Approximation with vectorial exponential functions of solutions of the P N model for the transport of particles

Christophe Buet, Bruno Despres, Guillaume Morel

Approximation with vectorial exponential functions of solutions of the P_N model for the transport of particles

Christophe Buet[1,5,6], Bruno Despres[2,3,6], Guillaume Morel[4,6]

[1] CEA, DAM, DIF, F-91297 Arpajon, France

[2] Sorbonne Université, UMR 7598, Laboratoire Jacques-Louis Lions, F-75005, Paris, France

[3] Institut Universitaire de France.

[4] IMT Atlantique, 29238 Brest, France.

[5] Université Paris-Saclay, CEA, Laboratoire en Informatique Haute Performance pour le Calcul et la simulation, 91680 Bruyères-le-Châtel, France.

Abstract

Trefftz discontinuous Galerkin (TDG) methods have recently shown potential [6, 27] for numerical approximation of transport equations with exponential modes. This paper focus on a proof of convergence in two-space dimension for the TDG method through the study of the approximation properties of the exponential solutions constructed in [6]. We show that these vectorial exponential functions can achieve high order convergence with a significant gain in term of the number of basis functions compare to more standard discontinuous Galerkin schemes. The fundamental part of the proof is based on discrete Fourier techniques conveniently adapted to the matrices of the problem.

1 Introduction

Following the literature in nuclear engineering [3, 2, 22, 1] and radiation transfer [23, 28, 15], the tridimensional linear Boltzmann equation in dimension 1+3+2

$$\partial_t f + \Omega \cdot \nabla f = -\sigma_a f + \sigma_s \left(\frac{1}{4\pi} \int f d\Omega - f \right) \tag{1}$$

projected on a finite number of spherical harmonics is called the P_N model. Here $t \geq 0$ is the time variable, $\mathbf{x} = (x, y, z) \in \mathbb{R}^3$ is the space variable and $\Omega = (\cos\theta\cos\psi, \cos\theta\sin\psi, \sin\theta) \in S^2$ is the direction variable. The absorption coefficient is $\sigma > 0$ and the scattering coefficient is $\sigma_s > 0$. The **index** $N \geq 1$ of a P_N model is related to the number of spherical harmonics and so is related to the number of unknowns (the size) of the model. Taking $N > 1$ large is a way to approximate the transport equation (1) with a satisfactory accuracy. For the calculation of numerical approximations of systems like (1), it is clear that the high number of dimensions of the transport equation induces important numerical difficulty. This is similar for P_N models with $N > 1$. That is why any theoretical possibility to reduce the computational burden of such calculations [10, 14] must be investigated.

In this work, we focus on stationary ($\partial_t = 0$) and bidimensional ($\partial_z = 0$) general P_N models. Our main contribution is to prove that the Trefftz Discontinuous Galerkin (TDG) method provides an accurate method which is asymptotically much better than the classical Finite Element Method (FEM) [29, 4] for the calculation of numerical solutions to P_N models.

Before stating the main result in Theorems 1.1 and 1.2, we review some recent material about Trefftz methods and compare their asymptotic convergence for the numerical approximation on a

[6] E-mail addresses: christophe.buet@cea.fr, despres@ann.jussieu.fr, guillaume.morel@imt-atlantique.fr

mesh of **characteristic length** $h > 0$ with the asymptotic convergence of FEM or discontinuous Galerkin (DG). The TDG method [7, 8, 12, 20] uses exact solutions of a given partial differential equation (PDEs) as basis functions. In [6], exact solutions to the spherical harmonic approximation of the transport equation (P_N model) have been constructed and several numerical results already show great behavior of the TDG method when using these particular basis functions. In the case of Trefftz methods, convergence results could be particularly interesting since high order convergence can generally be achieved with fewer degrees of freedoms compare to more standard DG methods [27, 8, 12, 16, 17, 19]. Indeed, usually when considering the standard DG [14, 13] or FEM method, the approximation properties of simple monomials (such as 1, x, y... for example) can be easily studied since they appear in the Taylor expansion of every regular functions. In space dimension 2, one gets estimates of convergence with respect to h and the number of basis functions per cell p which have have the general form: $\|u_{\text{exact}} - u_h\| \approx O(h^{\text{order}})$ where $p \approx O(\text{order}^2)$, for a norm $\|\cdot\|$ and regularity assumptions not discussed at this stage. The power 2 is because one needs to exhaust all monomials in the Pascal's triangle up to the order to obtain a local accuracy $O(h^{\text{order}})$. To understand the meaning of these estimates, we rewrite the numerical error as $\varepsilon_{\text{FEM/DG}} = \|u_{\text{exact}} - u_h\|$. Then one gets a general asymptotic formula between the total cost in terms of basis functions on a mesh and the numerical error

$$\varepsilon_{\text{FEM/DG}} \approx O(h^{\text{order}}) \approx O(h^{\sqrt{p}}). \tag{2}$$

The great promise of TDG methods is a strong improvement on this scaling. Indeed the law $p = O(\text{order}^2)$ is approximatively replaced by $p = O(\text{order})$ for Trefftz methods. So one gets

$$\varepsilon_{\text{TDG}} \approx O(h^{\text{order}}) \approx O(h^{p}). \tag{3}$$

In three-space dimension, the powers are different but the ordering is the same, indeed $\varepsilon_{\text{FEM/DG}} \approx O(h^{\sqrt[3]{p}})$ and $\varepsilon_{\text{TDG}} \approx O(h^{\sqrt{p}})$. For large $p >> 1$, the numerical error obtained with TDG is asymptotically much better than with FEM or DG. However this gain has the (human) cost that one needs to calculate specific basis functions adapted to the problem, because monomials cannot be used in general. The functions are exponential in our case. In a preliminary work [27], we showed an optimal convergence result for the P_1 model, by means of explicit manipulations. In this work we prove a similar result for a general case P_N model, where $N \in 2\mathbb{N}+1$ is odd since it corresponds to the engineering literature.

The structure of P_N models is borrowed from [15] and described in more details in Section 2. Let m be the total size of the system (equal to the number of spherical harmonics considered in the model), m_e be the number of even moments and m_o be the number of odd moments. In this work we follow the literature and take N odd. So $m = \frac{1}{2}(N+1)(N+2)$, $m_e = \frac{1}{4}(N+1)^2$ and $m_o = \frac{1}{4}(N+1)(N+3)$. The models can be written under the form of a Friedrichs system with linear relaxation [6, 27, 26]

$$\Big(\mathcal{A}\partial_x + \mathcal{B}\partial_y\Big)\mathbf{u}(t,\mathbf{x}) = -\mathcal{R}\mathbf{u}(t,\mathbf{x}), \quad \mathbf{x} = (x,y)^T, \tag{4}$$

where the vector of unknowns is $\mathbf{u} \in \mathbb{R}^m$ and the symmetric matrices $\mathcal{A}, \mathcal{B}, \mathcal{R} \in \mathbb{R}^{m\times m}$ are specific to the $2\text{D}\frac{1}{2}$ geometry. to the space variable. In our case, all matrices are constant for the simplicity of the presentation, and their exact definition of the matrices will be given later. With the order given in [15], the symmetric matrices $\mathcal{A}$ and $\mathcal{B}$ have a structure in rectangular blocks

$$\mathcal{A} = \begin{pmatrix} 0 & A \\ A^T & 0 \end{pmatrix} \in \mathbb{R}^{m\times m}, \quad \mathcal{B} = \begin{pmatrix} 0 & B \\ B^T & 0 \end{pmatrix} \in \mathbb{R}^{m\times m} \tag{5}$$

where $\sigma_a > 0$ is the absorption coefficient, $\sigma_s > 0$ is the scattering coefficient and the sub-matrices are $A, B \in \mathbb{R}^{m_e \times m_o}$. One has a diagonal block structure for $\mathcal{R}$ on the right hand side

$$\mathcal{R} = \begin{pmatrix} R_e & 0 \\ 0 & R_o \end{pmatrix} \in \mathbb{R}^{m\times m}, \tag{6}$$

where $R_e = \text{diag}(\sigma_a, \sigma_a + \sigma_s, ..., \sigma_a + \sigma_s) \in \mathbb{R}^{m_e \times m_e}$ and $R_o = (\sigma_a + \sigma_s) I_{m_o} \in \mathbb{R}^{m_o \times m_o}$ are diagonal matrices, with I_{m_o} the identity matrix of $\mathbb{R}^{m_o \times m_o}$. The problem made with the equation (5) supplemented with an outgoing boundary condition (33) is endowed with strong quadratic stability estimates, as shown in the core of the paper.

The general form of the vectorial exponential solutions can now be described. They are based on a unitary rotation matrix $\mathcal{U}(\theta) \in \mathbb{R}^{m \times m}$, see (21) below. Let $\mathbf{w}^t \in \mathbb{R}^m$ for $1 \leq t \leq m_e$ be exact solutions of a 1D problem precised later and let $\mathbf{d}_s = (\cos\theta_s, \sin\theta_s) \in \mathbb{R}^2$ be **equi-distributed** directions with angles

$$\theta_s = \frac{2\pi s}{2n+3}, \quad 1 \leq s \leq 2n+3.$$

We will show below that the following vectorial exponential functions

$$\mathbf{u}^{st}(\mathbf{x}) = \mathcal{U}(\theta_s)\mathbf{w}^t e^{\lambda_t \mathbf{d}_s \cdot \mathbf{x}}, \quad 1 \leq t \leq m_e, \ 1 \leq s \leq 2n+3, \tag{7}$$

are solutions to the P_N model (4), provided the $\lambda_t > 0$ and the $\mathbf{w}^t \in \mathbb{R}^m$ are correctly defined. These functions are called exponential solutions because of the exponential terms $e^{\lambda_t \mathbf{d}_s \cdot \mathbf{x}}$. Using the vectorial exponential functions as basis functions in TDG yields a number of basis functions per cell

$$p = m_e(2n+3) \tag{8}$$

where m_e depends on the model (that is on N) and n can be increased to use more and more basis functions per cell. This number p of basis functions per cell is the same that enters in the scaling laws (2-3).

The main contribution of this work is summarized in the following Theorems of convergence, where as usual, we consider only meshes which satisfy a condition of uniform regularity.

Theorem 1.1. *Take $N \in 2\mathbb{N}+1$. Take $m_e(2n+3)$* **equi-distributed** *basis functions with $n \geq N-1$ or $n = 0$. The TDG method satisfy the following estimate*

$$\|\mathbf{u} - \mathbf{u}_h\|_{L^2(\Omega)} \leq C h^{n+1/2} \|\mathbf{u}\|_{W^{n+1,\infty}(\Omega)}, \tag{9}$$

with $h = \max_{\Omega_j \in \mathcal{T}_h} h_j$, $h_j = \text{diam}(\Omega_j)$, $\mathbf{u}_h$ stands for the solution to the TDG method calculated along the basis of vectorial exponential functions and C is a universal constant.

Theorem 1.2. *Take $N = 3$. Then (9) holds for all $n \in \mathbb{N}$.*

The Theorem 1.1 combined with (8) shows once again the remarkable property (3) of the TDG method with respect to the FEM/DG (2). This result was proved for the P_1 model in [27, Theorem 1.2] by means of specific methods adapted to the low dimensionality of P_1. Note however that there is a restriction in our estimate of Theorem 1.1. This is due to purely technical difficulties which probably could be tackled, and we strongly believe that the law (9) holds for all $n \geq 0$ and all $N \in 2\mathbb{N}+1$. Indeed in the second Theorem 1.2 which deals with the P_3 model for which recent numerical results have been published [26, 6, 5], the gap between $n = 0$ is filled and $n = 3 - 1 = 2$, and we show that the same estimate holds for $n = 1$.

The organization of the work is as follows. In Section 2, we present preliminary material about the definitions and properties of the matrices $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{R}$. The matrices are specific of the P_N approximation in $2D\frac{1}{2}$ geometry, nevertheless we believe some of the material is common to most P_N models (2D, 3D, various symmetries) where the matrices come from moment approximations. In Section 3 we present the idea of expontial basis functions. Since we desire to be constructive, we take the example of a Trefftz Discontinuous Galerkin method (TDG) and explain why exponential basis functions yield a priori good approximation estimates. Some of the material in this part is standard, this is why we restrict the presentation to the minimum, in particular the stability estimate of TDG. This stability estimate is a kind of Cea's lemma or Strang's first lemma: it explains how a global error estimate between the exact solution and the numerical solution is bounded by local best error

approximation estimates. Next in Section 4, we prove local best error approximation estimate. We rely on a method (initially introduced in [8]) adapted to h-convergence. It consists of showing that a certain matrix of optimal rank. The difficulty is that the matrix is rectangular. The method of the proof is by linear algebra, discrete Fourier techniques and the use of a technical property proved in Section 2. A simple test of numerical convergence illustrates the theory. Additional material about real spherical harmonics is postponed to the Appendix.

2 Preliminary material

The definition of the matrices A and B of the P_N model (4-6) is based on the real spherical harmonics [15]. We also provide some information about the right and left kernels of the matrices, since it plays a key role for the establishment of the main Theorems.

2.1 Real spherical harmonics

We give preliminary minimal information about the real spherical harmonics $X_k^m(\psi,\mu) \in \mathbb{R}$ where the indices are $k \in \mathbb{N}$ and $m \in \mathbb{Z}$ with $|m| \leq k$. More is in the Appendix. Real spherical harmonics are orthonormal for the L^2 product on the sphere S^2 identified with $[0,2\pi] \times [-1,1]$

$$\int_{S^2} X_k^m(\psi,\mu) X_{k'}^{m'}(\psi,\mu) d\psi d\mu = \delta_{kk'}\delta_{mm'}, \qquad \text{where } \psi \in [0,2\pi] \text{ and } \mu = \cos\theta \in [-1,1]. \tag{10}$$

The even harmonics correspond to indices in

$$E(N) = \{(k,m) \in \mathbb{N} \times \mathbb{Z},\ |m| \leq k \leq N,\ k \text{ and } m \text{ even}\},\ \text{where } \dim(E(N)) = m_e = \frac{1}{4}(N+1)^2.$$

The odd harmonics correspond to indices in

$$O(N) = \{(k,m) \in \mathbb{N} \times \mathbb{Z},\ |m| \leq k \leq N,\ k \text{ and } m \text{ odd}\},\ \text{where } \dim(O(N)) = m_o = \frac{1}{4}(N+1)(N+3).$$

Here we introduce **an important convention** for this work. Any function

$$X(\psi,\mu) = \sum_{E(N)} \alpha_k^m X_k^m(\psi,\mu) + \sum_{O(N)} \beta_k^m X_k^m \text{ with } \alpha = (\alpha_k^m) \in \mathbb{R}^{m_e} \text{ and } \beta = (\beta_k^m) \in \mathbb{R}^{m_o}$$

is identified with the vector $\mathbf{X} \in \mathbb{R}^m$ of its coefficients

$$\mathbf{X} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \quad \alpha \in \mathbb{R}^{m_e},\ \beta \in \mathbb{R}^{m_o}.$$

Since real spherical harmonics are orthogonal, the quadratic product between two vectors $\mathbf{X}, \widetilde{\mathbf{X}} \in \mathbb{R}^m$ is equal to its counterpart with functions

$$(\mathbf{X}, \widetilde{\mathbf{X}}) = \sum_{E(N)} \alpha_k^m \widetilde{\alpha}_k^m + \sum_{O(N)} \beta_k^m \widetilde{\beta}_k^m = \int_{S^2} X(\psi,\mu) \widetilde{X}(\psi,\mu) d\psi d\mu.$$

That is why the orthogonality between vectors is also identified with the orthogonality between functions. This convention brings great simplifications in our notations throughout this work.

The following result states that an even (resp. odd) real spherical harmonics premultiplied by $\cos\psi\sqrt{1-\mu^2}$ or by $\sin\psi\sqrt{1-\mu^2}$ is transformed into a linear combination of odd (resp. even) real spherical harmonics.

Lemma 2.1. *For all admissible pairs $(k,m) \in E(N) \bigcup O(N)$, one has the identity*

$$\left\{ \begin{array}{l} \cos\psi \\ \sin\psi \end{array} \right. \times \sqrt{1-\mu^2} X_k^m(\psi,\mu) = \sum_{\epsilon=\pm 1} \sum_{\sigma=\pm 1} a_\epsilon^\sigma X_{k+\epsilon}^{m+\sigma}$$

for some real coefficients a_ϵ^σ.

Proof. The proof is based on the recursion relations (83). See also [15]. □

As a consequence and since N is odd, one has that

$$\cos\psi\sqrt{1-\mu^2}\underset{E(N)}{\mathrm{Span}}\{X_k^m\} \subset \underset{O(N)}{\mathrm{Span}}\{X_k^m\} \text{ and } \sin\psi\sqrt{1-\mu^2}\underset{E(N)}{\mathrm{Span}}\{X_k^m\} \subset \underset{O(N)}{\mathrm{Span}}\{X_k^m\}. \tag{11}$$

The reciprocal embeddings are wrong

$$\cos\psi\sqrt{1-\mu^2}\underset{O(N)}{\mathrm{Span}}\{X_k^m\} \not\subset \underset{E(N)}{\mathrm{Span}}\{X_k^m\} \text{ and } \sin\psi\sqrt{1-\mu^2}\underset{O(N)}{\mathrm{Span}}\{X_k^m\} \subset \underset{E(N)}{\mathrm{Span}}\{X_k^m\}. \tag{12}$$

Let us consider the space

$$\mathcal{O}(N) = \mathrm{Span}\left\{ X(\psi,\mu) = \sqrt{1-\mu^2} \times \mu^p \times \left\{ \begin{array}{l} \cos(2m+1)\psi \\ \sin(2m+1)\psi \end{array} \right. , \ 0 \le p,m,p+m \le \frac{N-1}{2} \right\}.$$

A graphical description of the basis functions in $\mathcal{O}(N)$ is in Figure 2.1. By construction, odd real

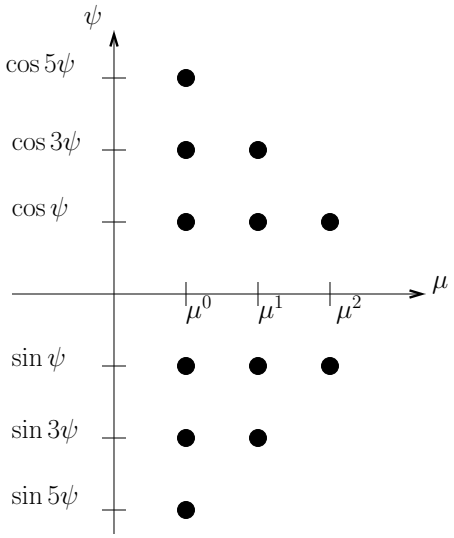

Figure 1: Depiction of the structure of $\mathcal{O}(N)$ for $N = 5$. A bullet represents the product of the function $\sqrt{1-\mu^2}$ with a monomial μ^p (in the horizontal axis) and with a trigonometric monomial (in the vertical axis).

spherical harmonics are linear combination of functions in $\mathcal{O}(N)$, so

$$\underset{O(N)}{\mathrm{Span}}\{X_k^m\} \subset \mathcal{O}(N).$$

Moreover a counting method shows that $\dim(\mathcal{O}(N)) = 2+4+\cdots+2\frac{N-1}{2} = \frac{1}{4}(N+1)(N+3)$. So the dimension of $\mathcal{O}(N)$ is equal to m_o which is also the dimension of odd real spherical harmonics. So the spaces have the same dimension and they are equal, that is

$$\underset{O(N)}{\mathrm{Span}}\{X_k^m\} = \mathcal{O}(N). \tag{13}$$

For $X \in \mathcal{O}(N)$, a convenient representation is

$$X(\psi,\mu) = \sqrt{1-\mu^2} \sum_{p=0}^{\frac{N-1}{2}} X_p(\psi)\mu^p \tag{14}$$

with an expansion of the function $\psi \mapsto X_p(\psi)$ which can be written either

$$X_p(\psi) = \sum_{s=0}^{\frac{N-1}{2}-p} \alpha_s^p \cos(2s+1)\psi + \beta_s^p \sin(2s+1)\psi, \quad \alpha_s^p, \beta_s^p \in \mathbb{R}, \tag{15}$$

or equivalently

$$X_p(\psi) = \sum_{s=p}^{N-p} \gamma_s^p e^{-i(N-2s)\psi}, \quad \gamma_s^p = \overline{\gamma_{N-s}^p} \in \mathbb{C}. \tag{16}$$

2.2 Definition of the matrices A and B

To construct a P_N model, one must calculate the transport matrices $\mathcal{A}$ and $\mathcal{B}$ (4-5). The transport matrices are defined by inserting the components of the direction $\Omega = (\cos\theta\cos\psi, \cos\theta\sin\psi, \sin\theta)$ which comes the transport equation (1) inside the formula (10) for the scalar product of two real spherical harmonics. Due to the structure (4-5), it is sufficient to calculate the extra-diagonal terms coming from the coupling of odd real spherical harmonics and even real spherical harmonics. Indeed Lemma 2.1 and the orthogonality of real spherical harmonics show that the diagonal blocks vanish.

The matrix $A \in \mathbb{R}^{m_e \times m_o}$ is defined by the weak product for all $\alpha = (\alpha_k^m) \in \mathbb{R}^{m_e}$ and $\beta = (\beta_k^m) \in \mathbb{R}^{m_o}$

$$(\alpha, A\beta) = \int_{S^2} \cos\psi\sqrt{1-\mu^2} \sum_{E(N)} \alpha_k^m X_k^m(\psi,\mu) \sum_{O(N)} \beta_{k'}^{m'} X_{k'}^{m'}(\psi,\mu) d\psi d\mu. \tag{17}$$

Similarly the matrix $B \in \mathbb{R}^{m_e \times m_o}$ is defined by the weak product for all $\alpha = (\alpha_k^m) \in \mathbb{R}^{m_e}$ and $\beta = (\beta_k^m) \in \mathbb{R}^{m_o}$

$$(\alpha, B\beta) = \int_{S^2} \sin\psi\sqrt{1-\mu^2} \sum_{E(N)} \alpha_k^m X_k^m(\psi,\mu) \sum_{O(N)} \beta_{k'}^{m'} X_{k'}^{m'}(\psi,\mu) d\psi d\mu \tag{18}$$

The nullity of the block-diagonal terms comes from the identity

$$\int_{S^2} \cos\psi\sqrt{1-\mu^2} \sum_{E(N)} \alpha_k^m X_k^m(\psi,\mu) \sum_{E(N)} \alpha_{k'}^{m'} X_{k'}^{m'}(\psi,\mu) d\psi d\mu = 0$$

for all $\alpha, \alpha' \in \mathbb{R}^{m_e}$ and from the identity

$$\int_{S^2} \cos\psi\sqrt{1-\mu^2} \sum_{O(N)} \beta_k^m X_k^m(\psi,\mu) \sum_{O(N)} \beta_{k'}^{m'} X_{k'}^{m'}(\psi,\mu) d\psi d\mu = 0$$

for all $\beta, \beta' \in \mathbb{R}^{m_o}$. These two relations come from Lemma 2.1.

Two main examples are the P_1 model and the P_3 model. For the P_1 model, one has $m_e = 1$, $m_o = 2$, $m = 3$, $A = \left(0, \frac{1}{\sqrt{3}}\right)$ and $B = \left(\frac{1}{\sqrt{3}}, 0\right)$. For the P_3 model [26, 6, 5], one has $m = 10$, $m_e = 4$, $m_o = 6$ and

$$A = \begin{pmatrix} 0 & \frac{1}{\sqrt{3}} & 0 & 0 & 0 & 0 \\ \frac{1}{\sqrt{5}} & 0 & \sqrt{\frac{3}{14}} & -\frac{1}{\sqrt{70}} & 0 & 0 \\ 0 & -\frac{1}{\sqrt{15}} & 0 & 0 & \sqrt{\frac{6}{35}} & 0 \\ 0 & -\frac{1}{\sqrt{5}} & 0 & 0 & -\frac{1}{\sqrt{70}} & \sqrt{\frac{3}{14}} \end{pmatrix} \tag{19}$$

The matrix B is

$$B = \begin{pmatrix} \frac{1}{\sqrt{3}} & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{\sqrt{5}} & 0 & 0 & -\frac{1}{\sqrt{70}} & -\sqrt{\frac{3}{14}} \\ -\frac{1}{\sqrt{15}} & 0 & 0 & \sqrt{\frac{6}{35}} & 0 & 0 \\ -\frac{1}{\sqrt{5}} & 0 & \sqrt{\frac{3}{14}} & \frac{1}{\sqrt{70}} & 0 & 0 \end{pmatrix}. \tag{20}$$

2.3 Properties of the matrices

In preparation of the main Theorems, we study the kernel of A and the kernel A^T, and we establish the rotational invariance properties of A and B. We will see that $\ker(A^T) \subset \mathbb{R}^{m_e}$ is trivial. On the other hand $\ker(A) \subset \mathbb{R}^{m_o}$ is non trivial. However the intersection of sufficiently many rotations the orthogonal of the kernel is trivial, and this property will the keystone of the proof of the main Theorem 1.1. Also the definition of the rotation operators is needed for the construction of our exponential basis functions (7).

A rotation of angle θ is defined by

$$\sum_{E(N)} \alpha_k^m X_k^m(\psi,\mu) + \sum_{O(N)} \beta_k^m X_k^m(\psi,\mu) \overset{\theta}{\longmapsto} \sum_{E(N)} \alpha_k^m X_k^m(\psi+\theta,\mu) + \sum_{O(N)} \beta_k^m X_k^m(\psi+\theta,\mu).$$

By definition (82) of real spherical harmonics, there exist coefficients $\widetilde{\alpha} \in \mathbb{R}^{m_e}$ and $\widetilde{\beta} \in \mathbb{R}^{m_o}$ such that

$$\sum_{E(N)} \widetilde{\alpha}_k^m X_k^m(\psi,\mu) = \sum_{E(N)} \alpha_k^m X_k^m(\psi+\theta,\mu) \text{ and } \sum_{O(N)} \widetilde{\beta}_k^m X_k^m(\psi,\mu) = \sum_{O(N)} \beta_k^m X_k^m(\psi+\theta,\mu)$$

for all possible ψ and μ. So one can write $\widetilde{\alpha} = U_e(\theta)\alpha$ and $\widetilde{\beta} = U_o(\theta)\beta$ where the unitary rotation matrices $U_e(\theta)$ and $U_o(\theta)$ are assembled in a larger rotation matrix

$$\mathcal{U}(\theta) = \begin{pmatrix} U_e(\theta) & 0 \\ & U_o(\theta) \end{pmatrix} \in \mathcal{M}_m(\mathbb{R}), \quad U_e(\theta) \in \mathbb{R}^{m_e \times m_e} \text{ and } U_o(\theta) \in \mathbb{R}^{m_o \times m_o}. \tag{21}$$

From the definition of real spherical harmonics in the Appendix (see also [26]), it is easy to show that the coefficients of the matrix $U_e(\theta)$ are even values of $\cos(r\theta)$ and $\sin r\theta$, that is $r \in 2\mathbb{N}$. Similarly the coefficients of the matrix $U_o(\theta)$ are odd values of $\cos(r\theta)$ and $\sin r\theta$, that is $r \in 2\mathbb{N}+1$. By construction one has

$$\begin{aligned} &\mathcal{U}(\theta)\mathcal{R} = \mathcal{R}\mathcal{U}(\theta), \\ &\mathcal{U}(\theta)\mathcal{A} = (\cos(\theta)\mathcal{A} + \sin(\theta)\mathcal{B})\mathcal{U}(\theta), \\ &\mathcal{U}(\theta)\mathcal{B} = (-\sin(\theta)\mathcal{A} + \cos(\theta)\mathcal{B})\mathcal{U}(\theta). \end{aligned} \tag{22}$$

It is easily deduced that $U_e(\frac{\pi}{2})A = BU_o(\frac{\pi}{2})$, which is unitary correspondence between A and B. Now we study properties of A, and similar properties follow for B from the unitary correspondence.

Lemma 2.2. *One has $ker(A^T) = \{0\} \subset \mathbb{R}^{m_e}$, which can be rewritten as $AA^T > 0$.*

Proof. This is evident for the P_1 model (it was proved first in Morel's thesis [26]). Let us take $(\alpha_{k,m}) \in E(N)$ in the kernel of A^T. By (17) one has that for all possible $(\beta_{k',m'}) \in O(N)$

$$\int_{S^2} \cos\psi\sqrt{1-\mu^2} \sum_{E(N)} \alpha_k^m X_k^m(\psi,\mu) \sum_{O(N)} \beta_{k'}^{m'} X_k^m(\psi,\mu) d\psi d\mu = 0.$$

Using Lemma 2.1 or embedding (11), take $\sum_{O(N)} \beta_{k'}^{m'} X_k^m(\psi,\mu) = \cos\psi\sqrt{1-\mu^2} \sum_{E(N)} \alpha_k^m X_k^m(\psi,\mu)$. So one gets $\int_{S^2} \left|\cos\psi\sqrt{1-\mu^2} \sum_{E(N)} \alpha_k^m X_k^m(\psi,\mu)\right|^2 d\psi d\mu = 0$. It yields $\sum_{E(N)} \alpha_k^m X_k^m(\psi,\mu) = 0$. The linear independence of the real spherical harmonics shows that $\alpha_k^m = 0$ for all indices. Then $\ker(A^T) = \{0\}$. Finally one has always $\ker(A^T) = \ker(AA^T)$ so the symmetric matrix AA^T is positive and the proof is ended. □

Next, our goal is to characterize the kernel of A, where the fundamental issue is that the reversing embeddings do not hold, see (12). It is also connected to the fact that A is a rectangular matrix with more columns than lines. Instead the rank theorem yields that $\dim(\ker(A)) = m_o - \text{rank}(A) \geq m_o - m_e > 0$. The characterization of this kernel is performed in a series of elementary results.

Lemma 2.3. *One has* $ker(A) = \left(\cos\psi\sqrt{1-\mu^2}\ \underset{E(N)}{Span}\{X_k^m\}\right)^\perp \subset \mathbb{R}^{m_o}$. *So* $dim\,(ker(A)) = m_o - m_e$.

Proof. The first identity is just a rewriting of (17) which states that $\sum_{O(N)} \beta_{k'}^{m'} X_{k'}^{m'} \in \ker(A)$ if and only if it is orthogonal to all functions $\cos\psi\sqrt{1-\mu^2}X_k^m$ for indices $(k,m) \in E(N)$. By orthogonality, one can also write

$$\ker(A)^\perp = \cos\psi\sqrt{1-\mu^2}\ \underset{E(N)}{\text{Span}}\{X_k^m\} \quad \Longrightarrow \quad \ker(A) = \left(\cos\psi\sqrt{1-\mu^2}\ \underset{E(N)}{\text{Span}}\{X_k^m\}\right)^\perp.$$

As $\text{range}(A) = \ker(A^T)^\perp$ and $\ker(A^T)$ is trivial by the previous Lemma, one gets $\text{range}(A) = \mathbb{R}^{m_e}$ and $\text{rank}(A) = m_e$. The rectangular matrix A spans $\mathbb{R}^{m_o}$ into $\mathbb{R}^{m_e}$. Finally the rank Theorem yields $\dim(\ker(A)) = m_o - \text{rank}(A) = m_o - m_e$. □

To continue we note the pure imaginary number $\mathbf{i}^2 = -1$ and define the kernel

$$K_N(\psi) = -\mathbf{i}\sum_{r=0}^{N}(-1)^r e^{\mathbf{i}(N-2r)\psi}. \tag{23}$$

A convenient sumation yields

$$K_N(\psi) = -\mathbf{i}e^{\mathbf{i}N\psi}\sum_{r=0}^{N}(-1)^r e^{-\mathbf{i}2r\psi} = -\mathbf{i}e^{\mathbf{i}N\psi}\frac{1-e^{-2\mathbf{i}(N+1)\psi}}{1+e^{-2\mathbf{i}i\psi}} = \frac{\sin(N+1)\psi}{\cos\psi}. \tag{24}$$

So the kernel K_N is real valued.

Lemma 2.4. *Let* $X \in \mathcal{O}(N)$. *One has that* $X \in ker(A)^\perp \Longleftrightarrow \int_0^{2\pi} K_N(\psi)X(\psi,\mu)d\psi = 0$ *for all* μ.

Proof. • For $(k,m) \in E(N)$, one remarks that

$$\int_0^{2\pi} K_N(\psi)\cos\psi\sqrt{1-\mu^2}X_k^m(\psi,\mu)d\psi = \sqrt{1-\mu^2}\int_0^{2\pi}\sin(N+1)\psi\ X_k^m(\psi,\mu)d\psi.$$

By construction, the function $X_k^m(\psi,\mu)$ can be expanded with respect to $e^{ir\psi}$ with $|r| \leq N$. Then the integral of $e^{ir\psi}$ against $\sin(N+1)\psi$ vanishes, that is $\int_0^{2\pi} K_N(\psi)\cos\psi\sqrt{1-\mu^2}X_k^m(\psi,\mu)d\psi = 0$. By linear combination, it yields that

$$\ker(A)^\perp \subset H, \qquad \text{where } H = \left\{X \in \mathcal{O}(N) \mid \int_0^{2\pi} K_N(\psi)X(\psi,\mu)d\psi = 0 \text{ for all } \mu\right\}.$$

• To prove the equality of two spaces, we show now that they have the same dimension. Take $X \in \mathcal{O}(N)$. Using the representation (14-16), one has the expansion

$$\int_0^{2\pi} K_N(\psi)X(\psi,\mu)d\psi = \sqrt{1-\mu^2}\sum_{p=0}^{\frac{N-1}{2}} m_p(X_p)\mu^p \tag{25}$$

where

$$\begin{aligned} m_p(X_p) &= \int_0^{2\pi} K_N(\psi)X_p(\psi)d\psi \\ &= -i\int_0^{2\pi}\sum_{r=0}^{N}(-1)^r e^{i(N-2r)\psi}\sum_{s=p}^{N-p} e^{i(N-2s)\psi}d\psi \\ &= -2i\pi\sum_{s=p}^{N-p}(-1)^s\gamma_s^p. \end{aligned} \tag{26}$$

We notice that the linear forms $(m_p)_{0\leq p\leq \frac{N-1}{2}}$ are linearly independent. Naturally, the condition $\int_0^{2\pi} K_N(\psi)\cos\psi\sqrt{1-\mu^2}X(\psi,\mu)d\psi = 0$ is equivalent to $m_p(X_p) = 0$ where $0 \leq p \leq \frac{N-1}{2}$: these conditions define a sub-space with co-dimension $\frac{N-1}{2}+1 = \frac{N+1}{2}$.
• Therefore one has

$$\dim H = \dim \mathcal{O}(N) - \frac{N+1}{2} = \frac{1}{4}(N+1)(N+3) - \frac{1}{2}(N+1) = \frac{1}{4}(N+1)^2 = m_e.$$

Since $\ker(A)^\perp$ is embedded in this space and has the same dimension, it is the same space. □

Let us now consider a finite set $S(g)$ of $g > 0$ equidistributed angles

$$S(g) = \left\{\frac{\pi r}{g}\right\}_{1\leq r\leq g} \subset [0,\pi) \tag{27}$$

and the vectorial subspace of $\mathcal{O}(N)$

$$\mathcal{A}(N,g) = \bigcap_{\theta\in S(g)} (U_o(\theta)\ker(A))^\perp_{\mathcal{O}(N)} = \bigcap_{\theta\in S(g)} U_o(\theta)\left(\ker(A)^\perp_{\mathcal{O}(N)}\right) \subset \mathcal{O}(N). \tag{28}$$

This vectorial subspace is the intersection of many different rotations of the same subspace $\ker(A)^\perp$. The next result can be seen as a certain condition of non degeneracy of the subspace $\ker(A)^\perp$. Indeed if $\ker(A)^\perp$ would be invariant with respect all rotations, then the result would be impossible. We notice that the angles are in $(0,\pi]$. Technically, it is compatible what is needed in the next Lemma, see (30). It is also compatible with the end of the proof of the main Theorem, see the definition of the angles μ_p in (78).

Lemma 2.5. *dim* $\mathcal{A}(N,g) = 0$ *for* $g \geq N+1$.

Proof. Take a vector $X \in \mathcal{A}(N,g) \subset \mathcal{O}(N)$, that is $X(\psi,\mu) = \sum_{O(N)} \beta_k^m X_k^m(\psi,\mu)$. Using (28), one has $\int_{S^2} X'(\psi+\theta,\mu)X(\psi,\mu)d\psi d\mu = 0$ for all $X' \in \ker(A)$ and all $\theta \in S(g)$. This is rewritten as

$$\int_{S^2} X'(\psi,\mu)X(\psi-\theta,\mu)d\psi d\mu = 0, \quad \forall X' \in \ker(A).$$

Lemma 2.4 yields $\int_0^{2\pi} K_N(\psi)X(\psi-\theta,\mu)d\psi = 0$ for all μ, rewritten as

$$\int_0^{2\pi} K_N(\psi-\theta)X(\psi,\mu)d\psi = 0, \text{ for all } \mu \text{ and } \theta \in S(g).$$

Using (23) one has $K_N(\psi-\theta) = -i\sum_{r=0}^N (-1)^r e^{i(N-2r)\psi}e^{-i(N-2r)\theta}$, so one gets

$$\sum_{r=0}^N \left(\int_0^{2\pi} e^{i(N-2r)\psi}X(\psi,\mu)d\psi\right)(-1)^r e^{-i(N-2r)\theta} = 0, \quad \forall\theta \in S(g). \tag{29}$$

Writing this expression for $N+1$ different values of $\theta \in [0,2\pi)$ yields a linear system with $N+1$ unknowns and $N+1$ equations. The right hand side is the null vector and the matrix is a Vandermonde matrix

$$V = \left(e^{2ir\theta_s}\right)_{0\leq r,s\leq N}.$$

Such a Vandermonde matrix is non singular if and only if $e^{2i\theta_s} \neq e^{2ir\theta_{s'}}$ for $0 \leq s,s' \leq N$ that is if

$$\theta_s - \theta_{s'} \notin \pi\mathbb{Z}, \quad 0 \leq s,s' \leq N. \tag{30}$$

Because the assumption (27), the condition (30) is fullfiled. So it yields the nullity of the unknown of the linear system

$$\int_0^{2\pi} e^{i(N-2r)\psi} X(\psi,\mu)d\psi = 0 \quad \text{for all } \mu \text{ and } 0 \le r \le N.$$

Since X is an odd moment of the P_N model (see also (16), it shows that $X = 0$ which ends the proof. □

For $2 \le g < N+1$, a similar method of analysis can be used. For $X \in \mathcal{A}(N,g)$, let us plug (14)-(16) into (29). It yields after simplification the equation

$$\sum_{s=p}^{N-p} (-1)^s \gamma_s^p e^{i2s\theta} = 0, \quad \theta \in S(g),\ 0 \le p \le \frac{N-1}{2}. \tag{31}$$

One obtains a rectangular linear system. The unknowns are $((-1)^s\gamma_s^p)_{p\le s\le N-p}$. The rectangular matrix is

$$M_p = \left(e^{\frac{i2\pi sr}{g}}\right)_{p\le s\le N-p}^{1\le r\le g}.$$

It is similar after normalisation to the rectangular matrix

$$N_p = \left(e^{\frac{i2\pi sr}{g}}\right)_{0\le s\le N-2p}^{0\le r\le g-1}.$$

Lemma 2.6. *rank* $N_p = \min(g, N-2p+1)$ *for* $0 \le p \le \frac{N-1}{2}$.

Proof. Necessarily the rank is less than the minimum of the number of rows and the number of lines, that is rank $N_p \le \min(g, N-2p+1)$. But there is always a block square sub-matrix of size $\min(g, N-2p+1)$ which is a Vandermonde matrix, with rank equal to its size. So the claim. □

The previous result has many consequences, displayed for example in the next two Lemmas.

Lemma 2.7. *dim* $\mathcal{A}(N,2) = 2m_e - m_o$.

Proof. Lemma 2.7 yields that rank $N_p = 2$ for all $\frac{N+1}{2}$ different values of p. So

$$\dim \mathcal{A}(N,2) = \dim \mathcal{O}(N) - 2\frac{N+1}{2} = \frac{1}{4}(N+1)(N+3) - (N+1) = \frac{1}{4}(N^2-1).$$

Now $2m_e - m_o = \frac{1}{2}(N+1)^2 - \frac{1}{4}(N+1)(N+3) = \frac{1}{4}(N^2-1)$, so the claim. □

Lemma 2.8. *dim* $\mathcal{A}(3,3) = 1$.

Proof. For $N = 3$, then $0 \le p \le 1 = \frac{N-1}{2}$. Then rank $N_0 = \min(3,4) = 3$ and rank $N_1 = \min(3,2) = 2$. So the linear equations (31) yield 5 linear independent equations and

$$\dim \mathcal{A}(3,3) = \dim \mathcal{O}(3) - 5 = 1.$$

□

3 Vectorial exponential basis functions in the context of TDG

This section is split in sub-sections. The first one explains a generic DG formulation on a mesh. The next sub-section constructs the Trefftz space of basis/shape functions which are vectorial exponential functions. Then, in sub-section 3.3, we insert the Trefftz space into the DG formulation to get our TDG method. The fundamental inequalities which allow the numerical analysis of a TDG method are provided in sub-section 3.4. The last sub-section 4 explains that if the exponential basis functions satisfy a certain fundament property, then the main Theorems 1.1 and 1.2 hold.

3.1 Mesh notation and generic Discontinuous Galerkin formulation

The partition or mesh of the space domain $\Omega \subset \mathbb{R}^2$ is denoted as $\mathcal{T}_h$. It is made of polyhedral non overlapping subdomains Ω_r, that is $\mathcal{T}_h = \bigcup_k \Omega_k$. The broken Sobolev space is

$$H^1(\mathcal{T}_h) := \{\mathbf{v} \in L^2(\Omega),\ \mathbf{v}_{|\Omega_k} \in H^1(\Omega_k)\ \forall \Omega_k \in \mathcal{T}_h\}$$

with the norm $\|\mathbf{w}\|^2_{1,\Omega} = \sum_k \|\mathbf{w}_k\|^2_{H^1(\Omega_k)}$ and the semi-norm $|\mathbf{w}|^2_{1,\Omega} = \sum_k \|\nabla \mathbf{w}_k\|^2_{L^2(\Omega_k)}$.

Let us make the assumption that the solution $\mathbf{u} \in H^1(\mathcal{T}_h)$ has some minimal regularity. We rewrite (4) under the form $L\mathbf{u} = \mathbf{0}$ and consider also the adjoint operator

$$L = \mathcal{A}\partial_x + B\partial_y + \mathcal{R}, \quad L^* = -\mathcal{A}\partial_x - B\partial_y + \mathcal{R} = -L + 2\mathcal{R}.$$

Multiplying the equation $L\mathbf{u} = \mathbf{0}$ by $\mathbf{v} \in H^1(\mathcal{T}_h)$ and integrating on Ω gives $\sum_k \int_{\Omega_k} \mathbf{v}_k \cdot L\mathbf{u}_k = 0$, where $\mathbf{v}_k = \mathbf{v}_{|\Omega_k}$ and $\mathbf{u}_k = \mathbf{u}_{|\Omega_k}$. An integration by parts yields

$$\sum_k \int_{\Omega_k} (L^*\mathbf{v}_k) \cdot \mathbf{u}_k + \sum_k \int_{\partial\Omega_k} \mathbf{v}_k \cdot \mathcal{M}_k(\mathbf{x})\mathbf{u}_k = 0, \tag{32}$$

where $\partial\Omega_k$ is the contour of the element Ω_k and the symmetric matrix $\mathcal{M}_k(\mathbf{x})$ in the last integral is defined on the boundary. Using the notations $\mathcal{M}(\mathbf{n}) = n_x\mathcal{A} + n_y\mathcal{B}$ and $\mathbf{n} = (n_x, n_y)$, and denoting the outward unit normal on the contour as $\mathbf{n}_k(\mathbf{x})$ for $\mathbf{x} \in \partial\Omega_k$, one has

$$\mathcal{M}_k(\mathbf{x}) = \mathcal{M}(\mathbf{n}_k(\mathbf{x})).$$

Since $\mathcal{M}_k(\mathbf{x})$ is symmetric, it can decomposed under the form $\mathcal{M}_k(\mathbf{x}) = \mathcal{M}^+_k(\mathbf{x}) + \mathcal{M}^-_k(\mathbf{x})$ where $\mathcal{M}^+_k$ is a non negative matrix, $\mathcal{M}^-_k$ is a non positive matrix and the matrices annihilate one the other $\mathcal{M}^+_k\mathcal{M}^-_k = \mathcal{M}^-_k\mathcal{M}^+_k = 0$. One can compute the eigenvectors $\mathcal{M}_k r = \lambda \mathbf{r}$, $\|\mathbf{r}\| = 1$, and set $\mathcal{M}^\pm_k = \sum_{\pm\lambda>0} \mathbf{r} \otimes \mathbf{r}$. We supplement the equation of the problem (4) with a very simple boundary condition

$$\mathcal{M}^-_k(\mathbf{x})\mathbf{u} = \mathbf{g} \quad \text{ on } \partial\Omega_k \cap \partial\Omega. \tag{33}$$

This boundary condition yields a global problem with good quadratic estimates, and it is mathematically convenient. Denoting Σ_{kj} the edge oriented from Ω_k to Ω_j when $k \neq j$ and Σ_{kk} the edges belonging to $\Omega_k \cap \partial\Omega$ (for simplicity we use the same notation whatever the number of edges in edge in $\Omega_k \cap \partial\Omega$), one can rewrite (32-33) as

$$\sum_k \int_{\Omega_k} (L^*\mathbf{v}_k) \cdot \mathbf{u}_k + \sum_k \sum_{j<k} \int_{\Sigma_{kj}} (\mathbf{v} \cdot \mathcal{M}(\mathbf{x})\mathbf{u})_k + (\mathbf{v} \cdot \mathcal{M}(\mathbf{x})\mathbf{u})_j \tag{34}$$

$$+ \sum_k \int_{\Sigma_{kk}} \mathbf{v}_k \cdot \mathcal{M}^+_k(\mathbf{x})\mathbf{u}_k = -\sum_k \int_{\Sigma_{kk}} \mathbf{v}_k \cdot \mathcal{M}^-_k(\mathbf{x})\mathbf{g}.$$

For a regular $\mathbf{u}$ satisfying the equation $L\mathbf{u} = 0$ in the whole domain, the normal flux is continuous at interfaces, that is $\mathcal{M}_k(\mathbf{x})\mathbf{u}_k(\mathbf{x}) = \mathcal{M}_k(\mathbf{x})\mathbf{u}_j(\mathbf{x}) = -\mathcal{M}_j(\mathbf{x})\mathbf{u}_j(\mathbf{x})$ for $\mathbf{x} \in \Sigma_{kj}$. This vectorial identity can be projected along the positive and negative eigenvectors of $\mathcal{M}_k = -\mathcal{M}_j$. Denoting $\mathcal{M}_{kj} = \mathcal{M}_{k|\Sigma_{kj}} = -\mathcal{M}_{j|\Sigma_{jk}} = -\mathcal{M}_{jk}$ on Σ_{kj}, one can write as well

$$\mathcal{M}_k\mathbf{u}_k = \mathcal{M}_{kj}\mathbf{u}_j = \mathcal{M}^+_{kj}\mathbf{u}_k + \mathcal{M}^-_{kj}\mathbf{u}_k = \mathcal{M}^+_{kj}\mathbf{u}_k + \mathcal{M}^-_{kj}\mathbf{u}_j$$

because the projection of $\mathcal{M}_k\mathbf{u}_k = \mathcal{M}_{kj}\mathbf{u}_j$ along the eigenvectors of the matrix $\mathcal{M}_k = \mathcal{M}_{kj}$ yields the continuity $\mathbf{r}_{kj} \cdot \mathbf{u}_k = \mathbf{r}_{kj} \cdot \mathbf{u}_j$ for $\lambda \neq 0$. One obtains the identity $(\mathbf{v} \cdot \mathcal{M}(\mathbf{x})\mathbf{u})_k + (\mathbf{v} \cdot \mathcal{M}(\mathbf{x})\mathbf{u})_j = (\mathbf{v}_k - \mathbf{v}_j) \cdot (\mathcal{M}^+_{kj}\mathbf{u}_k + \mathcal{M}^-_{kj}\mathbf{u}_j)$. So (34) can be recast as

$$\sum_k \int_{\Omega_k} (L^*\mathbf{v}_k) \cdot \mathbf{u}_k + \sum_k \sum_{j<k} \int_{\Sigma_{kj}} (\mathbf{v}_k - \mathbf{v}_j) \cdot (\mathcal{M}^+_{kj}(\mathbf{x})\mathbf{u}_k + \mathcal{M}^-_{kj}(\mathbf{x})\mathbf{u}_j) \tag{35}$$

$$+\sum_k \int_{\Sigma_{kk}} \mathbf{v}_k \cdot \mathcal{M}_k^+(\mathbf{x})\mathbf{u}_k = -\sum_k \int_{\Sigma_{kk}} \mathbf{v}_k \cdot \mathcal{M}_k^-(\mathbf{x})\mathbf{g}.$$

We define the bilinear form $a_{DG} : H^1(\mathcal{T}_h) \times H^1(\mathcal{T}_h) \to \mathbb{R}$ and the linear form $l : H^1(\mathcal{T}_h) \to \mathbb{R}$ as

$$\begin{aligned} a_{DG}(\mathbf{u},\mathbf{v}) =& \sum_k \int_{\Omega_k} (L^*\mathbf{v}_k)\cdot\mathbf{u}_k + \sum_k \sum_{j<k} \int_{\Sigma_{kj}} (\mathbf{v}_k - \mathbf{v}_j)\cdot(\mathcal{M}_{kj}^+(\mathbf{x})\mathbf{u}_k + \mathcal{M}_{kj}^-(\mathbf{x})\mathbf{u}_j) \\ &+ \sum_k \int_{\Sigma_{kk}} \mathbf{v}_k \cdot \mathcal{M}_k^+(\mathbf{x})\mathbf{u}_k, \quad \mathbf{u},\mathbf{v} \in H^1(\mathcal{T}_h), \\ l(\mathbf{v}) =& -\sum_k \int_{\Sigma_{kk}} \mathbf{v}_k \cdot \mathcal{M}_k^-(\mathbf{x})\mathbf{g}, \quad \mathbf{v} \in H^1(\mathcal{T}_h). \end{aligned} \tag{36}$$

One can rewrite (35) as $a_{DG}(\mathbf{u},\mathbf{v}) = l(\mathbf{v}),\ \forall \mathbf{v} \in H^1(\mathcal{T}_h)$.

3.2 Vectorial exponential functions and the Trefftz space

In this section, we explain how to create a Trefftz space made of vectorial exponential basis functions. Contrary to a classical Galerkin method with polynomial basis functions, a TDG method takes basis functions which are exact solutions in each cell to the main equation

$$V(\mathcal{T}_h) = \{\mathbf{v} \in H^1(\mathcal{T}_h), L\mathbf{v}_k = \mathbf{0} \quad \forall \Omega_k \in \mathcal{T}_h\} \subset H^1(\mathcal{T}_h),$$

where in our case $L = \mathcal{A}\partial_x + \mathcal{B}\partial_y + \mathcal{R}$. The space $V(\mathcal{T}_h)$ is a genuine subspace of $H^1(\mathcal{T}_h)$ except in the case $L = 0$. As usual with discontinuous methods, the basis functions have the same form in each cell. They are constructed with the exponential method and equi-distributed directions.

For $1 \le s \le 2n+3$ and $1 \le t \le m_e$, we consider $(2n+3)m_e$ vectorial exponential functions

$$\mathbf{u}^{st}(x,y) = e^{\lambda_t(\cos\theta_s x + \sin\theta_s y)}\mathcal{U}(\theta_s)\mathbf{w}_t \tag{37}$$

where $\lambda_t \in \mathbb{R}$, $\theta_s = 2\pi\frac{s}{2n+3}$ and $\mathbf{w}_t \in \mathbb{R}^m$. These vectors $\mathbf{w}^t$ have an even-odd decomposition

$$\mathbf{w}_t = \begin{pmatrix} \mathbf{w}_{te} \in \mathbb{R}^{m_e} \\ \mathbf{w}_{to} \in \mathbb{R}^{m_o} \end{pmatrix}. \tag{38}$$

Lemma 3.1 (Construction of vectorial exponential functions)**.** *Assume $(\lambda_t, \mathbf{w}_{te})$ for $1 \le t \le m_e$ is an eigenpair of the reduced equation*

$$R_e R_o \mathbf{w}_{te} = \lambda_t^2 (AA^T)\mathbf{w}_{te}, \qquad \lambda_t > 0. \tag{39}$$

Then $\mathbf{u}^{st}$ defined by (37-38) is a vectorial exponential solution to the P_N model (4).

Proof. Plug the representation (37) in (4). It yields $-\mathcal{R}\mathcal{U}(\theta_s)\mathbf{w}_t = \lambda_t(\cos\theta_s\mathcal{A} + \sin\theta_s\mathcal{B})\mathcal{U}(\theta_s)\mathbf{w}_t$. With the rotational invariance (22), it simplifies into $-\mathcal{R}\mathbf{w}_t = \lambda_t\mathcal{A}\mathbf{w}_t$. Next the decomposition (38) yields

$$\begin{cases} -R_e\mathbf{w}_{te} = \lambda_t A\mathbf{w}_{to}, \\ -R_o\mathbf{w}_{to} = \lambda_t A^T\mathbf{w}_{te}. \end{cases} \tag{40}$$

The matrix $R_o = \sigma_t I_{m_o}$ is diagonal. Multiply the first line by σ_t. It yields the eigenequation (39). The matrices are symmetric positive, that is $R_eR_o > 0$ and $AA^T > 0$ by Lemma 2.2, so there exists an eigendecomposition of (39) with real eigenvectors $\mathbf{w}_{te} \in \mathbb{R}^{m_e}$ and real positive eigenvalues $\lambda_t > 0$. Reciprocally take an eigenpair (39) and define $\mathbf{w}_{to} = -R_o^{-1}\lambda_t A^T\mathbf{w}_{te}$. It yields the first line of (40), which, with (37), ends the proof of the claim. □

Definition 3.2. *The Trefftz space that we study in this work is spanned by vectorial exponential functions*

$$V_h(\mathcal{T}_h) = \{\mathbf{v} \in H^1(\mathcal{T}_h), \mathbf{v}_k \in Span\left(\mathbf{u}^{st}\right)_{1\leq s\leq 2n+3}^{1\leq t\leq m_e} \quad \forall \Omega_k \in \mathcal{T}_h\}. \tag{41}$$

The dimension of the Trefftz space is dim $V_h(\mathcal{T}_h)$ = Number cells $\times$ $(2n+3) \times m_e$. One can also write

$$V_h(\mathcal{T}_h) = \oplus_k \left(\text{Span}\left(\mathbf{u}^{st}\right) 1_{\Omega_k}\right) = \text{Span}\left(\mathbf{u}^{st}\right) \otimes \text{Span}\left(1_{\Omega_k}\right)$$

where 1_{Ω_k} is the indicatrix function of the cell Ω_k.

3.3 Trefftz Discontinuous Galerkin formulation

One approximates general functions $\mathbf{u}, \mathbf{v} \in V(\mathcal{T}_h)$ by functions in the discrete Trefftz space $\mathbf{u}_h, \mathbf{v}_h \in V_h(\mathcal{T}_h)$. Starting from the bilinear form a_{DG} (36), one consider the volume term which can be written as

$$\left(L^*\mathbf{v}_k\right) \cdot \mathbf{u}_k = -\left(\mathcal{A}\partial_x + \mathcal{B}\partial_y\right)\mathbf{v}_k \cdot \mathbf{u}_k + \mathcal{R}\mathbf{v}_k \cdot \mathbf{u}_k$$
$$= -\left(\mathcal{A}\partial_x + \mathcal{B}\partial_y\right)\mathbf{v}_k \cdot \mathbf{u}_k - \mathbf{v}_k \cdot \left(\mathcal{A}\partial_x + \mathcal{B}\partial_y\right)\mathbf{v}_k = -\left(\mathcal{A}\partial_x + \mathcal{B}\partial_y\right)(\mathbf{v}_k \cdot \mathbf{u}_k).$$

With a direct integration of the first term in (36) , one gets a bilinear form $a_T(\cdot,\cdot)$

$$a_T(\mathbf{u},\mathbf{v}) = -\sum_k \sum_{j<k} \int_{\Sigma_{kj}} (\mathcal{M}_{kj}^-(\mathbf{x})\mathbf{v}_k + \mathcal{M}_{kj}^+(\mathbf{x})\mathbf{v}_j) \cdot (\mathbf{u}_k - \mathbf{u}_j) - \sum_k \int_{\Sigma_{kk}} \mathbf{v}_k \cdot \mathcal{M}_k^-(\mathbf{x})\mathbf{u}_k, \quad \mathbf{u}, \mathbf{v} \in V(\mathcal{T}_h). \tag{42}$$

The relaxation matrix $\mathcal{R}$ completely disappeared in the bilinear form. It might seem a paradox at first sight but it is not because, for a Trefftz method, information about the matrix $\mathcal{R}$ is encoded in the basis functions. Also there is no volume term in this formulation which may be easier to implement, even if it is perhaps more a matter of personal taste. The related bilinear form $l : V(\mathcal{T}_h) \to \mathbb{R}$ is the same as in (36), that is $l(\mathbf{v}) = -\sum_k \int_{\Sigma_{kk}} \mathbf{v}_k \cdot \mathcal{M}_k^-(\mathbf{x})\mathbf{g}$ for all $\mathbf{v} \in V(\mathcal{T}_h)$.

The Trefftz numerical analyzed in this work is as follows. We take $V_h(\mathcal{T}_h)$ the finite subspace of $V(\mathcal{T}_h)$ made with vectorial exponential functions. The Trefftz discontinuous Galerkin method is formulated as

$$\begin{cases} \text{Find } \mathbf{u}_h \in V_h(\mathcal{T}_h) \text{ such that} \\ a_T(\mathbf{u}_h, \mathbf{v}_h) = l(\mathbf{v}_h), \quad \forall \mathbf{v}_h \in V_h(\mathcal{T}_h). \end{cases} \tag{43}$$

3.4 Numerical analysis of TDG

TDG methods with exponential basis functions are not standard with respect to traditional DG methods, and it is useful to review basic results from the TDG theory [25, 11, 21, 27, 26] before explaining the accuracy offered by the vectorial exponential basis functions.

One defines two semi-norms on $H^1(\mathcal{T}_h)$

$$\|\mathbf{u}\|_{DG}^2 = \sum_k \int_{\Omega_k} \mathbf{u}_k \cdot R\mathbf{u}_k + \sum_k \sum_{j<k} \frac{1}{2} \int_{\Sigma_{kj}} (\mathbf{u}_k - \mathbf{u}_j) \cdot |M_{kj}|(\mathbf{u}_k - \mathbf{u}_j) + \sum_k \frac{1}{2} \int_{\Sigma_{kk}} \mathbf{u}_k \cdot |\mathcal{M}_k|\mathbf{u}_k,$$
$$\|\mathbf{u}\|_{DG^*}^2 = \sum_k \int_{\partial\Omega_k} -\mathbf{u}_k \cdot \mathcal{M}_k^- \mathbf{u}_k, \tag{44}$$

with $|\mathcal{M}_{kj}| = |\mathcal{M}_{jk}| = \mathcal{M}_{kj}^+ - \mathcal{M}_{kj}^-$. First steps are to show that these two semi-norms are in fact norms on the Trefftz space. All proves can be completed from [25, 11, 21, 27, 26].

Lemma 3.3. *One has the inequality* $\|\mathbf{v}\|_{DG} \leq c\|\mathbf{v}\|_{DG^*}$ *for* $\mathbf{v} \in V(\mathcal{T}_h)$*, with* $c = \sqrt{\frac{5}{2}}$.

Lemma 3.4. *The semi-norms* $\|\cdot\|_{DG}$ *and* $\|\cdot\|_{DG^*}$ *are norms on the Trefftz space* $V(\mathcal{T}_h)$.

Lemma 3.5 (Coercivity). *One has* $a_T(\mathbf{u},\mathbf{u}) = \|\mathbf{u}\|_{DG}^2$ *for* $\mathbf{u} \in V(\mathcal{T}_h)$.

Lemma 3.6 (Continuity). *The bound* $a_T(\mathbf{u},\mathbf{v}) \leq \sqrt{2}\|\mathbf{u}\|_{DG}\|\mathbf{v}\|_{DG^*}$ *holds for* $\mathbf{u},\mathbf{v} \in V(\mathcal{T}_h)$.

The classical quasi-optimality result is the following.

Lemma 3.7 (Quasi-optimality). *The TDG formulation (43) admits a unique solution* $\mathbf{u}_h \in V_h(\mathcal{T}_h)$ *which satisfies the estimate* $\|\mathbf{u}-\mathbf{u}_h\|_{DG} \leq \sqrt{2}\inf_{\mathbf{v}_h \in V_h(\mathcal{T}_h)} \|\mathbf{u}-\mathbf{v}_h\|_{DG^*}$.

Next we present some elementary estimates adapted to our problem. Proofs are in [27, 26].

Lemma 3.8. *One has the a priori bound* $\|\mathbf{w}\|_{L^2(\Omega)} \leq C\|\mathbf{w}\|_{DG}$ *where the constant* $C > 0$ *depends on the invertible matrice* $\mathcal{R}$.

Lemma 3.9. *One has the a priori bound* $\|\mathbf{w}\|_{DG^*}^2 \leq C\sum_j \|\mathbf{w}\|_{L^2(\Omega_j)}\left(\frac{1}{h_j}\|\mathbf{w}\|_{L^2(\Omega_j)} + |\mathbf{w}|_{1,\Omega_j}\right)$ *where* $h_j = diam(\Omega_j)$ *and the constant* $C > 0$ *depends on the matrices* $\mathcal{A}$ *and* $\mathcal{B}$.

Proposition 3.10. *One has the bound*

$$\|\mathbf{u}-\mathbf{u}_h\|_{L^2(\Omega)} \leq C \inf_{\mathbf{v}_h \in V_h}\left(h^{\frac{1}{2}}\,|\nabla\mathbf{u}-\nabla\mathbf{v}_h|_{L^2(\Omega)}^{\text{broken}} + h^{-\frac{1}{2}}\,\|\mathbf{u}-\mathbf{v}_h\|_{L^2(\Omega)}\right) \tag{45}$$

where the constant $C > 0$ *depends on the matrices of the problem.*

Proof. Plug the result of Lemma 3.8 in the right hand side of the inequality of Lemma 3.7, then plug the result of Lemma 3.9 in the left hand side. □

3.5 The fundamental property of vectorial exponential functions

Inequality (45) gives a bound of the L^2 norm of numerical error in function of the best error approximation in a weighted H^1 norm. It remains to show that this best error approximation is high order with respect to h. For this task we adapt a method that was proposed in [8]. The idea is consider an infinite expansion

$$\mathbf{u}(x,y) = \sum_{p,q\geq 0} \mathbf{u}_{pq}x^p y^q \tag{46}$$

for a smooth solution of $L\mathbf{u} = 0$.

Lemma 3.11. *One has the recurrence relations*

$$(p+1)\mathcal{A}\mathbf{u}_{p+1,q} + (q+1)\mathcal{B}\mathbf{u}_{p,q+1} = -\mathcal{R}\mathbf{u}_{pq}, \quad \forall p,q \geq 0. \tag{47}$$

Proof. Plug the expansion (46) in the equation (4) and identity the coefficient in front of $x^p y^q$. □

Formal expansions of the exponential basis functions (37) write as well

$$\mathbf{u}^{st}(x,y) = \sum_{p,q\geq 0} \mathbf{u}_{pq}^{st}x^p y^q, \qquad 1 \leq s \leq 2n+3,\ 1 \leq t \leq m_e. \tag{48}$$

Let us consider (46) and (48) in a generic cell $\omega \in \mathcal{T}_h$. Up to a translation, this generic cell contains the origin $O = (0,0)$. Both the main equation (4) and the family of exponential basis functions $\mathbf{u}^{st}$ are invariant with respect to translations, so this assumption is not a restriction.

Let us define the space $V_n \subset \mathbb{R}^{m\times\frac{(n+1)(n+2)}{2}}$ which corresponds to the truncation at order n of the formal series (46)

$$V_n = \{(\mathbf{u}_{pq})_{0\leq p+q\leq n},\ \text{where (47) is satisfied for } 0 \leq p+q \leq n\}.$$

For $p+q=n$, the condition means that there exists additional vectors $(\mathbf{u}_{p'q'})_{p'+q'=n+1}$ such that the condition $(p+1)\mathcal{A}\mathbf{u}_{p+1,q}+(q+1)\mathcal{B}\mathbf{u}_{p,q+1}=-\mathcal{R}\mathbf{u}_{pq}$ holds for all $p+q=n$. By definition the Taylor expansion of any sufficiently smooth solution $\mathbf{u}\in C^{n+1}$ of (4) generates an element in V_n.

A linear combination of the basis functions is a local $O(h^{n+1})$ approximation of $\mathbf{u}$ if and only if one can find coefficients α^{st} such that

$$\sum_{s=1}^{2n+3}\sum_{t=1}^{m_e}\mathbf{u}_{pq}^{st}\alpha^{st}=\mathbf{u}_{pq},\qquad 0\le p+q\le n,\quad p,q\ge 0. \tag{49}$$

This is a rectangular linear system with $(2n+3)m_e$ unknowns which are the α^{st} for $1\le s\le 2n+3$ and $1\le t\le m_e$, and with $\frac{(n+1)(n+2)}{2}$ linear equations. The number of linear equations is equal to the number of different pairs (p,q). The coefficients of the linear system are the Taylor coefficients of the expansions (48). The rectangular matrix of the linear system is

$$M_n=\left(\mathbf{u}_{pq}^{st}\right)_{0\le p,q\le n}^{1\le s\le 2n+3,\ 1\le t\le m_e}\in\mathbb{R}^{a\times b},\quad a=\frac{(n+1)(n+2)}{2},\ b=(2n+3)m_e. \tag{50}$$

By construction, all columns of the matrix M_n belong to V_n because they are made from vectorial exponential functions which satisfy all recurrence relations (48). Therefore one has by definition

$$\mathrm{range}(M_n)\subset V_n. \tag{51}$$

Definition 3.12. *What we call the fundamental property of the vectorial exponential functions is* $\mathrm{range}(M_n)=V_n$.

Assume the fundamental property. Then there exists $(2n+3)m_e$ coefficients $(\alpha^{st})_{1\le t\le m_e}^{1\le s\le 2n+3}$ solution of the linear system (49). This solution of the linear system may be non unique if the kernel of the matrix is not trivial. However it is an exercize in linear algebra to show that it is possible to determine at leat one particular solution which is bounded by in norm of the coefficients in the right hand side, that is

$$\max_{1\le s\le 2n+3}^{1\le t\le m_e}|\alpha^{st}|\le C\max_{p,q}|\mathbf{u}_{pq}|,\quad\forall(\mathbf{u}_{pq})\in V_n.$$

This property is proved by constructing a pseudo-inverse of M_n.

Proposition 3.13. *Assume the fundamental property and take a generic cell* $\omega\in\mathcal{T}_h$. *There exists a linear combination of vectorial exponential functions* $\mathbf{v}_h=\sum_{s=1}^{2n+3}\sum_{t=1}^{m_e}\alpha^{st}\mathbf{u}^{st}$ *with the bounds*

$$\|\mathbf{u}-\mathbf{v}_h\|_{L^\infty(\omega)}\le C\|\mathbf{u}\|_{W^{n+1}(\omega)}h^{n+1} \tag{52}$$

and

$$\|\nabla\mathbf{u}-\nabla\mathbf{v}_h\|_{L^\infty(\omega)}\le C\|\mathbf{u}\|_{W^{n+1}(\omega)}h^{n}. \tag{53}$$

Proof. Standard Taylor expansion at order $n+1$ are

$$\mathbf{u}(x,y)=\sum_{0\le p+q\le n}\mathbf{u}_{pq}x^py^q+O(h^{n+1})|\mathbf{u}|_{W^{n+1}(\omega)}$$

and

$$\mathbf{v}_h(x,y)=\sum_{s=1}^{2n+3}\sum_{t=1}^{m_e}\left(\sum_{0\le p+q\le n}\mathbf{u}_{pq}x^py^q\right)\alpha^{st}+O(h^{n+1})\max_{s,t}|\alpha^{st}|.$$

Subtracting the second expansion to the first one yields (52). A similar technique for the derivatives yields (53): for the x derivative for example, one make the subtraction of

$$\partial_x\mathbf{u}(x,y)=\sum_{0\le p+q\le n}\mathbf{u}_{pq}px^{p-1}y^q+O(h^{n})|\mathbf{u}|_{W^{n+1}(\omega)}$$

and

$$\partial_x \mathbf{v}_h(x,y) = \sum_{s=1}^{2n+3} \sum_{t=1}^{m_e} \left(\sum_{0\le p+q\le n} \mathbf{u}_{pq} p x^{p-1} y^q \right) \alpha^{st} + O(h^n) \max_{s,t} |\alpha^{st}|.$$

By subtraction all coefficients vanish. It is similar for the derivative with respect to y. □

Still assuming that the fundamental property holds, one obtains the main result of this work.

Proof of Theorem (1.1). Plug (52-53) in (45) and bound the number of cells using the uniformity of the mesh. □

4 Proof of the fundamental property

So far we have explained that vectorial exponential functions, provided the matrix of their Taylor coefficients satisfy the fundamental property, yield high order convergence. In this section, we prove the fundamental property by linear algebra and discrete Fourier techniques.

4.1 Upper bound on $\dim(V_n)$

In order to study the structure of the space V_n, it is valuable to introduce another space called W_n which is simpler to examine. This is done by elimination of the linear equations in (47) for $p+q<n$, and elimination of corresponding variables $\mathbf{u}_{pq}$ for $p+q<n$. Lemmas 4.1 and 4.2 yields also an upper on the rank of the matrix M.

Let us define the linear subspace $W_n \subset \mathbb{R}^{m\times(n+1)}$

$$W_n = \{(\mathbf{u}_{pq})_{p+q=n},\ \text{there exists } (\mathbf{v}_{pq}) \in V_n \text{ such that } \mathbf{u}_{pq} = \mathbf{v}_{pq} \text{ for } p+q=n.\} \tag{54}$$

Lemma 4.1. *One has $dim(V_n) = dim(W_n)$.*

Proof. Take a basis in W_n. Then, with a descending recurrence based on (47), all these basis vectors $\in R^{m\times(n+1)}$ can be completed as vectors $\in \mathbb{R}^{m\times\frac{(n+1)(n+2)}{2}}$. It yields a basis in V_n, so the proof. □

Lemma 4.2. *One has $dim(W_n) \le (n+1)m_e + \min\left((n+2)m_e, (n+1)m_o\right)$.*

Proof. Write

$$\mathbf{u}_{pq} = \begin{pmatrix} \alpha_{pq} \in \mathbb{R}^{m_e} \\ \beta_{pq} \in \mathbb{R}^{m_o} \end{pmatrix} \in \mathbb{R}^m. \tag{55}$$

The idea is to evaluate separately the dimension of the (α_{pq}) and the dimension of the (β_{pq}). One has immediately that $\dim\left(\text{Span}\left\{(\alpha_{pq})_{p+q=n}\right\}\right) \le (n+1)m_e$.

One notes that

$$\beta_{pq} = -R_o^{-1}\left((p+1)A^T\alpha_{p+1,q} + (q+1)B^T\alpha_{p,q+1}\right).$$

So $\dim\left(\text{Span}\left\{(\beta_{pq})_{p+q=n}\right\}\right) \le \dim\left(\text{Span}\left\{(\alpha_{pq})_{p+q=n+1}\right\}\right) \le (n+2)m_e$. But one has also that $\dim\left(\text{Span}\left\{(\beta_{pq})_{p+q=n}\right\}\right) \le (n+1)m_o$. Therefore $\dim\left(\text{Span}\left\{(\beta_{pq})_{p+q=n}\right\}\right) \le \min((n+2)m_e, (n+1)m_o)$ which induces the claim. □

Lemma 4.3. *The inequality of Lemma 4.2 is an equality for the P_1 model.*

Proof. For the P_1 system the recurrence relations write

$$p+q=n: \begin{cases} (p+1)\begin{pmatrix} 0 \\ \frac{1}{\sqrt{3}} \end{pmatrix}\cdot\beta_{p+1,q} + (q+1)\begin{pmatrix} \frac{1}{\sqrt{3}} \\ 0 \end{pmatrix}\cdot\beta_{p,q+1} & = -\sigma_a\alpha_{pq}, \\ \qquad\qquad \begin{pmatrix} (p+1)\alpha_{p+1,q} \\ (q+1)\alpha_{p,q+1} \end{pmatrix} & = -\sigma_t\beta_{pq}. \end{cases} \tag{56}$$

Here $m_e = 1$ and $m_o = 2$, that is $\alpha_{pq} \in \mathbb{R}$ and $\beta_{pq} \in \mathbb{R}^2$. For the first line, whatever are the $(\alpha_{pq})_{p+q=n}$, it is possible to find $(\beta_{p'q'})_{p'+q'=n+1}$ which satisfy the constraints. So $\dim\left(\text{Span}\left\{(\alpha_{pq})_{p+q=n+1}\right\}\right) = n+1$. For the second line, the $n+1$ vectors β_{pq} (that is a priori $2n+2$ scalar quantities) depend linearly on the $n+2$ scalar quantities $\alpha_{p'q'}$ for $p'+q' = n+1$. Here $\dim\left(\text{Span}\left\{(\beta_{pq})_{p+q=n+1}\right\}\right) = n+2$. It yields a space W_n with $\dim(W_n) = (n+1)+(n+2) = 2n+3$. Since $m_e = 1$, it is the claim. □

4.2 Matrix transformations

We come to the heart of the matter, which is to analyze the matrix M_n to show the fundamental property $\text{range}(M_n) = V_n$. From the expansion (46) the coefficients of the columns of the matrix M_n are

$$\mathbf{u}_{pq}^{st} = \frac{\lambda_t^{p+q}\cos^p\theta_s \sin^q\theta_s}{p!q!}\,\mathcal{U}(\theta_s)\mathbf{w}_t = \frac{\cos^p\theta_s \sin^q\theta_s}{p!q!}\begin{pmatrix} \lambda_t^{p+q} U_e(\theta_s)\mathbf{w}_{te} \\ -\lambda_t^{p+q+1} U_o(\theta_s) R_o^{-1} A^T \mathbf{w}_{te}\end{pmatrix}. \tag{57}$$

The collection of the orthonormal eigenvectors $\mathbf{w}_{te}$ of the reduced problem (39) is assembled in the matrix

$$H = \left(\begin{array}{c|c|c|c} \mathbf{w}_{1,e} & \mathbf{w}_{2,e} & \dots & \mathbf{w}_{m_e,e}\end{array}\right) \in \mathbb{R}^{me\times m_e}.$$

This is a unitary matrix by construction, that is

$$H^T H = I_e. \tag{58}$$

The corresponding collection of eigenvalues is assembled in the square matrix

$$D = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_{m_e}) \in \mathbb{R}^{me\times m_e}.$$

We note that the right multiplication of H by a rth-power of D is

$$HD^r = \left(\begin{array}{c|c|c|c} \lambda_1^r\mathbf{w}_{1,e} & \lambda_2^r\mathbf{w}_{2,e} & \dots & \lambda_{m_e}^r\mathbf{w}_{m_e,e}\end{array}\right) \in \mathbb{R}^{me\times m_e}.$$

Starting from the matrix M_n and keeping only the lines which correspond to $p+q = n$, one gets a reduced matrix denoted as

$$N = \left(\begin{array}{c|c|c|c} \cos^n\theta_1 \begin{pmatrix} U_e(\theta_1)HD^n \\ -U_o(\theta_1)R_o^{-1}A^THD^{n+1}\end{pmatrix} & \dots & \dots & \dots \\ \hline \dots & \dots & \dots & \dots \\ \hline \cos^p\theta_1\sin^q\theta_1 \begin{pmatrix} U_e(\theta_1)HD^n \\ -U_o(\theta_1)R_o^{-1}A^THD^{n+1}\end{pmatrix} & \dots & \dots & \dots \\ \hline \dots & \dots & \dots & \dots \\ \hline \sin^n\theta_1 \begin{pmatrix} U_e(\theta_1)HD^n \\ -U_o(\theta_1)R_o^{-1}A^THD^{n+1}\end{pmatrix} & \dots & \dots & \dots \\ \hline\hline \theta_1 & \theta_1 \leftarrow \theta_2 & \dots & \theta_1 \leftarrow \theta_{2n+3}\end{array}\right). \tag{59}$$

The block lines in (59) correspond to the indices p and q such that $p+q = n$, that is $n+1$ block lines from $(p,q) = (n,0)$ to $(p,q) = (0,n)$. The block columns in (59) correspond to different angles, that is $2n+3$ block columns from θ_1 to θ_{2n+3}. Each block is a rectangular matrix in $\mathbb{R}^{m\times m_e}$ decomposed in square matrix in $\mathbb{R}^{m_e\times m_e}$ on top of a rectangular matrix in $\mathbb{R}^{m_o\times m_e}$. The different angles used in the different block columns are written under the triple line. The dimension is $N \in \mathbb{R}^{a\times b}$ with $a = (n+1)\times m$ and $b(2n+3)\times m_e$. Notice that we drop the index n in the notation of the matrix N because it plays no role.

Lemma 4.4. *rank*$(N) =$ *rank*$(M_n) \leq$ *dim*(W_n).

Proof. It is by construction of the matrix N and corollary of (51) and Lemma 4.1. □

To analyze the rank of N, the method is by successive transformations which are based on multiplication on the right by non singular matrices. For mathematical convenience, we make linear combinations of the block-lines to replace $\cos^n\theta$, $\cos^{n-1}\sin\theta$, ..., $\sin^n\theta$ with $e^{in\theta}$, $e^{i(n-2)\theta}$, ..., $e^{-in\theta}$. For this operation, multiplication on the left by a convenient non singular matrix is enough. It is very standard so we do not develop the algebra. We obtain a complex valued rectangular matrix $N_1 \in \mathbb{C}^{a\times b}$ with the same size and same rank

$$\text{rank}(N_1) = \text{rank}(N). \tag{60}$$

It can be written as

$$N_1 = \left(\begin{array}{c|c|c|c} e^{in\theta_1}\begin{pmatrix} U_e(\theta_1)HD^n \\ -U_o(\theta_1)R_o^{-1}A^THD^{n+1}\end{pmatrix} & \dots & \dots & \dots \\ \hline \dots & \dots & \dots & \dots \\ \hline e^{i(n-2p)\theta_1}\begin{pmatrix} U_e(\theta_1)HD^n \\ -U_o(\theta_1)R_o^{-1}A^THD^{n+1}\end{pmatrix} & \dots & \dots & \dots \\ \hline \dots & \dots & \dots & \dots \\ \hline e^{-in\theta_1}\begin{pmatrix} U_e(\theta_1)HD^n \\ -U_o(\theta_1)R_o^{-1}A^THD^{n+1}\end{pmatrix} & \dots & \dots & \dots \\ \hline\hline \theta_1 & \theta_1 \leftarrow \theta_2 & \dots & \theta_1 \leftarrow \theta_{2n+3} \end{array}\right). \tag{61}$$

To continue the transformations, we define two aditionnal matrices. The first one is a square matrix $E_1 \in \mathbb{R}^{b\times b}$ with $b = (2n+3)\times me$

$$E_1 = -\sigma_t \begin{pmatrix} D^{-n-1}H^{-1}U_e(-\theta_1) & 0 & \dots & 0 \\ 0 & D^{-n-1}H^{-1}U_e(-\theta_2) & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & D^{-1}H^{-1}U_e(-\theta_{2n+3}) \end{pmatrix}. \tag{62}$$

The second matrice is also a square matrix, but with a different size

$$Q = HD^{-1}H^{-1} \in \mathcal{M}_{m_e}(\mathbb{C}). \tag{63}$$

One also has two algebraic relations. The first one comes from (21-22)

$$U_o(\theta)A^T = (\cos\theta A^T + \sin\theta B^T)U_e(\theta). \tag{64}$$

The second one is simply

$$\sigma_t R_o^{-1} = I_o. \tag{65}$$

Now we use the structures (61-65) to calculate the matrix $N_2 = N_1E_1 \in \mathbb{C}^{a\times b}$

$$N_2 = \left(\begin{array}{c|c|c|c} e^{in\theta_1}\begin{pmatrix} -\sigma_t U_e(\theta_1)QU_e(-\theta_1) \\ \cos\theta_1 A^T + \sin\theta_1 B^T\end{pmatrix} & \dots & \dots & \dots \\ \hline \dots & \dots & \dots & \dots \\ \hline e^{i(n-2q)\theta_1}\begin{pmatrix} -\sigma_t U_e(\theta_1)QU_e(-\theta_1) \\ \cos\theta_1 A^T + \sin\theta_1 B^T\end{pmatrix} & \dots & \dots & \dots \\ \hline \dots & \dots & \dots & \dots \\ \hline e^{-in\theta_1}\begin{pmatrix} -\sigma_t U_e(\theta_1)QU_e(-\theta_1) \\ \cos\theta_1 A^T + \sin\theta_1 B^T\end{pmatrix} & \dots & \dots & \dots \\ \hline\hline \theta_1 & \theta_1 \leftarrow \theta_2 & \dots & \theta_1 \leftarrow \theta_{2n+3} \end{array}\right). \tag{66}$$

We consider that the mathematical structure of the matrix N_1 is more amenable for mathematical analysis than (59) for 3 reasons. Firstly the matrices $U_e(\theta_s)QU_e(-\theta_s)$ are positive hermitian for all θ_s because H is unitary (58) and so Q is also positive hermitian. Secondly the matrices $\cos\theta_s A^T + \sin\theta_s B^T$

have a spectral decomposition which can be established using (22). Thirdly complex exponentials show up in the (block)lines, so it suggests to make a discrete Fourier transform to obtain more decoupling.

Let us now consider that the angles are equidistributed, so that a discrete Fourier transform is possible. To simplify the notations, we note

$$\mu = \theta_1 \text{ and } \theta_s = s\mu \text{ for } 1 \leq s \leq 2n+3. \tag{67}$$

The matrix N_2 can be written with the block structure $N_2 = (N_{2,pq})$ with

$$N_{2,pq} = e^{i(n-2(p-1))q\mu} \begin{pmatrix} -\sigma_t U_e(q\mu) Q U_e(-q\mu) \\ \cos q\mu A^T + \sin q\mu B^T \end{pmatrix}, \qquad 1 \leq p \leq n+1,\ 1 \leq q \leq 2n+3.$$

To perform the discrete Fourier transform, let us define the block diagonal square matrix $E_2 \in \mathbb{C}^{b\times b}$ with $b = (2n+3) \times m_e$

$$E_2 = \left(e^{ipq\mu} I_e\right)_{1\leq p,q\leq 2n+3}. \tag{68}$$

It is a non singular block diagonal VanderMonde matrix. The product is the matrix

$$N_3 = N_2 E_2 \in \mathbb{C}^{a\times b}$$

with a block structure $N_3 = (N_{3,pq})_{1\leq p\leq n+1, 1\leq q\leq 2n+3}$ with

$$N_{3,pq} = \sum_{k=1}^{2n+3} N_{2,pk} e^{ikq\mu} = \sum_{k=1}^{2n+3} e^{i(n-2(p-1)+q)k\mu} \begin{pmatrix} -\sigma_t U_e(k\mu) Q U_e(-k\mu) \\ \cos k\mu A^T + \sin k\mu B^T \end{pmatrix}, \qquad 1 \leq p \leq n+1,\ 1 \leq q \leq 2n+3. \tag{69}$$

One again N_3 and N have the same rank

$$\operatorname{rank}(N_3) = \operatorname{rank}(N).$$

The result below explains that this way of transforming the matrices yields important simplifications because certain coefficients vanish. We write $N_{3,pq} = \begin{pmatrix} N^e_{3,pq} \in \mathbb{C}^{m_e\times m_e} \\ N^o_{3,pq} \in \mathbb{C}^{m_o\times m_e} \end{pmatrix}$.

Lemma 4.5. *Consider the block representation (69) of the matrix N_3 for a pair (p,q) such that $1 \leq p \leq n+1$ and $1 \leq q \leq 2n+3$. Three cases arise.*

- *Assume $q \neq 2(p-1) - n \pm 1 \pmod{2n+3}$. Then $M^o_{3,pq} = 0$.*
- *Assume $q = 2(p-1) - n - 1 \pmod{2n+3}$. Then $M^o_{3,pq} = \frac{1}{2}(2n+3)(A-iB)^T$.*
- *Assume $q = 2(p-1) - n + 1 \pmod{2n+3}$. Then $M^o_{3,pq} = \frac{1}{2}(2n+3)(A+iB)^T$.*

Proof. • First case: $q \neq 2(p-1) - n \pm 1 \pmod{2n+3}$. One can write $\cos\theta A^T + \sin\theta B^T = \frac{1}{2}e^{i\theta}(A - iB)^T + \frac{1}{2}e^{-i\theta}(A+iB)^T$ so

$$\sum_{k=1}^{2n+3} e^{i(n-2(p-1)+q)k\mu} \left(\cos k\mu A^T + \sin k\mu B^T\right)$$

$$= \frac{1}{2}(A-iB)^T \sum_{k=1}^{2n+3} e^{i(n-2(p-1)+q+1)k\mu} + \frac{1}{2}(A+iB)^T \sum_{k=1}^{2n+3} e^{i(n-2(p-1)+q-1)k\mu}$$

$$= \frac{1}{2}(A-iB)^T \sum_{k=1}^{2n+3} u^k + \frac{1}{2}(A+iB)^T \sum_{k=1}^{2n+3} v^k \tag{70}$$

where $u = e^{i(n-2(p-1)+q+1)\mu}$ and $v = e^{i(n-2(p-1)+q-1)\mu}$.

One has that $u \neq 1$ because $q \neq 2(p-1) - n - 1 \pmod{2n+3}$ and $u^{2k+3} = 1$ because $\mu = \frac{2\pi}{2n+3}$. So u is a non trivial root of unity. Therefore

$$\sum_{k=1}^{2n+3} u^k = u \sum_{k=0}^{2n+2} u^k = u \frac{1-u^{2n+3}}{1-u} = 0.$$

Similarly $\sum_{k=1}^{2n+3} v^k = 0$. So the result of (70) is zero.
• Second case: $q = 2(p-1) - n - 1 \pmod{2n+3}$. Then $u = 1$ in in (70), with v still a non trivial root of unity. So the result of (70) is $\frac{1}{2}(2n+3)(A - iB)^T$.
• Third case: $q = 2(p-1) - n + 1 \pmod{2n+3}$. Then $v = 1$ in in (70), with u a non trivial root of unity. So the result of (70) is $\frac{1}{2}(2n+3)(A + iB)^T$. □

Let us consider a block-line with index $1 \leq p \leq n+1$ of the matrix N_3. In view of the three cases of Lemma 4.5, the sub-block $N_{3,pq}$ is non zero for exactly two values of the index of the block-column $1 \leq q \leq 2n+3$. It strongly suggests to reorder the matrix N_3 by permutations.

That is why we permute the block-columns of M_3 so that the $n+1$ different values of q for which the first condition of Lemma 4.5 holds

$$q \in \{-n, -n+2, \ldots, n-2, n\}_{\text{mod } 2n+3}.$$

are ordered first. The other values

$$q \notin \{-n, -n+2, \ldots, n-2, n\}_{\text{mod } 2n+3}$$

are ordered after.

After the permutations of the block-lines, we also perform a permutation of the block columns so that all $N^e_{3,pq}$ show up on top of all $N^o_{3,pq}$.

These two permutations of the matrix N_3 can be characterized with 2 real non singular permutations matrices

$$P_1 \in \mathbb{R}^a, \quad P_1^T P_1 = I_a, \quad a = (n+1)m_o$$

and

$$P_2 \in \mathbb{R}^b, \quad P_2^T P_2 = I_b, \quad b = (2n+3)m_e.$$

It sets a new matrix

$$N_4 = P_1 N_3 P_2$$

It has the structure

$$N_4 = \begin{pmatrix} N_4^{11} & N_4^{12} \\ 0 & N_4^{22} \end{pmatrix} \in \mathbb{C}^{a \times b} \tag{71}$$

where the global structure is given by

$$\begin{array}{rll} N_4^{11} = \left(Y_{p,-n+2(q-1)}\right)_{1 \leq p,q \leq n+1} & \in \mathbb{C}^{c \times c}, & c = (n+1) \times m_e, \\ N_4^{12} & \in \mathbb{C}^{c \times d}, & d = (n+2) \times m_e, \\ 0 & \in \mathbb{C}^{e \times c}, & e = (n+1)m_o, \\ N_4^{22} & \in \mathbb{C}^{e \times d}. & \end{array} \tag{72}$$

One again the same is unchanged

$$\text{rank}(N_4) = \text{rank}(N).$$

The point is of course that the diagonal structure of N_4 simplifies the study of its rank since it is sufficient to study the rank of the diagonal blocks. The top left block N_4^{11} being a square matrix, it is not difficult to show it is invertible. On the contrary the bottom right block N_4^{22} is still rectangular, so its study will be a little more involved.

Lemma 4.6. *The square matrix N_4^{11} is invertible with rank$(N_4^{11}) = (n+1)m_e$.*

Proof. Indeed (69) implies that

$$N_{4,pq}^{11} = -\sigma_t \sum_{k=1}^{2n+3} e^{i2(q-p)k\mu} U_e(k\mu) Q U_e(-k\mu) \in \mathbb{C}^{m_e \times m_e}.$$

Take $X = (X_1, \dots, X_{n+1}) \in \mathbb{C}^{(n+1)m_e}$. With the standard notation for the sesquilinear product in complex algebra, one has

$$(N_4^{11} X, X) = \sum_{p=1}^{n+1}\sum_{q=1}^{n+1} (N_{4,pq}^{11} X_p, X_q) = -\sigma_t \sum_{p=1}^{n+1}\sum_{q=1}^{n+1}\sum_{k=1}^{2n+3} e^{i2(q-p)k\mu} \left(U_e(k\mu) Q U_e(-k\mu) X_p, X_q\right)$$

$$= -\sigma_t \sum_{k=1}^{2n+3} \left(U_e(k\mu) Q U_e(-k\mu) \left(\sum_{p=1}^{n+1} e^{-i2pk\mu} X_p \right), \left(\sum_{q=1}^{n+1} e^{-i2qk\mu} X_q \right) \right)$$

$$= -\sigma_t \sum_{k=1}^{2n+3} \left(Q U_e(-k\mu) \left(\sum_{p=1}^{n+1} e^{-i2pk\mu} X_p \right), U_e(-k\mu) \left(\sum_{p=1}^{n+1} e^{-i2pk\mu} X_q \right) \right).$$

The matrix Q is hermitian positive by construction (63), that is $Q = Q^* > 0$. Take $X \in \ker(N_4^{11})$. Then $\sum_{p=1}^{n+1} e^{-i2pk\mu} X_p = 0$ for all $1 \le k \le 2n+3$. To end the proof, let us take an integer $1 \le p' \le 2n+3$. If $p' = 2p-1$ is odd, then we set $Y_{p'} = 0$. If $p' = 2p$ is even, then we set $Y_{p'} = X_p$. So one can write $\sum_{p'=1}^{2n+3} e^{-ip'k\mu} Y_{p'} = 0$. A discrete Fourier transform yields $Y_{p'} = 0$ for all p'. So $X = 0$. More generally $\ker(N_4^{11}) = \{0\}$, so the claim is proved. □

Let us now turn to the remaining rectangular bottom right block.

Lemma 4.7. *By construction, the rectangular matrix $N_4^{22} \in \mathbb{C}^{e \times d}$ has a sparse explicit structure*

$$N_4^{22} = \frac{2n+3}{2} \begin{pmatrix} A^T - iB^T & A^T + iB^T & 0 & \dots & 0 & 0 \\ 0 & A^T - iB^T & A^T + iB^T & \dots & 0 & \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & A^T - iB^T & A^T + iB^T & 0 \\ 0 & 0 & \dots & \dots & A^T - iB^T & A^T + iB^T \end{pmatrix}$$

where the number of (block)lines is $e = (n+1)m_o$ and the number of (block)columns is $d = (n+2)m_e$.

Proof. Consequence of Lemma 4.5 and the definition of the permutation matrices P_1 and P_2. □

Let us define the space $Y \subset \mathbb{C}^{m_e}$

$$Y = \bigcap_{p=1}^{n+2} \text{range}\left(\cos\mu_p A^T - \sin\mu_p B^T\right). \tag{73}$$

It will appear that this space is equal the the space $\mathcal{A}(N, g)$ previously defined in (28) and it is the keystone of the proof.

Lemma 4.8. *One has rank$(N_4^{22}) = (n+2)m_e - dim(Y)$*

Proof. • Since N_4^{22} goes from $\mathbb{C}^{(n+2)m_e}$ into $\mathbb{C}^{(n+1)m_o}$, the claim is proved provided that $\ker(N_4^{22}) = \dim(Y)$. Then the claim will follow by the rank Theorem since

$$(n+2)m_e = \text{rank}(N_4^{22}) + \dim\ker(N_4^{22}) = \text{rank}(N_4^{22}) + \dim(Y),$$

• So let us study the kernel $\ker(N_4^{22})$. Take $X = (x_1, \dots, x_{n+2})^t \in \ker(N_4^{22})$, with $x_q \in \mathbb{C}^{m_e}$ for $q = 1, \dots, n+2$. A discrete Fourier transform is performed with respect to the index q

$$\begin{pmatrix} x_1 \\ \dots \\ x_q \\ \dots \\ x_{n+2} \end{pmatrix} = \sum_{p=1}^{n+2} \begin{pmatrix} e^{i2\pi p\frac{1}{n+2}}\alpha_p \\ \dots \\ e^{i2\pi p\frac{q}{n+2}}\alpha_p \\ \dots \\ e^{i2\pi p\frac{n+2}{n+2}}\alpha_p \end{pmatrix}, \quad \alpha_q \in \mathbb{C}^{m_e},\ 1 \le q \le n+2. \tag{74}$$

Then

$$0 = N_4^{22}X = \frac{2n+3}{2}\sum_{p=1}^{n+2} \begin{pmatrix} e^{i2\pi p\frac{1}{n+2}}\left((A^T - iB^T) + e^{i2\pi p\frac{1}{n+2}}(A^T + iB^T)\right)\alpha_p \\ \dots \\ e^{i2\pi p\frac{q}{n+2}}\left((A^T - iB^T) + e^{i2\pi p\frac{1}{n+2}}(A^T + iB^T)\right)\alpha_p \\ \dots \\ e^{i2\pi p\frac{n+1}{n+2}}\left((A^T - iB^T) + e^{i2\pi p\frac{1}{n+2}}(A^T + iB^T)\right)\alpha_p \end{pmatrix} \in \mathbb{C}^{(n+1)m_o}. \tag{75}$$

Still a consequence of the rectangular structure, one cannot directly perform an inverse Fourier transform because the vector is made of only $n+1$ vectors of size m_o. For a Fourier technique, the coefficients must consider $e^{i2\pi p\frac{n+2}{n+2}}$ which is not in (75). One more quantity is needed. So a possibility is to add one line in (75) and to write for some unknown vector $z \in \mathbb{C}^{m_o}$

$$\frac{2n+3}{2}\sum_{p=1}^{n+2} \begin{pmatrix} e^{i2\pi p\frac{1}{n+2}}\left((A^T - iB^T) + e^{i2\pi p\frac{1}{n+2}}(A^T + iB^T)\right)\alpha_p \\ \dots \\ e^{i2\pi p\frac{q}{n+2}}\left((A^T - iB^T) + e^{i2\pi p\frac{1}{n+2}}(A^T + iB^T)\right)\alpha_p \\ \dots \\ e^{i2\pi p\frac{n+1}{n+2}}\left((A^T - iB^T) + e^{i2\pi p\frac{1}{n+2}}(A^T + iB^T)\right)\alpha_p \\ e^{i2\pi p\frac{n+2}{n+2}}\left((A^T - iB^T) + e^{i2\pi p\frac{1}{n+2}}(A^T + iB^T)\right)\alpha_p \end{pmatrix} = \begin{pmatrix} 0 \\ \dots \\ 0 \\ \dots \\ 0 \\ z \end{pmatrix} \in \mathbb{C}^{(n+2)m_e}. \tag{76}$$

Now one can perform a discrete inverse Fourier transform more easily. One obtains

$$\frac{(2n+3)(n+2)}{2}\left((A^T - iB^T) + e^{i2\pi p\frac{1}{n+2}}(A^T + iB^T)\right)\alpha_p = z, \quad 1 \le p \le n+2. \tag{77}$$

It is actually evident by direct insertion that (77) is the solution of (76). That is

$$\frac{(2n+3)(n+2)e^{i\pi p\frac{1}{n+2}}}{2}\left(\cos\mu_p A^T - \sin\mu_p B^T\right)\alpha_p = z, \quad \mu_p = \pi p\frac{1}{n+2}, \quad 1 \le p \le n+2. \tag{78}$$

By definition $z \in \text{range}\left(\cos\mu_p A^T - \sin\mu_p B^T\right)$. Therefore $z \in Y = \bigcap_{p=1}^{n+2}\text{range}\left(\cos\mu_p A^T - \sin\mu_p B^T\right)$. In summary of the construction of this paragraph, all $X \in \ker(N_4^{22})$ generate a $z \in Y$.

• Reciprocally, take $z \in Y$ so that there exists at least one family $(\alpha_p)_{1\le p\le n+2}$ which solves (78). For this family, thanks to the Fourier formula (74) one $X \in \ker(N_4^{22})$. Let us show that X is actually unique.

Consider two different families α'_p and α''_p for $1 \le p \le n+2$, both satisfying (78) for the same $z \in Y$. Then $\alpha_p = \alpha'_p - \alpha''_p$ satisfies (78) for $z = 0$. By (64), one has $\left(\cos\mu_p A^T - \sin\mu_p B^T\right) = U_o(-\mu_p)A^T U_e(\mu_p)$. So

$$A^T U_e(\mu_p)z = 0 \Longleftrightarrow (AA^T)U_e(\mu_p)\alpha_p = 0.$$

By Lemma (2.2), the matrix is non singular so $U_e(\mu_p)\alpha_p = 0$ and $\alpha_p = 0$ for $1 \le p \le n+2$. Therefore the solution $(\alpha_p)_{1\le p\le n+2}$ to (78) is unique, which in turn yields that $X \in \ker(N_4^{22})$ is unique.

• So Y is in bijection with $\ker(M_4^{22})$ and their dimensions are equal. The proof of the claim is ended. □

4.3 The space Y

Lemma 4.9. $Y = \mathcal{A}(N, n+2)$.

Proof. One has that range $\left(\cos \mu_p A^T - \sin \mu_p B^T\right) = [\ker (\cos \mu_p A - \sin \mu_p B)]^\perp$. The rotational invariance identities (22) yield $\cos \mu_p A - \sin \mu_p B = U_e(-\mu_p) A U_o(\mu_p)$ so

$$\ker (\cos \mu_p A - \sin \mu_p B) = U_o(-\mu_p) \ker(A).$$

Finally $U_o(-\mu_p)$ is a unitary matrix, so

$$\ker (\cos \mu_p A - \sin \mu_p B)^\perp = U_o(-\mu_p) \ker(A)^\perp$$

Therefore one has the equivalent definition of the space

$$Y = \bigcap_{p=1}^{n+2} U_o(-\mu_p) \left(\ker(A)^\perp\right). \tag{79}$$

The μ_p are defined in (78), that is $\mu_p = \pi \frac{p}{n+2}$ for $1 \leq p \leq n+2$. Comparison of (27-28) and (79) shows that $Y = \mathcal{A}(N, g)$ with $g = n+2$. □

Lemma 4.10. *Take $n \geq N-1$. Then $Y = \{0\}$.*

Proof. The claim follows from Lemma 2.5. □

Proposition 4.11. *Take $n \geq N-1$. One has rank(N_4) = rank(M_n) = dim$(W_n) = (2n+3)m_e$.*

Proof. By Lemmas 4.6 and 4.8, one gets that $\operatorname{rank}(N_4) = (n+1)m_e + (n+2)m_e = (2n+3)m_e$. By Lemma 4.2 and 4.4, one gets

$$(2n+3)m_e = \operatorname{rank}(N_4) = \operatorname{rank}(M_n) \leq \dim(W_n) \leq (2n+3)m_e.$$

So all inequalities are equalities. □

Proposition 4.12. *Take $n = 0$. One has rank(N_4) = rank(M_n) = dim$(W_n) = m$.*

Proof. By Lemma 2.7 one has that $\dim Y = \dim (\mathcal{A}(N, 2)) = 2m_e - m_o$. So

$$m_e + 2m_e - (2m_e - m_o) = m_e + m_o = m = \operatorname{rank}(N_4) = \operatorname{rank}(M_n)$$

By Lemma 4.2 and 4.4 one has

$$\operatorname{rank}(M_n) \leq \dim(W_n) \leq m_e + m_o = m.$$

So all inequalities are equalities. □

Proposition 4.13. *Take $N = 3$ and $n = 1$. One has rank$(M_n) = 19$.*

Proof. For $N = 3$, then $m_e = 4$ and $m_o = 6$, ad also $Y = \mathcal{A}(3, 3)$. By Lemma 2.8 $\operatorname{rank}(N_4) = \operatorname{rank}(M_n) = (2n+3)m_e - \dim Y = 20 - 1 = 19$. □

4.4 Final proofs of Theorem 1.1 and Theorem 1.2

Proof of Theorem 1.1. By Lemmas 4.11 and 4.12, one gets that the columns vectors of M_n (which all belong to V_n) span a space which has the same dimension as V_n. So the fundamental property 3.12 is established. Then the bounds (52-53) are inserted in (45), and the proof is ended. □

Proof of Theorem 1.2. One considers $N = 3$ and $n = 1$. On the one hand one has $\operatorname{rank} M_1 = 19$ by proposition 4.12. On the other hand one has the upper bound of Lemma 4.2 with $m_e = 4$ and $m_o = 6$

$$\dim V_1 \leq 2m_e + \min(3m_e, 2m_o) = 8 + \min(12, 12) = 20.$$

There is a mismatch because $19 < 20$, so one cannot conclude without a sharper upper bound. At inspection, it appears that the bound $\dim V_1 \leq 20$ is not optimal. We now show that a sharper bound is possible. We use a method similar to the one of Lemma 4.3 which is by direct examination of the relations (47) for $p + q = n = 1$. Of course, this is possible only because the dimension is low.

One has $\dim V_1 = \dim W_1 \leq 8 + \operatorname{rank} N$ where the square matrix $N = \begin{pmatrix} 2A^T & B^T & 0 \\ 0 & A^T & 2B^T \end{pmatrix} \in \mathcal{M}_{12}(\mathbb{R})$ come from (47) for $p + q = 1$ and the sub-matrices are given by (19-20). One has $\operatorname{rank} N = \operatorname{rank} N^T = 12 - \dim \ker N^T$. Since $N^T = \begin{pmatrix} 2A & 0 \\ B & A \\ 0 & 2B \end{pmatrix}$, then $\ker N^T$ is made of vectors $(\beta, \gamma) \in \mathbb{R}^6 \times \mathbb{R}^6$ such that

$$\beta \in \ker A, \quad \gamma \in \ker B, \quad B\beta + A\gamma = 0.$$

With natural notations, one has from (19-20)

$$\begin{aligned} &\beta \in \ker A \Longleftrightarrow \beta_2 = \beta_5 = \beta_6 = \tfrac{1}{\sqrt{5}}\beta_1 + \sqrt{\tfrac{3}{14}}\beta_3 - \tfrac{1}{\sqrt{70}}\beta_4 = 0, \\ &\gamma \in \ker B \Longleftrightarrow \gamma_1 = \gamma_3 = \gamma_4 = \tfrac{1}{\sqrt{5}}\gamma_2 - \tfrac{1}{\sqrt{70}}\gamma_5 - \sqrt{\tfrac{3}{14}}\gamma_6 = 0. \end{aligned} \tag{80}$$

It yields 8 linearly independent linear constraints. Inspection of the equation $B\beta + A\gamma = 0$ shows that the second line vanishes identically because of (80). It shows that one of the four linear equations in $B\beta + A\gamma = 0$ is redundant with the ones in (80). So $\dim \ker N^T \geq 1$. Therefore $\operatorname{rank} N \leq 11$ and finally $\dim V_1 \leq 8 + 11 = 19$. One gets the equality with $\dim M_1 = \dim V_1 = 19$. So the fundamental property is proved and the proof is ended. □

5 Numerical illustrations

In this section we show that the theoretical estimates of convergence are observed in TDG calculations. We consider the TDG method with the solutions (7) for the particular cases of the P_1 and P_3 models see [6, Section 4] or [26, Chapter 5] for explicit formula of these solutions. In the following the scheme is tested with equi-distributed directions starting with $\mathbf{d}_1 = (1, 0)$.

We consider the stationary P_1 model for which $m_e = 1$. Let $\mathbf{x} = (x, y)^T, \Omega = [0, 1]^2$, $\sigma_a = 1/\sqrt{3}, \sigma_s = 1/\sqrt{3}$. The exact solution we consider here is

$$\mathbf{u}_{ex}(\mathbf{x}) = \left(\cos(y)e^{\sqrt{3}x}, -(\sqrt{3}/2)\cos(y)e^{\sqrt{3}x}, 0.5\sin(y)e^{\sqrt{3}x} \right)^T.$$

Results obtained with $n = 0$, 1 and 2, that is with $2n + 3 = 3$, 5 and 7 basis functions are displayed on the left of Figure 2. As stated in Theorem 1.1 for the particular case $N = 1$, one only needs two additional basis functions to increase the order by a factor 1. Note however that the orders obtained here are slightly better than those predicted in Theorem 1.1: with $n = 3$, 5 and 7 basis functions, one gets respectively order 0.8, 1.5 and 2.5.

We also consider the stationary P_3 model for which $m_e = 4$. Let $\mathbf{x} = (x, y)^T, \Omega = [0, 1]^2$, $\sigma_a = 0.2, \sigma_s = 0.3$. The exact solution we consider is taken from the solution (7) and has for eigenvalue

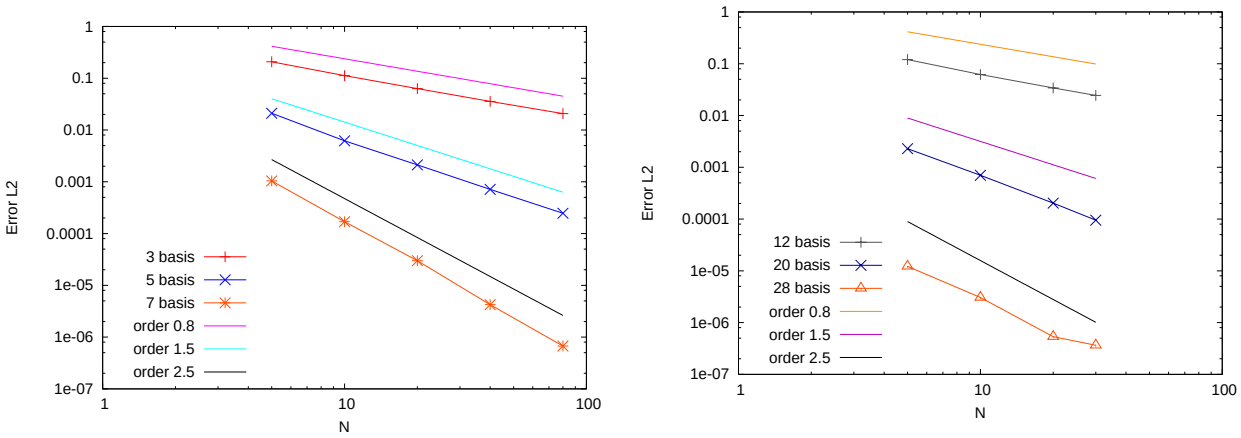

Figure 2: Order depending on the number of basis functions. On the left P_1 model and on the right P_3 model. L^2 error in logarithmic scale and random meshes. Here the number N on the horizontal axis refers to the number of cells. More precisely $h \approx 1/N$ and the numbers of cells in the TDG simulation scales as $\#\text{cells} \approx 1/h^2 \approx N^2$.

$\sqrt{7}/\sqrt{3}$ with a direction $\mathbf{d} = (\cos \pi/4, \sin \pi/4)^T$. Of course this solution does not belong to our basis functions.
Results obtained with $n = 0$, 1 and 2, that is $3, 5$ and 7 directions (for a total of 12, 20 and 28 basis functions) are displayed on the right of Figure 2. The order for $n = 2$ and $n = 0$ recovers the estimate Theorem 1.1. For $n = 1$, we also observe convergence at the rate predicted by the second main Theorem 1.2. We remind the reader that the proof of this Theorem was obtained in Section 4.4 by means of refinement of the method used for the general case.
Note that the tests for the P_3 model are displayed on much coarser meshes than for the P_1 model. This comes from the bad condition number of the matrix which is a well known drawback of the TDG method [8, 18, 26] and occurs when increasing the number of basis functions on fine meshes. Since we do not want the condition number to interfere with the error study we choose not to refine the meshes too much. Still, the bad conditioning of the matrix can probably be seen on the last point of the curve representing 28 basis functions which is not completely aligned with the other points. Using better preconditionner could solve this issue.

6 Conclusions

The core techniques, developped in this work for the analysis of the h-convergence of the vectorial exponential functions adapted to the P_N model, rely on the matrix M_n. It appears that this matrix is rectangular in the general case. This sole fact generates most of the technical mathematical difficulties that we have encountered. It must be emphasized on that the same methodology for the Helmholtz equation [8, 24] is much more simple due to the fact that the matrices are square Vandermonde matrices. It is also the case for the P_1 model with the even-odd simplification studied in one of our previous work [27]. In this work, we also use the properties of Vandermonde matrices, but in a much more indirect way.

With respect to the results in [26], our results are more powerful in the sense that we use the minimal number of basis functions and the proof does not rely on the Bezout theorem for roots of systems of multivariate polynomial equations. Nevertheless there is the restriction: the directions of

the vectorial exponential functions must be equi-distributed, which is not the case in the more general situation considered in [26].

Equi-distributed directions which are natural in two-space dimension do not generalize in three-space dimensions. This fundamental three-dimensional difficulty has been also encountered in [9] for the approximation of Maxwell's equations and in [24] for the approximation of the Helmholtz equation. It is possible also that the approach developped in [24] for three-space dimension could be applied to show h-convergence in two-space dimension for non equi-distributed directions almost everywhere in the space of admissible angles.

An open problem remains at the end of this study, which is to obtain optimal results in the gap $0 < n < N-1$ for general $N \in 2\mathbb{N}+1$. For $N = 1$ the gap is empty, and for $N = 3$ the gap is covered by Theorem 1.2. To cover general $N \geq 5$ will require the development of new techniques for the analysis of the rank of M_n where the rank is stricly less than the number of rows and the number of columns of the matrix.

A Spherical harmonics

A.1 Legendre functions

The spherical harmonics are based on the Legendre functions P_k^l which read

$$P_k^m(\mu) = \begin{cases} \dfrac{1}{2^k k!}(1-\mu^2)^{m/2}\dfrac{d^{k+m}}{d\mu^{k+m}}((\mu^2-1)^k), & m \geq 0, \\ (-1)^m \dfrac{(k+m)!}{(k-m)!} P_k^{-m}(\mu), & m < 0. \end{cases} \tag{81}$$

The Legendre polynomials satisfy orthogonal relations such as $\frac{1}{2}\int_{-1}^{1} P_k^0 d\mu = \delta_{k,0}$ and $\frac{1}{2}\int_{-1}^{1} P_k^m P_{k'}^m d\mu = \frac{1}{(a_k^m)^2}\delta_{k,k'}$ where the normalization factor is $a_k^m = \sqrt{(2k+1)\frac{(k-m)!}{(k+m)!}}$. They also satisfy recursion relations which are very useful to construct the matrices of the P_N model

$$\begin{cases} \sqrt{1-\mu^2}P_k^m = \frac{1}{2k+1}\Big(P_{k+1}^{m+1} - P_{k-1}^{m+1}\Big), \\ \sqrt{1-\mu^2}P_k^m = \frac{1}{2k+1}\Big(-(k-m+1)(k-m+2)P_{k+1}^{m-1} + (k+m-1)(k+m)P_{k-1}^{m-1}\Big), \\ \mu P_k^m = \frac{1}{2k+1}\Big((k-m+1)P_{k+1}^m + (k+m)P_{k-1}^m\Big). \end{cases}$$

A.2 Real spherical harmonics

The **complex** spherical harmonics are $\widehat{X}_k^m(\psi,\phi) := (-1)^m a_k^m P_k^m(\cos\phi)e^{\mathbf{i}m\psi}$ for $|m| \leq k$, where the pure imaginary number is $\mathbf{i}^2 = -1$. Let us note $\mu = \cos\phi$. The **real** spherical harmonics X_k^m are

$$. \begin{cases} X_k^m(\psi,\mu) = a_k^m P_k^m(\mu), & m = 0, \\ X_k^m(\psi,\mu) = a_k^m\sqrt{2}\cos(m\psi)P_k^m(\mu), & 0 < m \leq k, \\ X_k^m(\psi,\mu) = a_k^{|m|}\sqrt{2}\sin(|m|\psi)P_k^{|m|}(\mu), & -k \leq m < 0. \end{cases} \tag{82}$$

They satisfy the recursion relations

$$\begin{cases} \cos\psi\sqrt{1-\mu^2}X_k^m &= \varepsilon^m(A_k^m X_{k+1,m+1} - B_k^m X_{k-1}^{m+1}) - \zeta^m(C_k^m X_{k+1}^{m-1} - D_k^m X_{k-1}^{m-1}), \\ \sin\psi\sqrt{1-\mu^2}X_k^m &= \eta^m(A_k^m X_{k+1}^{-m-1} - B_k^m X_{k-1}^{-m-1}) + \phi^m(C_k^m X_{k+1}^{-m+1} - D_k^m X_{k-1}^{-m+1}), \\ \mu X_k^m &= E_k^m X_{k+1}^m + F_{k,m} X_{k-1}^m, \end{cases} \tag{83}$$

where

$$\begin{cases} A_k^m = \sqrt{\dfrac{(k+m+1)(k+m+2)}{(2k+1)(2k+3)}}, & B_k^m = \sqrt{\dfrac{(k-m-1)(k-m)}{(2k-1)(2k+1)}}, \\ C_k^m = \sqrt{\dfrac{(k-m+1)(k-m+2)}{(2k+1)(2k+3)}}, & D_k^m = \sqrt{\dfrac{(k+m-1)(k+m)}{(2k-1)(2k+1)}}, \\ E_k^m = \sqrt{\dfrac{(k-m+1)(k+m+1)}{(2k+1)(2k+3)}}, & F_k^m = \sqrt{\dfrac{(k-m)(k+m)}{(2k-1)(2k+1)}}, \end{cases}$$

and the other coefficients are given in Table 1.

	$m<-1$	$m=-1$	$m=0$	$m=1$	$m>1$
ε^m	$-\frac{1}{2}$	0	$\frac{\sqrt{2}}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
ζ^m	$-\frac{1}{2}$	$-\frac{1}{2}$	0	$\frac{\sqrt{2}}{2}$	$\frac{1}{2}$
η^m	$-\frac{1}{2}$	$-\frac{\sqrt{2}}{2}$	$\frac{\sqrt{2}}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
ϕ^m	$-\frac{1}{2}$	$-\frac{1}{2}$	0	0	$\frac{1}{2}$

Table 1: Coefficients of the equations (83)

They also satisfy orthogonality relations like $\frac{1}{4\pi}\int_{S^2} X_k^m d\psi d\mu = \delta_{k,0}\delta_{m,0}$ and $\frac{1}{4\pi}\int_{S^2} X_k^m X_{k',m'} d\psi d\mu = \delta_{k,k'}\delta_{m,m'}$.

References

[1] A. V. Avvakumov, V. F. Strizhov, P. N. Vabishchevich, and A. O. Vasilev, *Numerical modeling of neutron transport in sp3 approximation by finite element method, arxiv*, 2019, Arxiv: 1903.11502v1.

[2] Y. Azmy and E. Sartori, *Nuclear computational science: a century in review*, Springer, 2010.

[3] G. Bell and G. S., *Nuclear Reactor theory*, Van Nostrand Reinhold Company, 1970.

[4] S. C. Brenner and L. Scott, *The mathematical theory of finite element methods. 3rd ed.*, New York, NY: Springer, 3rd ed. ed., 2008.

[5] C. Buet, B. Després, and G. Morel, *Discretization of the pn model for 2d transport of particles with a trefftz discontinuous galerkin method, hal preprint*, 2019, https://hal.sorbonne-universite.fr/hal-02372279/document.

[6] C. Buet, B. Despres, and G. Morel, *Trefftz Discontinuous Galerkin basis functions for a class of Friedrichs systems coming from linear transport*, ACOM, 4 (2020).

[7] A. Buffa and P. Monk, *Error estimates for the Ultra Weak Variational Formulation of the Helmholtz equation*, ESAIM: Mathematical Modelling and Numerical Analysis, 42 (2008), pp. 925–940.

[8] O. Cessenat and B. Després, *Application of an Ultra Weak Variational Formulation of Elliptic PDEs to the Two-Dimensional Helmholtz Problem*, SIAM J. Numer. Anal., 35 (1998).

[9] O. Cessenat and B. Després, *Using plane waves as base functions for solving time harmonic equations with the ultra weak variational formulation*, vol. 11, 2003, pp. 227–238. Medium-frequency acoustics.

[10] M. M. Crockatt, A. J. Christlieb, C. K. Garrett, and C. D. Hauck, *Hybrid methods for radiation transport using diagonally implicit Runge-Kutta and space-time discontinuous Galerkin time integration*, J. Comput. Phys., 376 (2019), pp. 455–477.

[11] A. Ern and J. Guermond, *Discontinuous galerkin methods for friedrichs' systems. i. general theory*, SIAM J. Numerical Analysis, 44 (2006), pp. 753–778.

[12] C. J. Gittelson, R. Hiptmair, and I. Perugia, *Plane wave discontinuous Galerkin methods: Analysis of the h-version.*, ESAIM, Math. Model. Numer. Anal., 43 (2009), pp. 297–331.

[13] J.-L. Guermond and G. Kanschat, *Asymptotic analysis of upwind discontinuous Galerkin approximation of the radiative transport equation in the diffusive limit.*, SIAM J. Numer. Anal., 48 (2010), pp. 53–78.

[14] V. Heningburg and C. D. Hauck, *A hybrid finite-volume, discontinuous Galerkin discretization for the radiative transport equation*, Multiscale Model. Simul., 19 (2021), pp. 1–24.

[15] F. Hermeline, *A discretization of the multigroup P_N radiative transfer equation on general meshes.*, J. Comput. Phys., 313 (2016), pp. 549–582.

[16] R. Hiptmair, A. Moiola, and I. Perugia, *Plane wave discontinuous Galerkin methods for the 2D Helmholtz equation: analysis of the p-version*, SIAM J. Numer. Anal., 49 (2011), pp. 264–284.

[17] R. Hiptmair, A. Moiola, and I. Perugia, *Plane wave discontinuous Galerkin methods: exponential convergence of the hp-version.*, Found. Comput. Math., 16 (2016), pp. 637–675.

[18] T. Huttunen, P. Monk, and J. P. Kaipio, *Computational aspects of the ultra-weak variational formulation.*, J. Comput. Phys., 182 (2002), pp. 27–46.

[19] L.-M. Imbert-Gérard, *Interpolation properties of generalized plane waves*, Numer. Math., 131 (2015), pp. 683–711.

[20] L.-M. Imbert-Gérard and B. Després, *A generalized plane-wave numerical method for smooth nonconstant coefficients*, IMA J. Numer. Anal., 34 (2014), pp. 1072–1103.

[21] F. Kretzschmar, A. Moiola, I. Perugia, and S. M. Schnepp, *A priori error analysis of space–time Trefftz discontinuous Galerkin methods for wave problems* , IMA Journal of Numerical Analysis, 36 (2016), p. 1599.

[22] R. G. McClarren, *Theoretical aspects of the simplified pn equations*, Transport Theory and Statistical Physics, (2010), pp. 73–109.

[23] D. Mihalas and B. W. Mihalas, *Foundations of radiation hydrodynamics*, Oxford University Press, New York, 1984.

[24] A. Moiola, R. Hiptmair, and I. Perugia, *Plane wave approximation of homogeneous Helmholtz solutions*, Z. Angew. Math. Phys., 62 (2011), pp. 809–837.

[25] P. Monk and G. R. Richter, *A discontinuous Galerkin method for linear symmetric hyperbolic systems in inhomogeneous media.*, J. Sci. Comput., 22-23 (2005), pp. 443–477.

[26] G. Morel, *Asymptotic-preserving and well-balanced schemes for transport models using Trefftz discontinuous Galerkin method*, theses, Sorbonne Université, Sept. 2018, https://hal.archives-ouvertes.fr/tel-01911872.

[27] G. Morel, C. Buet, and B. Després, *Trefftz discontinuous Galerkin method for Friedrichs systems with linear relaxation: application to the P_1 model*, Computational Methods in Applied Mathematics, (2018), pp. 521–557.

[28] G. C. Pomraning, *The equations of radiation hydrodynamics*, International Series of Monographs in Natural Philosophy, Oxford: Pergamon Press, 1973.

[29] O. Zienkiewicz, *Origins, milestones and directions of the finite element method? a personal view*, Archives of Computational Methods in Engineering, (1995), pp. 1–48.