,1)$, we will compare the solutions of the following problems

1. f the solution of the variational problem (119) in H_0^1,

$$\mathsf{f} \in H_0^1 \quad : \quad \mathsf{a}_w(\mathsf{f},\mathsf{h}) = L_\mathsf{g}\mathsf{h} \quad , \forall\, \mathsf{h} \in H_0^1.$$

This solution also uniquely solves the Dirichlet BVP,

$$-\mathsf{f}'' - w\mathsf{f}' = \mathsf{g} \quad ; \quad \mathsf{f}(1) = \mathsf{f}(1) = 0. \tag{96}$$

2. $f_h \in P_{r,D}$ the solution of the variational problem (119) in $P_{r,D}$,

$$f_h \in P_{r,D} \quad : \quad \mathsf{a}_w(f_h, \mathsf{h}) = L_{\mathsf{g}}\mathsf{h} \quad , \forall\, \mathsf{h} \in P_{r,D}. \tag{97}$$

3. The discrete problem whose solution is given by $[p]_{\rm int}$ with $p \in P_{r,D}$,

$$\mathcal{M}_{\rm int}\,[p]_{\rm int} = \hat{M}_{{\rm int},\star}[\mathsf{g}] \tag{98}$$

Remark 9. *For the given form of the matrix problem* (98)*, we have used the isomorphism between $P_{r,D}$ and $\mathbb{R}^{r-1}$ discussed in Subsection B.4. Note also that at $w = 0$ the $\mathcal{M}_{int}$ becomes $\hat{S}_{int}$. The invertibility of $\mathcal{M}_{int}$ and $\hat{S}_{int}$, for $0 \leq w < \pi^2$, are proved in Subsection B.5.* △

Relation between the VP (97) **and the discrete problem** : We first establish the relation between the VP (97) and the discrete problem(98) for both $w = 0$ and $0 < w < \pi^2$.

Lemma 27. *For $\mathsf{g} \in P_r$, the variational solution f_h to VP* (97) *gives the solution to* (98)*; explicitly, we have*

$$\hat{S}_{int}\,[f_h]_{int} = \hat{M}_{int,\star}\,[\mathsf{g}] \qquad ; \qquad \mathcal{M}_{int}\,[f_h]_{int} = \hat{M}_{int,\star}\,[\mathsf{g}]\,, \quad w < \pi^2.$$

Proof. For $h \in P_{r,D}$, using Corollary 23, we can write the RHS of (97) the variational problem solved by f_h as

$$L_{\mathsf{g}}(\mathsf{h}) = \mathsf{a}_{\mathsf{M}}(\mathsf{g}, h) = [h] \cdot \hat{M}[g] = [h]_{\rm int} \cdot \hat{M}_{{\rm int},\star}[g].$$

We also use Corollary 23, together with the isomorphism between $\mathbb{R}^{n-1}$ and $P_{r,D}$ to rewrite the LHS of (97). As a result, we can write (97) in matrix form,

$$v \cdot \hat{S}_{\rm int}\,[f_h]_{\rm int} - v \cdot \hat{M}_{\rm int}\,[f_h]_{\rm int} = v \cdot \hat{M}_{{\rm int},\star}[g] \quad , \quad \forall\, v \in \mathbb{R}^{r-1}.$$

$$\Rightarrow \mathcal{M}_{\rm int}[f_h]_{\rm int} = \hat{M}_{{\rm int},\star}[g].$$

□

Relation between the discrete problem (97) **and the continuous problem** : This is done for $w = 0$ and $0 < \omega < \pi^2$.

Case 1 : $w = 0$ For $p \in P_{r-2}$, we will show that the VP solution f_h in $P_{r,D}$ concides with the solution $\mathsf{f} = \Delta_D^{-1}\mathsf{g}$ to (122).

Remark 10. *If g is polynomial, then so is $\mathsf{f} = \Delta_D^{-1}\mathsf{g}$; furthermore, we have,*

$$\Delta_D^{-k}\mathsf{g} \in P_{\deg \mathsf{g}+2k,D} \quad ; \quad \frac{d^{2k}}{d\,x^{2k}}(\Delta_D^{-k} f) = (-1)^k f. \tag{99}$$

This can be seen from the explicit form of Δ_D^{-1},

$$\Delta_D^{-1}\mathsf{g} = -\int_0^x \int_0^y \mathsf{g}(s)\,ds\,dy + x\int_0^1 \int_0^y \mathsf{g}(s)\,ds\,dy. \tag{100}$$

Implicitly hidden in the above statement is that the kernel of $-f'' = 0$ is polynomial, which is not the case for $-f'' - wf = g$ when $w > 0$. △

Proposition 28. *For* $\mathsf{g} \in P_{r-2}$*, we have*

$$\hat{S}_{int}^{-1}\,\hat{M}_{int,\star}\,[\mathsf{g}] \quad = \quad \left[\Delta_D^{-1}\,\mathsf{g}\right]_{int}.$$

In general, if $\deg \mathsf{g} + 2k \leq r$ *then*

$$\left(\hat{S}_{int}^{-1}\hat{M}_{int}\right)^{k-1}\,\hat{S}_{int}^{-1}\,\hat{M}_{int,\star}\,[\mathsf{g}] \quad = \quad \left[\Delta_D^{-k}\,\mathsf{g}\right]_{int}. \tag{101}$$

Proof. **Property 1 :** We have shown in Lemma 27 that

$$\hat{S}_{\rm int}^{-1}\,\hat{M}_{{\rm int},\star}\,[\mathsf{g}] = [f_h]_{\rm int}\,.$$

On the other hand, as we have discussed above, since $p \in P_{r-2}$, $\mathsf{f} = \Delta_D^{-1}\mathsf{g} \in P_{r,D}$. At the same time, being the solution in H_0^1, it satisfies

$$\mathsf{a_S}(\mathsf{f},\mathsf{h}) = L_{\mathsf{g}}\mathsf{h} \quad , \quad \mathsf{h} \in P_{r,D}.$$

By uniqueness of solution of the VP in $P_{r,D}$, this means that

$$f_h = \mathsf{f} = \Delta_D^{-1}\mathsf{g}.$$

Property 2 : We prove (101) by induction for $1 \leq k \leq k_0 := [\frac{r-\deg p}{2}] - 1$. We have shown that (101) holds for $k = 1$ in Property 1. Suppose the statement is true for $k-1$ with $k-1 < k_0$, i.e.

$$\left(\hat{S}_{\rm int}^{-1}\,\hat{M}_{\rm int}\right)^{k-1}\,\hat{S}_{\rm int}^{-1}\,\hat{M}_{{\rm int},\star}\,[p] = \left[\Delta_D^{-k}p\right]_{\rm int}\,.$$

We next verify for k. We first note that, for $k-1 < k_0$, $k \leq k_0$, and as a result

$$\deg \Delta_D^{-k}\,p = \deg p + 2k \leq \deg p + 2k_0 = \deg p + 2[\frac{r-\deg p}{2}] - 2 \leq r-2.$$

As a result, $\Delta_D^{-k}\,p \in P_{r-2,D}$.

$$\begin{aligned}\left(\hat{S}_{\rm int}^{-1}\,\hat{M}_{\rm int}\right)^{k}\,\hat{S}_{\rm int}^{-1}\,\hat{M}_{{\rm int},\star}\,[p] &= \hat{S}_{\rm int}^{-1}\hat{M}_{\rm int}\left[\left(\hat{S}_{\rm int}^{-1}\,\hat{M}_{\rm int}\right)^{k-1}\,\hat{S}_{\rm int}^{-1}\,\hat{M}_{{\rm int},\star}\,[p]\right]\\ &\overset{\text{induction hypothesis}}{=} \hat{S}_{\rm int}^{-1}\hat{M}_{\rm int}\,\left[\Delta_D^{-k}\,p\right]_{\rm int}\\ &\overset{(84)}{=} \hat{S}_{\rm int}^{-1}\hat{M}_{{\rm int},\star}\,\left[\Delta_D^{-k}\,p\right].\end{aligned}$$

Since $\Delta_D^{-k}p \in P_{r-2,D}$ we can apply Property 1 and obtain

$$\left(\hat{S}_{\rm int}^{-1}\,\hat{M}_{\rm int}\right)^{k}\,\hat{S}_{\rm int}^{-1}\,\hat{M}_{{\rm int},\star}\,[p] = \left[\Delta_D^{-1}\,(\Delta_D^{-k}\,p)\right]_{\rm int} = \left[\Delta_D^{-(k+1)}p\right]_{\rm int}\,.$$

This finishes the proof.

□

Remark 11. *We can also prove Step 1 directly. Applying Proposition 22 to* $p \in P_r$*, we obtain*

$$1 \leq j \leq r-1: \quad \left(\hat{M}_{int,\star}\,[p]\right)_j = \mathsf{a_M}\left(\hat{\phi}_{j+1}\,,\,p\right) \overset{(99)}{=} \mathsf{a_M}\left(\hat{\phi}_{j+1}\,,\,-(\Delta_D^{-1}p)''\right).$$

On the other hand, since $p \in P_r$ *with* $\deg p + 2 \leq r$*,* $\Delta_D^{-1}p \in P_{r,D}$*. Hence we can apply* (84)*, c.f. Proposition 22, to* $\Delta_D^{-1}p$*, and obtain*

$$1 \leq j \leq r-1: \quad \left(\hat{S}_{int}\,[\Delta_D^{-1}\,p]_{int}\right)_j = \mathsf{a_M}\left(\hat{\phi}_{j+1}\,,\,-(\Delta_D^{-1}\,p)''\right). \qquad \triangle$$

Case 2 : $0 < w < \pi^2$ When $w > 0$ and for any $g \in L^2$, due to the existence of the of a non-polynomial kernel of the continuous problem $-f'' - w'' = g$ generated by $\sin(w^{1/2}x)$ and $\cos(w^{1/2}x)$, the exact solution f is no longer polynomial, which is the case when $w = 0$, c.f Remark 10. Nevertheless, we still have analogous statement to Proposition 28, but this time we incur an error term but with well-known bound.

Proposition 29. *For* g *and* $\tilde{\mathsf{g}} \in P_r$*, with* $0 < w < \pi^2$*, we have*

$$\hat{M}_{int,\star}[\mathsf{g}] \cdot \mathcal{M}_{int}^{-1} \hat{M}_{int,\star}[\tilde{\mathsf{g}}] = \mathsf{a}_w\left(\tilde{\mathsf{f}}\,, \mathsf{f}\right) + \mathsf{a}_w\left(\tilde{f}_h - \tilde{\mathsf{f}}\,, \mathsf{f} - f_h\right).$$

where $\mathsf{f}, \tilde{\mathsf{f}}$ *are the exact solutions to* (122) *with inhomogeneous term* g, $\tilde{\mathsf{g}}$*, respectively, and* $f_h, \tilde{f}_h \in P_{r,D}$ *the solutions to the variational problem* (97) *in* $P_{r,D}$ *with inhomogeneous term* g, $\tilde{\mathsf{g}}$*, respectively. Specifically,*

$$\begin{array}{l} -f'' - wf = g\ ,\ f(0) = f(1) = 0 \\ -\tilde{f}'' - w\tilde{f} = \tilde{g}\ ,\ \tilde{f}(0) = \tilde{f}(1) = 0 \end{array} \quad ; \quad \begin{array}{l} a_w(f_h, h) = L_g h \\ a_w(\tilde{f}_h, h) = L_{\tilde{g}} h \end{array} \quad ,\forall\, h \in P_{r,D}.$$

Each term in the above expression is analytic in w*, with the errorl term having the bound,*

$$\mathsf{a}_w\left(\tilde{f}_h - \tilde{\mathsf{f}}\,, \mathsf{f} - f_h\right) \leq \frac{(\pi^2 + |w|)^2}{(\pi^2 - |w|)^3}\, \pi^{-2\mathsf{n}(\mathsf{g}) - 2\mathsf{n}(\tilde{\mathsf{g}})}\, w^{\mathsf{n}(\mathsf{g}) + \mathsf{n}(\tilde{\mathsf{g}}) + 2}\, \|\mathsf{g}\|_{L^2}\, \|\tilde{\mathsf{g}}\|_{L^2}.$$

where the $\mathsf{n}(g)$ *is defined in terms of* r *and the degree* $\deg g$ *of polynomial* g *,*

$$\mathsf{n}(g) := \left[\frac{r - \deg g}{2}\right] - 1. \tag{102}$$

Proof. Lemma 27 relates the solution of the variational problem in $P_{r,D}$ to the matrix problem (98), explicitly we have

$$\mathcal{M}_{\mathrm{int}}^{-1} \hat{M}_{\mathrm{int},\star}[\tilde{\mathsf{g}}] = \left[\tilde{f}_h\right]_{\mathrm{int}} \quad ; \quad \mathcal{M}_{\mathrm{int}}^{-1} \hat{M}_{\mathrm{int},\star}[\mathsf{g}] = [f_h]_{\mathrm{int}}\,.$$

And we can write

$$\hat{M}_{\mathrm{int},\star}[g] \;=\; \mathcal{M}_{\mathrm{int}}\, [f_h]_{\mathrm{int}}.$$

Since f_h and $\tilde{f}_h \in P_{r,D}$, we can apply Proposition 22 to obtain

$$\hat{M}_{\mathrm{int},\star}[\mathsf{g}] \cdot \mathcal{M}_{\mathrm{int}}^{-1} \hat{M}_{\mathrm{int},\star}[\tilde{\mathsf{g}}] = \left[\tilde{f}_h\right]_{\mathrm{int}} \cdot \hat{M}_{\mathrm{int},\star}[\mathsf{g}] = \left[\tilde{f}_h\right]_{\mathrm{int}} \cdot \mathcal{M}_{\mathrm{int}}[f_h]_{\mathrm{int}} = \mathsf{a}\left(\tilde{f}_h\,,\, f_h\right). \tag{103}$$

Using the a-orthogonality between $\mathsf{f} - f_h$, $\tilde{\mathsf{f}} - \tilde{f}_h$ with $P_{r,D}$,

$$\mathsf{a}\left(\mathsf{f} - f_h\,, \tilde{f}_h\right) = \mathsf{a}\left(\tilde{\mathsf{f}} - \tilde{f}_h\,, f_h\right) = 0,$$

we replace the last expression of (103) with

$$\begin{aligned} \mathsf{a}\left(\tilde{f}_h\,,\, f_h\right) &= \mathsf{a}\left(\tilde{f}_h\,,\, f_h - \mathsf{f} + \mathsf{f}\right) \quad = \mathsf{a}\left(\tilde{f}_h\,, \mathsf{f}\right) \quad = \mathsf{a}\left(\tilde{f}_h - \tilde{\mathsf{f}} + \tilde{\mathsf{f}}\,, \mathsf{f}\right) \\ &= \mathsf{a}\left(\tilde{f}_h - \tilde{\mathsf{f}}\,, \mathsf{f}\right) + \mathsf{a}\left(\tilde{\mathsf{f}}\,, \mathsf{f}\right) \quad = \mathsf{a}\left(\tilde{f}_h - \tilde{\mathsf{f}}\,, \mathsf{f} - f_h + f_h\right) + \mathsf{a}\left(\tilde{\mathsf{f}}\,, \mathsf{f}\right) \\ &= \mathsf{a}\left(\tilde{\mathsf{f}}\,, \mathsf{f}\right) + \mathsf{a}\left(\tilde{f}_h - \tilde{\mathsf{f}}\,, \mathsf{f} - f_h\right). \end{aligned}$$

Hence we have arrived at

$$\hat{M}_{\mathrm{int},\star}[\mathsf{g}] \cdot \mathcal{M}_{\mathrm{int}}^{-1}\, \hat{M}_{\mathrm{int},\star}[\tilde{\mathsf{g}}] \quad = \quad \mathsf{a}\left(\tilde{\mathsf{f}}\,, \mathsf{f}\right) + \mathsf{a}\left(\tilde{f}_h - \tilde{\mathsf{f}}\,, \mathsf{f} - f_h\right).$$

Part 2 : We next obtain a bound for $\mathsf{a}\left(\tilde{f}_h - \tilde{\mathsf{f}}\,, \mathsf{f} - f_h\right)$. By the H_0^1 boundedness of a_w, c.f (92), we have an initial bound,

$$\mathsf{a}_w\left(\tilde{f}_h - \tilde{\mathsf{f}}\,, \mathsf{f} - f_h\right) \;\leq\; \left(1 + |w|\,\pi^{-2}\right) \left\| \left(\tilde{f}_h - \tilde{\mathsf{f}}\right)' \right\|_{L^2} \left\| (f_h - \mathsf{f})' \right\|_{L^2}. \tag{104}$$

We next use Céa 's Lemma (95) to bound the difference $f_h - \mathsf{f}$, and $\tilde{f}_h - \tilde{\mathsf{f}}$; since the two bounds are similar, we only write that for f_h,

$$\left\| (\mathsf{f} - f_h)' \right\|_{L^2} \leq \left(\frac{1 + |w|\pi^{-2}}{1 - |w|\pi^{-2}} \right)^{1/2} \|(\mathsf{f} - \mathsf{h})'\|_{L^2} \quad , \quad \forall\, \mathsf{h} \in P_{r,D}. \tag{105}$$

We are going choose a candidate in $P_{r,D}$ to bound the RHS of (105). For this purpose, we turn to Proposition 34, which gives that, for $|w| < \pi^2$, f is analytic in w and has the follwing expansion, for any $N \geq 0$,

$$\mathsf{f} = H_N \mathsf{g} + w^{N+1} E_N \mathsf{g} \quad ; \quad H_N \mathsf{g} := \sum_{k=0}^{N} w^k\, \Delta_D^{-(k+1)}\, \mathsf{g}. \tag{106}$$

We also recall the bound on the error term,

$$\|(E_N \mathsf{g})'\|_{L^2} \leq \frac{\pi^{1-2N}}{\pi^2 - w} \|\mathsf{g}\|_{L^2}.$$

By Remark 10, the approximating function $H_N \mathsf{g} \in P_{\deg g + 2(N+1),D}$. By its definition, $\mathsf{n}(g)$ is the largest index N so that $H_N \mathsf{g} \in P_{r,D}$. In fact, the degree of $H_{\mathsf{n}(g)} g$ is equal to r when $r - \deg g$ is even, and is to $r-1$ when $r - \deg g$ odd. This means we can choose $H_{\mathsf{n}(\mathsf{g})}$ to bound the distance between f and $P_{r,D}$, which is in the RHS of (105),

$$\left\| (\mathsf{f} - f_h)' \right\|_{L^2} \;\leq\; w^{\mathsf{n}(\mathsf{g})+1} \left(\frac{1 + |w|\pi^{-2}}{1 - |w|\pi^{-2}} \right)^{1/2} \frac{\pi^{1-2\mathsf{n}(\mathsf{g})}}{\pi^2 - w} \|\mathsf{g}\|_{L^2}.$$

We use the above inequality in (104) and obtain the stated bound .

□

Remark 12 (The difference between the exact solution and the approximate Galerkin solution)**.** *The current notation remains the same as in previous discussion. By Proposition 26 and 34, for* $|w| < \pi^2$*, both* f *and* f_h *are analytic in* w*. By Céa's Lemma, we have the bound* (105)*, which implies that first* $\mathsf{n}(\mathsf{g}) - 1$ *terms in their expansions in* w *coincide. This can be seen independenly as follows.*

By Proposition 26, for $0 < w < \pi^2$*,* $\mathcal{M}_{int}^{-1}$ *is analytic in* w*, and for any* $N \geq 0$*, we can write*

$$[f_h]_{int} = \mathcal{M}_{int}^{-1} \hat{M}_{int,\star}[g] = \sum_{k=0}^{N} w^k \left(\hat{S}_{int}^{-1}\, \hat{M}_{int} \right)^k \hat{S}_{int}^{-1}) \hat{M}_{int,\star}[g] + w^{N+1}\, [e_N g]_{int};$$

with

$$e_N g \in P_{r,D} \quad \textit{and} \quad [e_N g]_{int} := \left(\hat{S}_{int}^{-1}\hat{M}_{int}\right)^N \mathcal{M}_{int}^{-1}\hat{M}_{int,\star}[g].$$

We can apply Proposition 28, to the first $\mathsf{n}(g)$ *terms and obtain*

$$\left(\hat{S}_{int}^{-1}\,\hat{M}_{int}\right)^k \hat{S}_{int}^{-1}\,\hat{M}_{int,\star}[g] = \left[\,\Delta_D^{-(k+1)}\,g\right]_{int} \quad , \quad 0 \le k \le \mathsf{n}(g). \tag{107}$$

The sum of these terms are in fact $H_{\mathsf{n}(g)}g$*, c.f* (106)*. As a result, we can write*

$$[f_h]_{int} = \left[H_{\mathsf{n}(\mathsf{g})}\mathsf{g}\right]_{int} + w^{\mathsf{n}(\mathsf{g})+1}\left[e_{\mathsf{n}(g)}\mathsf{g}\right]_{int}.$$

$$\Longrightarrow \quad f_h = H_{\mathsf{n}(g)}g + w^{\mathsf{n}(g)+1}e_{\mathsf{n}(g)}g = \mathsf{f} + H_{\mathsf{n}(g)}g - \mathsf{f} + w^{\mathsf{n}(g)+1}e_{\mathsf{n}(g)}g.$$

In short, the difference between $f_h \in P_{r,D}$ *and* f *is given by*

$$P_{r,D} \ni f_h = \mathsf{f} - w^{\mathsf{n}(g)+1}E_{\mathsf{n}(g)}g + w^{\mathsf{n}(g)+1}e_{\mathsf{n}(g)}g. \qquad \triangle$$

C.3 Applications (Important technical lemmas)

Lemma 30. *We have the identities*

$$\hat{S}_{int}^{-1}\mathfrak{a}_0 = \left[\hat{\phi}_r - x\right]_{int} \quad ; \quad \hat{S}_{int}^{-1}\mathfrak{b}_0 = \left[\hat{\phi}_0 + x - 1\right]_{int}.$$

$$\mathfrak{a}_0 \cdot \hat{S}_{int}^{-1}\mathfrak{b}_0 = \hat{S}_{r0} + 1 \quad ; \quad \mathfrak{b}_0 \cdot \hat{S}_{int}^{-1}\mathfrak{b}_0 = \hat{S}_{00} - 1.$$

Proof. The components of $\mathfrak{a}_0$ are given by, c.f. (21),

$$1 \le j \le r-1 : \quad (\mathfrak{a}_0)_j = \mathsf{a}_\mathsf{S}(\hat{\phi}_j, \hat{\phi}_r) = \mathsf{a}_\mathsf{M}(\hat{\phi}_j, -\hat{\phi}_r'').$$

The last equality is obtained by integrating by parts and noting that $\hat{\phi}_j(0) = \hat{\phi}_j(1) = 0$ for $1 \le j \le r-1$. By Proposition 22, we can then write $\mathfrak{a}_0$ as

$$\mathfrak{a}_0 = \hat{M}_{\text{int},\star}\left[-\hat{\phi}_r''\right].$$

Applying Proposition 28, we obtain

$$\hat{S}_{\text{int}}^{-1}\mathfrak{a}_0 = \hat{S}_{\text{int}}^{-1}\hat{M}_{\text{int},\star}\left[-\hat{\phi}_r''\right] = [g]_{\text{int}} \ ,$$

where g solves the BVP $-g'' = -\hat{\phi}_r''$, $g(0) = g(1) = 0$, whose unique solution is given by $\hat{\phi}_r - x$.

The same reasoning applies to obtain the result with $\mathfrak{b}_0$, which can be written as $\mathfrak{b}_0 = \hat{M}_{\text{int},\star}\left[-\hat{\phi}_0''\right]$, c.f. (21). As a result, $\hat{S}_{\text{int}}^{-1}\mathfrak{b}_0 = [g]_{\text{int}}$ where g solves the BVP $-g'' = -\hat{\phi}_0$, $g(0) = g(1) = 0$, whose unique solution is given by $\hat{\phi}_0 + x - 1$.

Using the above calculations, we obtain directly

$$\mathfrak{a}_0 \cdot \hat{S}_{\text{int}}^{-1}\mathfrak{b}_0 = \mathsf{a}_\mathsf{S}(\hat{\phi}_r, \hat{\phi}_0 + x - 1) = \hat{S}_{r0} + \int_0^1 \hat{\phi}_r'\,dx = \hat{S}_{r0} + \hat{\phi}_r(1) - \hat{\phi}_r(0) = \hat{S}_{r0} + 1.$$

$$\mathfrak{b}_0 \cdot \hat{S}_{\text{int}}^{-1}\mathfrak{b}_0 = \mathsf{a}_\mathsf{S}\left(\hat{\phi}_0, \hat{\phi}_0 + x - 1\right) = \hat{S}_{00} + \int_0^1 \phi_0'\,dx = \hat{S}_{00} + \hat{\phi}_0(1) - \hat{\phi}_0(0) = \hat{S}_{00} - 1.$$

□

We next apply of Proposition 29 to obtain explicit description of the coefficients of the first r terms in the expansion in w of $\kappa \cdot \mathcal{M}_{\rm int}^{-1}\kappa$. This plays a crucial role in the results of Section 4. We recall the definition of κ,

$$\kappa = (\kappa_1, \ldots, \kappa_{r-1}) \quad ; \quad \kappa_j = \mathsf{a}_{\mathsf{M}}\big(\hat{\phi}_j\,,\,1\big) \quad , \quad 1 \le j \le r-1.$$

Here 1 denotes the constant function taking value 1.

Proposition 31. *For $|w| < \pi^2$, the Taylor expansion of the rational function $\kappa \cdot \mathcal{M}_{int}^{-1}(w)\kappa$ converges and is given by,*

$$\kappa \cdot \mathcal{M}_{int}^{-1}(w)\,\kappa = -w^{-1} + \frac{2 - 2\cos w^{1/2}}{w^{3/2}\sin w^{1/2}} + w^{2[r/2]}\,\mathsf{e}(\kappa),$$

With ***const***(w) *defined in* (102)*, the analytic error term $\mathsf{e}(\kappa)$ is bounded by,*

$$|\mathsf{e}(\kappa)| \le \pi^{4(1-[r/2])}\,\frac{(\pi^2 + |w|)^2}{(\pi^2 - |w|)^3}.$$

Proof. By Proposition 22, we can express κ as $\kappa = \hat{M}_{1,\star}\,[1]$, and then apply Lemma 29 to evaluate

$$\kappa \cdot \mathcal{M}_{\rm int}^{-1}\kappa = \hat{M}_{{\rm int},\star}\,[1] \cdot \mathcal{M}_{\rm int}^{-1}\hat{M}_{{\rm int},\star}\,[1] \quad \overset{\text{Lemma 29}}{=} \mathsf{a}_w\,(f, f) + \mathsf{O}(w^{2[r/2]}).$$

Here $f = (\Delta_D - w)^{-1}1$ is the unique solution to the BVP

$$-f'' - wf = 1 \quad ; \quad f(0) = f(1).$$

Since $f \in H_0^1$ we do an integration by parts and get

$$\kappa \cdot \mathcal{M}_{\rm int}^{-1}\kappa = \int_0^1 f(-f'' - wf)\,dx + \mathsf{O}(w^{2[r/2]}) = \int_0^1 f\,dx + \mathsf{O}(w^{2[r/2]}).$$

It remains to evaluate $\int_0^1 f\,dx$,

$$wf = -1 - f'' \quad \Rightarrow \quad w\int_0^1 f\,dx = -\int_0^1 dx - \int_0^1 f''\,dx = -1 + f'(0) - f'(1).$$

We can solve explicitly for f,

$$f = \frac{\cos(w^{1/2}x) - 1}{w} - \frac{\cos w^{1/2} - 1}{w\sin w^{1/2}}\sin(w^{1/2}x)$$

$$\Rightarrow f' = -\frac{\sin(w^{1/2}x)}{w^{1/2}} - \frac{\cos w^{1/2} - 1}{w^{1/2}\sin w^{1/2}}\cos(w^{1/2}x).$$

In particular, we have

$$f'(0) = -\frac{\cos w^{1/2} - 1}{w^{1/2}\sin w^{1/2}} \quad ; \quad f'(1) = -\frac{\sin(w^{1/2})}{w^{1/2}} - \frac{\cos w^{1/2} - 1}{w^{1/2}\sin w^{1/2}}\cos(w^{1/2}). \tag{108}$$

Inserting in the values of $f'(0)$ and $f'(1)$ give in (108), we obtain

$$w\int_0^1 f\,dx = -1 + \frac{\sin w^{1/2}}{w^{1/2}} + \frac{(\cos w^{1/2} - 1)^2}{w^{1/2}\sin w^{1/2}}.$$

The RHS simplifies to give the leading term stated in the lemma.

It remains to obtain the bound for the error term, which is also given in Proposition 29. We note that the corresponding inhomogeneous term for our case is the constant polynomial -1, with

$$\mathsf{n}(-1) = [r/2] - 1 \quad ; \quad \|-1\|_{L^2} = 1\,.$$

As a result, we obtain the bound as stated. □

We next carry out the same calculation for $\mathfrak{a} \cdot \mathcal{M}_{\rm int}^{-1}\mathfrak{b}$. This will play a crucial role in the results of Lemma 4. We first define the following vectors, c.f. Remark 13 for a dicussion of their introduction,

$$\mathsf{Y} := \hat{M}_{\rm int}\hat{S}_{\rm int}^{-1}\mathfrak{b}_0 - \mathfrak{b}_1 \quad ; \quad \mathsf{X} := \hat{M}_{\rm int}\hat{S}_{\rm int}^{-1}\mathfrak{a}_0 - \mathfrak{a}_1. \tag{109}$$

As a result, we have

$$\mathfrak{b}_1 = \hat{M}_{\rm int}\hat{S}_{\rm int}^{-1}\mathfrak{b}_0 - \mathsf{Y} \quad ; \quad \mathfrak{a}_1 = \hat{M}_{\rm int}\hat{S}_{\rm int}^{-1}\mathfrak{a}_0 - \mathsf{X}.$$

We insert these above expressions into the definition of $\mathfrak{b}$ and $\mathfrak{a}$ to obtain,

$$\begin{aligned} \Rightarrow \mathfrak{b} = \mathfrak{b}_0 - w\mathfrak{b}_1 &= \hat{S}_{\rm int}\hat{S}_{\rm int}^{-1}\mathfrak{b}_0 + w\left(\mathsf{Y} - \hat{M}_{\rm int}\hat{S}_{\rm int}^{-1}\mathfrak{b}_0\right) = \mathcal{M}_{\rm int}\hat{S}_{\rm int}^{-1}\mathfrak{b}_0 + w\mathsf{Y}; \\ \mathfrak{a} = \mathfrak{a}_0 - w\mathfrak{a}_1 &= \mathcal{M}_{\rm int}\hat{S}_{\rm int}^{-1}\mathfrak{a}_0 + w\mathsf{X}. \end{aligned} \tag{110}$$

The advantage of using these new vectors can be seen as follows. Using Corollary 30 we have

$$\hat{S}_{\rm int}^{-1}\mathfrak{a}_0 = \Big[\big(\hat{\phi}_{r+1} - x\big)\Big]_{\rm int} \quad ; \quad \hat{S}_{\rm int}^{-1}\mathfrak{b}_0 = \mathcal{V}\big(\hat{\phi}_0 + x - 1\big).$$

and then Proposition 22 to write the components of X and Y as : for $1 \le j \le r-1$,

$$\begin{aligned} \pi_j\mathsf{X} &= \pi_j\hat{M}_{\rm int}\hat{S}_{\rm int}^{-1}\mathfrak{a}_0 - \pi_j\mathfrak{a}_1 = \mathsf{a}_{\mathsf{M}}(\hat{\phi}_j, \hat{\phi}_r - x) - \mathsf{a}_{\mathsf{M}}(\hat{\phi}_j, \hat{\phi}_r) = \mathsf{a}_{\mathsf{M}}(\hat{\phi}_j, -x); \\ \pi_j\mathsf{Y}_j &= \pi_j\hat{M}_{\rm int}\hat{S}_{\rm int}^{-1}\mathfrak{b}_0 - \pi_j\mathfrak{b}_1 = \mathsf{a}_{\mathsf{M}}(\hat{\phi}_j, \hat{\phi}_0 + x - 1) - \mathsf{a}_{\mathsf{M}}(\hat{\phi}_j, \hat{\phi}_0) = \mathsf{a}_{\mathsf{M}}(\hat{\phi}_j, x - 1). \end{aligned}$$

This means that

$$\mathsf{X} = \hat{M}_{{\rm int},\star}\,[-x] \quad ; \quad \mathsf{Y} = \hat{M}_{{\rm int},\star}\,[x-1]\,. \tag{111}$$

Proposition 32. *For $|w| < \pi^2$, the Taylor expansion of the rational function $\mathfrak{a}(w)\cdot\mathcal{M}_{int}^{-1}(w)\mathfrak{b}(w)$ converges and is given by,*

$$\mathfrak{a}(w)\cdot\mathcal{M}_{int}^{-1}(w)\,\mathfrak{b}(w) = \hat{S}_{r0} - w\hat{M}_{r0} + \frac{w^{1/2}}{\sin w^{1/2}} + w^{2[\frac{r-1}{2}]+2}\,\mathsf{e}(\mathfrak{a},\mathfrak{b})\,.$$

*With **const**(w) defined in* (102)*, the analytic error term $\mathsf{e}(\mathfrak{a},\mathfrak{b})$ is bounded by*

$$|\mathsf{e}(\mathfrak{a},\mathfrak{b})| \le 3\,\pi^{4(1-[\frac{r-1}{2}])}\frac{(\pi^2+|w|)^2}{(\pi^2-|w|)^3}.$$

Proof. **Step 1 :** We first rewrite $\mathfrak{a}$ and $\mathfrak{b}$ by using their equivalent expressions (110) in terms of X and Y,

$$\begin{aligned} \mathfrak{a}\cdot\mathcal{M}_{\rm int}^{-1}\mathfrak{b} &= \Big(\mathcal{M}_{\rm int}\hat{S}_{\rm int}^{-1}\mathfrak{a}_0 + w\mathsf{X}\Big)\cdot\mathcal{M}_{\rm int}^{-1}\Big(\mathcal{M}_{\rm int}\hat{S}_{\rm int}^{-1}\mathfrak{b}_0 + w\mathsf{Y}\Big) \\ &= \Big(\mathcal{M}_{\rm int}\hat{S}_{\rm int}^{-1}\mathfrak{a}_0 + w\mathsf{X}\Big)\cdot\Big(\hat{S}_{\rm int}^{-1}\mathfrak{b}_0 + w\mathcal{M}_{\rm int}^{-1}\mathsf{Y}\Big) \\ &= \Big(\hat{S}_{\rm int} - w\hat{M}_{\rm int}\Big)\hat{S}_{\rm int}^{-1}\mathfrak{a}_0\cdot\hat{S}_{\rm int}^{-1}\mathfrak{b}_0 + w\Big(\hat{S}_{\rm int}^{-1}\mathfrak{a}_0\cdot\mathsf{Y} + \mathsf{X}\cdot\hat{S}_{\rm int}^{-1}\mathfrak{b}_0\Big) + w^2\mathsf{X}\cdot\mathcal{M}_{\rm int}^{-1}\mathsf{Y} \\ &= \mathfrak{a}_0\cdot\hat{S}_{\rm int}^{-1}\mathfrak{b}_0 + w\Big(-\hat{M}_{\rm int}\hat{S}_{\rm int}^{-1}\mathfrak{a}_0\cdot\hat{S}_{\rm int}^{-1}\mathfrak{b}_0 + \hat{S}_{\rm int}^{-1}\mathfrak{a}_0\cdot\mathsf{Y} + \mathsf{X}\cdot\hat{S}_{\rm int}^{-1}\mathfrak{b}_0\Big) + w^2\mathsf{X}\cdot\mathcal{M}_{\rm int}^{-1}\mathsf{Y}\,, \end{aligned}$$

As a result, at the end of Step 1, we have obtained

$$\mathfrak{a}\cdot\mathcal{M}_{\rm int}^{-1}\mathfrak{b}=\mathfrak{a}_0\cdot\hat{S}_{\rm int}^{-1}\mathfrak{b}_0+w\left(\hat{S}_{\rm int}^{-1}\mathfrak{a}_0\cdot\mathsf{Y}-\mathfrak{a}_1\cdot\hat{S}_{\rm int}^{-1}\mathfrak{b}_0\right)+w^2\mathsf{X}\cdot\mathcal{M}_{\rm int}^{-1}\mathsf{Y}. \tag{112}$$

Step 2 : It remains to evaluate the three terms in the RHS. The first term is given by Corollary 30,

$$\mathfrak{a}_0\cdot\hat{S}_{\rm int}^{-1}\mathfrak{b}_0=\hat{S}_{(r+1)1}+1. \tag{113}$$

The second term can be computed readily from the form (111) of Y,

$$\begin{aligned}\hat{S}_{\rm int}^{-1}\mathfrak{a}_0\cdot\mathsf{Y}-\mathfrak{a}_1\cdot\hat{S}_{\rm int}^{-1}\mathfrak{b}_0&=\mathsf{a}_{\mathsf{M}}(\hat{\phi}_r-x,x-1)-\mathsf{a}_{\mathsf{M}}(\hat{\phi}_r,\hat{\phi}_0+x-1)\\&=-\int_0^1 x(x-1)\,dx-\hat{M}_{r0}=\tfrac{1}{6}-\hat{M}_{r0}.\end{aligned} \tag{114}$$

We calculate the third term by first replace X and Y by their equivalent expressions given by (111); after this step, we can apply Proposition 29,

$$\begin{aligned}\mathsf{X}\cdot\mathcal{M}_{\rm int}^{-1}\mathsf{Y}&=\hat{M}_{{\rm int},\star}\,[-x]\cdot\mathcal{M}_{\rm int}^{-1}\hat{M}_{{\rm int},\star}\,[x-1]\\&=\mathsf{a}_w\left(\tilde{f},f\right)+\mathsf{a}_w\left(\tilde{f}_h-\tilde{f},f-f_h\right)\\&=\mathsf{a}_M\left(\tilde{f},-f''-wf\right)+\mathsf{a}_w\left(\tilde{f}_h-\tilde{f},f-f_h\right),\end{aligned}$$

where f and $\tilde{f}$ are the unique solution to the following BVP-s

$$-\tilde{f}''-w\tilde{f}=-x\,,\ \tilde{f}(0)=\tilde{f}(1)=0\quad;\quad -f''-wf=x-1\,,\ f(0)=f(1)=0.$$

We can thus compute f and $\tilde{f}$ explicitly,

$$\tilde{f}=\frac{x}{w}-\frac{\sin(w^{1/2}x)}{w\sin w^{1/2}}\quad;\quad f=\frac{1-x}{w}-\frac{\sin(w^{1/2}(1-x))}{w\sin w^{1/2}}.$$

Using this, we calculate $\mathsf{a}_{\mathsf{M}}(\tilde{f},-f''-wf)$

$$\begin{aligned}\mathsf{a}_{\mathsf{M}}(\tilde{f},-f''-wf)&=\int_0^1\left[\frac{x}{w}-\frac{\sin(w^{1/2}x)}{w\sin w^{1/2}}\right](x-1)\,dx\\&=-\frac{1}{w^2}-\frac{1}{6w}+\frac{1}{w^{3/2}\sin w^{1/2}}.\end{aligned}$$

It remains to obtain the bound for the error term $\mathsf{a}\left(\tilde{f}_h-\tilde{f},f-f_h\right)$, which can be obtain directly from Proposition 29. In our case, the corresponding inhomogeneous term is the polynomials $-x$ and $x-1$, with

$$\mathsf{n}(-x)=\mathsf{n}(x-1)=[\tfrac{r-1}{2}]-1\quad;\quad \|x-1\|_{L^2(0,1)}=\|x\|_{L^2(0,1)}=\tfrac{1}{\sqrt{3}}.$$

As a result, we obtain

$$\mathsf{X}\cdot\mathcal{M}_{\rm int}^{-1}\mathsf{Y}=-\frac{1}{w^2}-\frac{1}{6w}+\frac{1}{w^{3/2}\sin w^{1/2}}+w^{2[\frac{r-1}{2}]}\mathsf{e}(\mathfrak{a},\mathfrak{b}). \tag{115}$$

The proof is finished by inserting (113), (114) and (115) into (112). After simplification, we obtain the stated expression in the lemma. □

Remark 13. *We will discuss why a direct application Proposition 29, without the rewrite with* X *and* Y *as we did, will introduce problems. We first write* $\mathfrak{a}$ *and* $\mathfrak{b}$ *so that they satisfy the hypothesis of Proposition 29,*

$$\begin{aligned}\mathfrak{a}(w) = \mathfrak{a}_0 - w\mathfrak{a}_1 &= \hat{S}_{int,\star}\left[\hat{\phi}_r\right] - w\hat{M}_{int,\star}\left[\hat{\phi}_r\right] \\ &= \hat{M}_{int,\star}\left[-\hat{\phi}_r'' - w\hat{\phi}_r\right] = \hat{M}_{int,\star}\left[(\Delta - w)\hat{\phi}_r\right].\end{aligned}$$

Similarly, we have

$$\mathfrak{b}(w) = \mathfrak{b}_0 - w\mathfrak{b}_1 = \hat{M}_{int,\star}\left[(\Delta_D - w)\hat{\phi}_0\right].$$

Application of Proposition 29 gives

$$\mathfrak{a}\cdot\mathcal{M}_{int}^{-1}\mathfrak{b} = \mathsf{a}_{\mathsf{M}}\left(\hat{\phi}_0 + c_1\sin(w^{1/2}x) + c_2\cos(w^{1/2}x)\,,(\Delta - w)\hat{\phi}_r\right) \;+\; \textit{Error Term.}$$

The error term is only $\mathsf{O}(w^2)$ *since* $\mathsf{n}\left((\Delta - w)\hat{\phi}_r\right) = \left[\frac{r-(r-2)}{2}\right] - 1 = 0$. *Thus without further work, we do not have information of the coefficients of first* r *terms in the expansion in* w *of the error for* $r > 2$. *As a result of this we introduced vectors* X *and* Y *which are associated with polynomials of lower degrees and with simpler definitions, c.f* (109) *and* (111). △

D Properties of the continuous problem

D.1 Optimal Poincare constant in dimension 1

Consider $f \in H_0^1(0,1)$.

1. The family of functions $\sqrt{2}\sin(n\pi x)$, $n = 1, 2, \ldots$ are eigenfunctions corresponding eigenvalues $n^2\pi^2$, $n = 1, 2, \ldots$ of symmetric operators $-\frac{d^2}{dx^2}$ with Dirichlet boundary conditions.

2. The family of functions $1\,,\ \sqrt{2}\cos(n\pi x)$, $n = 1, 2, \ldots$, are eigenfunctions corresponding eigenvalues $n^2\pi^2$, $n = 0, 1, 2, \ldots$ of symmetric operators $-\frac{d^2}{dx^2}$ with Neumann boundary conditions.

As a result, each of these families of functions forms an orthonormal basis for $L^2(0,1)$ equipped with the usual L^2 inner product. Using this fact, for $f \in H_0^1(0,1)$, we write

$$f = \sum_{n=1}^{\infty} f_n\sqrt{2}\sin(n\pi x) \quad \Rightarrow f' = \sum_{n=1}^{\infty} n\pi\,\sqrt{2}f_n\cos(n\pi x)\,.$$

With Bessel's equality, we have

$$\|f\|_{L^2}^2 = \sum_{n=1}^{\infty} f_n^2 \quad ; \quad \|f'\|_{L^2}^2 = \pi^2\sum_{n=1}^{\infty} n^2 f_n^2.$$

$$\Rightarrow \sum_{n=1}^{\infty} f_n^2 = \sum_{n=1}^{\infty}(nf_n)(\frac{1}{n}f_n) \overset{\text{Cauchy Schwartz}}{\leq} \left(\sum_{n=1}^{\infty}\frac{1}{n^2}f_n^2\right)^{1/2}\left(\sum_{n=1}^{\infty} n^2 f_n^2\right)^{1/2}.$$

As a result

$$\|f\|_{L^2}^2 \leq \frac{1}{\pi}\|f'\|_{L^2}\|f\|_{L^2} \quad \Rightarrow \|f\|_{L^2} \leq \frac{1}{\pi}\|f'\|_{L^2}.$$

Thus $1/\pi$ is a Poincaré constant. In fact, it is the smallest Poincaré constant,

$$\frac{1}{\pi} = \frac{\|\sin(\pi x)\|_{L^2}}{\|(\sin(\pi x))'\|_{L^2}} \leq \sup_{\substack{f\in H_0^1 \\ f\neq 0}} \frac{\|f\|_{L^2}}{\|f'\|_{L^2}} \leq \frac{1}{\pi}.$$

Remark 14 (Relation with the discrete problem)**.** *We recall that for $p \in P_{r,D}$, c.f Proposition B.4, where $P_{r,D}$ denotes the set of polynomials of degree r with Dirichlet boundary conditions,*

$$\|f\|_{L^2}^2 = [f]_{int} \cdot \hat{M}_{int}[f]_{int} \quad ; \quad \|f'\|_{L^2}^2 = [f]_{int} \cdot \hat{S}_{int}[f]_{int}.$$

Since $P_{r,D} \subset H_0^1$ we have

$$\pi^2 = \inf_{\substack{f\in H_0^1 \\ f\neq 0}} \frac{\|f'\|_{L^2}^2}{\|f\|_{L^2}^2} \leq \inf_{\substack{f\in P_{r,D} \\ f\neq 0}} \frac{\|f'\|_{L^2}^2}{\|f\|_{L^2}^2} = \inf_{\substack{f\in P_{r,D} \\ f\neq 0}} \frac{[f]_{int} \cdot \hat{S}_{int}[f]_{int}}{[f]_{int} \cdot \hat{M}_{int}[f]_{int}}.$$

As a result, for all $v \in \mathbb{R}^{r-1}$ we have

$$\pi^2\, v \cdot \hat{M}_{int} v \leq v \cdot \hat{S}_{int} v\,. \tag{116}$$

D.2 Apriori Estimates

Lemma 33. *With $0 \leq w < \pi^2$, for $f \in H_0^1$ a solution of $-f'' - wf = g$ with $g \in L^2(0,1)$, we have the following apriori estimates*

$$\|f\|_{L^2} \leq \frac{1}{\pi^2 - w}\|u\|_{L^2}. \tag{117}$$

$$\|f'\|_{L^2} \leq \frac{\pi}{\pi^2 - w}\|u\|_{L^2}. \tag{118}$$

Proof. Since $f \in H_0^1$ a solution of $-f''-wf = g$, we multiply both sides by f and do a integration by parts on the LHS to obtain,

$$\int_0^1 (f')^2\, ds - w\int_0^1 f^2\, ds = \int_0^1 gf\, ds. \tag{119}$$

Using Poincaré Lemma to bound from below of LHS and Cauchy-Schwartz to bound from above RHS we obtain

$$(\pi^2 - w)\|f\|_{L^2}^2 \leq \int_0^1 (f')^2\, ds - w\int_0^1 f^2\, ds = \int_0^1 uf\, ds \leq \|u\|_{L^2}\,\|f\|_{L^2}.$$

As a result, we obtain (117). Next, we use the L^2 bound in (119) to give

$$\begin{aligned}\int_0^1 (f')^2\, ds &= \int_0^1 uf\, ds + w\int_0^1 f^2\, ds \\ &\overset{(117)}{\leq} \frac{1}{\pi^2 - w}\|u\|_{L^2}^2 + \frac{w}{(\pi^2-w)^2}\|u\|_{L^2}^2 \quad = \frac{\pi^2}{(\pi^2-w)^2}\|u\|_{L^2}^2\,.\end{aligned}$$

This finishes the proof. □

D.3 Properties of Solution of the Dirichlet BVP

For $g \in L^2(0,1)$, the boundary value problem (BVP),

$$-f'' = g \quad , \quad f(0) = f(1) = 0. \tag{120}$$

has a unique solution in $H_0^1(0,1)$, which we denote by $\Delta_D^{-1} g$. By Lemma (33) with $w = 0$, we have that $(\Delta_D)^{-1}$ is a bounded operator on $L^2(0,1)$ with norm bounded by

$$\|\Delta_D^{-1}\|_{L^2 \to L^2} \leq \pi^{-2}. \tag{121}$$

For $w > 0$, the boundary value problem

$$-f'' - wf = g \quad ; \quad f(0) = f(1) = 0. \tag{122}$$

has a unique solution, which we denote by $(-\Delta_D - w)^{-1} g$. The following proposition shows that this solution is analytic in w in a neighborhood of $w = 0$. See also Remark 15 for another point of view.

Proposition 34. *For $g \in L^2(0,1)$ independent of w, with $0 < w < \pi^2$, $(\Delta_D - w)^{-1} g$ is analytic in w and can be written as*

$$\left[(\Delta_D - w)^{-1} g\right](w,x) \;=\; F_N(w,x) + w^{N+1} E_N(w,x) \quad , \quad \forall\, N \geq 0\,; \tag{123}$$

where

$$F_N(w,x) := \sum_{n=0}^{N} w^n \, \Delta_D^{-(n+1)} g \quad ; \quad E_N(w,x) := \Delta_D^{-(N+1)} (\Delta_D - w)^{-1} \, g;$$

$$\|F_N'\|_{L^2} \leq \pi \frac{1 - (\frac{w}{\pi^2})^{N+1}}{\pi^2 - w} \|g\|_{L^2} \quad ; \quad \|E_N'\|_{L^2} \leq \frac{\pi^{1-2N}}{\pi^2 - w} \|g\|_{L^2}.$$

Proof. The Dirichlet operator Δ_D with domain $H_0^1 \cap H^2$ is a self-adjoint extension of an unbounded operator $(-\frac{d^2}{dx^2}, \mathcal{C}_0^\infty(0,1))$. Using that Δ_D has an bounded inverse Δ_D^{-1} in L^2, we can factor $\Delta_D - w$ as,

$$\Delta_D - w \, \mathrm{Id}_{H_0^1 \cap H^2 \to L^2} = (\mathrm{Id}_{L^2 \to L^2} - w \, \Delta_D^{-1}) \circ \Delta_D.$$

Since $\|\Delta_D^{-1}\|_{L^2 \to L^2} \leq \pi^{-2}$, c.f (121), if $w < \pi^2$, the Neumann series of $w\Delta_D^{-1}$ converges in the operator norm $\|\cdot\|_{L^2 \to L^2}$, and hence $\mathrm{Id} - w\Delta_D^{-1}$ is invertible as a bounded operator on $L^2(0,1)$. In this case, we have

$$(\Delta_D - w)^{-1} = \Delta_D^{-1} (\mathrm{Id} - w\Delta_D^{-1})^{-1} = \Delta_D^{-1} \sum_{k=0}^{\infty} w^k \Delta_D^{-k} = \sum_{k=0}^{\infty} w^k \Delta_D^{-(k+1)}.$$

As a result of this, we obtain the expansion (123) of $(\Delta_D - w)^{-1} g$ for $g \in L^2(0,1)$, for any $N \geq 0$.

We next obtain a bound for the approximating term F_N and the error term E_N using the apriopri estimates in Lemma 33. For F_N,

$$F_N = -\Delta_D \sum_{n=0}^{N} w^n \Delta_D^{-n} g$$

$$\Rightarrow \|F_N'\|_{L^2} \leq \pi^{-1} \sum_{n=0}^{N} w^n \|\Delta_D^{-n} g\|_{L^2} \leq \pi^{-1} \sum_{n=0}^{N} w^n \pi^{-2n} \|g\|_{L^2}$$

$$\Rightarrow \|F_N'\|_{L^2} \leq \pi^{-1} \frac{1 - w^{N+1}\,\pi^{-2(N+1)}}{1 - w\pi^{-2}} \|g\|_{L^2}.$$

For E_N, we first rewrite to get rid of the infinite sum notation,

$$\begin{aligned} w^{N+1} E_N &= \sum_{k=N+1}^{\infty} w^k \Delta_D^{-(k+1)}\, g = w^{N+1} \Delta_D^{N+1} \sum_{k=0}^{\infty} w^k \Delta_D^{-(k+1)}\, g \\ &= w^{N+1} \Delta_D^{-(N+1)} (\Delta_D - w)^{-1}\, g \\ \Rightarrow \|E_N'\|_{L^2} &\leq \frac{\pi}{\pi^2 - w} \|\Delta_D^{-(N+1)} g\|_{L^2} \leq \frac{\pi}{\pi^2 - w} \pi^{-2(N+1)} \|g\|_{L^2}. \end{aligned}$$

□

Remark 15. *Since we are dealing with ODE, we can write out readily the exact form of* $(\Delta_D - w)^{-1} g$. *A basis for solutions of* $-f'' - wf = 0$ *is the pair* $\cos(w^{1/2}x)$ *and* $\sin(w^{1/2}x)$, *the Wronskian of which* $W(w,x) = w^{1/2}$. *By variation of parameters method, a particular solution to* $-f'' - wf = u$ *is given by*

$$f_{particular} := -\cos(w^{1/2}x) \int_0^x \frac{\sin(w^{1/2}s)}{w^{1/2}s}\, s\, u(s)\, ds + x \frac{\sin(w^{1/2}x)}{w^{1/2}x} \int_0^x \cos(w^{1/2}s)\, u(s)\, ds.$$

The above function is analytic in w *with infinite radius of convergence. Incorporating the boundary conditions at* 0 *and* 1, *we obtain the solution to the Dirichlet problem* (122),

$$(\Delta_D - w)^{-1} g = f_{particular}(w,x) - f_{particular}(w,1) \frac{\sin(w^{1/2}x)}{\sin(w^{1/2})}. \tag{124}$$

In the above expression, $f_{particular}(w,x)$, *and thus* $f_{particular}(w,1)$ *is analytic in* w *with infinite radius of analyticity, and the last factor, which is*

$$\frac{\sin(w^{1/2}x)}{\sin(w^{1/2})} = \frac{w^{1/2}x\big(1 + \sum_{n=1}^{\infty} c_n w^n x^{2n}\big)}{w^{1/2}\big(1 + \sum_{n=1}^{\infty} c_n w^n\big)}, \tag{125}$$

is analytic in w *in a neighborhood of* $w = 0$ *with the radius of analyticity* π^2.

E Discrete Z-transform

We refer the readers to [4] and the references therein for a brief history on z-transform and for a list of fields of applications using this transform, e.g. signal theory, harmonic analysis and quantum physics, among others. Our working definition of z-transform is the following.

Definition 5. *For a scalar sequence* $u = \{u_n\}_{n \in \mathbb{Z}}$,

$$[Zu](z) := \sum_{n=-\infty}^{\infty} u_n\, z^n, \tag{126}$$

if the RHS converges. We will also work with the following version

$$[Zu](\theta) := \sum_{n=-\infty}^{\infty} u_n\, e^{2\pi i n \theta}. \tag{127}$$

Recall in Subsection 3.2, we have denoted by $\mathfrak{M}(\mathbb{Z}, S)$ the set of sequences taking value in S and indexed by $\mathbb{Z}$; its elements are denoted by $u = \{u_J\}_{J\in\mathbb{Z}}$, and by $\mathfrak{M}(\mathbb{Z}\times\mathbb{Z}_r, S)$ the set of sequences taking value in S and indexed by $\mathbb{Z}\times\mathbb{Z}_r$; its elements are denoted by $u = \{u_{J,k}\}_{J\in\mathbb{Z},0\le k<r}$.

Definition 5 extends readily to vector-valued sequences, upon noting that for a vector-valued sequence $U \in \mathfrak{M}(\mathbb{Z}, \mathbb{C}^r)$, c.f for the, $\pi_i U$ is a scalar sequence for $1 \le i \le r$. Here, π_i denotes the projection onto the i-th component,

Definition 6. *For a vector-valued sequence $U \in \mathfrak{M}(\mathbb{Z}, \mathbb{C}^r)$, $[Z_{\mathfrak{B}}U](z)$ is a vector of dimension r with components given by*

$$\pi_i\, Z_{\mathfrak{B}}\, U = Z\, \pi_i\, U \quad , 1 \le i \le r, \tag{128}$$

if the RHS makes sense. Here there are also two versions corresponding to whichever definition of scalar Z used.

In the case where U is obtained from by a scalar sequence u by the 'blocking' map $\mathfrak{b}_r$, c.f. (16)-(17), i.e $U = \mathfrak{b}_r u$,

$$[\pi_k Z_{\mathfrak{B}}\, U](z) = \sum_{J=-\infty}^{\infty} u_{Jr+k}\, z^J\,,\ 1 \le k \le r.$$

Remark 16. *The corresponding continuous version to our discrete transform is the following: for a function u*

$$[Zu](x,z) := \sum_{n=-\infty}^{\infty} u(x+n)\, z^n \quad ; \quad [Zu](x,\theta) := \sum_{n=-\infty}^{\infty} u(x+n)\, e^{2\pi i n\theta}.$$

Let $l^2(\mathbb{Z})$ denote the set of square summable sequences,

$$l^2(\mathbb{Z}) = \left\{ u \in \mathfrak{M}(\mathbb{Z},\mathbb{C}) : \sum_{k\in\mathbb{Z}} u_k^2 < \infty \right\}.$$

We recall that $\{e^{2\pi i n x}\}_{n\in\mathbb{Z}}$ is complete orthonormal set for $L^2(0,1)$ equipped with the usual complex L^2 scalar product,

$$\langle f, g\rangle_{L^2(0,1)} = \int_0^1 f\, \overline{g}\, dx.$$

As a result, if sequence $u \in l^2(\mathbb{Z})$, then $\sum_{k\in\mathbb{Z}} u_k e^{2\pi i k\theta}$ converges in $L^2_\theta(0,1)$. Thus, for $u \in l^2(\mathbb{Z})$, Zu is well-defined as an $L^2(0,1)$ function in θ, with u_n being the n-th Fourier Series coefficient of Zu,

$$u_n = \langle [Zu](\theta)\,,\, e^{2\pi i n\theta}\rangle_{L^2(0,1)} = \int_0^1 [Zu](\theta)\, e^{-2\pi\, i\, n\, \theta}\, d\theta.$$

In addition, by the Plancherel's identity, we have

$$\|Zu\|_{L^2_\theta(0,1)} = \left\| \sum_{k\in\mathbb{Z}} u_k\, e^{2\pi i k\theta} \right\|_{L^2(0,1)} = \|u\|_{l^2(\mathbb{Z})}. \tag{129}$$

Proposition 35. *For a scalar sequence $u \in l^2(\mathbb{Z})$, Zu defined by (126) is well-defined as a L^2 function on $(0,1)$ in θ, with the following additional properties:*

1. *Zu is periodic in θ .*
2. *By* (129), *$Z : l^2(\mathbb{Z}) \longrightarrow L^2_\theta(0,1)$ is an isometry.*
3. *u_n is the n-th Fourier coefficient of Zu, thus can be retrieved by*

$$u_n = \left\langle [Zu](\theta)\,,\, e^{2\pi i n\theta} \right\rangle_{L^2(0,1)} = \int_0^1 [Zu](\theta)\, e^{-2\pi\, i\, n\, \theta}\, d\theta.$$

If $z \in \mathbb{C}$ with $|z| = 1$, we can write $z = e^{2\pi i\theta}$ for some $\theta \in [0,1)$, thus these above results can be transferred to the second version of Z, c.f. (127).

Proposition 36. *For a scalar sequence $u \in l^2(\mathbb{Z})$, Zu defined by* (127) *makes sense as an L^2 function in z on the unit disc $\mathbf{S}^1 \subset \mathbb{C}$.*

$$Z : l^2(\mathbb{Z}) \longrightarrow L^2(\mathbf{S}^1) \text{ is an isometry.}$$

In this case, u_n can be obtained from Zu by

$$u_n = \frac{1}{2\pi i} \oint_{\mathbf{C}_1} [Zu](z)\ \frac{dz}{z^{n+1}}.$$

Here $\mathbf{C}_s$ is a simple closed curve counter-clockwise oriented and lying in the circle centered at 0 *with radius s.*

The above results extend readily to the vector version of the Z transform.

Proposition 37. *For a vector-valued sequence $U \in (l^2(\mathbb{Z}))^r$, $Z_{\mathfrak{B}}U$ has the following properties.*

1. *With definition of Z given by* (126), *$[Z_{\mathfrak{B}}U](\theta) \in (L^2_\theta(0,1))^r$.*
2. *With definition of Z given by* (127), *$[Z_{\mathfrak{B}}U](z) \in (L^2_z(\mathbf{S}^1))^r$.*
3. *We have the inverse formula,*

$$U_J \;=\; \int_0^1 [Z_{\mathfrak{B}}U](\theta)\, e^{-2\pi\, i\, J\, \theta}\, d\theta \;=\; \frac{1}{2\pi i} \oint_{\mathbf{C}_1} [Z_{\mathfrak{B}}U](z)\ \frac{dz}{z^{J+1}}. \tag{130}$$

Define the shift operator τ acting on a sequence U

$$(\tau_k U)_J := U_{J+k}. \tag{131}$$

We will make use of the following relation between the translation operator and $Z_{\mathfrak{B}}$ operator to convert an recurrence relation with constant coefficients into an algebraic one.

Proposition 38. 1. *For scalar sequence u,*

$$[Z\, \tau_{\pm k}\, u](z) = z^{\mp k}\, [Zu](z) \quad , \quad k \in \mathbb{Z}^+.$$

2. *For the vector-valued sequence sequence U of the form $U = \mathfrak{b}_r u$, c.f* (16)-(17), *for some scalar sequence u, we have*

$$[Z_{\mathfrak{B}}\, \tau_{\pm k} U](z) = z^{\mp k}\, [Z_{\mathfrak{B}} U](z) \quad , \quad k \in \mathbb{Z}^+.$$

Proof. **Property 1 :**

$$[Z\tau_k u](z) = \sum_{n=-\infty}^{\infty} u(n+k)\, z^k = \sum_{n=-\infty}^{-1} u(n+k)\, z^k + \sum_{n=0}^{\infty} u(n+k)\, z^k.$$

Let $j = n + k$ then $n = j - k$

$$[Z\tau_k u](z) = z^{-k} \sum_{j=-\infty}^{-1+k} u(j)\, z^j + z^{-k} \sum_{j=k}^{\infty} u(j) z^j = z^{-k} \sum_{j=-\infty}^{\infty} u(j)\, z^j.$$

Property 2 : Working from the definition of $Z_{\mathfrak{B}}$

$$[\pi_i\, Z_{\mathfrak{B}}\, \tau_k\, U](z) = (Z\pi_i\tau_k U)\,,\ 1 \le i \le r.$$

we obtain the following series of equality.

$$\begin{aligned}\big[\pi_i Z_{\mathfrak{B}}\, \tau_k U)\big](z) = \big[Z\, \pi_i\, \tau_k U\big](z) \quad &\overset{\star}{=} \quad \sum_{J=-\infty}^{\infty} u_{(J+k)N+i}\, z^J \\ = z^{-k} \sum_{J=-\infty}^{\infty} u_{JN+i}\, z^J \quad &= \quad z^{-k}\big[Z_{\mathfrak{B}}\, \pi_i\, U\big](z).\end{aligned}$$

Equality $(\star)$ comes from the fact that $(\tau_k U)_J = U_{J+k}$, so for $U = \mathfrak{b}_r u$,

$$(\pi_i\, \tau_k\, U)_J = u_{(J+k)N+i}\,,\ 1 \le i \le r.$$

As a result, we have

$$[Z_{\mathfrak{B}}\, \tau_k\, U](z) = z^{-k}[\pi_i Z_{\mathfrak{B}}\, U](z) = z^{-k}[\mathtt{Z}U](z).$$

□

Remark 17. *In literature, the type of Z transform used to solve recurrence relation with constant coefficients are usually one sided rather two-sided as currently used in our paper,*

$$Z_+ u = \sum_{n=0}^{\infty} u_n\, z^n \quad ; \quad Z_- u = \sum_{n=-\infty}^{n} u_n\, z^n.$$

Their interaction with the translation operator is

$$[Z_+\tau_k u](z) = z^{-k}[Z_+ u](z) - \sum_{n=0}^{k-1} u(n) z^{n-k};$$

$$[Z_+\tau_{-k} u](z) = z^k Z_+ u + \sum_{n=-k}^{-1} u(n) z^{k+n}.$$

In particular if u is such that $u(n) = 0$ for $n < 0$ then

$$[Z_+\tau_{-k} u](K) = z^k Z_+ u.$$

F Miscellaneous Facts

F.1 Some calculations of determinant of matrices

We consider a matrix C of size $n \times n$ of type

$$C = \begin{pmatrix} 0 & v^t \\ w & C_{\text{int}} \end{pmatrix},$$

with v, w are row vectors of size $n-1$.

Lemma 39. *Suppose C_{int} is invertible*

$$\det C = -\det C_{int} \times w \cdot C_{int}^{-1} v.$$

Proof. We first write C as the following product,

$$C = \begin{pmatrix} 1 & 0_{1\times(n-1)} \\ 0_{(n-1)\times 1} & C_{\text{int}} \end{pmatrix} \begin{pmatrix} 0 & v^t \\ C_{\text{int}}^{-1} w & \text{Id}_{(n-1)\times(n-1)} \end{pmatrix}.$$

The proof is finished if we can show that the determinant of the second matrix in the factor is $w \cdot C_{\text{int}}^{-1} v$.

Denote by $[\text{Id}_{(n-1)\times(n-1)}]_j$ the matrix obtained from the square identity matrix of size $n-1$ by replacing j-th row by v^t . Since

$$v^t = \sum_{j=1}^{n-1} v_j e_j^t,$$

using the multilinearity of det, we obtain that

$$\det [\text{Id}_{(n-1)\times(n-1)}]_j = v_j.$$

As a result of this,

$$\begin{aligned} \det \begin{pmatrix} 0 & v^t \\ C_{\text{int}}^{-1} w & \text{Id} \end{pmatrix} &= \sum_{j=1}^{n-1} (-1)^{(j+2)} (C_{\text{int}}^{-1} w)_j \times (-1)^{j-1} \det[\text{Id}_{(n-1)\times(n-1)}]_j \\ &= -\sum_{i=1}^{n-1} (C_{\text{int}}^{-1} w)_i \, v_j \quad = -v \cdot C^{-1} w. \end{aligned}$$

The factor $(-1)^{j-1}$ comes from having to permute v^t from the first row to the j-th one, for $1 \le j \le n-1$.

□

F.2 Asymptotics with arccos

Proposition 40. *The following statement is equivalent*

1. $\phi = w^{1/2}(1 + \mathsf{O}(w^r))$.
2. $\cos\phi = \cos w^{1/2} + w\mathsf{O}(w^r)$.

Proof. (1) ⇒ (2) **:** Suppose $\phi - w^{1/2} = w^{1/2}\mathsf{O}(w^r)$, then we have

$$\cos(\phi - w^{1/2}) - 1 = w\mathsf{O}(w^{2r}) \quad ; \text{ and } \quad \sin(\phi - w^{1/2}) = w^{1/2}\mathsf{O}(w^r).$$

Hence

$$\begin{aligned}\cos(\phi) - \cos w^{1/2} &= \cos(\phi - w^{1/2} + w^{1/2}) - \cos w^{1/2} \\ &= \Big(\cos(\phi - w^{1/2}) - 1\Big)\cos w^{1/2} - \sin(\phi - w^{1/2})\sin(w^{1/2}) \\ &= w\mathsf{O}(w^{2r})\mathsf{O}(1) - w\mathsf{O}(w^r)\mathsf{O}(1) = w\mathsf{O}(w^r).\end{aligned}$$

(2) ⇒ (1) **:** This direction follows from Lemma 41 and 42 . □

Lemma 41. *Suppose* $\cos\phi = \cos w^{1/2} + w\mathsf{O}(w^r)$ *then*

$$\cos(\phi - w^{1/2}) = 1 + w^{1/2}\mathsf{O}(w^r).$$

Proof. Suppose $\cos\phi = \cos w^{1/2} + w\mathsf{O}(w^r)$ then

$$\begin{aligned}1 - \cos^2\phi &= \sin^2(w^{1/2}) - 2w\,\mathsf{O}(w^r)\cos w^{1/2} - w^2\,\mathsf{O}(w^r)\mathsf{O}(w^r) \\ &= \sin^2(w^{1/2}) - w\mathsf{O}(w^r)\Big(2\cos w^{1/2} - w\mathsf{O}(w^r)\Big).\end{aligned}$$

Note that since $\sin^2(w^{1/2}) = w(1 + \mathsf{O}(1))$ we can write the above expression as

$$1 - \cos^2\phi = \sin^2(w^{1/2})\Big(1 - \mathsf{O}(w^r)\Big).$$

$$\Rightarrow \quad \sin(\phi) = \sqrt{1 - \cos^2\phi} = \sin(w^{1/2})\Big(1 - \mathsf{O}(w^r)\Big).$$

As a result of this,

$$\begin{aligned}\cos(\phi - w^{1/2}) &= \cos(\phi)\cos w^{1/2} - \sin(\phi)\sin w^{1/2} \\ &= \cos^2 w^{1/2} + w\mathsf{O}(w^r)\cos w^{1/2} - \sin^2 w^{1/2} + \sin w^{1/2}\mathsf{O}(w^r).\end{aligned}$$

The RHS simplies to give

$$\cos(\phi - w^{1/2}) = 1 + w^{1/2}\mathsf{O}(w^r).$$

□

Lemma 42. *Suppose* $\cos(\phi - w^{1/2}) = 1 + w^{1/2}\mathsf{O}(w^r)$ *then*

$$\phi = w^{1/2}(1 + \mathsf{O}(w^r)).$$

Proof. We recall the Taylor expansion of arccos and arcsin

$$\arccos x = \tfrac{\pi}{2} - \arcsin x;$$

$$\arcsin x = \sum_{n=0}^{\infty} \frac{(2n)!}{4^n (n!)^2 (2n+1)} x^{2n+1} \quad , \quad |x| \le 1.$$

As a result,

$$\begin{aligned}
\phi - w^{1/2} &= \tfrac{\pi}{2} - \arcsin\left(1 + w^{1/2}\mathsf{O}(w^r)\right) \\
&= \tfrac{\pi}{2} - \sum_{n=0}^{\infty} \frac{(2n)!}{4^n (n!)^2 (2n+1)} \left(1 + w^{1/2}\mathsf{O}(w^r)\right)^{2n+1} \\
&= \tfrac{\pi}{2} - \sum_{n=0}^{\infty} \frac{(2n)!}{4^n (n!)^2 (2n+1)} \left(1 + w^{1/2}\mathsf{O}(w^r)\right) \\
&= \tfrac{\pi}{2} - \arcsin(1) + w^{1/2}\mathsf{O}(w^r).
\end{aligned}$$

As a result, we have

$$\phi - w^{1/2} = w^{1/2}\mathsf{O}(w^r).$$

□

Contents

References

[1] Mark Ainsworth. Discrete dispersion relation for hp-version finite element approximation at high wave number. *SIAM Journal on Numerical Analysis*, 42(2):553–575, 2004.

[2] Bernard Brooks. The coefficients of the characteristic polynomial in terms of eigenvalues and the elements of an $n \times n$ matrix. *Applied Mathematics Letters*, 19, 2006.

[3] Gary C. Cohen. *Higher-order numerical methods for transient wave equations*. Springer-Verlag, Berlin, Heidelberg, New York, 2002. Springer Series in Scientific Computation.

[4] Alexander D. Poularikas (editor in chief). *Transform and Applications Handbook*. CRC Press, 3rd edition, 2010.

[5] Philippe Guillaume. Nonlinear eigenproblems. *Siam J. Matrix Anal. Appl.*, 20(3), 1999.

[6] Isaac Harari and Thomas JR Hughes. Finite element methods for the helmholtz equation in an exterior domain: model problems. *Computer Methods in Applied Mechanics and Engineering*, 87(1):59–96, 1991.

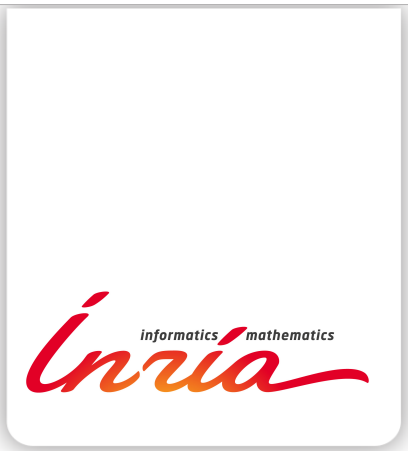
[7] Frank Ihlenburg and Ivo Babuška. Finite element solution to the helmholtz equation with high wavenumber - part 1: the h-version of the fem. *Compt. Math. Appl.*, 30, 1995.

[8] Frank Ihlenburg and Ivo Babuška. Finite element solution to the helmholtz equation with high wavenumber - part 2 : the h-p version of the fem. *Siam J. Numer. Anal.*, 34(1), 1997.

[9] Patrick Joly. Introduction à l'analyse mathématique de la propagation d'ondes en régime harmonique. Polycopié du Cours de Master 2 - Mathématiques et Applications . Université Pierre et Marie Curie, 2006-2007.

[10] L. Thompson and P.M. Pinksy. Complex wavenumber fourier analysis of the p-version finite element method. *Comp. Mech.*, 13, 1994.

[11] Lloyd N Trefethen. Group velocity in finite difference schemes. *SIAM review*, 24(2):113–136, 1982.



RESEARCH CENTRE
BORDEAUX – SUD-OUEST

200 avenue de la Vieille Tour
33405 Talence Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399