Trefftz discontinuous Galerkin method for Friedrichs systems with linear relaxation: application to the P 1 model

Guillaume Morel, Christophe Buet, Bruno Després

Trefftz discontinuous Galerkin method for Friedrichs systems with linear relaxation: application to the P_1 model

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

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

[2] Sorbonne Universités, UPMC Univ Paris 06, UMR 7598, Laboratoire Jacques-Louis Lions, F-75005, Paris, France

[3] Institut Universitaire de France.

December 7, 2017

Abstract

This work deals with the first Trefftz Discontinuous Galerkin (TDG) scheme for a model problem of transport with relaxation. The model problem is written as a P_N or S_N model, and we study in more details the P_1 model in dimension 1 and 2. We show that TDG method provides natural well-balanced (WB) and asymptotic preserving (AP) discretization since exact solutions are used locally in the basis functions. High order convergence with respect to the mesh size in two dimensions is proved together with the asymptotic property for P_1 model in dimension one. Numerical results in dimensions 1 and 2 illustrate the theoretical properties.

1 Introduction

This work deals with the design and analysis of a new Trefftz Discontinuous Galerkin (TDG) method proposed for the P_N (spherical harmonic expansion) and S_N (discrete ordinate method) approximation of the transport equation of photons, neutrons or other types of particles

$$\partial_t I(t,\mathbf{x},\mathbf{\Omega}) + \mathbf{\Omega}\cdot\nabla I(t,\mathbf{x},\mathbf{\Omega}) = -\sigma_a(\mathbf{x})I(t,\mathbf{x},\mathbf{\Omega}) + \sigma_s(\mathbf{x})\left(|I| - I(t,\mathbf{x},\mathbf{\Omega})\right), \tag{1}$$

where I is the distribution function, t the time variable, $\mathbf{x}\in\mathbb{R}^d$ the space variable, $\mathbf{\Omega}$ the direction and $|I| = \frac{1}{4\pi}\int_{S^2} I(t,\mathbf{x},\mathbf{\Omega}')d\mathbf{\Omega}'$ is the mean of I. Absorption and scattering coefficient are denoted as

$$\sigma_a(\mathbf{x}) \geq 0 \text{ and } \sigma_s(\mathbf{x}) \geq 0.$$

We adopt the common strategy which is to use write the P_N and S_N reduced models [8, 17] in the form of Friedrichs system with relaxation, as in to (2).

Numerical approximation of the transport equation and related reduced models is challenging because of the two spatially dependent coefficients σ_a and σ_s. It is known that boundary layers may occur when σ_a, σ_s vary significantly and that the transport equation tends to a diffusion limit when σ_s is high. Standard schemes fail to correctly capture both of these two phenomena. To capture the diffusion limit with reasonable computational time, the idea of so called asymptotic preserving schemes has been introduced [26] and applied to transport problems, see [4, 25, 33] and reference therein. To capture boundary layers it may be a good idea to use well-balanced schemes which preserve, for example, the stationary states of the model (state of the art can be found in [15]). For recent works on boundary layers see for example [29, 35]. Schemes which are both asymptotic preserving and well balanced have been designed and studied in one dimension [16, 27]. However, direct extension in higher

[4]E-mail addresses: guillaume.morel.ocre@cea.fr, christophe.buet@cea.fr, despres@ann.jussieu.fr

dimensions may fail to capture boundary layer [34]. In general and except in some particular cases, two dimensional asymptotic preserving schemes are not well balanced.

The goal of this work is to discretize P_N and S_N models with TDG schemes which are both asymptotic preserving and well balanced (in a sense that will be defined later). We will restrict the study to homogeneous coefficients which may nevertheless be stiff. Given a system of partial differential equations (PDE), TDG method are discontinuous Galerkin type schemes that use solutions to the model as basis functions. The name comes from the seminal 1926 paper of E. Trefftz which has been recently translated in English [30]. Trefftz method has been widely used and studied for wave propagation problems [6, 7, 14, 18, 28] see also the review [20] and reference therein. TDG method have their pros and cons.

- **Pros**:
  - Incorporate a priori knowledge in the basis functions which are therefore well adapted to multiscale problems.
  - Often need less degrees of freedom to reach a given accuracy. A typical example for the 2D version of the P_1 model (3) in the dominant absorption regime $\sigma_a > 0$ (with $c = \varepsilon = 1$) is illustrated in the table below, where we compare the number p of basis functions needed to achieve a given fractional order. The first line is for our TDG method. One gets $p_{\text{TDG}} = 2(\text{order} + 1)$ which is a rephrasing of the result of proposition 7.12. The second line is the optimal number of basis function for a general DG method $p_{\text{DG}} = \frac{3}{2}(\text{order} + \frac{1}{2})(\text{order} + \frac{3}{2})$.

order	1/2	3/2	5/2	7/2	9/2
p_{TDG}	3	5	7	9	11
p_{DG}	3	9	18	30	45

    In particular the number of basis functions is the same to get order $= 1/2$. One always gets $p_{\text{TDG}} \leq p_{\text{DG}}$.
  - Is easy to incorporate in DG codes since one only needs to change the basis functions.
- **Cons**:
  - May suffer ill-conditioning due to poor linear independence of the basis functions [7, 21]. For wave problems, some remedies exist in the literature [14].
  - The practical calculation of the basis functions adds to the computational burden. If one can calculate the basis functions analytically, the computational burden is moderate. If it is not the case, the computational burden is heavier: several options could be consider such as computing numerically the basis functions or relying on the general procedure [23, 22, 24].

In this work we adapt the TDG formalism to a general first order PDE with linear relaxation which encompasses the P_N and S_N models with homogeneous coefficients. For first order PDE the adjoint equations may differ from the direct equations, and therefore one can construct two kinds of basis functions: using adjoint solutions or using direct solutions. It turns out that using adjoint solutions is not an efficient method in our case and we will therefore focus on TDG method with direct solutions. Another possibility is to adopt a Petrov-Galerkin approach choosing test functions as adjoint solutions and trial functions as direct solutions [12, 13]. However, we have noticed stability issues with this method for time dependent problem. Therefore the Petrov-Galerkin method will not be studied hereafter.

We will present the method in a general framework to consider both stationary and time dependent problems. Let Ω_S be a bounded polygonal/polyhedral Lipschitz space domain in $\mathbb{R}^d$ and consider a time interval $[0, T]$, $T > 0$. We denote $\Omega = \Omega_S$ for stationary problems and $\Omega = \Omega_S \times [0, T]$ for time dependent problems. We first apply the method to Friedrichs systems [11] with linear relaxation

$$\begin{cases} \sum_{i=0}^{d} A_i \partial_i \mathbf{u} = -R(\mathbf{x})\mathbf{u}, & \text{in } \Omega, \\ M^- \mathbf{u} = M^- \mathbf{g}, & \text{in } \partial\Omega, \end{cases} \tag{2}$$

the dependent variable is $\mathbf{u} \in \mathbb{R}^m$, $\mathbf{x} = (x_1, ..., x_d) \in \mathbb{R}^d$ is the space variable and t is the time variable. The coefficients σ_a and σ_s in (1) are contained in the relaxation matrix R. Recalling that the problem can be stationary or time dependent one may write $\mathbf{u}(t, \mathbf{x})$ or just $\mathbf{u}(\mathbf{x})$ depending on the situation. The matrices $A_i, R(\mathbf{x}) \in \mathbb{R}^{m\times m}$ are symmetric and we assume $R(\mathbf{x}) \in \mathbb{R}^{m\times m}$ is a non negative matrix, i.e. $(R(\mathbf{x})\mathbf{v}, \mathbf{v}) \geq 0$ for all $\mathbf{v} \in \mathbb{R}^m, \mathbf{x} \in \mathbb{R}^d$. We use the notation $\partial_0 = \partial_t$, $\partial_i = \partial_{x_i}$ for $i = 1, ..., d$ and we will therefore take $A_0 = I_m$ even if it is possible to consider more general non negative matrices for A_0. The outward normal unit vector is $\mathbf{n}(t, \mathbf{x}) = (n_t, n_{x_1}, ..., n_{x_d})$ for $x \in \partial\Omega$ and of course for stationary problems $n_t = 0$ for all $x \in \partial\Omega$. We set $M(\mathbf{n}) = A_0 n_t + \sum_{i=1}^d A_i n_{x_i}$, on $\partial\Omega$. Since M is symmetric one has the standard decomposition $M(\mathbf{n}) = M^+(\mathbf{n}) + M^-(\mathbf{n})$ where M^+ is a non negative matrix and M^- is a non positive matrix. We use the matrix M^- to write the boundary conditions with $\mathbf{g} \in L^2(\partial\Omega)$. Finally we assume the problem (2) admits a unique solution. A fundamental example of Friedrichs system in one dimension that we desire to treat is the P_1 model

$$\begin{cases} \partial_t p + \frac{c}{\varepsilon}\partial_x v = -\sigma_a p, \\ \partial_t v + \frac{c}{\varepsilon}\partial_x p = -(\sigma_a + \frac{\sigma_s}{\varepsilon^2})v, \end{cases} \tag{3}$$

here $1/\varepsilon$ represents the speed of light. The associated asymptotic model when $\varepsilon \to 0$ is

$$\begin{cases} \partial_t p - \frac{c^2}{\sigma_s}\partial_{xx} p = -\sigma_a p, \\ v = -\frac{c\varepsilon}{\sigma_s}\partial_x p. \end{cases}$$

One of our goal is to show that TDG method naturally captures those kinds of asymptotic regimes. To see if the scheme approaches correctly this one dimensional limit model we write the TDG method as a finite difference scheme. Under this form one can formally show that this scheme is asymptotic preserving and new compared to other popular one dimensional schemes [16]. The asymptotic result can be stated as follows (all the hypotheses needed to make the theorem rigorous are given in Section 4).

Proposition 1.1 (Time dependent $1D$ case)**.** *Assume $c = 1$, $\sigma_a = 0$. When $\varepsilon \to 0$ the formal limit of the scheme (33) with two basis functions in dimension one is an asymptotic scheme consistent with the P_1 model limit.*

The main convergence result about the stationary P_1 model in two dimensions can be stated as follows (all the hypotheses needed to make the theorem rigorous are given in Section 5).

Theorem 1.2 (Stationary 2D case)**.** *Assume $c = 1$, $\varepsilon = 1$ and $\sigma_a + \sigma_s > 0$ which is the general regime. Consider the stationary two dimensional P_1 model and a basis of $2n + 1$ shape functions (not necessarily equi-distributed). One has the h-convergence estimate*

$$\|\mathbf{u} - \mathbf{u}_h\|_{L^2(\Omega)} \leq Ch^{n-1}\|\mathbf{u}\|_{W^{n+1,\infty}(\Omega)},$$

where $\mathbf{u}$ stands for the exact solution and $\mathbf{u}_h$ for the approximate solution calculated by the TDG method.

For technical reasons, this L^2 convergence estimate in the general regime looses one half order of convergence compared with the one obtained in the absorption regime (Proposition 7.12). Nevertheless Theorem 1.2 clearly shows one of the well-known advantages of the TDG method compared to other more traditional schemes. Whereas the number of basis functions for the TDG method is linear with respect to the sought order, it becomes quadratic when considering, for example, the finite element method. The TDG method may therefore be computationally more efficient than the FEM at least in the $2D$ case. Moreover and as it is often the case with discontinuous Galerkin method, numerical results actually show better order of convergence than the one displayed in theorem 1.2. The estimate is sub-optimal since the error is measured in quadratic norm and the right hand side is measured in maximum norm. The convergence order $n - 1$ is the worst case allowed by the physical hypothesis

$\sigma_a+\sigma_s > 0$. The proof shows that it corresponds to vanishing absorption $\sigma_a = 0$ and positive scattering $\sigma_s > 0$, which results in vanishing damping of the first variable p, see (3). Therefore the main point of the proof in the general regime is to get L^2 control of the first variable p using the properties of the TDG method.

This paper is organized as follows: in Section 2 we present the TDG method for Friedrichs systems. Section 3 is devoted to the analysis the method, in particular we give in this section a quasi-optimality result and the well-balanced property of the scheme. Section 4 and 5 give some applications to the P_1 model in one and two dimensions. In Section 4, we focus on the one dimensional P_1 model, show how to construct the basis functions and study formally the asymptotic behavior of the scheme. In Section 5, we focus on the two dimensional P_1 model, show how to construct the basis functions. Numerical results are given in one and two dimensions in Section 6. In particular some numerical results bring evidence that TDG methods naturally capture internal boundary layers and so are well adapted to multiscale problems. The proof of the main Theorem 1.2 about the h convergence of the method for the stationary case is in Chapter 7. The appendix gathers various technical results.

2 Presentation of the method

All the vector will be noted in bold. For $\mathbf{v}(\mathbf{x}) \in \mathbb{R}^m$ we will also use the simplified notation $\mathbf{v} \in L^2(\Omega)$ instead of $\mathbf{v} \in L^2(\Omega)^m$. Moreover we may write $\mathbf{v} = (v_1, ..., v_m)^T$ where ^{T} denotes the transpose and denote $\mathbf{v}^2 = \mathbf{v}^T\mathbf{v}$ to facilitate the distinction with other types of norms or semi-norms.

2.1 Mesh notation and generic discontinuous Galerkin formulation

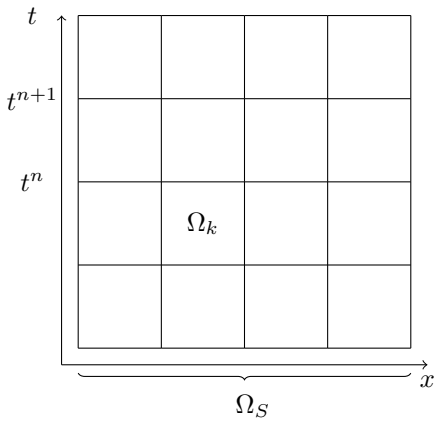


Figure 1: Illustration of the partition $\mathcal{T}_h$ for a time dependent problem.

The partition or mesh of the space domain $\Omega = \Omega_S \subset \mathbb{R}^d$ is denoted as $\mathcal{T}_h$. It is made of polyhedral non overlapping subdomains $\Omega_{S,r}$, that is $\mathcal{T}_h = \cup_r \Omega_{S,r}$. For a space time problem we first split the time interval into smaller time intervals (t_n, t_{n+1}) with $0 = t_0 < t_1 < ... < t_N = T$. Making an abuse of notation, the mesh of the space time domain $\Omega = \Omega_S \times [0, T] \subset \mathbb{R}^{d+1}$ is still denoted as $\mathcal{T}_h = \cup_{r,n} \Omega_{S,r} \times (t_n, t_{n+1})$. One must therefore be careful that $\mathcal{T}_h$ denotes either a purely spatial mesh for stationary models or a space-time mesh for time dependent models. Moreover the cells or subdomains will be referred to with the same notation, that is $\Omega_k = \Omega_{S,r}$ or $\Omega_k = \Omega_{S,r} \times (t_n, t_{n+1})$. In summary one can write in both cases $\mathcal{T}_h = \cup_k \Omega_k$ and the context makes these notations non ambiguous.

The broken Sobolev space is

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

In the following we assume $\mathbf{u} \in H^1(\mathcal{T}_h)$. For convenience we may rewrite (2) under the form $L\mathbf{u} = \mathbf{0}$ and consider also the adjoint operator

$$L = \sum_i A_i \partial_i + R, \quad L^* = -\sum_i A_i \partial_i + R.$$

All matrices are constant (do not depend either on the time variable or on the space variables). Multiplying (2) by $\mathbf{v} \in H^1(\mathcal{T}_h)$ and integrating on Ω gives

$$\sum_k \int_{\Omega_k} \mathbf{v}_k^T L \mathbf{u}_k = 0, \tag{4}$$

where $\mathbf{v}_k = \mathbf{v}_{|\Omega_k}$, $\mathbf{u}_k = \mathbf{u}_{|\Omega_k}$. Integrating by parts one gets

$$\sum_k \int_{\Omega_k} \left(L^* \mathbf{v}_k\right)^T \mathbf{u}_k + \sum_k \int_{\partial\Omega_k} \mathbf{v}_k^T M_k \mathbf{u}_k = 0,$$

where $\partial\Omega_k$ is the contour of the element Ω_k with an outward unit normal $\mathbf{n}_k = (n_t, n_{x_1}, ..., n_{x_d})^T$, $M = A_0 n_t + \sum_i A_i n_i$ and $M_k = M(\mathbf{n}_k)$. 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 even if there is more than one edge in $\Omega_k \cap \partial\Omega$), one can write

$$\begin{aligned}\sum_k \int_{\Omega_k} \left(L^* \mathbf{v}_k\right)^T \mathbf{u}_k + \sum_k \sum_{j<k} \int_{\Sigma_{kj}} (\mathbf{v}^T M \mathbf{u})_k + (\mathbf{v}^T M \mathbf{u})_j \\ + \sum_k \int_{\Sigma_{kk}} \mathbf{v}_k^T M_k^+ \mathbf{u}_k = -\sum_k \int_{\Sigma_{kk}} \mathbf{v}_k^T M_k^- \mathbf{g}.\end{aligned}$$

For $\mathbf{u}$ satisfying the equation (2), the normal flux is

$$M_k \mathbf{u}_k = -M_j \mathbf{u}_j = f_{kj}(\mathbf{u}_k, \mathbf{u}_j), \quad \text{on } \Sigma_{kj} \tag{5}$$

where f_{kj} is a numerical flux on Σ_{kj} defined below. One has

$$\sum_k \sum_{j<k} \int_{\Sigma_{kj}} (\mathbf{v}^T M \mathbf{u})_k + (\mathbf{v}^T M \mathbf{u})_j = \sum_k \sum_{j<k} \int_{\Sigma_{kj}} (\mathbf{v}_k - \mathbf{v}_j)^T f_{kj}(\mathbf{u}_k, \mathbf{u}_j).$$

Because M is symmetric one can decompose M under the form $M = M^+ + M^-$ where M^+ is a non negative matrix and M^- is a non positive matrix. In the following we will consider the upwind flux $f_{kj}(\mathbf{u}_k, \mathbf{u}_j) = M_{kj}^+ \mathbf{u}_k + M_{kj}^- \mathbf{u}_j$, where $M_{kj} = M_{k|\Sigma_{kj}}$. Finally one has

$$\sum_k \int_{\Omega_k} \left(L^* \mathbf{v}_k\right)^T \mathbf{u}_k + \sum_k \sum_{j<k} \int_{\Sigma_{kj}} (\mathbf{v}_k - \mathbf{v}_j)^T (M_{kj}^+ \mathbf{u}_k + M_{kj}^- \mathbf{u}_j) \tag{6}$$

$$+ \sum_k \int_{\Sigma_{kk}} \mathbf{v}_k^T M_k^+ \mathbf{u}_k = -\sum_k \int_{\Sigma_{kk}} \mathbf{v}_k^T M_k^- \mathbf{g}. \tag{7}$$

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

$$\begin{aligned}a_{DG}(\mathbf{u}, \mathbf{v}) = &\sum_k \int_{\Omega_k} (L^* \mathbf{v}_k)^T \mathbf{u}_k + \sum_k \sum_{j<k} \int_{\Sigma_{kj}} (\mathbf{v}_k - \mathbf{v}_j)^T (M_{kj}^+ \mathbf{u}_k + M_{kj}^- \mathbf{u}_j) \\ &+ \sum_k \int_{\Sigma_{kk}} \mathbf{v}_k^T M_k^+ \mathbf{u}_k, \quad \mathbf{u}, \mathbf{v} \in H^1(\mathcal{T}_h), \\ l(\mathbf{v}) = &-\sum_k \int_{\Sigma_{kk}} \mathbf{v}_k^T M_k^- \mathbf{g}, \quad \mathbf{v} \in H^1(\mathcal{T}_h).\end{aligned} \tag{8}$$

One can rewrite (6) as $a_{DG}(\mathbf{u},\mathbf{v}) = l(\mathbf{v}),\ \forall \mathbf{v} \in H^1(\mathcal{T}_h)$. We can now define the classic discontinuous Galerkin method for Friedrichs systems with polynomial basis functions [9, 32]. Define $\mathbb{P}_q^d$ the space of polynomials of d variables, of total degree at most q and the broken polynomial space

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

Definition 2.1. *Assume $P_m(\mathcal{T}_h)$ is a finite subspace of $H^1(\mathcal{T}_h)$, for example $P_m(\mathcal{T}_h) = \mathbb{P}_q^d(\mathcal{T}_h)$. The standard upwind discontinuous Galerkin method for Friedrichs systems is formulated as follows*

$$\begin{cases} \text{find } \mathbf{u}_h \in P_m(\mathcal{T}_h) \text{ such that} \\ a_{DG}(\mathbf{u}_h, \mathbf{v}_h) = l(\mathbf{v}_h), \quad \forall \mathbf{v}_h \in P_m(\mathcal{T}_h). \end{cases} \tag{9}$$

Note that because of the conservation equation (5), the exact solution to (2) also verify

$$a_{DG}(\mathbf{u}, \mathbf{v}_h) = l(\mathbf{v}_h), \quad \forall \mathbf{v}_h \in H^1(\mathcal{T}_h). \tag{10}$$

2.2 Trefftz Discontinuous Galerkin formulation

Since our goal is to use Trefftz method we take basis functions which are solutions to (2) in each cell

$$V(\mathcal{T}_h) = \{\mathbf{v} \in H^1(\mathcal{T}_h), L\mathbf{v}_k = \mathbf{0} \quad \forall \Omega_k \in \mathcal{T}_h\} \subset H^1(\mathcal{T}_h). \tag{11}$$

The space $V(\mathcal{T}_h)$ is a genuine subspace of $H^1(\mathcal{T}_h)$ except in the case $L = 0$ which is of no interest. Starting from the bilinear form a_{DG} from (8), one notices that the volume term can be written for all functions in $V(\mathcal{T}_h)$ as

$$\int_{\Omega_k} \left(L^*\mathbf{v}_k\right)^T \mathbf{u}_k = 2\int_{\Omega_k} \mathbf{v}_k^T R \mathbf{u}_k, \quad \forall \mathbf{u}, \mathbf{v} \in V(\mathcal{T}_h). \tag{12}$$

One can therefore define a bilinear form $a_T : V(\mathcal{T}_h) \times V(\mathcal{T}_h) \to \mathbb{R}$ as

$$\begin{aligned} a_T(\mathbf{u},\mathbf{v}) = &\sum_k 2\int_{\Omega_k} \mathbf{v}_k^T R\mathbf{u}_k + \sum_k \sum_{j<k} \int_{\Sigma_{kj}} (\mathbf{v}_k - \mathbf{v}_j)^T (M_{kj}^+ \mathbf{u}_k + M_{kj}^- \mathbf{u}_j) \\ &+ \sum_k \int_{\Sigma_{kk}} \mathbf{v}_k^T M_k^+ \mathbf{u}_k, \quad \mathbf{u}, \mathbf{v} \in V(\mathcal{T}_h). \end{aligned} \tag{13}$$

Thanks to an integration by part for functions $\mathbf{v} \in V(\mathcal{T}_h)$ which are piecewise homogeneous solutions of the equation, one gets an equivalent formulation of the bilinear form $a_T(\cdot,\cdot)$

$$a_T(\mathbf{u},\mathbf{v}) = -\sum_k \sum_{j<k} \int_{\Sigma_{kj}} (M_{kj}^- \mathbf{v}_k + M_{kj}^+ \mathbf{v}_j)^T (\mathbf{u}_k - \mathbf{u}_j) - \sum_k \int_{\Sigma_{kk}} \mathbf{v}_k^T M_k^- \mathbf{u}_k, \quad \mathbf{u}, \mathbf{v} \in V(\mathcal{T}_h). \tag{14}$$

The relaxation term R completely disappeared in the formulation (14). It might seem a paradox at first sight but it is not because, for a Trefftz method, some information about R is encoded in the basis functions. Since there is no volume term in the formulation (14) compared to (13) it may be easier to implement. The related bilinear form $l : V(\mathcal{T}_h) \to \mathbb{R}$ is the same as in (8), that is $l(\mathbf{v}) = -\sum_k \int_{\Sigma_{kk}} \mathbf{v}_k^T M_k^- \mathbf{g}$ for all $\mathbf{v} \in V(\mathcal{T}_h)$.

Definition 2.2. *Assume $V_m(\mathcal{T}_h)$ is a finite subspace of $V(\mathcal{T}_h)$. The upwind Trefftz discontinuous Galerkin method for the model problem (2) is formulated as follows*

$$\begin{cases} \text{find } \mathbf{u}_h \in V_m(\mathcal{T}_h) \text{ such that} \\ a_T(\mathbf{u}_h, \mathbf{v}_h) = l(\mathbf{v}_h), \quad \forall \mathbf{v}_h \in V_m(\mathcal{T}_h). \end{cases} \tag{15}$$

We give some examples of subspace $V_m(\mathcal{T}_h)$.

- **Example 1:** the P_1 model in one dimension reads

$$\begin{cases} \partial_t p + \frac{c}{\varepsilon}\partial_x v = -\sigma_a p, \\ \partial_t v + \frac{c}{\varepsilon}\partial_x p = -\sigma_t v, \end{cases}$$

the dependent variable is $\mathbf{u} = (p, v)^T$ and $c, \sigma_a, \sigma_s \in \mathbb{R}^+$, $\varepsilon \in \mathbb{R}_*^+$, $\sigma_t = \sigma_a + \frac{\sigma_s}{\varepsilon^2}$. Assuming solutions are under the form $\mathbf{z}e^{\lambda x}$ one gets λ by solving $\det(A_1\lambda + R) = 0$ and then study the kernel of the matrix $A_1\lambda + R$ to find the vector $\mathbf{z}$. A possible choice for V_m is then $Span(V_m) = \{\mathbf{e}_1, \mathbf{e}_2\}$ with $\mathbf{e}_1(x) = \begin{pmatrix} -\sqrt{\sigma_t} \\ \sqrt{\sigma_a} \end{pmatrix} e^{\frac{\varepsilon}{c}\sqrt{\sigma_a\sigma_t}x}$, $\mathbf{e}_2(x) = \begin{pmatrix} \sqrt{\sigma_t} \\ \sqrt{\sigma_a} \end{pmatrix} e^{-\frac{\varepsilon}{c}\sqrt{\sigma_a\sigma_t}x}$.

- **Example 2:** consider the one dimensional case $A_1\partial_x\mathbf{u} = -R\mathbf{u}$. If the matrix A_1 is non singular one can write V under the form $V = \{\mathbf{v}(x) \text{ s.t. } \mathbf{v}(x) = e^{-A_1^{-1}Rx}\mathbf{c}\}$. For a two dimensional model $A_1\partial_{x_1}\mathbf{u} + A_2\partial_{x_2}\mathbf{u} = -R\mathbf{u}$, a general principle is that one can make the rotation $x' = x_1\cos(\theta) + x_2\sin(\theta)$, $\theta \in [0, 2\pi[$. Assuming the solution depends only on x' one gets $A_1'\partial_{x'}\mathbf{u}' = -R'\mathbf{u}'$ which can be solved in an identical way as the one dimensional case if the matrix A_1' is non singular.

- **Example 3:** however most of the time when considering physical models the matrix A_1 will be singular. For example the hyperbolic heat equation in two dimensions is

$$\begin{cases} \partial_t p + \frac{c}{\varepsilon}\operatorname{div}\mathbf{v} = 0, \\ \partial_t \mathbf{v} + \frac{c}{\varepsilon}\nabla p = -\frac{\sigma_s}{\varepsilon^2}\mathbf{v}, \end{cases}$$

the unknown is $\mathbf{u} = (p, \mathbf{v})^T \in \mathbb{R}^3$ and $c, \sigma_s \in \mathbb{R}^+$, $\varepsilon \in \mathbb{R}_*^+$. For simplicity we consider stationary solutions. Deriving the second equation and inserting in the first equation, one gets $\Delta p = 0$. Therefore denoting the harmonic polynomials in two dimensions as $q_k(\mathbf{x})$ for $k \in \mathbb{N}$, a possible choice for V_m is $Span(V_m) = \{\mathbf{e}_i,\ i = 1, ..., m\}$ with $\mathbf{e}_i = \begin{pmatrix} \frac{\sigma_s}{\varepsilon}q_i \\ -c\nabla q_i \end{pmatrix}$.

Remark 2.3. *In case of a time dependent problem, even if the classic upwind discontinuous Galerkin formulation (9) and the upwind Trefftz discontinuous Galerkin formulation (15) are posed on the whole space-time domain Ω, they still can be decoupled time step after time step. It comes from the fact that the matrix A_0 is definite positive and therefore $M^-(\mathbf{n}) = 0$ if $\mathbf{n} = (1, 0, ..., 0)$. Define $a_T^n : V(\mathcal{T}_h) \times V(\mathcal{T}_h) \to \mathbb{R}$ (related to the general bilinear form (14)) and $l^n : V(\mathcal{T}_h) \to \mathbb{R}$ as*

$$\begin{aligned} a_T^n(\mathbf{u}, \mathbf{v}) = &- \sum_k \sum_{j<k} \int_{\Sigma_{k^n j^n}} (M^-_{k^n j^n}\mathbf{v}_k^n + M^+_{k^n j^n}\mathbf{v}_j^n)^T(\mathbf{u}_k^n - \mathbf{u}_j^n) - \sum_k \int_{\partial\Omega_S \cap \partial\Omega_{k^n}} (\mathbf{v}_k^n)^T M^-_{k^n}\mathbf{u}_k^n \\ &- \sum_k \int_{\Sigma_{k^n k^{n-1}}} (\mathbf{v}_k^n)^T M^-_{k^n k^{n-1}}\mathbf{u}_k^n, \quad \mathbf{u}, \mathbf{v} \in V(\mathcal{T}_h), \\ l^n(\mathbf{v}) = &- \sum_k \int_{\partial\Omega_S \cap \partial\Omega_{k^n}} (\mathbf{v}_k^n)^T \mathbf{g}_S - \sum_k \int_{\Sigma_{k^n k^{n-1}}} (\mathbf{v}_k^n)^T M^-_{k^n k^{n-1}}\mathbf{u}_k^{n-1}, \quad \mathbf{v} \in V(\mathcal{T}_h), \end{aligned} \tag{16}$$

where we used the convention $\Sigma_{k^1 k^0} = \partial\Omega_{k^1} \cap (\partial\Omega \times \{0\})$ and $\Sigma_{k^{N+1} k^N} = \partial\Omega_{k^N} \cap (\partial\Omega \times \{T\})$. The formulation (15) is equivalent to the series of space problems

$$\begin{cases} \text{find } \mathbf{u}_h^n,\ n = 1, ..., N, \text{ such that} \\ a_T^n(\mathbf{u}_h^n, \mathbf{v}_h^n) = l^n(\mathbf{v}_h^n), \quad \forall \mathbf{v}_h^n \in V_m(\mathcal{T}_h). \end{cases} \tag{17}$$

A fully different choice of basis functions is also possible using the adjoint operator L^*. Assume $V^*(\mathcal{T}_h) = \{\mathbf{v} \in H^1(\mathcal{T}_h),\ L^*\mathbf{v}_k = \mathbf{0}\ \forall \Omega_k \in \mathcal{T}_h\} \subset H^1(\mathcal{T}_h)$, define $a_{AT} : V^*(\mathcal{T}_h) \times V^*(\mathcal{T}_h) \to \mathbb{R}$ as

$$a_{AT}(\mathbf{u}, \mathbf{v}) = \sum_k \sum_{j<k} \int_{\Sigma_{kj}} (\mathbf{v}_k - \mathbf{v}_j)^T (M^+_{kj}\mathbf{u}_k + M^-_{kj}\mathbf{u}_j) + \sum_k \int_{\Sigma_{kk}} \mathbf{v}_k^T M^+_k \mathbf{u}_k, \tag{18}$$

and consider $V_m^*(\mathcal{T}_h)$ a finite subspace of $V^*(\mathcal{T}_h)$. The upwind adjoint Trefftz discontinuous Galerkin method for the model problem (2) reads

$$\begin{cases} \text{find } \mathbf{u}_h \in V_m^*(\mathcal{T}_h) \text{ such that} \\ a_{AT}(\mathbf{u}_h, \mathbf{v}_h) = l(\mathbf{v}_h), \quad \forall \mathbf{v}_h \in V_m^*(\mathcal{T}_h), \end{cases} \tag{19}$$

with l a linear form as in (8). Even if when $R = 0$ these two approaches coincide, the problems we are interested in are such that $R = R^T \neq 0$, so these two methods are different in our case. Therefore the final solution will be in the space $V^* \neq V$ and it is not clear if a finite subspace of V^* can give a good approximation of V using standard norms. Another possibility is to adopt a Petrov-Galerkin approach choosing trial functions in $V(\mathcal{T}_h)$ and test functions in $V^*(\mathcal{T}_h)$ [12, 13]. However, we have noticed some stability issue with this method for time dependent problem. Therefore these methods will not be studied further.

3 Analysis of the Trefftz Discontinuous Galerkin method

3.1 Well posedness and quasi-optimality

In this section we show well posedness of (15) and a quasi-optimality bound in mesh-dependent norms. Our analysis follows some results of [28] where special case with $R = 0$ was studied. We define two semi-norms on $H^1(\mathcal{T}_h)$

$$\begin{aligned} \|\mathbf{u}\|_{DG}^2 &= \sum_k \int_{\Omega_k} \mathbf{u}_k^T R \mathbf{u}_k + \sum_k \sum_{j<k} \frac{1}{2} \int_{\Sigma_{kj}} (\mathbf{u}_k - \mathbf{u}_j)^T |M_{kj}| (\mathbf{u}_k - \mathbf{u}_j) + \sum_k \frac{1}{2} \int_{\Sigma_{kk}} \mathbf{u}_k^T |M_k| \mathbf{u}_k, \\ \|\mathbf{u}\|_{DG^*}^2 &= \sum_k \int_{\partial\Omega_k} -\mathbf{u}_k^T M_k^- \mathbf{u}_k, \end{aligned} \tag{20}$$

with $|M_{kj}| = |M_{jk}| = M_{kj}^+ - M_{kj}^-$. First we show that these two semi-norms are in fact norms on the Trefftz space. We will need the following lemmas.

Lemma 3.1. *One has the inequality* $\|\mathbf{v}\|_{DG} \leq c\|\mathbf{v}\|_{DG^*}$ *for all* $\mathbf{v} \in V(\mathcal{T}_h)$*, with* $c = \sqrt{\frac{5}{2}}$.

Proof. Assume $\mathbf{v} \in V(\mathcal{T}_h)$ then $L\mathbf{v}_k = \mathbf{0}$, $\forall \Omega_k \in \mathcal{T}_h$. Multiplying by $\mathbf{v}_k$ and integrating over Ω_k one gets

$$\frac{1}{2} \int_{\partial\Omega_k} \mathbf{v}_k M_k \mathbf{v}_k + \int_{\Omega_k} \mathbf{v}_k R \mathbf{v}_k = 0. \tag{21}$$

Therefore one has

$$\sum_k \int_{\Omega_k} \mathbf{v}_k R \mathbf{v}_k \leq -\frac{1}{2} \sum_k \int_{\partial\Omega_k} \mathbf{v}_k M_k^- \mathbf{v}_k = \frac{1}{2} \|\mathbf{v}\|_{DG^*}^2, \tag{22}$$

which is a bound for the first term in the definition of the DG norm (20). Moreover because R is non negative one also finds using (21) $\int_{\partial\Omega_k} \mathbf{v}_k M_k \mathbf{v}_k \leq 0$ that is $\int_{\partial\Omega_k} \mathbf{v}_k M_k^+ \mathbf{v}_k \leq -\int_{\partial\Omega_k} \mathbf{v}_k M_k^- \mathbf{v}_k$ and consequently

$$\int_{\partial\Omega_k} \mathbf{v}_k |M_k| \mathbf{v}_k \leq -2 \int_{\partial\Omega_k} \mathbf{v}_k M_k^- \mathbf{v}_k. \tag{23}$$

An elementary inequality gives $\frac{1}{2}(\mathbf{v}_k - \mathbf{v}_j)^T |M_{kj}| (\mathbf{v}_k - \mathbf{v}_j) \leq \mathbf{v}_k^T |M_{kj}| \mathbf{v}_k + \mathbf{v}_j^T |M_{kj}| \mathbf{v}_j$ thus

$$\sum_k \sum_{j<k} \frac{1}{2} \int_{\Sigma_{kj}} (\mathbf{v}_k - \mathbf{v}_j)^T |M_{kj}| (\mathbf{v}_k - \mathbf{v}_j) + \sum_k \frac{1}{2} \int_{\Sigma_{kk}} \mathbf{v}_k |M_k| \mathbf{v}_k \leq \sum_k \int_{\partial\Omega_k} \mathbf{v}_k^T |M_k| \mathbf{v}_k,$$

and therefore using (23)

$$\sum_k \sum_{j<k} \frac{1}{2} \int_{\Sigma_{kj}} (\mathbf{v}_k - \mathbf{v}_j)^T |M_{kj}| (\mathbf{v}_k - \mathbf{v}_j) + \sum_k \frac{1}{2} \int_{\Sigma_{kk}} \mathbf{v}_k |M_k| \mathbf{v}_k \leq -2 \sum_k \int_{\partial\Omega_k} \mathbf{v}_k M_k^- \mathbf{v}_k = 2\|\mathbf{v}\|_{DG^*}^2, \tag{24}$$

which is a bound for the second and third terms in the definition of the DG norm (20). Finally combining (22) and (24) with the definition of the DG norm (20) one gets $\|\mathbf{v}\|_{DG}^2 \leq \frac{5}{2}\|\mathbf{v}\|_{DG^*}^2$. □

Lemma 3.2. *Assume $M \in \mathbb{R}^{n\times n}$ is a symmetric matrix. Then one has*

$$\mathbf{z}^T M^2 \mathbf{z} \leq C \mathbf{z}^T |M| \mathbf{z}, \quad \forall \mathbf{z} \in \mathbb{R}^n,$$

where we have used the decomposition of $M = M^+ + M^-$, M^+ is a non negative matrix, M^- is a non positive matrix and $|M| = M^+ - M^-$.

Proof. First we notice that $\mathbf{z}^T|M|\mathbf{z} = \mathbf{z}^T M^+ \mathbf{z} - \mathbf{z}^T M^- \mathbf{z}$ and $\mathbf{z}^T M^2 \mathbf{z} = \mathbf{z}^T (M^+)^2 \mathbf{z} + \mathbf{z}^T (M^-)^2 \mathbf{z}$.

Let λ^+ be the maximum eigenvalue of M^+. Denoting λ_i and $\mathbf{r}_i$ the eigenvalue and eigenvector of M^+ one has $\lambda^+ \mathbf{z}^T M^+ \mathbf{z} = \lambda^+ \sum_{\lambda_i \geq 0} \lambda_i (\mathbf{z}, \mathbf{r}_i)^2 \geq \sum_{\lambda_i \geq 0} \lambda_i^2 (\mathbf{z}, \mathbf{r}_i)^2 = \mathbf{z}^T (M^+)^2 \mathbf{z}$. A similar inequality applies to the matrix M^- gives finally $\mathbf{z}^T |M| \mathbf{z} \geq \frac{1}{\rho(M)+1} \mathbf{z}^T M^2 \mathbf{z}$, $\forall \mathbf{z} \in \mathbb{R}^n$. This completes the proof. □

We can now show that the two semi-norms $\|\cdot\|_{DG}$ and $\|\cdot\|_{DG^*}$ are in fact norms on the Trefftz space $V(\mathcal{T}_h)$.

Proposition 3.3. *The semi-norms $\|\cdot\|_{DG}$ and $\|\cdot\|_{DG^*}$ are norms on the Trefftz space $V(\mathcal{T}_h)$.*

Proof. Assume $\mathbf{u} \in V(\mathcal{T}_h)$ and $\|\mathbf{u}\|_{DG} = \mathbf{0}$. Lemma 3.2 imply that $M\mathbf{u}$ has vanishing jump across each edge of $\mathcal{T}_h$. Thus $\mathbf{u}$ is a solution to the general problem $L\mathbf{u} = \mathbf{0}$ in Ω. Moreover $\int_{\partial\Omega} \mathbf{u}^T |M| \mathbf{u} = 0$. Therefore $\mathbf{u}$ is solution of

$$\begin{cases} L\mathbf{u} = \mathbf{0}, & \text{in } \Omega, \\ M^- \mathbf{u} = \mathbf{0}, & \text{on } \partial\Omega. \end{cases}$$

We conclude $\mathbf{u} = \mathbf{0}$ in Ω using the uniqueness of the solution. Thus $\|\cdot\|_{DG}$ is a norm on $V(\mathcal{T}_h)$. Thanks to lemma 3.1 we also conclude that $\|\cdot\|_{DG^*}$ is also a norm on $V(\mathcal{T}_h)$. This completes the proof. □

Next, we study the coercivity and the continuity of the bilinear form $a(\cdot,\cdot)$ regarding the norms $\|\cdot\|_{DG}$ and $\|\cdot\|_{DG^*}$.

Proposition 3.4 (Coercivity)**.** *For all $\mathbf{u} \in H^1(\mathcal{T}_h)$ one has $a_{DG}(\mathbf{u},\mathbf{u}) = \|\mathbf{u}\|_{DG}^2$. For all $\mathbf{u} \in V(\mathcal{T}_h)$ one has $a_{DG}(\mathbf{u},\mathbf{u}) = a_T(\mathbf{u},\mathbf{u})$.*

Proof. The proof is taken from [32]. Let $\mathbf{u}, \mathbf{v} \in H^1(\mathcal{T}_h)$. The bilinear form (8) reads

$$\begin{aligned} a_{DG}(\mathbf{u},\mathbf{v}) = &\sum_k \int_{\Omega_k} \big([-\sum_i A_i \partial_i + R]\mathbf{v}_k\big)^T \mathbf{u}_k + \sum_k \sum_{j<k} \int_{\Sigma_{kj}} (\mathbf{v}_k - \mathbf{v}_j)^T (M_{kj}^+ \mathbf{u}_k + M_{kj}^- \mathbf{u}_j) \\ &+ \sum_k \int_{\Sigma_{kk}} \mathbf{v}_k^T M_k^+ \mathbf{u}_k. \end{aligned}$$

Integrating by part and using $M_{kj} = -M_{jk}$ one has

$$\begin{aligned} a_{DG}(\mathbf{u},\mathbf{v}) = &\sum_k \int_{\Omega_k} \mathbf{v}_k^T (\sum_i A_i \partial_i + R)\mathbf{u}_k + \sum_k \sum_{j<k} \int_{\Sigma_{kj}} -\mathbf{v}_k^T M_{kj} \mathbf{u}_k + \mathbf{v}_j^T M_{kj} \mathbf{u}_j \\ &+ (\mathbf{v}_k - \mathbf{v}_j)^T (M_{kj}^+ \mathbf{u}_k + M_{kj}^- \mathbf{u}_j) + \sum_k \int_{\Sigma_{kk}} \mathbf{v}_k^T M_k^+ \mathbf{u}_k - \mathbf{v}_k^T M_{kj} \mathbf{u}_k. \end{aligned}$$

Using $M = M^+ + M^-$ one finds

$$a_{DG}(\mathbf{u},\mathbf{v}) = \sum_k \int_{\Omega_k} \mathbf{v}_k^T L \mathbf{u}_k - \sum_k \sum_{j<k} \int_{\Sigma_{kj}} (M_{kj}^- \mathbf{v}_k + M_{kj}^+ \mathbf{v}_j)^T (\mathbf{u}_k - \mathbf{u}_j) - \sum_k \int_{\Sigma_{kk}} \mathbf{v}_k^T M_k^- \mathbf{u}_k.$$

Since $L = -L^* + 2R$ one gets

$$\begin{aligned} a_{DG}(\mathbf{u},\mathbf{v}) = &- \sum_k \int_{\Omega_k} \mathbf{v}_k^T L^* \mathbf{u}_k + \sum_k 2 \int_{\Omega_k} \mathbf{v}_k^T R \mathbf{u}_k \\ &- \sum_k \sum_{j<k} \int_{\Sigma_{kj}} (M_{kj}^- \mathbf{v}_k + M_{kj}^+ \mathbf{v}_j)^T (\mathbf{u}_k - \mathbf{u}_j) - \sum_k \int_{\Sigma_{kk}} \mathbf{v}_k^T M_k^- \mathbf{u}_k. \end{aligned}$$

Summing the above expression of $a(\cdot,\cdot)$ and the one in (8) one gets with $\mathbf{v} = \mathbf{u}$ the equality $2a_{DG}(\mathbf{u},\mathbf{u}) = 2\|\mathbf{u}\|_{DG}^2$. Moreover from (12) one deduces $a_{DG}(\mathbf{u},\mathbf{u}) = a_T(\mathbf{u},\mathbf{u})$, $\forall \mathbf{u} \in V(\mathcal{T}_h)$. This completes the proof. □

Proposition 3.5 (Continuity)**.** *The continuity bound* $a_T(\mathbf{u},\mathbf{v}) \leq \sqrt{2}\|\mathbf{u}\|_{DG}\|\mathbf{v}\|_{DG^*}$ *holds for all* $\mathbf{u},\mathbf{v} \in V(\mathcal{T}_h)$.

Proof. Using $-M_{jk}^- = M_{kj}^+$, the norm DG^* can be recast into the form

$$\|\mathbf{u}\|_{DG^*}^2 = \sum_k \sum_{j<k} \int_{\Sigma_{kj}} -\mathbf{u}_k^T M_{kj}^- \mathbf{u}_k + \mathbf{u}_j^T M_{kj}^+ \mathbf{u}_j - \sum_k \int_{\Sigma_{kk}} \mathbf{u}_k^T M_k^- \mathbf{u}_k. \tag{25}$$

Since $|M^-| = -M^-$ and M^+, M^- are respectively non negative and non positive symmetric matrices, the bilinear form a_T (14) can be written as

$$\begin{aligned} a_T(\mathbf{u},\mathbf{v}) = &\sqrt{2}\Big[\sum_k \sum_{j<k} \int_{\Sigma_{kj}} \Big(\sqrt{|M_{kj}^-|}\mathbf{v}_k\Big)^T \sqrt{|M_{kj}^-|}\Big(\frac{\mathbf{u}_k - \mathbf{u}_j}{\sqrt{2}}\Big) + \Big(-\sqrt{M_{kj}^+}\mathbf{v}_j\Big)^T \sqrt{M_{kj}^+}\Big(\frac{\mathbf{u}_k - \mathbf{u}_j}{\sqrt{2}}\Big) \\ &+ \sum_k \int_{\Sigma_{kk}} \Big(\sqrt{|M_k^-|}\mathbf{v}_k\Big)^T \Big(\sqrt{|M_k^-|}\frac{\mathbf{u}_k}{\sqrt{2}}\Big)\Big]. \end{aligned}$$

Using the Cauchy-Schwartz inequality, one sees that the first term of each scalar product is bounded by $\|\mathbf{v}\|_{DG^*}$ and the second term by $\|\mathbf{u}\|_{DG}$. This completes the proof. □

We can now give the following classical quasi-optimality result.

Proposition 3.6 (Quasi-optimality)**.** *For any finite dimensional space* $V_m(\mathcal{T}_h) \subset V(\mathcal{T}_h)$*, the TDG formulation (15) admits a unique solution* $\mathbf{u}_h \in V_m(\mathcal{T}_h)$*. Moreover, the following quasi-optimality bounds holds*

$$\|\mathbf{u} - \mathbf{u}_h\|_{DG} \leq \sqrt{2} \inf_{\mathbf{v}_h \in V_m(\mathcal{T}_h)} \|\mathbf{u} - \mathbf{v}_h\|_{DG^*},$$

where $\mathbf{u}$ *stands for the exact solution to (2).*

Proof. From propositions 3.3 and 3.4 one deduces the uniqueness of the discrete solution $\mathbf{u}_h$. Existence of $\mathbf{u}_h$ follows from uniqueness. Moreover $\forall \mathbf{v}_h \in V_m(\mathcal{T}_h)$ one has

$$\|\mathbf{u} - \mathbf{u}_h\|_{DG}^2 = a_T(\mathbf{u} - \mathbf{u}_h, \mathbf{u} - \mathbf{u}_h) = a_T(\mathbf{u} - \mathbf{u}_h, \mathbf{u} - \mathbf{v}_h) \leq \sqrt{2}\|\mathbf{u} - \mathbf{u}_h\|_{DG}\|\mathbf{u} - \mathbf{v}_h\|_{DG^*},$$

thanks to propositions 3.4 and 3.5, to the consistency equality (10) and to (15). □

Using the quasi-optimality proposition one has the well-balanced property of the scheme. However there is an important difference between the one-dimensional case and higher dimensions. In one dimension a scheme is well-balanced if it captures all the stationary states of a hyperbolic system. This is possible because, in one dimension, the number of linearly independent stationary solutions is finite. However in two dimensions the space of stationary solutions becomes infinite. It has a huge impact on what is a well-balanced scheme in space dimensions higher than one. One must choose a finite subset of solutions for which the scheme is supposed to be exact. This is our practical definition of a well-balanced scheme and that's why it is immediately deduce from the quasi-optimality result of proposition 3.6. Of course a standard DG scheme has the same quasi-optimality result, but it can be well-balanced only for some particular polynomial functions. On the contrary a TDG method can be well-balanced for more general solutions which contain for example exponential factors as in Example 1 in Section 2.2 for which $\sigma_a > 0$.

Proposition 3.7 (Well-balanced scheme)**.** *The scheme (15) is well-balanced in the sense that if the solution* $\mathbf{u} \in H^1(\Omega)$ *of (2) is locally (in each cell) a linear combination of the basis functions (which are by construction exact solutions), then* $\mathbf{u}_h = \mathbf{u}$.

Proof. One can take $\mathbf{v}_h = \mathbf{u}$ in proposition 3.6. Therefore one has $\|\mathbf{u}-\mathbf{u}_h\|_{DG} = 0$. Since $\mathbf{u}-\mathbf{u}_h \in V(\mathcal{T}_h)$ one concludes using proposition 3.3. □

3.2 Estimate in standard norms

In the previous section, the error is bounded in terms of DG-norm. It is of course desirable to have estimates in a more standard norm. In this section we present some elementary L^2 lower bounds of the DG norm which take advantage of the relaxation matrix R and an L^2 upper bound of the DG^* norm.

Proposition 3.8. *Assume* $\Omega_k \in \mathcal{T}_h$, $R_k = R(\mathbf{x})_{|\Omega_k}$, *and* $\forall k$ R_k *is definite positive. One has*

$$\frac{1}{\sup_{k\in\mathcal{T}_h} \|\sqrt{R_k}^{-1}\|^2} \|\mathbf{u}\|_{L^2(\Omega)} \leq \|\mathbf{u}\|_{DG}, \quad \forall \mathbf{u} \in H^1(\mathcal{T}_h).$$

Proof. A basic inequality is $\mathbf{v}^2 \leq \|\sqrt{R_k}^{-1}\|^2 \Big(\mathbf{v}^T R_k \mathbf{v}\Big)$. Let $\mathbf{v} \in H^1(\mathcal{T}_h)$. Integrating over Ω_k, summing over all cells and using the definition of the DG-norm (20), one gets the assertion. □

This inequality holds when R is definite positive but degenerates when $R \to 0$. For non stationary problems, one can give a L^2 lower bound at the final time that does not depend on R.

Proposition 3.9. *For time dependent problems one has*

$$\|\mathbf{u}\|_{L^2(\Omega_S\times\{T\})} \leq \|\mathbf{u}\|_{DG}, \quad \forall \mathbf{u} \in H^1(\mathcal{T}_h).$$

Proof. Consider $\mathbf{n}(t,\mathbf{x})$ on $\partial\Omega$ with $\mathbf{n}(t,\mathbf{x}) = (n_t, n_{x_1}, ..., n_{x_d})^T = (1,0,...,0)^T$ one has $|M|((1,0,...,0)^T) = A_0 = I$. So

$$\sum_k \int_{\Omega_{S,k}\times\{T\}} \mathbf{u}_k^2 \leq \sum_k \frac{1}{2} \int_{\Omega_{S,k}\times\{T\}} \mathbf{u}_k^T A_0 \mathbf{u}_k \leq \sum_k \frac{1}{2} \int_{\Sigma_{kk}} \mathbf{u}_k^T |M_{kj}| \mathbf{u}_k, \quad \forall \mathbf{u} \in H^1(\mathcal{T}_h),$$

and the assertion follows from the definition of the DG-norm. □

Let us define the semi-norm $|\mathbf{u}|^2_{1,\Omega} := \int_\Omega \sum_{i=1}^n \sum_{j=1}^d (\partial_j \mathbf{u}_i)^2$.

Proposition 3.10. *One has*

$$\|\mathbf{u}\|^2_{DG^*} \leq C \sum_k \|\mathbf{u}\|_{L^2(\Omega_k)} \Big(\frac{1}{h_k}\|\mathbf{u}\|_{L^2(\Omega_k)} + |\mathbf{u}|_{1,\Omega_k}\Big), \quad \forall \mathbf{u} \in H^1(\mathcal{T}_h), \tag{26}$$

where $h_k = diam(\Omega_k)$ *and the constant* C *depends on the* A_i.

More precisely if one A_i is in $O(\frac{1}{\varepsilon})$ with respect to ε, the constant C scales like $\frac{1}{\varepsilon}$.

Proof. Let $\mathbf{u} \in \mathcal{T}_h$ one has $\|\mathbf{u}\|^2_{DG^*} = \sum_k \int_{\partial\Omega_k} -\mathbf{u}_k^T M^-_{kj} \mathbf{u}_k$ and therefore $\|\mathbf{u}\|^2_{DG^*} \leq C \sum_k \int_{\partial\Omega_k} \mathbf{u}_k^2$. We now use the trace inequality from theorem 1.6.6 in [3] in each cell Ω_k on each component of the vector $\mathbf{u}$

$$\|\mathbf{u}\|^2_{L^2(\partial\Omega_k)} \leq C \|\mathbf{u}\|_{L^2(\Omega_k)} \Big(\frac{1}{h_k} \|\mathbf{u}\|_{L^2(\Omega_k)} + |\mathbf{u}|_{1,\Omega_k} \Big), \quad \forall \mathbf{u} \in H^1(\Omega_k).$$

Summing over all cells one finally gets the equation (26). This completes the proof. □

4 Application in one dimension

We consider a concrete example, the P_1 model which is a first simple approximation of the transport equation using spherical harmonic expansion of the solution. An interesting property of the P_1 model is that like the transport equation it admits a diffusive limit when $\varepsilon \to 0$. The time dependent version of the P_1 model in one dimension reads

$$\begin{cases} \partial_t p + \frac{c}{\varepsilon} \partial_x v = -\sigma_a(x) p, \\ \partial_t v + \frac{c}{\varepsilon} \partial_x p = -\sigma_t(x) v, \end{cases} \tag{27}$$

the unknown is $\mathbf{u} = (p, v)^T$ and $c, \sigma_a, \sigma_s \in \mathbb{R}^+$, $\varepsilon \in \mathbb{R}^+_*$, $\sigma_t = \sigma_a + \frac{\sigma_s}{\varepsilon^2}$. The reader should be aware that σ_t depends on ε and behave as $\frac{1}{\varepsilon^2}$ when $\sigma_s > 0$ and $\varepsilon \to 0$. When $\varepsilon \to 0$ the variable p of the system (27) follows a diffusion equation.

Proposition 4.1. *When $\varepsilon \to 0$, the variable p and v of (27) behave formally as*

$$\left| \begin{aligned} &\partial_t p - \frac{c^2}{\sigma_s} \partial_{xx} p = -\sigma_a p, \\ &v = -\frac{c\varepsilon}{\sigma_s} \partial_x p. \end{aligned} \right. \tag{28}$$

Proof. Multiplying the second equation of (31) by ε^2 and neglecting the term in ε^2 one gets $v = -\frac{c\varepsilon}{\sigma_s} \partial_x p$. Inserting this expression in the first equation of (31) one finds $\partial_t p - \frac{c^2}{\sigma_s} \partial_{xx} p = -\sigma_a p$. □

4.1 Construction of the basis functions for high order time dependent scheme

In order to use the Trefftz method (15) one needs to find solutions to the model (27). In particular we would like to give a general procedure to increase the number of basis functions in order to get high order of convergence, if needed. In the following we search for particular solutions to (27) under the form $\mathbf{u}(t,x) = \mathbf{q}(t,x) e^{\lambda x}$ where $\mathbf{q}(t,x)$ is a polynomial in space and time. For simplicity we consider a polynomial of degree at most one in space and time. For brevity the proofs of this section are postponed in the appendix.

Proposition 4.2. *Assume constant coefficients σ_a and σ_t, $c \neq 0$ and $\sigma_a \neq 0$. The P_1 model (27) admits the following solutions*

$$\begin{aligned} \mathbf{e}_1^{\pm}(x) &= \begin{pmatrix} \sqrt{\sigma_t} \\ \mp\sqrt{\sigma_a} \end{pmatrix} e^{\pm \frac{\varepsilon}{c} \sqrt{\sigma_a \sigma_t} x}, \\ \mathbf{e}_2^{\pm}(t,x) &= \begin{pmatrix} -c \frac{\sigma_t - \sigma_a}{4\sigma_a \sqrt{\sigma_t}} \pm \varepsilon \frac{\sigma_a + \sigma_t}{2\sqrt{\sigma_a}} x + c\sqrt{\sigma_t} t \\ \mp c \frac{\sigma_t - \sigma_a}{4\sigma_t \sqrt{\sigma_a}} - \varepsilon \frac{\sigma_a + \sigma_t}{2\sqrt{\sigma_t}} x \mp c\sqrt{\sigma_a} t \end{pmatrix} e^{\pm \frac{\varepsilon}{c} \sqrt{\sigma_a \sigma_t} x}. \end{aligned} \tag{29}$$

Proof. The proof is given in appendix A. □

Because the basis functions (29) are solutions to (27), one can use them in the case $\sigma_a \neq 0$. The problem with such basis comes from the limit cases. Indeed they degenerate to the same limit as $\sigma_a \to 0$ which cause some numerical instability. However, one can construct new solutions which remain stable in the limit case $\sigma_a \to 0$.

Proposition 4.3. *There exists $\tilde{\mathbf{e}}_1, \tilde{\mathbf{e}}_2, \tilde{\mathbf{e}}_3, \tilde{\mathbf{e}}_4$, linear combinations of the solutions (29) such that*

$$\begin{aligned}
\tilde{\mathbf{e}}_1(x) &\underset{\sigma_a \to 0}{\to} \begin{pmatrix} \frac{\varepsilon\sigma_t}{c} x \\ -1 \end{pmatrix}, \\
\tilde{\mathbf{e}}_2(x) &\underset{\sigma_a \to 0}{\to} \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \\
\tilde{\mathbf{e}}_3(t,x) &\underset{\sigma_a \to 0}{\to} \begin{pmatrix} -\frac{\varepsilon^2\sigma_t}{2c} x^2 - ct \\ \varepsilon x \end{pmatrix}, \\
\tilde{\mathbf{e}}_4(t,x) &\underset{\sigma_a \to 0}{\to} \begin{pmatrix} -\frac{\varepsilon^3\sigma_t^2}{6c^2} x^3 - \varepsilon\sigma_t t x - \varepsilon x \\ \frac{\varepsilon^2\sigma_t}{2c} x^2 + ct \end{pmatrix}.
\end{aligned} \tag{30}$$

Proof. The proof is given in appendix A. □

Remark 4.4. *Note that the solutions (29) are only defined in the case $c \neq 0$. However, up to a multiplication by c or c^2 if needed, the limit solutions (30) can also be used in the case $c = 0$.*

4.2 Asymptotic preserving properties

In this section we study the behavior of the scheme when $\varepsilon \to 0$. One cannot use directly the L^2 estimates of the previous section mainly because the parameter ε appears in the $\|\cdot\|_{DG}$ and $\|\cdot\|_{DG^*}$ norms. Here we choose to interpret the scheme (15) as a finite difference scheme which has several advantages. Under this form we observe that the scheme is new compared to other popular one dimensional finite difference schemes [16]. Moreover one can study, at least formally, the asymptotic behavior of a finite difference scheme by means of a Hilbert expansion. We consider the P_1 model with no absorption

$$\begin{cases} \partial_t p + \frac{c}{\varepsilon}\partial_x v = 0, \\ \partial_t v + \frac{c}{\varepsilon}\partial_x p = -\frac{\sigma_s}{\varepsilon^2} v, \end{cases} \tag{31}$$

with $\varepsilon \in \mathbb{R}_*^+$, $\sigma_s, c \in \mathbb{R}^+$. For the sake of simplicity assume that σ_s is constant in the domain. We use the stationary basis functions e_1 and e_2 defined in each cells as

$$\mathbf{e}_{k,1}(t,x) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \mathbf{e}_{k,2}(t,x) = \begin{pmatrix} -\frac{\sigma_s}{c\varepsilon}(x - x_k) \\ 1 \end{pmatrix}, \tag{32}$$

where x_k is the abscissa of the center of the cell k. For simplicity assume the step space $h = x_{k+1} - x_k$ is constant for all k. Using the basis functions (32) in (9) and considering $x = x_k$ with periodic boundary conditions one gets the following scheme (see appendix C for details)

$$\begin{cases} \dfrac{p_k^{n+1} - p_k^n}{\Delta t} + \dfrac{c}{2\varepsilon h}\Big[- p_{k+1} + 2p_k - p_{k-1} + (1-a)(v_{k+1} - v_{k-1})\Big]^{n+1} = 0, \\ (1 + \dfrac{a^2}{3})\dfrac{v_k^{n+1} - v_k^n}{\Delta t} + \dfrac{c}{2\varepsilon h}\Big[a^2(v_{k+1} + 2v_k + v_{k-1}) + (-v_{k+1} + 2v_k - v_{k-1}) \\ \quad +(1+a)(p_{k+1} - p_{k-1})\Big]^{n+1} = -\dfrac{\sigma_s}{\varepsilon^2} v_k^{n+1}, \end{cases} \tag{33}$$

with $a = \frac{\sigma_s h}{2c\varepsilon}$.

Remark 4.5. *One can interpret the first component of the basis function* $\mathbf{e}_{k,2}(t,x)$ *in* (32) *as a correction to the standard finite volume method. Indeed the standard finite volume method is equivalent to consider the formulation* (9) *with the two basis functions* $\mathbf{e}_{k,1} = (1,0)^T$, $\mathbf{e}_{k,2} = (0,1)^T$. *The scheme is then* (33) *with* $a = 0$. *This scheme is not asymptotic preserving when* $\varepsilon \to 0$.

Proposition 4.6. *When* $\varepsilon \to 0$ *the scheme* (33) *admits the formal limit*

$$\left|\begin{array}{l} (v^0_{k+1} + v^0_k)^{n+1} = 0, \\ \Big(\dfrac{v^1_{k+1} + 2v^1_k + v^1_{k-1}}{4}\Big)^{n+1} = -\dfrac{c}{\sigma_s}\Big(\dfrac{p^0_{k+1} - p^0_{k-1}}{2h}\Big)^{n+1}, \\ \dfrac{(\bar{p}^0_k)^{n+1} - (\bar{p}^0_k)^n}{\Delta t} - \dfrac{c^2}{\sigma_s}\Big(\dfrac{p^0_{k+2} - 2p^0_k + p^0_{k-2}}{4h^2}\Big)^{n+1} = 0, \end{array}\right. \tag{34}$$

with $\bar{p}^0_k = (\frac{2}{3}p^0_{k+2} + 4p^0_{k+1} + \frac{20}{3}p^0_k + 4p^0_{k-1} + \frac{2}{3}p^0_{k-2})/16$ *a local mean value of* p^0_k. *The limit scheme* (34) *is consistent with the limit model* (28) *and therefore the scheme is asymptotic preserving.*

Proof. We adopt the notations $\{\!\{f\}\!\}_{k+\frac{1}{2}} = \frac{f_{k+1}+f_k}{2}$, $[\![f]\!]_{k+\frac{1}{2}} = \frac{f_{k+1}-f_k}{2}$ and $\delta_t f = \frac{f^{n+1}-f^n}{\Delta t}$. With these notations the scheme (33) can be written under the form

$$\delta_t p_k + \frac{c}{\varepsilon h}\Big[-([\![p]\!]_{k+\frac{1}{2}} - [\![p]\!]_{k-\frac{1}{2}}) + (1-a)(\{\!\{v\}\!\}_{k+\frac{1}{2}} - \{\!\{v\}\!\}_{k-\frac{1}{2}})\Big]^{n+1} = 0, \tag{35}$$

$$\begin{aligned}(1+\frac{a^2}{3})\delta_t v_k + \frac{c}{\varepsilon h}\Big[a^2(\{\!\{v\}\!\}_{k+\frac{1}{2}} + \{\!\{v\}\!\}_{k-\frac{1}{2}}) + 2av_k - ([\![v]\!]_{k+\frac{1}{2}} - [\![v]\!]_{k-\frac{1}{2}}) \\ +(1+a)([\![p]\!]_{k+\frac{1}{2}} + [\![p]\!]_{k-\frac{1}{2}})\Big]^{n+1} = 0.\end{aligned} \tag{36}$$

Let $p = \sum_{i\geq 0} p^i\varepsilon^i$ and $v = \sum_{i\geq 0} v^i\varepsilon^i$. We inject these expressions in (35) and (36) and we expand all coefficients and variables with respect to ε. The important step is to expand a with respect to ε using the definition $a = \frac{\sigma_s h}{2c\varepsilon}$. The terms $O(\frac{1}{\varepsilon^2})$ in (35) and $O(\frac{1}{\varepsilon^3})$ in (36) are

$$\{\!\{v\}\!\}^0_{k+\frac{1}{2}} - \{\!\{v\}\!\}^0_{k-\frac{1}{2}} = 0,$$

$$\{\!\{v\}\!\}^0_{k+\frac{1}{2}} + \{\!\{v\}\!\}^0_{k-\frac{1}{2}} = 0.$$

These two equations together give

$$\{\!\{v\}\!\}^0_{k+\frac{1}{2}} = 0, \forall k. \tag{37}$$

Now we study the terms in $O(\frac{1}{\varepsilon})$ in (35) and in $O(\frac{1}{\varepsilon^2})$ in (36). Using (37) one has

$$-([\![p]\!]^0_{k+\frac{1}{2}} - [\![p]\!]^0_{k-\frac{1}{2}}) - \frac{\sigma_s h}{2c}(\{\!\{v\}\!\}^1_{k+\frac{1}{2}} - \{\!\{v\}\!\}^1_{k-\frac{1}{2}}) = 0,$$

$$\frac{\sigma_s h}{6c}\delta_t v^0_k + \frac{c}{h}\Big[[\![p]\!]^0_{k+\frac{1}{2}} + [\![p]\!]^0_{k-\frac{1}{2}} + \frac{\sigma_s h}{2c}(\{\!\{v\}\!\}^1_{k+\frac{1}{2}} + \{\!\{v\}\!\}^1_{k-\frac{1}{2}}) + 2v^0_k\Big] = 0.$$

Therefore, subtracting these two equations, one finds

$$\frac{h}{3c}\delta_t v^0_k + \{\!\{v\}\!\}^1_{k+\frac{1}{2}} + \frac{4c}{\sigma_s h}v^0_k = -\frac{2c}{\sigma_s h}[\![p]\!]^0_{k+\frac{1}{2}}, \forall k.$$

Adding this equality for k and $k-1$ and using (37) one deduces

$$\{\!\{v\}\!\}^1_{k+\frac{1}{2}} + \{\!\{v\}\!\}^1_{k-\frac{1}{2}} = -\frac{2c}{\sigma_s h}([\![p]\!]^0_{k+\frac{1}{2}} + [\![p]\!]^0_{k-\frac{1}{2}}), \forall k. \tag{38}$$

Finally with the terms in $O(1)$ for (35) and in $O(\frac{1}{\varepsilon})$ for (36)

$$\delta_t p_k^0 + \frac{c}{h}\left[-([\![p]\!]^1_{k+\frac{1}{2}} - [\![p]\!]^1_{k-\frac{1}{2}}) + (\{\!\{v\}\!\}^1_{k+\frac{1}{2}} - \{\!\{v\}\!\}^1_{k-\frac{1}{2}}) - \frac{\sigma_s h}{2c}(\{\!\{v\}\!\}^2_{k+\frac{1}{2}} - \{\!\{v\}\!\}^2_{k-\frac{1}{2}})\right]^{n+1} = 0,$$

$$\frac{\sigma_s^2 h^2}{12c^2}\delta_t v_k^1 + \frac{c}{h}\left[\frac{\sigma_s h}{2c}(2v_k^1 + [\![p]\!]^1_{k+\frac{1}{2}} + [\![p]\!]^1_{k-\frac{1}{2}}) + [\![p]\!]^0_{k+\frac{1}{2}} + [\![p]\!]^0_{k-\frac{1}{2}} - ([\![v]\!]^0_{k+\frac{1}{2}} - [\![v]\!]^0_{k-\frac{1}{2}}) + \frac{\sigma_s^2 h^2}{4c^2}(\{\!\{v\}\!\}^2_{k+\frac{1}{2}} + \{\!\{v\}\!\}^2_{k-\frac{1}{2}})\right]^{n+1} = 0.$$

Dividing the first equation by σ_s, using (37), (38) and multiplying by $\frac{2c}{\sigma_s^2 h}$ the second equation one gets

$$\frac{1}{\sigma_s}\delta_t p_k^0 + \left[\frac{c}{\sigma_s h}\Big(- ([\![p]\!]^1_{k+\frac{1}{2}} - [\![p]\!]^1_{k-\frac{1}{2}}) + \{\!\{v\}\!\}^1_{k+\frac{1}{2}} - \{\!\{v\}\!\}^1_{k-\frac{1}{2}}\Big) - \frac{\{\!\{v\}\!\}^2_{k+\frac{1}{2}} - \{\!\{v\}\!\}^2_{k-\frac{1}{2}}}{2}\right]^{n+1} = 0,$$

$$\frac{h}{6c}\delta_t v_k^1 + \left[\frac{c}{\sigma_s h}\Big(- \{\!\{v\}\!\}^1_{k+\frac{1}{2}} + 2v_k^1 - \{\!\{v\}\!\}^1_{k-\frac{1}{2}} + [\![p]\!]^1_{k+\frac{1}{2}} + [\![p]\!]^1_{k-\frac{1}{2}} + \frac{4c}{\sigma_s h}v_k^0\Big) + \frac{\{\!\{v\}\!\}^2_{k+\frac{1}{2}} + \{\!\{v\}\!\}^2_{k-\frac{1}{2}}}{2}\right]^{n+1} = 0.$$

Adding and subtracting these two equations one finds

$$\{\!\{v\}\!\}^2_{k-\frac{1}{2}} + \frac{2c}{\sigma_s h}[\![p]\!]^1_{k-\frac{1}{2}} + \frac{4c^2}{\sigma_s^2 h^2}v_k^0 = -\frac{1}{\sigma_s}\delta_t p_k^0 - \frac{h}{6c}\delta_t v_k^1 - \frac{2c}{\sigma_s h}(v_k^1 - \{\!\{v\}\!\}^1_{k-\frac{1}{2}})^{n+1}, \tag{39}$$

and

$$\{\!\{v\}\!\}^2_{k+\frac{1}{2}} + \frac{2c}{\sigma_s h}[\![p]\!]^1_{k+\frac{1}{2}} + \frac{4c^2}{\sigma_s^2 h^2}v_k^0 = \frac{1}{\sigma_s}\delta_t p_k^0 - \frac{h}{6c}\delta_t v_k^1 - \frac{2c}{\sigma_s h}(v_k^1 - \{\!\{v\}\!\}^1_{k+\frac{1}{2}})^{n+1}. \tag{40}$$

Using (39) in $k+1$ and subtracting (40) to (39) one gets

$$\frac{1}{\sigma_s}\delta_t(p_{k+1}^0 + p_k^0) + \frac{h}{6c}\delta_t(v_{k+1}^1 - v_k^1) + \frac{2c}{h\sigma_s}(v_{k+1}^1 - v_k^1)^{n+1} = \frac{4c^2}{\sigma_s^2 h^2}(v_k^0 - v_{k+1}^0).$$

Adding this equation for k and $k-1$ and using (37) one has

$$\frac{1}{\sigma_s}\delta_t(p_{k+1}^0 + 2p_k^0 + p_{k-1}^0) + \frac{h}{3c}\delta_t(\{\!\{v\}\!\}^1_{k+\frac{1}{2}} - \{\!\{v\}\!\}^1_{k-\frac{1}{2}}) - \frac{4c}{\sigma_s h}(\{\!\{v\}\!\}^1_{k+\frac{1}{2}} - \{\!\{v\}\!\}^1_{k-\frac{1}{2}})^{n+1} = 0.$$

Summing this equation for k and $k+1$ one gets

$$\frac{1}{\sigma_s}\delta_t(p_{k+2}^0 + 3p_{k+1}^0 + 3p_k^0 + p_{k-1}^0) + \frac{h}{3c}\delta_t(\{\!\{v\}\!\}^1_{k+\frac{3}{2}} - \{\!\{v\}\!\}^1_{k-\frac{1}{2}}) - \frac{4c}{\sigma_s h}(\{\!\{v\}\!\}^1_{k+\frac{3}{2}} - \{\!\{v\}\!\}^1_{k-\frac{1}{2}})^{n+1} = 0.$$

Summing this equation for k and $k-1$ one finally finds

$$\frac{1}{\sigma_s}\delta_t(p_{k+2}^0 + 4p_{k+1}^0 + 6p_k^0 + 4p_{k-1}^0 + p_{k-2}^0) + \frac{h}{3c}\delta_t(\{\!\{v\}\!\}^1_{k+\frac{3}{2}} + \{\!\{v\}\!\}^1_{k+\frac{1}{2}} - \{\!\{v\}\!\}^1_{k-\frac{1}{2}} - \{\!\{v\}\!\}^1_{k-\frac{3}{2}}) - \frac{4c}{\sigma_s h}(\{\!\{v\}\!\}^1_{k+\frac{3}{2}} + \{\!\{v\}\!\}^1_{k+\frac{1}{2}} - \{\!\{v\}\!\}^1_{k-\frac{1}{2}} - \{\!\{v\}\!\}^1_{k-\frac{3}{2}})^{n+1} = 0.$$

Using (38) one deduces

$$(\{\!\{v\}\!\}^1_{k+\frac{3}{2}} + \{\!\{v\}\!\}^1_{k+\frac{1}{2}}) - (\{\!\{v\}\!\}^1_{k-\frac{1}{2}} + \{\!\{v\}\!\}^1_{k-\frac{3}{2}}) = -\frac{c}{\sigma_s h}(p^0_{k+2} - 2p_k + p^0_{k-2}).$$

Therefore one finally has

$$\delta_t(\frac{2}{3}p^0_{k+2} + 4p^0_{k+1} + \frac{20}{3}p^0_k + 4p^0_{k-1} + \frac{2}{3}p^0_{k-2}) - \frac{4c^2}{\sigma_s}\left(\frac{p^0_{k+2} - 2p^0_k + p^0_{k-2}}{h^2}\right)^{n+1} = 0.$$

This equality is consistent with the first equation of the limit model (28). Moreover the equality (38) is also consistent with the second equation of (28). This completes the proof. □

5 Application to the P_1 model in two dimensions

In the previous section we have studied the well balanced and asymptotic preserving properties of the TDG method in one dimension for the P_1 approximation of the transport equation. Other schemes which satisfy these two properties have already been designed in one dimension (see for example [16]) but fundamentals difficulties arise when trying to extend those schemes to higher dimensions (unstructured mesh, infinite dimensional stationary state space, ...). One advantage of the TDG method (15) is that, given the approximation space $V(\mathcal{T}_h)$, it can be directly extended to the two dimensional case. The scheme will be well balanced (in the sense of proposition 3.7) and one can hope the asymptotic behavior of the two dimensional scheme will come naturally from the basis functions: numerical evidence shows it is indeed the case. In the following we consider the P_1 model in two dimensions

$$\begin{cases} \partial_t p(t,\mathbf{x}) + \frac{c}{\varepsilon}\operatorname{div}\mathbf{v}(t,\mathbf{x}) = -\sigma_a(\mathbf{x})p(t,\mathbf{x}), \\ \partial_t \mathbf{v}(t,\mathbf{x}) + \frac{c}{\varepsilon}\nabla\mathbf{p}(t,\mathbf{x}) = -\sigma_t(\mathbf{x})\mathbf{v}(t,\mathbf{x}), \end{cases} \tag{41}$$

with the unknown $\mathbf{u} = (p,\mathbf{v})^T \in \mathbb{R}^3$. The coefficients $\sigma_t = \sigma_a + \frac{\sigma_s}{\varepsilon^2}, \sigma_a, \sigma_s \in \mathbb{R}^+$ depend on $\mathbf{x}$ while $\varepsilon \in \mathbb{R}^+_*, c \in \mathbb{R}^+$ are constants. We write $\mathbf{x} = (x,y)^T$. The system (41) can be recast into the form (2) with $d = 2$, $n = 3$ and

$$A_0 = I_m,\ A_1 = \frac{c}{\varepsilon}\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},\ A_2 = \frac{c}{\varepsilon}\begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix},\ R(\mathbf{x}) = \begin{pmatrix} \sigma_a(\mathbf{x}) & 0 & 0 \\ 0 & \sigma_t(\mathbf{x}) & 0 \\ 0 & 0 & \sigma_t(\mathbf{x}) \end{pmatrix}.$$

The reader should be aware that σ_t depends on ε and behave as $\frac{1}{\varepsilon^2}$ when $\sigma_s > 0$ and $\varepsilon \to 0$.

Stationary solutions to the P_1 model (41) with constant coefficients are candidates to be basis functions.

Proposition 5.1 (A first family of basis functions)**.** *Take* $\mathbf{d}_k = (\cos(\phi_k), \sin(\phi_k))^T \in \mathbb{R}^2, c \neq 0$ *and assume constant coefficients* σ_a, σ_t*. The functions*

$$\mathbf{e}_k = \begin{pmatrix} \sqrt{\sigma_t} \\ -\sqrt{\sigma_a}\mathbf{d}_k \end{pmatrix} e^{\frac{\varepsilon}{c}\sqrt{\sigma_a\sigma_t}(\mathbf{d}_k,\mathbf{x})}, \tag{42}$$

are solution to the model problem (41)*.*

Proof. Assume the solution of (41) is under the form $\mathbf{e}_k(\mathbf{x}) = \mathbf{z}_k e^{\lambda(\mathbf{d}_k,\mathbf{x})}$ for some $\mathbf{z}_k \in \mathbb{R}^3$. Setting $M_\lambda = \lambda(A_1\cos(\phi_k) + A_2\sin(\phi_k)) + R$, one obtains the eigenproblem $M_\lambda \mathbf{z}_k = \mathbf{0}$. The values of λ such that $\det(M_\lambda) = 0$ are $\lambda = \pm\frac{\varepsilon}{c}\sqrt{\sigma_a\sigma_t}$. Taking $\lambda = \frac{\varepsilon}{c}\sqrt{\sigma_a\sigma_t}$ one has $Ker(M_{\frac{\varepsilon}{c}\sqrt{\sigma_a\sigma_t}}) = Span(\mathbf{w})$ with $\mathbf{w} = (\sqrt{\sigma_t}, -\sqrt{\sigma_a}\cos(\phi_k), -\sqrt{\sigma_a}\sin(\phi_k))^T$. Taking $\mathbf{z}_k = \mathbf{w}$ and $\mathbf{e}_k(\mathbf{x}) = \mathbf{z}_k e^{\lambda(\mathbf{d}_k,\mathbf{x})}$, one finds a non trivial solution to the model (41). This ends the proof. □

Proposition 5.2 (A second family of basis functions)**.** *Assume $\sigma_a = 0$ and σ_t is constant. Denote $q_k(\mathbf{x})$, $k \in \mathbb{N}$, the scaled harmonic polynomial in two dimensions*

$$q_1 = 1,\ q_{2l} = \frac{2^{1-l}}{l!}\Re(x+iy)^l \text{ and } q_{2l+1} = \frac{2^{1-l}}{l!}\Im(x+iy)^l \text{ for } l \in \mathbb{N}^*. \tag{43}$$

The following functions are solutions to the P_1 model (41)

$$\mathbf{e}_k = \begin{pmatrix} \frac{\sigma_s}{\varepsilon} q_k \\ -c\nabla q_k \end{pmatrix}, \quad k = 1, ..., m. \tag{44}$$

Proof. Consider the stationary version of (41). Deriving the second and third equations and inserting it in the first equation, one sees that p follows a second order equation $\Delta p = 0$. By definition the scaled harmonic polynomials $q_k(\mathbf{x})$ are solutions and one gets the first component of the solution. It is then easy to deduce the second and third components of the solution. This completes the proof. □

Because the basis functions (42) are solution to (41), one can use them in the case $\sigma_a \neq 0$. The problem with such basis comes from the limit cases. Indeed the vectors degenerate to the same limit as $\sigma_a \to 0$. Our goal is to show there exist a stable basis which degenerates correctly when $\sigma_a \to 0$. We proceed as in [14, Section 3.1] and consider the matrix $M_{2n+1} := M_{\theta_1,\theta_2,...,\theta_{2n+1}} \in \mathbb{R}^{2n+1\times 2n+1}$ defined as

$$M_{2n+1} := M_{\theta_1,\theta_2,...,\theta_{2n+1}} := \begin{pmatrix} 1 & 1 & ... & 1 \\ \cos(\theta_1) & \cos(\theta_2) & ... & \cos(\theta_{2n+1}) \\ \sin(\theta_1) & \sin(\theta_2) & ... & \sin(\theta_{2n+1}) \\ \cos(2\theta_1) & \cos(2\theta_2) & ... & \cos(2\theta_{2n+1}) \\ \sin(2\theta_1) & \sin(2\theta_2) & ... & \sin(2\theta_{2n+1}) \\ \vdots & \vdots & ... & \vdots \\ \cos(n\theta_1) & \cos(n\theta_2) & ... & \cos(n\theta_{2n+1}) \\ \sin(n\theta_1) & \sin(n\theta_2) & ... & \sin(n\theta_{2n+1}) \end{pmatrix}, \tag{45}$$

This matrix is invertible under general conditions.

Proposition 5.3. *Let $\theta_1, ..., \theta_{2n+1} \in [0, 2\pi[$ with $\theta_i \neq \theta_j$ if $i \neq j$. Then the matrix M_{2n+1} is invertible.*

Proof. We take the proof given in [14]. Assume $\boldsymbol{\psi} = (\psi_0, ..., \psi_{2n+1})^T$ and $M_{2n+1}^T \boldsymbol{\psi} = 0$ then

$$\psi_0 + \sum_{l=1}^{n} \psi_{2l-1}\cos(l\theta_k) + \psi_{2l}\sin(l\theta_k) = 0, \quad \text{for } k = 1, ..., 2n+1.$$

Therefore, $\boldsymbol{\psi}$ is the coefficient vector for a real valued trigonometric polynomial of degree n with $2n+1$ different zeros θ_k, $k = 1, ..., 2n+1$. This polynomial is zero everywhere and one can conclude $\boldsymbol{\psi} = \mathbf{0}$. This completes the proof. □

We give a new family of basis functions which degenerate correctly when $\sigma_a \to 0$.

Definition 5.4 (A third family of basis functions)**.** *Let $n \in \mathbb{N}$ and consider $2n+1$ solutions to the P_1 model $\mathbf{e}_i$ $i = 1, ..., 2n+1$. We set $a_{k,j} = (M_{2n+1})^{-1}_{k,j}$ and define*

$$\tilde{\mathbf{e}}_j = \sqrt{\sigma_s}\left(\frac{\varepsilon}{c}\sqrt{\sigma_a\sigma_t}\right)^{-\lfloor \frac{j}{2} \rfloor} \sum_{k=1}^{2n+1} a_{k,j}\mathbf{e}_k, \quad j = 1, ..., 2n+1. \tag{46}$$

Proposition 5.5. *The functions $\tilde{\mathbf{e}}_i$ from (46) remains stable when $\sigma_a \to 0$. More precisely, denoting by $q_i(\mathbf{x})$ the scaled harmonic polynomial (43), one has*

$$\tilde{\mathbf{e}}_k \underset{\sigma_a \to 0}{\to} \begin{pmatrix} \frac{\sigma_s}{\varepsilon} q_k \\ -c\nabla q_k \end{pmatrix} \text{ for } k = 1, ..., 2n+1.$$

Proof. The proof is based on the stable basis argument used for the Helmoltz equation in [14]. Deriving the second and third equations of (41) and inserting it in the first equation, one sees that for stationary solutions, the variable p follows a second order equation $\Delta p = \frac{\varepsilon^2}{c^2}\sigma_t\sigma_a p$. This equality is satisfied by the scaled harmonic polynomials (43) in the case $\sigma_a = 0$. Following [14] and using the definition of the coefficients a_{kj}, one can show that the first component of the functions $\tilde{\mathbf{e}}_i$ tends to the scaled harmonic polynomial times $\frac{\sigma_s}{\varepsilon}$. Because these functions are still solutions to (41) one can write them when $\sigma_a = 0$ under the form

$$\tilde{\mathbf{e}}_k = \begin{pmatrix} \frac{\sigma_s}{\varepsilon} q_k \\ -c\nabla q_k \end{pmatrix}, \quad k = 1, ..., m.$$

This completes the proof. □

The proof of the main Theorem of convergence 1.2 attached to these basis functions is postponed in Chapter 7.

6 Numerical results

The goal of this section is to validate the convergence and asymptotic behavior of the scheme on some numerical examples in one and two dimensions for stationary and time dependent problems. We will consider two regimes: the case $\varepsilon = 1$ and the case $\varepsilon << 1$.

6.1 One dimensional time dependent tests

We use random meshes made of N nodes, where the vertices are moved randomly around their initial position by a factor of at most 33%.

6.1.1 Study of the order

For the time dependent P_1 model in one dimension (27) consider the case $\Omega_S = [0, 1]$, $\varepsilon = 1, c = 1, \sigma_a = 1, \sigma_s = 1$, $h = 1/N$ for $N = 20, 40, 60, 80, 100$, $T = 0.024$ and $dt = T/N$. The exact solution is $\mathbf{u}_{ex} = (e^{-t}, e^{-2t})$ and we set $M^-\mathbf{u} = M^-\mathbf{u}_{ex}$ on the boundary. The functions (29) are used as basis functions.

We study two cases: a first one with only the two stationary basis functions e_1^-, e_1^+ and a second one with four basis functions $e_1^-, e_1^+, e_2^-, e_2^+$. Figure 2 shows that the scheme is convergent with the two basis functions e_1^-, e_1^+ and that one increases the order by adding the basis functions e_2^-, e_2^+. More precisely, order 1 is achieved with the two basis functions e_1^-, e_1^+ whereas order 2 is achieved with the four basis functions $e_1^-, e_1^+, e_2^-, e_2^+$.

6.1.2 Asymptotic preserving regime

We test the asymptotic behavior of the scheme (33) for the model problem (31). Naive schemes need many degrees of freedom, and therefore an important computational time, to be able to capture the correct diffusion limit when $\varepsilon \to 0$. The so called asymptotic preserving schemes have been designed [16, 26] to get the correct limit with a reasonable amount of degrees of freedom. We have shown in Section 4 that the TDG method leads to a new asymptotic preserving scheme and we can now illustrate this property. To this end we take $\Omega_S = [0, 1]$, $\varepsilon = 0.001$, $\sigma_s = 1$, $c = 1$ and $T = 0.01$. Consider p_0 the fundamental solution to the heat equation and the variable v_0 associated in the limit $\varepsilon \to 0$

$$p_0(x, t) = \frac{1}{2\sqrt{\pi t}} e^{\frac{-(x-0.5)^2}{4t}}, \quad v_0(x, t) = -\varepsilon\partial_x p_0(x, t).$$

Finally $M^-(p, v)^T = M^-(p_0, v_0)^T$ is imposed on the boundary. Figure 3 compares the numerical solution with $(p_0, v_0)^T$. One sees that even with few degrees of freedom the solution is correctly approximated.

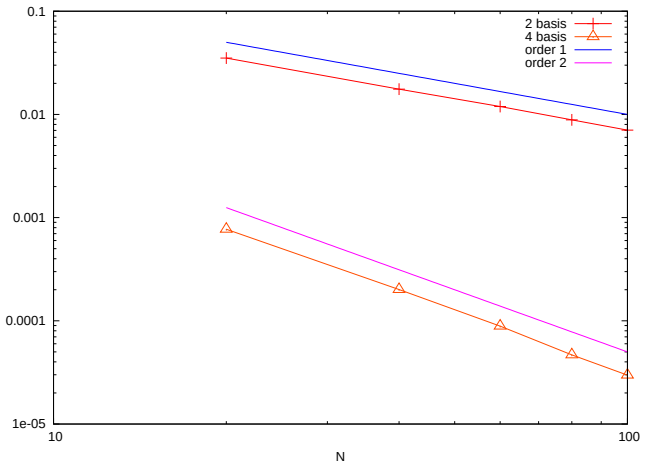


Figure 2: Study of the L^2 error on the final time step in logarithmic scale for temporal one dimensional model. Error with the two stationary basis functions and the four basis functions. Random meshes.

6.2 Two dimensional tests

We now consider two dimensional model. Meshes made of random quads are using. A random quad mesh is made of $N \times N$ quads, $N \in \mathbb{N}^*$, where the vertices are move randomly around their initial position by a factor of at most 33%.

6.2.1 $2D$ convergence with absorption

Consider the stationary P_1 model in two dimensions (61). Let $\mathbf{x} = (x, y)^T, \Omega_S = [0, 1]^2$, $\varepsilon = 1, c = 1, \sigma_a = 1, \sigma_s = 1$. The exact solution we consider here is

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

We assume $M^-\mathbf{u} = M^-\mathbf{u}_{ex}$ is imposed on the boundary and consider $m \in \mathbb{N}$ basis functions as in (42)

$$\mathbf{e}_k(\mathbf{x}) = (\sqrt{2}, \mathbf{d}_k)e^{\sqrt{2}(\mathbf{d}_k, \mathbf{x})}, \quad k = 1, ..., m,$$

with $\mathbf{d}_k = (\cos(\phi_k), \sin(\phi_k))^T, \phi_k = 2(k-1)\pi/m$.

Results obtained with $3, 5$ and 7 basis functions are displayed on the left of Figure 4. As stated in proposition 7.12, one only needs two additional basis functions to increase the order by a factor 1. Note however that the orders obtain here are slightly better than those predicted in proposition 7.12: with 3, 5 and 7 basis functions one gets respectively order 0.8, 1.5 and 2.5.

6.2.2 $2D$ convergence without absorption

Consider the stationary P_1 model in two dimensions (61). Consider the same parameters as before but without absorption: $\mathbf{x} = (x, y)^T, \Omega_S = [0, 1]^2$, $\varepsilon = 1, c = 1, \sigma_a = 0, \sigma_s = 1$. The exact solution is

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

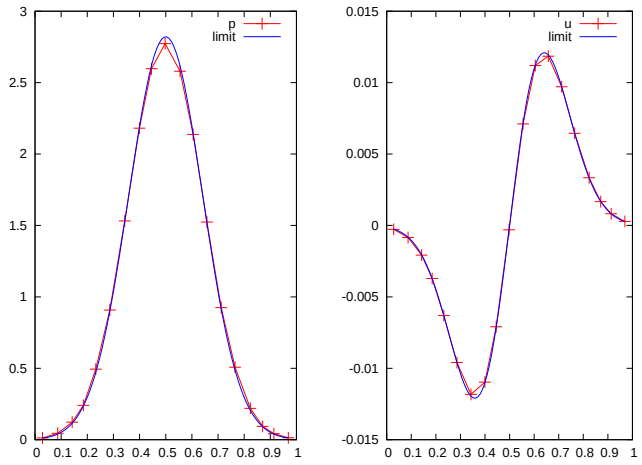


Figure 3: Numerical solution obtained with the numerical scheme (33) with $\varepsilon = 0.001$ to p_0 (left) and v_0 (right). Random mesh with 20 nodes and $dt = 0.01/20$. Good accuracy illustrate the AP properties of the TDG scheme.

Again assume $M^- \mathbf{u} = M^- \mathbf{u}_{ex}$ is imposed on the boundary and consider $m \in \mathbb{N}$ basis functions as in (42)

$$\mathbf{e}_k(\mathbf{x}) = (\sqrt{2}, \mathbf{d}_k) e^{\sqrt{2}(\mathbf{d}_k, \mathbf{x})}, \quad k = 1, ..., m,$$

with $\mathbf{d}_k = (\cos(\phi_k), \sin(\phi_k))^T, \phi_k = 2(k-1)\pi/m$.

Results obtained with $3, 5$ and 7 basis are displayed on the right of Figure 4. The orders are very close to those obtained in the case $\sigma_a \neq 0$ (left of the Figure 4): with 3, 5 and 7 basis functions one respectively gets order 0.5, 1.5 and 2.5.

6.2.3 Boundary layers in two dimensions

We study the stationary P_1 model in two dimensions with discontinuous coefficients. The domain is $\Omega = [0, 1]^2$ and we define Ω_1 (resp. Ω_2) as $\Omega_1 = [0.35, 0.65]^2$ (resp. $\Omega_2 = \Omega \setminus \Omega_1$). The geometry is represented in Figure 5. We take $\epsilon = 1$ and $c = \frac{1}{\sqrt{3}}$. The absorption coefficient $\sigma_a = 2 \times \mathbf{1}_{\Omega_1}(\mathbf{x})$ has compact support in Ω_1. The scattering coefficient $\sigma_s = 2 \times \mathbf{1}_{\Omega_2}(\mathbf{x}) + 10^5 \times \mathbf{1}_{\Omega_1}(\mathbf{x})$ is discontinuous and takes a high value in Ω_1. Even if we consider a random mesh, the interface between Ω_1 and Ω_2 is a straight line.

To show why it can be challenging for standard schemes to capture boundary layers we compare the TDG method with the DG scheme with affine basis functions (that is 1, x, y). Since the P_1 model has 3 components this gives us a total of 9 basis functions per cell for the DG scheme. For the TDG scheme we take only 3 or 5 basis functions per cell. Note that for the TDG method one must choose the directions of the basis functions in Ω_1 since $\sigma_a > 0$. As we will see this choice plays an important role to correctly capture the boundary layers and it seems essential to locally get the one dimensional direction perpendicular to the interface.

Both DG and TDG converge to the same asymptotic solution for thinner and thinner meshes. The 2D asymptotic solution represented in Figure 6 is calculated on a 200×200 mesh with the TDG method with 5 basis functions per cell (except at the interface see below). The default equi-distributed

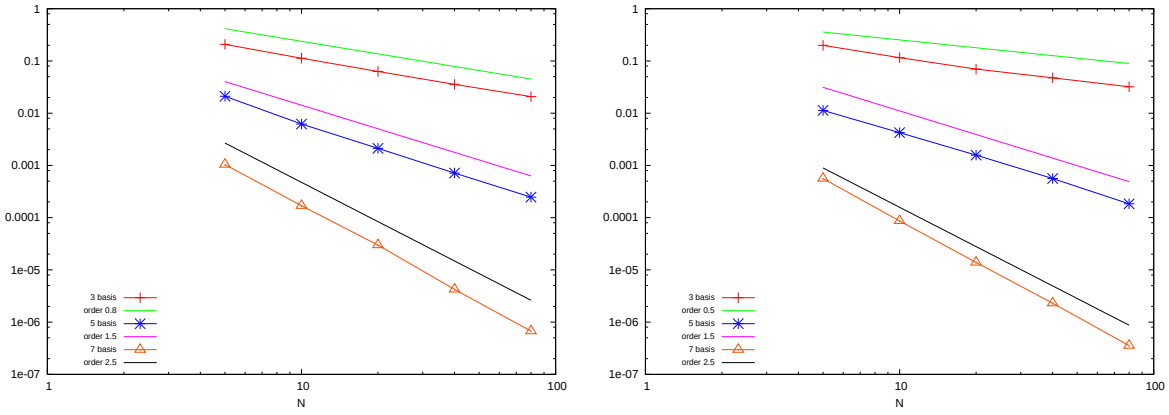


Figure 4: Case $\sigma_a = 1$ on the left and $\sigma_a = 0$ on the right. L^2 error in logarithmic scale of the TDG method for the stationary two dimensional P_1 model. 3 basis functions (red), 5 basis functions (blue) and 7 basis functions (orange). Random meshes.

directions in Ω_1 are

$$\begin{aligned} \mathbf{d}_1 = (1,0)^T, \quad \mathbf{d}_2 = (\cos\frac{2\pi}{5}, \sin\frac{2\pi}{5})^T, \quad \mathbf{d}_3 = (\cos\frac{4\pi}{5}, \sin\frac{4\pi}{5})^T, \\ \mathbf{d}_4 = (\cos\frac{6\pi}{5}, \sin\frac{6\pi}{5})^T, \quad \mathbf{d}_5 = (\cos\frac{8\pi}{5}, \sin\frac{8\pi}{5})^T. \end{aligned} \tag{47}$$

At the interface in Ω_1 we make a special choice of directions

$$\mathbf{d}_1 = (1,0)^T, \quad \mathbf{d}_2 = (0,1)^T, \quad \mathbf{d}_3 = (-1,0)^T, \quad \mathbf{d}_4 = (0,-1)^T. \tag{48}$$

These directions are well adapted if one considers a one dimensional problem at the interface. For example on a 20×20 mesh there are 36 cells in Ω_1 and, among those 36 cells, there are 20 cells with at least an edge which belongs to the interface. The directions (48) are taken in those 20 cells and the directions (47) everywhere else. We will also study the TDG method with only 3 basis functions per cell. With 3 basis functions per cell we consider the following equi-distributed directions

$$\mathbf{d}_1 = (1,0)^T, \quad \mathbf{d}_2 = (\cos\frac{2\pi}{3}, \sin\frac{2\pi}{3})^T, \quad \mathbf{d}_3 = (\cos\frac{4\pi}{3}, \sin\frac{4\pi}{3})^T. \tag{49}$$

We compare the DG and TDG methods on a coarse 20×20 mesh.

In Figure 6, we represent the variable p. For the TDG method we take either 3 or 5 basis functions except at the interface in Ω_1 where we use the 4 directions (48). One observes that the boundary layer is not correctly captured by the DG scheme. The approximation given by the TDG scheme seems more accurate.

In Figure 7, we take a one dimensional cut at $y = 0.5$ to compare more precisely the numerical results. The graphic on the left shows that the TDG gives indeed a much better approximation than the DG method especially with 5 basis functions per cell. Our interpretation is that it is because the boundary layer is correctly captured by TDG but poorly captured by DG.

The graphic on the right of Figure 7 illustrates why it is very important to use the directions (48) at the interface to obtained a satisfactory discretization of the boundary layer on a coarse mesh. We consider the TDG method with 5 basis functions per cell and compare two cases. In the first one the directions are (47) in all cells of Ω_1. In the second one the directions (48) are used at the interface. The

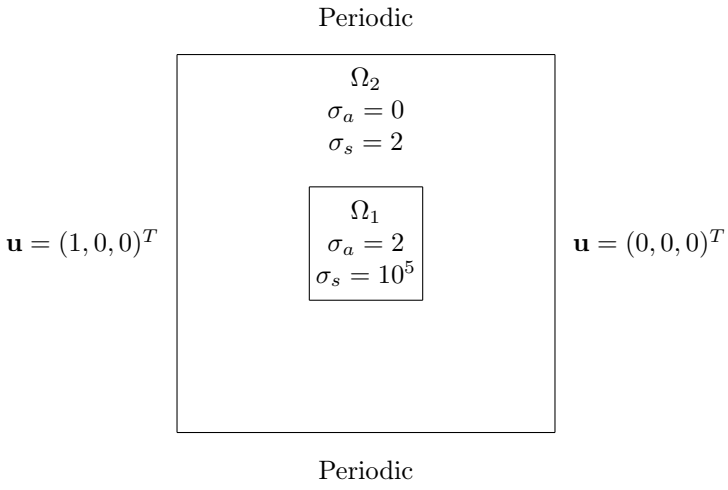


Figure 5: Domain and boundary condition for the two dimensional boundary layers test.

graphic shows that the TDG method gives a non correct approximation with only the directions (47). However if one locally adapts the directions at the interface the TDG method recovers a very good accuracy. Once again, our interpretation is that it is because the boundary layer is correctly captured with these parameters.

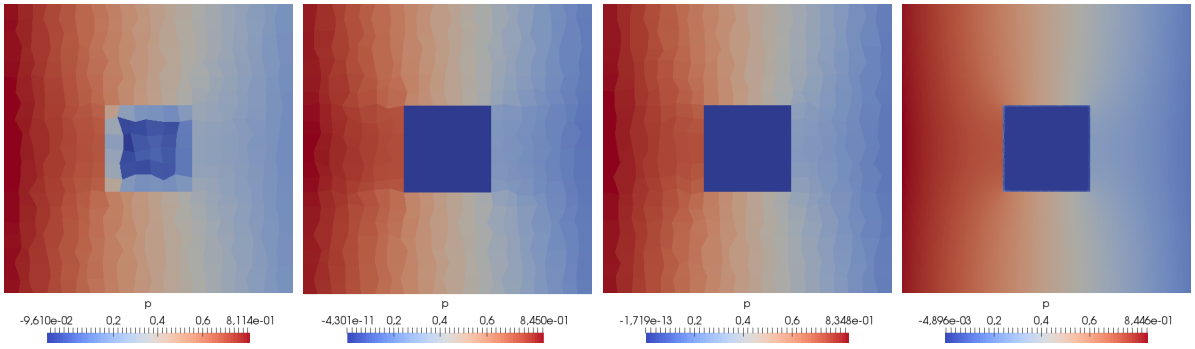


Figure 6: Representation of the variable p for the test case 6.2.3. From left to right: DG scheme with 9 basis functions per cell, TDG scheme with 3 basis functions per cell, TDG scheme with 5 basis functions per cell and reference solution. For the TDG method the directions at the interface in Ω_1 are locally adapted into the 4 directions (48).

6.2.4 Asymptotic preserving study for time dependent model

We study here the asymptotic behavior of the TDG method in the case $\sigma_a = 0$ and consider the test case from [5] for the time dependent P_1 model (41). Let $\mathbf{x} = (x, y)^T, \Omega_S = [0,1]^2$, $T = 0.036$, $\sigma_a = 0, \sigma_s = 1, c = 1$, and consider the solution

$$p_0 = f + \frac{\varepsilon^2}{\sigma_s}\partial_t f, \quad \mathbf{u}_0 = -\frac{\varepsilon}{\sigma_s}\nabla f,$$

with

$$f(t, \mathbf{x}) = \alpha(t)\cos(2\pi x)\cos(2\pi y),$$

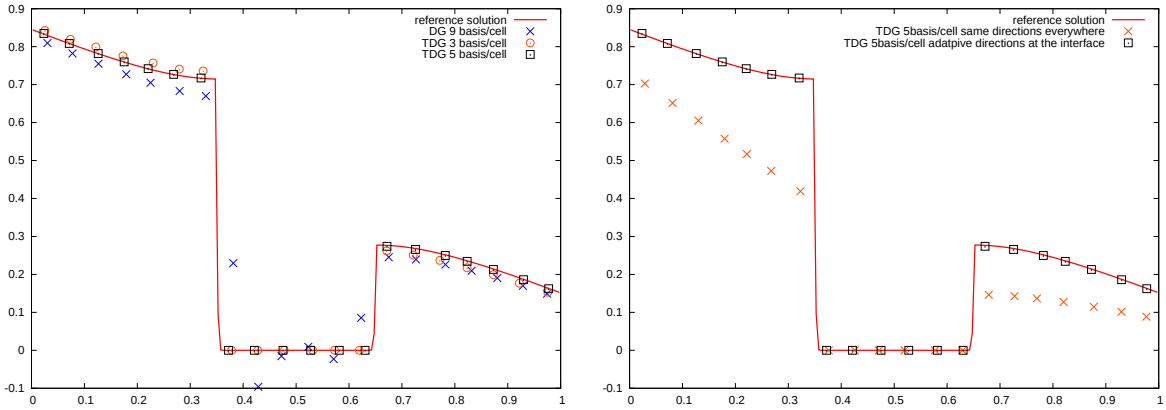


Figure 7: One dimensional representation of the variable p at $y = 0.5$ for the test case 6.2.3. Left: comparison between the DG method with 9 basis/cell (cross), the TDG method with 3 basis/cell (circle) and the TDG method with 5 basis/cell (square). In both cases the directions at the interface in Ω_1 are locally adapted into the 4 directions (48). Right: comparison between the TDG method with directions (47) only (cross) and the TDG method where the directions at the interface in Ω_1 are locally adapted into the 4 directions (48) (square).

with $\alpha(t)$ defined as

$$\alpha(t) = \frac{\lambda_2}{\lambda_2 - \lambda_1} e^{\lambda_1 t} - \frac{\lambda_1}{\lambda_2 - \lambda_1} e^{\lambda_2 t},$$

$$\lambda_1 = -\frac{\sigma_s\left(\sqrt{1 - \frac{\varepsilon^2}{\sigma_s^2} 32\pi^2} + 1\right)}{2\varepsilon^2}, \quad \lambda_2 = -\frac{\sigma_s\left(\sqrt{1 - \frac{\varepsilon^2}{\sigma_s^2} 32\pi^2} - 1\right)}{2\varepsilon^2}.$$

One can check that $(p_0, \mathbf{u}_0)$ is indeed a solution to (41) with $\sigma_a = 0$, see [5] for details. An exact relation is enforced between ε and the space step $h = \frac{1}{N}$. The relation between ε and h reads $\varepsilon = 0.01(40h)^\tau$ for $\tau \in \{0, \frac{1}{4}, \frac{1}{2}, 1, 2\}$. The error between the exact solution and the numerical solution is computed numerically in function of h for different values of τ. The result is displayed in Figure 8 for 3 stationary basis functions (44) and $dt = 0.36h^2$. One observes the convergence of the solution even for small values of ε.

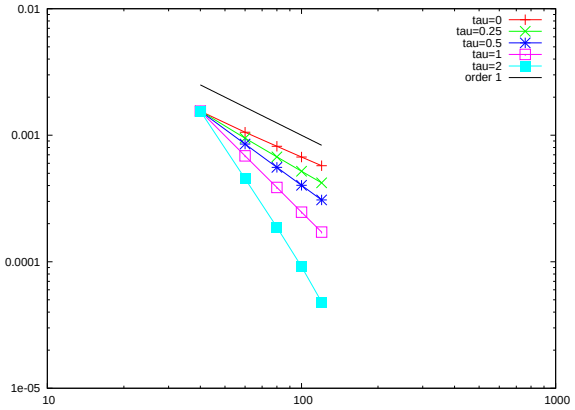
7 Proof of theorem 1.2 and h-convergence

First we consider the simpler case of the particular second order equation $\Delta u = \omega u$ which is closely related to the Helmholtz equation. This will then be generalized to study the approximation properties of stationary solutions to the P_1 model. Approximation results using solutions to the Helmholtz equation has already been studied in different ways. For the h version see [7] for the case $\omega < 0$ and [14] for the case $\omega \leq 0$ with a source term and more explicit constants. For p version estimate using Vekua theory see [18, 31] and [19] for the hp version.

7.1 Technical material

Let $u \in H^1(\Omega)$. We consider the following auxiliary second order equation

$$\Delta u = \omega u, \tag{50}$$



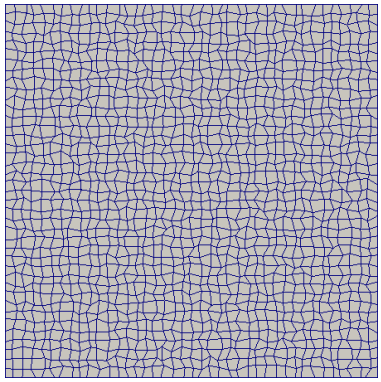


Figure 8: On the left: study of the L^2 error at the final time in logarithmic scale. TDG method for $\varepsilon = 0.01(40h)^\tau$ with $\tau = 0$ (red), $\tau = 0.25$ (green), $\tau = 0.5$ (dark blue) $\tau = 1$ (purple) and $\tau = 2$ (light blue). On the right: an example of random mesh in $2D$.

with $\omega \in \mathbb{R}$ which may take positive or negative values and our goal is to write a simplified Taylor expansion for regular solutions to this equation. Let $\mathbf{x} = (x, y)^T$ and fix $n \in \mathbb{N}$ and $\mathbf{x}_0 = (x_0, y_0)^T \in \Omega$. We note $T_k^p(\mathbf{x}) := \frac{C_k^p}{k!}(x - x_0)^p (y - y_0)^{k-p}$ for $0 \leq p \leq k$ and $T_k^p(\mathbf{x}) := 0$ in other cases. Every function $u \in C^{n+1}(\Omega)$ can be written under the form of a usual Taylor-Cauchy expansion which comes from [10, page 94]

$$u(\mathbf{x}) = \sum_{k=0}^{n} \sum_{p=0}^{k} \partial_x^p \partial_y^{k-p} u(\mathbf{x}_0) T_k^p(\mathbf{x}) + \sum_{p=0}^{n+1} \partial_x^p \partial_y^{n+1-p} u(\mathbf{x}_s) T_{n+1}^p(\mathbf{x}), \tag{51}$$

where $\mathbf{x}_s = (x_s, y_s)^T$, $x_s = (1 - s)x_0 + sx$, $y_s = (1 - s)y_0 + sy$, $s \in [0, 1]$. There is of course a double sum in the Taylor expansion, but for Trefftz methods it is possible to reduce the complexity using the fact that u is a solution to the model equation (50). This is classical [7, 19, 28] see also [24, 22, 23] with a different approach to the coefficients reduction procedure. In our analysis, we need intermediate quantities named α_k^p and β_k^p.

Definition 7.1. *Consider an integer $n \geq 0$. The functions α_k^p and β_k^p are defined in the range $0 \leq p \leq k \leq n$ by a decreasing recursion from $k = n$ to $k = 0$. The recursion reads:*

- *by convention set $\beta_{n+1}^p(\mathbf{x}) = \beta_{n+2}^p(\mathbf{x}) = 0$, $\beta_k^{-1}(\mathbf{x}) = \beta_k^{-2}(\mathbf{x}) = 0$, $\forall p, k$*
- *for $k = n$ to $k = 0$, do*
  - *for $p = 0$ to $p = k$, do*

$$\alpha_k^p(\mathbf{x}) := T_k^p(\mathbf{x}) + \omega \beta_{k+2}^p(\mathbf{x}), \tag{52}$$

$$\beta_k^p(\mathbf{x}) := \alpha_k^p(\mathbf{x}) - \beta_k^{p-2}(\mathbf{x}), \tag{53}$$

A graphical illustration of the procedure is provided in Figure 9.

Since $\beta_{n+1}^p(\mathbf{x}) = \beta_{n+2}^p(\mathbf{x}) = 0$, thus $\alpha_{n-1}^p(\mathbf{x}) = T_{n-1}^p(\mathbf{x})$, $\alpha_n^p(\mathbf{x}) = T_n^p(\mathbf{x})$. Also because $\beta_k^{-2} = \beta_k^{-1} = 0$ the equality (53) implies

$$\beta_k^0 = \alpha_k^0, \quad \beta_k^1 = \alpha_k^1, \quad 0 \leq k \leq n. \tag{54}$$

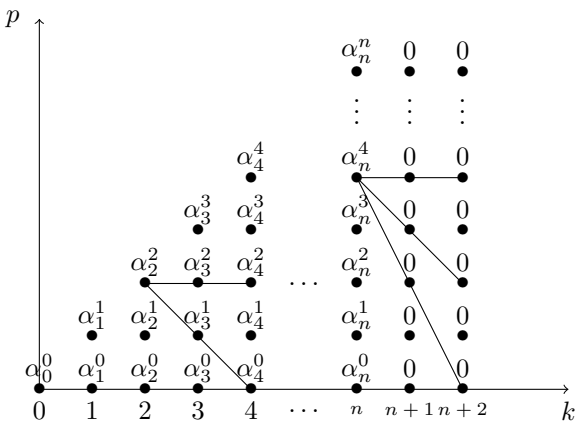


Figure 9: Dependence of the coefficients α_2^2 and α_n^4 in terms of the coefficients α_k^p for $\omega \neq 0$. The figure shows that α_k^p depends only on some coefficients α_{k+2}^j for $k \leq n$, $0 \leq j \leq k+2$.

In the case $\omega \neq 0$, the functions $\alpha_k^p(\mathbf{x})$ and $\beta_k^p(\mathbf{x})$ are polynomials of degree n if both n and k are even or odd and of degree $n-1$ otherwise. If $\omega = 0$ the functions $\alpha_k^p(\mathbf{x})$ and $\beta_k^p(\mathbf{x})$ are polynomials of degree k for $0 \leq k \leq n$. Note that in order to use simple notation we do not explicitly write the dependence of these functions in n and $\mathbf{x}_0$.

Proposition 7.2. *Assume $u \in C^{n+1}(\Omega)$ is solution to (50). Then the double sum Taylor expansion (51) can be recast as a simple sum with only zero or first order derivatives with respect to y*

$$\begin{aligned} u(\mathbf{x}) = u(\mathbf{x}_0)\beta_0^0(\mathbf{x}) + \sum_{k=1}^{n} \left[\partial_x^k u(\mathbf{x}_0)\beta_k^k(\mathbf{x}) + \partial_x^{k-1}\partial_y u(\mathbf{x}_0)\beta_k^{k-1}(\mathbf{x})\right] \\ + \sum_{p=0}^{n+1} \partial_x^p \partial_y^{n+1-p} u(\mathbf{x}_s) T_{n+1}^p(\mathbf{x}), \quad \forall \mathbf{x} \in \Omega, \end{aligned} \tag{55}$$

where $\mathbf{x}_s = (x_s, y_s)^T$, $x_s = (1-s)x_0 + sx$ and $y_s = (1-s)y_0 + sy$.

By symmetry, a similar result holds with high order derivative with respect to y and only zero and first order derivatives for respect to x. The proof which is purely technical is postponed to the appendix.

7.2 Approximation properties of auxiliary solutions to the equation (50)

To study the approximation properties of solutions to the equation (50) we will need the following matrix. Let $n \in \mathbb{N}$ and consider $2n+1$ functions $e_1, e_2, ..., e_{2n+1} \in W^{n,\infty}(\Omega)$. We define $S_{2n+1} := S_{e_1,e_2,...,e_{2n+1}} \in \mathbb{R}^{2n+1 \times 2n+1}$ such that

$$S_{2n+1} := S_{e_1,e_2,...,e_{2n+1}} := \begin{pmatrix} e_1 & e_2 & ... & e_{2n+1} \\ \partial_x e_1 & \partial_x e_2 & ... & \partial_x e_{2n+1} \\ \partial_y e_1 & \partial_y e_2 & ... & \partial_y e_{2n+1} \\ \partial_x^2 e_1 & \partial_x^2 e_2 & ... & \partial_x^2 e_{2n+1} \\ \partial_x\partial_y e_1 & \partial_x\partial_y e_2 & ... & \partial_x\partial_y e_{2n+1} \\ \vdots & \vdots & ... & \vdots \\ \partial_x^n e_1 & \partial_x^n e_2 & ... & \partial_x^n e_{2n+1} \\ \partial_x^{n-1}\partial_y e_1 & \partial_x^{n-1}\partial_y e_2 & ... & \partial_x^{n-1}\partial_y e_{2n+1} \end{pmatrix}.$$

For Θ a generic open set we will use the norm $\|u\|_{W^{n,\infty}(\Theta)} = \sum_{k=0}^{n} \sum_{p=0}^{k} \sup_{\mathbf{x}\in\Theta} |\partial_x^p \partial_y^{k-p} u(\mathbf{x})|$. In the vectorial case it is $\|\mathbf{u}\|_{W^{n,\infty}(\Theta)} = \sum_{j=1}^{m} \|u_j\|_{W^{n,\infty}(\Theta)}$.

Proposition 7.3. *Let $n \in \mathbb{N}$, $\mathbf{x}_0 \in \mathbb{R}^2$, assume $e_1, e_2, ..., e_{2n+1} \in W^{n+1,\infty}(\Omega)$ and $u \in W^{n+1,\infty}(\Omega)$ are solutions to the equation (50). If the matrix $S_{2n+1}(\mathbf{x}_0)$ is invertible then there exists real numbers $\mathbf{a} = (a_1, a_2, ..., a_{2n+1})^T \in \mathbb{R}^{2n+1}$ and a constant $C > 0$ such that*

$$\left\| \sum_{i=1}^{2n+1} a_i e_i - u \right\|_{L^\infty(\Omega_k)} \leq C h^{n+1} \|u\|_{W^{n+1,\infty}(\Omega)}, \qquad h = diam(\Omega_k).$$

and

$$\left\| \nabla \left(\sum_{i=1}^{2n+1} a_i e_i - u \right) \right\|_{L^\infty(\Omega_k)} \leq C h^{n} \|u\|_{W^{n+1,\infty}(\Omega)}, \qquad h = diam(\Omega_k).$$

Proof. Because the solutions $e_i, 1 \leq i \leq 2n+1$ and u are in $W^{n+1,\infty}(\Omega)$, one can write them under the form (55). Let

$$\mathbf{b} = (u(\mathbf{x}_0), \partial_x u(\mathbf{x}_0), \partial_y u(\mathbf{x}_0), ..., \partial_x^n u(\mathbf{x}_0), \partial_x^{n-1} \partial_y u(\mathbf{x}_0))^T \tag{56}$$

and consider the solution of the linear system $S_{2n+1}(\mathbf{x}_0)\mathbf{a} = \mathbf{b}$. The expansion (55) implies

$$\sum_{i=1}^{2n+1} a_i e_i(\mathbf{x}) - u(\mathbf{x}) = \sum_{p=0}^{n+1} \partial_x^p \partial_y^{n+1-p} w(\mathbf{x}_s) T_{n+1}^p(\mathbf{x}), \quad w = \sum_{i=1}^{2n+1} a_i e_i - u. \tag{57}$$

Since T_{n+1}^p is a difference to the power $n+1$, one immediately gets

$$\left\| \sum_{i=1}^{2n+1} a_i e_i - u \right\|_{L^\infty(\Omega_k)} \leq C \|w\|_{W^{n+1}(\Omega_k)} h^{n+1}.$$

Additionally the triangular inequality yields $\|w\|_{W^{n+1,\infty}(\Omega_k)} \leq \sum_{i=1}^{2n+1} |a_i| \|e_i\|_{W^{n+1,\infty}(\Omega_k)} + \|u\|_{W^{n+1,\infty}(\Omega_k)}$ where the coefficients a_i are bounded by $\|u\|_{W^{n+1,\infty}(\Omega_k)}$ as a consequence of (56) and the basis functions $\mathbf{e}_i$ are bounded by a constant. So $\|w\|_{W^{n+1,\infty}(\Omega_k)} \leq C \|u\|_{W^{n+1,\infty}(\Omega_k)}$ up to the redefinition of the constant. From (57) one deduces the second inequality. This completes the proof. □

We now consider some specific cases with non negative constant ω and study the invertibility of the matrix S_{2n+1}. First assume $\omega > 0$.

Proposition 7.4. *Let $n \in \mathbb{N}$, $\omega > 0$, ω constant, and consider the functions $e_1, ..., e_{2n+1}$*

$$e_i(\mathbf{x}) = e^{\sqrt{\omega}(\mathbf{d}_i, \mathbf{x})}, \quad i = 1, ..., 2n+1, \tag{58}$$

with $\mathbf{d}_i = (\cos(\theta_i), \sin(\theta_i))^T$, $\theta_i \in [0, 2\pi[$ and $\theta_i \neq \theta_j \ \forall i \neq j$. The functions e_i are solutions to the equation (50) and the matrix $S_{2n+1}(\mathbf{x})$ is invertible for all $\mathbf{x} \in \mathbb{R}^2$.

Proof. It is easy to check that the functions (58) are solutions to the equation (50) when ω is constant and positive. It remains to show that the matrix S_{2n+1} is invertible. For simplicity we consider centered solutions

$$e_i(\mathbf{x}) = e^{\sqrt{\omega}(\mathbf{d}_i, \mathbf{x}-\mathbf{x}_0)}, \tag{59}$$

with $\mathbf{x}_0 \in \mathbb{R}^2$. Multiplying each columns of S_{2n+1} by a positive constant does not change whether the determinant of S_{2n+1} is null or not. Doing so the matrix S_{2n+1} is recast with slight abuse of notation

as

$$S_{2n+1} = \begin{pmatrix} 1 & 1 & ... & 1 \\ \omega^{\frac{1}{2}}\cos(\theta_1) & \omega^{\frac{1}{2}}\cos(\theta_2) & ... & \omega^{\frac{1}{2}}\cos(\theta_{2n+1}) \\ \omega^{\frac{1}{2}}\sin(\theta_1) & \omega^{\frac{1}{2}}\sin(\theta_2) & ... & \omega^{\frac{1}{2}}\sin(\theta_{2n+1}) \\ \vdots & \vdots & ... & \vdots \\ \omega^{\frac{n}{2}}\cos^n(\theta_1) & \omega^{\frac{n}{2}}\cos^n(\theta_2) & ... & \omega^{\frac{n}{2}}\cos^n(\theta_{2n+1}) \\ \omega^{\frac{n}{2}}\sin(\theta_1)\cos^{n-1}(\theta_1) & \omega^{\frac{n}{2}}\sin(\theta_2)\cos^{n-1}(\theta_2) & ... & \omega^{\frac{n}{2}}\sin(\theta_{2n+1})\cos^{n-1}(\theta_{2n+1}) \end{pmatrix}.$$

We recall the equalities

$$\cos^n(x) = \frac{1}{2^{n-1}}\Big(\sum_{k=0}^{\lfloor\frac{n}{2}\rfloor} C_n^k \cos\big((n-2k)x\big) - \frac{C_n^{\lfloor\frac{n}{2}\rfloor}}{2}\big(\frac{n+1}{2} - \lfloor\frac{n+1}{2}\rfloor\big)\Big), \quad n \in \mathbb{N}^*,$$

$$\sin(x)\cos(nx) = \frac{1}{2}\Big(\sin((n+1)x) - \sin((n-1)x)\Big), \quad n \in \mathbb{N}.$$

Therefore each row of S_{2n+1} can be written as the corresponding row of $M_{2n+1} = M_{\theta_1,...,\theta_{2n+1}}$ multiplied by a positive coefficient and a linear combination of its previous rows. Since the matrix M_{2n+1} is invertible (proposition 5.3), the matrix S_{2n+1} is also invertible. This completes the proof. □

Now consider the case $\omega = 0$.

Proposition 7.5. *Let $n \in \mathbb{N}$, $\omega = 0$ and consider the functions $e_l = q_l$ for $l \geq 1$. These functions are solutions to the equation (50) and the matrix $S_{2n+1}(\mathbf{x})$ is invertible for all $\mathbf{x} \in \mathbb{R}^2$.*

Proof. By definition harmonic polynomials are solutions to the equation (50) when $\omega = 0$. For these solutions

$$S_{2n+1} = \begin{pmatrix} 1 & \Re(x+iy)^1 & \Im(x+iy)^1 & ... & \frac{2^{1-n}}{n!}\Re(x+iy)^n & \frac{2^{1-n}}{n!}\Im(x+iy)^n \\ 0 & \partial_x\Re(x+iy)^1 & \partial_x\Im(x+iy)^1 & ... & \frac{2^{1-n}}{n!}\partial_x\Re(x+iy)^n & \frac{2^{1-n}}{n!}\partial_x\Im(x+iy)^n \\ 0 & \partial_y\Re(x+iy)^1 & \partial_y\Im(x+iy)^1 & ... & \frac{2^{1-n}}{n!}\partial_y\Re(x+iy)^n & \frac{2^{1-n}}{n!}\partial_y\Im(x+iy)^n \\ \vdots & \vdots & \vdots & ... & \vdots & \vdots \\ 0 & \partial_x^n\Re(x+iy)^1 & \partial_x^n\Im(x+iy)^1 & ... & \frac{2^{1-n}}{n!}\partial_x^n\Re(x+iy)^n & \frac{2^{1-n}}{n!}\partial_x^n\Im(x+iy)^n \\ 0 & \partial_x^{n-1}\partial_y\Re(x+iy)^1 & \partial_x^{n-1}\partial_y\Im(x+iy)^1 & ... & \frac{2^{1-n}}{n!}\partial_x^{n-1}\partial_y\Re(x+iy)^n & \frac{2^{1-n}}{n!}\partial_x^{n-1}\partial_y\Im(x+iy)^n \end{pmatrix}.$$

One has $(x+iy)^k = \sum_{p=0}^k C_k^p (i)^{k-p} x^p y^{k-p}$, thus

$$\partial_x^k\Re(x+iy)^k = k!, \quad \partial_x^{k+1+l}\Re(x+iy)^k = 0, \quad \partial_x^{k-1+l}\partial_y\Re(x+iy)^k = 0, \quad \text{for all } l \in \mathbb{N},$$

and

$$\partial_x^{k-1}\partial_y\Im(x+iy)^k = C_k^1(k-1)!, \quad \partial_x^{k+l}\Im(x+iy)^k = 0, \quad \partial_x^{k+l}\partial_y\Re(x+iy)^k = 0, \quad \text{for all } l \in \mathbb{N}.$$

One deduces that the matrix S_{2n+1} is an upper triangular matrix with positive diagonal coefficients and is therefore invertible. This completes the proof. □

We can also proceed as in [14] and study stable basis that degenerate correctly when $\omega \to 0$.

Definition 7.6. *Let $n \in \mathbb{N}$, $\omega > 0$, ω constant, and $a_{k,j} = (M_{\theta_1,...,\theta_{2n+1}})^{-1}_{k,j}$. We define the following functions*

$$\tilde{e}_j = (\sqrt{\omega})^{-\lfloor\frac{j}{2}\rfloor} \sum_{k=1}^{2n+1} a_{k,j} e_k, \quad j = 1, ..., 2n+1, \tag{60}$$

with e_k defined as in (58).

These functions are stable in the sense that they tend to harmonic polynomials when $\sigma_a \to 0$.

Proposition 7.7. *We denote q_j, $j = 1, ..., 2n+1$ the first $2n+1$ scaled harmonic polynomials (43). One has*

$$\tilde{e}_j(\mathbf{x}) \underset{\omega \to 0}{\to} q_j(\mathbf{x}), \quad j = 1, ..., 2n+1.$$

Proof. The convergence is uniform on compact sets, see [14, Section 3.1]. □

By continuity one can therefore write $\tilde{e}_j = q_j$ if $\omega = 0$. With the solutions $\tilde{e}_i$, we study the invertibility of S in the case $\omega \geq 0$.

Proposition 7.8. *Let $n \in \mathbb{N}$, $\omega \geq 0$, ω constant and consider the functions $\tilde{e}_1, ..., \tilde{e}_{2n+1}$ in (60). The matrix $S_{\tilde{e}_1,...,\tilde{e}_{2n+1}}$ is invertible in $\mathbb{R}^2$.*

Proof. Let $\tilde{M} \in \mathbb{R}^{2n+1 \times 2n+1}$ be defined as $\tilde{M}_{k,j} = (\omega)^{-\lfloor \frac{j}{2} \rfloor} a_{k,j}$ where $a_{k,j}$ are the coefficients of the matrix $(M_{2n+1})^{-1}$. Since the matrix $(M_{2n+1})^{-1}$ is invertible the matrix $\tilde{M}$ is also invertible for all $\omega > 0$. From the definition of the functions $\tilde{e}_1, ..., \tilde{e}_{2n+1}$ and the definition of the matrix S one has

$$S_{\tilde{e}_1,...,\tilde{e}_{2n+1}} = S_{e_1,...,e_{2n+1}} \tilde{M}.$$

The matrix $S_{e_1,...,e_{2n+1}}$ and $\tilde{M}$ are both invertible for $\omega > 0$ therefore $S_{\tilde{e}_1,...,\tilde{e}_{2n+1}}$ is also invertible for all $\omega > 0$. Moreover for $\omega = 0$ the solutions $\tilde{e}_j$ are the scaled harmonic polynomials. From proposition 7.5 one gets the invertibility of the matrix $S_{\tilde{e}_1,...,\tilde{e}_{2n+1}}$ when $\omega = 0$. This completes the proof. □

7.3 Proof of theorem 1.2

We study the approximation properties of solutions to the stationary P_1 model. For simplicity we take $c = 1$ and assume the coefficients σ_a and σ_s are constants. The stationary P_1 model (41) reads

$$\begin{cases} \partial_x v_1 + \partial_y v_2 = -\varepsilon \sigma_a p, \\ \partial_x p = -\frac{1}{\varepsilon} \widetilde{\sigma}_t^{\varepsilon} v_1, \\ \partial_y p = -\frac{1}{\varepsilon} \widetilde{\sigma}_t^{\varepsilon} v_2 \end{cases} \tag{61}$$

where we note $\widetilde{\sigma}_t^{\varepsilon} = \varepsilon^2 \sigma_a + \sigma_s$, which still depends on ε, and assume $\sigma_a + \sigma_s > 0$. For convenience the unknown will be rewritten as $\mathbf{u} = (u_1, u_2, u_3)^T$. The system (61) can be recast into the form

$$(\partial_{xx} + \partial_{yy}) p = \sigma_a \widetilde{\sigma}_t^{\varepsilon} p.$$

One has the inequality

$$\|u_2\|_{W^{n,\infty}(\Omega_k)} + \|u_3\|_{W^{n\infty}(\Omega_k)} \leq C \|u_1\|_{W^{n+1,\infty}(\Omega_k)}, \quad C = \frac{1}{\sigma_a + \sigma_s}. \tag{62}$$

We assume the mesh quasi uniformity: there exists a constant C uniform with respect to the mesh sequence such that

$$\max_{\Omega_k \in \mathcal{T}_h} h_k \leq C \min_{\Omega_k \in \mathcal{T}_h} h_k. \tag{63}$$

We study the TDG scheme obtained by writing the equations under the form of a Friedrichs system (2) with

$$A_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad R = \begin{pmatrix} \varepsilon \sigma_a & 0 & 0 \\ 0 & \widetilde{\sigma}_t^{\varepsilon} & 0 \\ 0 & 0 & \widetilde{\sigma}_t^{\varepsilon} \end{pmatrix}.$$

For the stationary P_1 model (61) the matrix M reads $M(\mathbf{n}) = \begin{pmatrix} 0 & n_x & n_y \\ n_x & 0 & 0 \\ n_y & 0 & 0 \end{pmatrix}$, and we will use the decomposition

$$M^{\pm}(\mathbf{n}) = \frac{1}{2}\begin{pmatrix} \pm 1 & n_x & n_y \\ n_x & \pm n_x^2 & \pm n_x n_y \\ n_y & \pm n_x n_y & \pm n_y^2 \end{pmatrix}. \tag{64}$$

Our main goal is to obtain a proof of convergence in the case $\varepsilon = 1$. We will discuss the case $\varepsilon \to 0$ in a second stage.

Proposition 7.9. *Let $n \in \mathbb{N}$, $\Omega_k \in \mathcal{T}_h$, $\mathbf{x}_0 \in \Omega_k$, $\varepsilon = 1$ and $\sigma_a + \sigma_s > 0$. Consider $\mathbf{u} = (u_1, u_2, u_3)^T \in W^{n+1,\infty}(\Omega_k)$ solution to the P_1 model (61). Consider $\mathbf{e}_1, ..., \mathbf{e}_{2n+1} \in W^{n+1,\infty}(\Omega_k)$ specific solutions to the P_1 model, which can be either (42), or (44) or (46). There exists $\mathbf{a} = (a_1, ..., a_{2n+1})^T \in \mathbb{R}^{2n+1}$ such that*

$$\left\| \sum_{i=1}^{2n+1} a_i \mathbf{e}_i - \mathbf{u} \right\|_{L^\infty(\Omega_k)} \leq C h^n \|\mathbf{u}\|_{W^{n+1,\infty}(\Omega_k)}$$

and

$$\left\| \nabla \left(\sum_{i=1}^{2n+1} a_i \mathbf{e}_i - \mathbf{u} \right) \right\|_{L^\infty(\Omega_k)} \leq C h^{n-1} \|\mathbf{u}\|_{W^{n+1,\infty}(\Omega_k)}.$$

Proof. This is a direct consequence of proposition 7.3 applied to $u = u_1 = p$ combined with (62). □

We can now give an approximation result in terms of the $\|\cdot\|_{DG^*}$ norm.

Proposition 7.10. *Under the assumptions of proposition 7.9, there exists $\mathbf{v}_h \in V_m := Span\{\mathbf{e}_1, ..., \mathbf{e}_{2n+1}\}$ such that*

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

with $h = \max_{\Omega_k \in \mathcal{T}_h} h_k$, $h_k = diam(\Omega_k)$.

Proof. From proposition 7.9 one deduces that there exist $\mathbf{v}_h \in V_m$ such that $\forall \Omega_k$

$$\|\mathbf{u} - \mathbf{v}_h\|^2_{L^2(\Omega_k)} \leq C h_k^{2n+2} \|\mathbf{u}\|^2_{W^{n+1,\infty}(\Omega_k)},$$
$$|(\mathbf{u} - \mathbf{v}_h)|^2_{1,\Omega_k} \leq C h_k^{2n} \|\mathbf{u}\|^2_{W^{n+1,\infty}(\Omega_k)},$$

therefore

$$\|\mathbf{u} - \mathbf{v}_h\|_{L^2(\Omega_k)} \left(\frac{1}{h_k} \|\mathbf{u} - \mathbf{v}_h\|_{L^2(\Omega_k)} + |(\mathbf{u} - \mathbf{v}_h)|_{1,\Omega_k} \right) \leq C h_k^{2n+1} \|\mathbf{u}\|^2_{W^{n+1,\infty}(\Omega_k)}, \quad \forall \Omega_k.$$

Summing over all Ω_k and using that for a regular mesh of size h, the total number of elements is bounded by C/h^2 one has

$$\sum_k \|\mathbf{u} - \mathbf{v}_h\|_{L^2(\Omega_k)} \left(\frac{1}{h_k} \|\mathbf{u} - \mathbf{v}_h\|_{L^2(\Omega_k)} + |(\mathbf{u} - \mathbf{v}_h)|_{1,\Omega_k} \right) \leq C h^{2n-1} \|\mathbf{u}\|^2_{W^{n+1,\infty}(\Omega)}, \quad \forall \Omega_k.$$

One concludes using proposition 3.10. □

Combining the previous proposition with the results of Section 3 one can now give an estimation of the DG norm of the error.

Proposition 7.11. *Under the assumptions of proposition 7.9, consider the TDG method (15) with the decomposition (64). One has*

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

with $h = \max_{\Omega_k \in \mathcal{T}_h} h_k$, $h_k = diam(\Omega_k)$, where $\mathbf{u}_h$ stands for the solution to the TDG method.

Proof. Use proposition 7.10 and conclude with the quasi-optimality result from proposition 3.6. □

One can now easily study the convergence in quadratic norm using various physical assumptions on the coefficients.

Proposition 7.12 (Convergence in the dominant absorption regime: $\varepsilon = 1$, $\sigma_a > 0$, $\sigma_s \geq 0$)**.** *Consider $2n+1$ basis functions. Under the assumptions of proposition 7.11, one has*

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

with $h = \max_{\Omega_k \in \mathcal{T}_h} h_k$, $h_k = diam(\Omega_k)$ and where $\mathbf{u}_h$ stands for the solution to the TDG method.

Proof. Since $\sigma_a > 0$, $\widetilde{\sigma}_t^{\varepsilon} > 0$ and $\varepsilon = 1$, the matrix R is positive definite and one can give an L^2 lower bound of the DG norm with proposition 3.8. One concludes with proposition 7.11. □

Next case is the dominant scattering regime with $\sigma_s > 0$ and $\sigma_a = 0$. We will need the following technical lemmas.

Lemma 7.13. *Assume $w \in H^1(\mathcal{T}_h)$. One has*

$$\|w\|^2_{L^2(\Omega)} \leq C\Big(\|\partial_x w\|^2_{L^2(\Omega)} + \|\partial_y w\|^2_{L^2(\Omega)} + \frac{1}{h}\sum_k \sum_{j<k} \|[\![w]\!]\|^2_{L^2(\Sigma_{kj})} + \sum_k \|w\|^2_{L^2(\Sigma_{kk})}\Big),$$

with $h = \max_{\Omega_k \in \mathcal{T}_h} h_k$, $h_k = diam(\Omega_k)$ and where $[\![w]\!]$ denotes the jump of the function across a face.

Proof. We use (63) and the proof given in [2] (see also [1] for a weaker result). □

Lemma 7.14. *Assume $\mathbf{w} = (w_1, w_2, w_3)^T \in V(\mathcal{T}_h)$, $\varepsilon = 1$ and $\sigma_a + \sigma_s > 0$. One has*

$$\|\mathbf{w}\|_{L^2(\Omega)} \leq \frac{C}{\sqrt{h}} \|\mathbf{w}\|_{DG},$$

with $h = \max_{\Omega_k \in \mathcal{T}_h} h_k$, $h_k = diam(\Omega_k)$ and where the constant C is independent of h.

Proof. Using the definition of the DG norm (20) with $\sigma_a + \sigma_s > 0$ one gets

$$\|w_2\|^2_{L^2(\Omega)} \leq C\|\mathbf{w}\|^2_{DG}, \quad \|w_3\|^2_{L^2(\Omega)} \leq C\|\mathbf{w}\|^2_{DG}. \tag{65}$$

It remains to show $\|w_1\|_{L^2(\Omega)} \leq \frac{C}{\sqrt{h}}\|\mathbf{w}\|_{DG}$. The matrix $|M|$ reads

$$|M| = \begin{pmatrix} 1 & 0 & 0 \\ 0 & n_x^2 & n_x n_y \\ 0 & n_x n_y & n_y^2 \end{pmatrix}. \tag{66}$$

Since $\mathbf{w} \in V(\mathcal{T}_h)$ and $\sigma_a + \sigma_s > 0$, the L^2 generalization of the inequality (62) yields $\|\partial_x w_1\|^2_{L^2(\Omega)} \leq C\|w_2\|^2_{L^2(\Omega)}$ and $\|\partial_y w_1\|^2_{L^2(\Omega)} = C\|w_3\|^2_{L^2(\Omega)}$, $C \neq 0$. Therefore from the inequality (65), the definition (66) of the matrix $|M|$ and the definition of the DG norm (20) one deduces

$$\|\partial_x w_1\|^2_{L^2(\Omega)} + \|\partial_y w_1\|^2_{L^2(\Omega)} + \sum_k \sum_{j<k} \|[\![w_1]\!]\|^2_{L^2(\Sigma_{kj})} + \sum_k \|w_1\|^2_{L^2(\Sigma_{kk})} \leq C\|\mathbf{w}\|^2_{DG}.$$

One concludes using $V(\mathcal{T}_h) \subset H^1(\mathcal{T}_h)$ and lemma 7.13. □

Final proof of Theorem 1.2. The case $\sigma_a > 0$ is already treated in proposition 7.12. To treat the remaining case $\sigma_s > 0$ one can combine lemma 7.14 and proposition 7.11. The guaranteed order of convergence is the worst case, that is $n - 1$. This completes the proof. □

Theorem 1.2 illustrates one of the well known advantage of the Trefftz method: in dimension two, one needs only to add two basis functions to increase the order by one. On the contrary the number of basis functions is quadratic with respect to the order for standard DG methods.

Remark 7.15 (Case $\varepsilon \to 0^+$)**.** *It would be of course desirable to get uniform estimate in the case $\varepsilon \to 0^+$. The theorem 1.2 in particular could be very helpful since the cases $\varepsilon \to 0^+$ and $\sigma_a \to 0$ are closely related. However dependence in ε arises through the basis functions $\mathbf{e}_i$ and the solution $\mathbf{u}$ and this dependence must therefore be carefully studied when using the results of the previous sections. Whereas it is possible to easily study this limit regime for the basis functions $\mathbf{e}_i$, it is much harder for the solution $\mathbf{u}$ mostly because boundary layers may occur depending on the boundary values. We note that initial boundary layers can also arise for time dependent problems. These theoretical issues are left for future research.*

A Time dependent solutions to the P_1 model in one dimension

We give the proofs of the propositions in Section 4.1 and provide more material on how to construct the stationary and time dependent solutions (29) for the one dimensional P_1 model (27). First we recast (27) as in (2) with $d = 1$, $n = 2$, which reads

$$\partial_t \mathbf{u} + A_1 \partial_x \mathbf{u} = -R\mathbf{u}, \tag{67}$$

with

$$A_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad A_1 = \begin{pmatrix} 0 & \frac{c}{\varepsilon} \\ \frac{c}{\varepsilon} & 0 \end{pmatrix}, \quad R = \begin{pmatrix} \sigma_a & 0 \\ 0 & \sigma_t \end{pmatrix}.$$

In order to find the solutions (73) we search for particular solutions to (67) under the form

$$\mathbf{u}(x,t) = \mathbf{q}(x,t)e^{\lambda x} \tag{68}$$

with $\lambda \in \mathbb{R}$ and where $\mathbf{q} \in \mathbb{R}^n$ is a polynomial in x and t. For example we consider

$$\mathbf{q}(x,t) = \mathbf{q}_0 + x\mathbf{q}_1 + t\mathbf{q}_2 + xt\mathbf{q}_3. \tag{69}$$

Using (68) in (67) and dropping the exponential term one has

$$(\partial_t + A_1\partial_x + R)\mathbf{u} = 0 \Leftrightarrow (\partial_t + A_1\partial_x + (A_1\lambda + R))\,\mathbf{q}(x,t) = \mathbf{0}.$$

Extending $\mathbf{q}$ one finds

$$((A_1\lambda + R)\mathbf{q}_0 + A_1\mathbf{q}_1 + \mathbf{q}_2) + x\left((A_1\lambda + R)\mathbf{q}_1 + \mathbf{q}_3\right) + t\left((A_1\lambda + R)\mathbf{q}_2 + A_1\mathbf{q}_3\right) + xt(A_1\lambda + R)\mathbf{q}_3 = \mathbf{0}.$$

This equality holds for all x and t, thus one gets the following system

$$\begin{cases} (A_1\lambda + R)\mathbf{q}_3 = \mathbf{0} \\ (A_1\lambda + R)\mathbf{q}_1 = -\mathbf{q}_3 \\ (A_1\lambda + R)\mathbf{q}_2 = -A_1\mathbf{q}_3 \\ (A_1\lambda + R)\mathbf{q}_0 = -A_1\mathbf{q}_1 - \mathbf{q}_2. \end{cases} \tag{70}$$

Therefore the solutions to (67) under the form (68) with $\mathbf{q}$ given by (69) satisfy the system (70). We can now write the conditions (70) for the P_1 model.

Lemma A.1. *The conditions (70) read*

$$\begin{cases} \mathbf{q}_3 = \mathbf{0}, \\ (A_1\lambda + R)\mathbf{q}_2 = \mathbf{0}, \\ (A_1\lambda + R)\mathbf{q}_1 = \mathbf{0}, \\ (A_1\lambda + R)\mathbf{q}_0 = -A_1\mathbf{q}_1 - \mathbf{q}_2, \end{cases} \tag{71}$$

with $\lambda = \pm\frac{\varepsilon}{v}\sqrt{\sigma_a\sigma_t}$.

Proof. First, a necessary condition for (70) to admits a solution is $\det(A\lambda - R) = 0$. Since

$$A\lambda + R = \begin{pmatrix} \sigma_a & \frac{c}{\varepsilon}\lambda \\ \frac{c}{\varepsilon}\lambda & \sigma_t \end{pmatrix},$$

one deduces $\det(A_1\lambda + R) = 0 \Leftrightarrow \lambda = \pm\frac{\varepsilon}{c}\sqrt{\sigma_a\sigma_t}$. With this choice for λ, the matrix $A_1\lambda + R$ reads

$$A_1\lambda + R = \begin{pmatrix} \sigma_a & \pm\sqrt{\sigma_a\sigma_t} \\ \pm\sqrt{\sigma_a\sigma_t} & \sigma_t, \end{pmatrix}.$$

and one notices that

$$(A_1\lambda + R)^2 = (\sigma_a + \sigma_t)(A_1\lambda + R). \tag{72}$$

Thanks to the first and the second equations of (70) one has

$$\begin{cases} (A_1\lambda + R)\mathbf{q}_3 = \mathbf{0} \\ (A_1\lambda + R)\mathbf{q}_1 = -\mathbf{q}_3 \end{cases} \Rightarrow (A_1\lambda + R)^2\mathbf{q}_1 = (\sigma_a + \sigma_t)(A_1\lambda + R)\mathbf{q}_3 = \mathbf{0}.$$

From (72) one gets $(A_1\lambda + R)\mathbf{q}_1 = \mathbf{0}$, therefore $\mathbf{q}_3 = \mathbf{0}$. This completes the proof. □

Proposition A.2. *The P_1 model (67) admits the following four solutions*

$$\begin{aligned} \mathbf{e}_1^{\pm}(x) &= \begin{pmatrix} \sqrt{\sigma_t} \\ \mp\sqrt{\sigma_a} \end{pmatrix} e^{\pm\frac{\varepsilon}{c}\sqrt{\sigma_a\sigma_t}x}, \\ \mathbf{e}_2^{\pm}(x) &= \begin{pmatrix} -c\frac{\sigma_t-\sigma_a}{4\sigma_a\sqrt{\sigma_t}} \pm \varepsilon x\frac{\sigma_a+\sigma_t}{2\sqrt{\sigma_a}} + ct\sqrt{\sigma_t} \\ \mp c\frac{\sigma_t-\sigma_a}{4\sigma_t\sqrt{\sigma_a}} - \varepsilon x\frac{\sigma_a+\sigma_t}{2\sqrt{\sigma_t}} \mp ct\sqrt{\sigma_a} \end{pmatrix} e^{\pm\frac{\varepsilon}{c}\sqrt{\sigma_a\sigma_t}x}. \end{aligned} \tag{73}$$

Proof. One notices $Ker(A_1\lambda + R) = Span((\sqrt{\sigma_t}, \mp\sqrt{\sigma_a})^T)$. Thus with $\mathbf{w} = (\sqrt{\sigma_t}, \mp\sqrt{\sigma_a})^T$ and the relations (71) one gets

$$\mathbf{q}_1 = \alpha\mathbf{w}, \quad \mathbf{q}_2 = \beta\mathbf{w}, \quad \alpha, \beta \in \mathbb{R}.$$

From the last equality of (71) one sees that $-A_1\mathbf{q}_1 - \mathbf{q}_2 \in Im(A_1\lambda + R)$ which implies $-A_1\mathbf{q}_1 - \mathbf{q}_2 \in Ker((A_1\lambda+R)^T)^{\perp}$. Since the matrices A_1 and R are symmetric, $Ker((A_1\lambda+R)^T) = Ker(A_1\lambda+R) = Vect(\mathbf{w})$. A necessary condition is then $(-A_1\mathbf{q}_1 - \mathbf{q}_2, \mathbf{w}) = 0$ which is equivalent to

$$\alpha = \pm\frac{\sigma_a + \sigma_t}{2\sqrt{\sigma_a\sigma_t}}\frac{\varepsilon}{c}\beta.$$

Finally let $\mathbf{q}_0 = (q_0^1, q_0^2)^T$. From the fourth equation of (71) one gets $q_0^1 = \frac{1}{\sqrt{\sigma_a}}(\beta\frac{\sigma_a-\sigma_t}{2\sqrt{\sigma_t\sigma_a}} \mp \sqrt{\sigma_t}q_0^2)$. Thus one can choose $\mathbf{q}_0$ under the form

$$\mathbf{q}_0 = \beta\left(-\frac{\sigma_t - \sigma_a}{4\sigma_a\sqrt{\sigma_t}}, \mp\frac{\sigma_t - \sigma_a}{4\sigma_t\sqrt{\sigma_a}}\right)^T + \gamma\mathbf{w},$$

with $\gamma \in \mathbb{R}$. To sum up one has the following relations

$$\left|\begin{aligned} &\mathbf{q}_3 = \mathbf{0}, \\ &\mathbf{q}_2 = \beta\left(\sqrt{\sigma_t}, \mp\sqrt{\sigma_a}\right)^T, \\ &\mathbf{q}_1 = -\frac{\sigma_a + \sigma_t}{2\sqrt{\sigma_a\sigma_t}}\frac{\varepsilon}{c}\beta\left(\mp\sqrt{\sigma_t}, \sqrt{\sigma_a}\right)^T, \\ &\mathbf{q}_0 = \beta\left(-\frac{\sigma_t - \sigma_a}{4\sigma_a\sqrt{\sigma_t}}, \mp\frac{\sigma_t - \sigma_a}{4\sigma_t\sqrt{\sigma_a}}\right)^T + \gamma\left(\sqrt{\sigma_t}, \pm\sqrt{\sigma_a}\right)^T, \end{aligned}\right. \tag{74}$$

$\beta, \gamma \in \mathbb{R}$. Because the solutions are under the form $\mathbf{u}(x,t) = (\mathbf{q}_0 + x\mathbf{q}_1 + t\mathbf{q}_2 + xt\mathbf{q}_3)e^{\lambda x}$, with $\lambda = \pm\frac{\varepsilon}{c}\sqrt{\sigma_a\sigma_t}$, one finds the four basis functions (73). This completes the proof. □

Now we construct linear combinations of the solutions (73) that remain stable in the case $\sigma_a \to 0$. To make these solutions more convenient to read, we use the notations $z_x = \frac{\varepsilon}{c}\sqrt{\sigma_a\sigma_t}x$ and $\cosh(x) = \frac{e^x+e^{-x}}{2}$, $\sinh(x) = \frac{e^x-e^{-x}}{2}$.

Lemma A.3. *The following four functions are linear combinations of the solutions* (73)

$$\begin{aligned}
\tilde{\mathbf{e}}_1(x) &= \begin{pmatrix} \sqrt{\frac{\sigma_t}{\sigma_a}}\sinh(z_x) \\ -\cosh(z_x) \end{pmatrix}, \\
\tilde{\mathbf{e}}_2(x) &= \begin{pmatrix} \cosh(z_x) \\ -\sqrt{\frac{\sigma_a}{\sigma_t}}\sinh(z_x) \end{pmatrix}, \\
\tilde{\mathbf{e}}_3(t,x) &= \begin{pmatrix} -\varepsilon\frac{\sigma_t+\sigma_a}{2\sqrt{\sigma_a\sigma_t}}x\sinh(z_x) - ct\cosh(z_x) \\ c\frac{\sigma_t-\sigma_a}{2\sigma_t\sqrt{\sigma_a\sigma_t}}\sinh(z_x) + \varepsilon\frac{\sigma_t+\sigma_a}{2\sigma_t}x\cosh(z_x) + c\sqrt{\frac{\sigma_a}{\sigma_t}}t\sinh(z_x) \end{pmatrix}, \\
\tilde{\mathbf{e}}_4(t,x) &= \begin{pmatrix} c\frac{\sigma_t-\sigma_a}{2\sigma_a\sqrt{\sigma_a\sigma_t}}\sinh(z_x) - \varepsilon\frac{\sigma_t+\sigma_a}{2\sigma_a}x\cosh(z_x) - c\sqrt{\frac{\sigma_t}{\sigma_a}}t\sinh(z_x) \\ \varepsilon\frac{\sigma_t+\sigma_a}{2\sqrt{\sigma_a\sigma_t}}x\sinh(z_x) + ct\cosh(z_x) \end{pmatrix}.
\end{aligned} \tag{75}$$

Proof. One defines the following linear combinations of the functions (73)

$$\begin{aligned}
\mathbf{l}_1^{\pm}(x,t) &= \mathbf{e}_2^{\pm}(x,t) + c\frac{\sigma_t-\sigma_a}{4\sigma_a\sigma_t}\mathbf{e}_1^{\pm}(x,t), \\
\mathbf{l}_2^{\pm}(x,t) &= \mathbf{e}_2^{\pm}(x,t) - c\frac{\sigma_t-\sigma_a}{4\sigma_a\sigma_t}\mathbf{e}_1^{\pm}(x,t).
\end{aligned}$$

Then defining the four solutions

$$\begin{aligned}
\tilde{\mathbf{e}}_1(x,t) &= \frac{1}{2\sqrt{\sigma_a}}\left(\mathbf{e}_1^+(x,t) - \mathbf{e}_1^-(x,t)\right), \\
\tilde{\mathbf{e}}_2(x,t) &= \frac{1}{2\sqrt{\sigma_t}}\left(\mathbf{e}_1^+(x,t) + \mathbf{e}_1^-(x,t)\right). \\
\tilde{\mathbf{e}}_3(x,t) &= \frac{-1}{2\sqrt{\sigma_t}}\left(\mathbf{l}_1^+(x,t) + \mathbf{l}_1^-(x,t)\right), \\
\tilde{\mathbf{e}}_4(x,t) &= \frac{-1}{2\sqrt{\sigma_a}}\left(\mathbf{l}_2^+(x,t) - \mathbf{l}_2^-(x,t)\right),
\end{aligned}$$

one gets the functions (75). □

We show that these solutions remain stable in the limit case $\sigma_a \to 0$.

Proposition A.4. *When $\sigma_a \to 0$ ($\sigma_t \to \frac{\sigma_s}{\varepsilon^2}$), the solutions* (75) *tend to the following functions*

$$\begin{aligned}
\tilde{\mathbf{e}}_1(x,t) &\underset{\sigma_a\to 0}{\to} \begin{pmatrix} \frac{\varepsilon\sigma_t}{c}x \\ -1 \end{pmatrix}, \\
\tilde{\mathbf{e}}_2(x,t) &\underset{\sigma_a\to 0}{\to} \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \\
\tilde{\mathbf{e}}_3(x,t) &\underset{\sigma_a\to 0}{\to} \begin{pmatrix} -\frac{\varepsilon^2\sigma_t}{2c}x^2 - ct \\ \varepsilon x \end{pmatrix}, \\
\tilde{\mathbf{e}}_4(x,t) &\underset{\sigma_a\to 0}{\to} \begin{pmatrix} -\frac{\varepsilon^3\sigma_t^2}{6c^2}x^3 - \varepsilon\sigma_t t x - \varepsilon x \\ \frac{\varepsilon^2\sigma_t}{2c}x^2 + ct \end{pmatrix}.
\end{aligned}$$

Proof. One notices that

$$\cosh(z_x) \underset{\sigma_a \to 0}{\to} 1, \quad \frac{\sinh(z_x)}{\sqrt{\sigma_a \sigma_t}} \underset{\sigma_a \to 0}{\to} \frac{\varepsilon}{c} x. \tag{76}$$

The limit of $\tilde{\mathbf{e}}_1(x,t)$, $\tilde{\mathbf{e}}_2(x,t)$ and $\tilde{\mathbf{e}}_3(x,t)$ are simply obtained by using the expressions (76) in (75). The limit of the second component of $\tilde{\mathbf{e}}_4(x,t)$ can be obtained in a similar way. It remains to study the first component of $\tilde{\mathbf{e}}_4(x,t)$. One has

$$\begin{aligned}
&\frac{c(\sigma_t - \sigma_a)}{2\sigma_a\sqrt{\sigma_a\sigma_t}} \sinh(z_x) - \varepsilon x \frac{\sigma_t + \sigma_a}{2\sigma_a} \cosh(z_x) \\
&= \frac{c(\sigma_t - \sigma_a)}{2\sigma_a} \Big(\frac{\varepsilon}{c} x + \frac{\varepsilon^3 x^3 \sigma_a \sigma_t}{3! c^3} + o(\sigma_a^2)\Big) \quad - \varepsilon x \frac{\sigma_t + \sigma_a}{2\sigma_a} \Big(1 + \frac{\varepsilon^2 \sigma_a \sigma_t x^2}{2! c^2} + o(\sigma_a^2)\Big) \\
&= -\varepsilon x + \frac{\varepsilon^3 \sigma_t^2 x^3}{2c^2} \Big(-\frac{1}{6} + \frac{1}{2}\Big) + o(\sigma_a) = -\varepsilon x - \frac{\varepsilon^3 \sigma_t^2}{6c^2} x^3 + o(\sigma_a).
\end{aligned}$$

Because $-ct\sigma_t \frac{\sinh(z_x)}{\sqrt{\sigma_a\sigma_t}} \underset{\sigma_a \to 0}{\to} -\varepsilon \sigma_t t x$, one gets the expression of the limit of $\tilde{\mathbf{e}}_4(x,t)$. This completes the proof. □

B Proof of proposition 7.2

Lemma B.1. *Assume that the hypotheses of proposition 7.2 are satisfied. Then for all $0 \le l \le n-2$ one has the identity*

$$\begin{aligned}
&\sum_{p=0}^{l} \partial_x^p \partial_y^{l-p} u(\mathbf{x}_0) T_l^p(\mathbf{x}) + \sum_{p=0}^{l+2} \partial_x^p \partial_y^{l+2-p} u(\mathbf{x}_0) \alpha_{l+2}^p(\mathbf{x}) = \\
&\quad \sum_{p=0}^{l} \partial_x^p \partial_y^{l-p} u(\mathbf{x}_0) \alpha_l^p(\mathbf{x}) + \partial_x^{l+2} u(\mathbf{x}_0) \beta_{l+2}^{l+2}(\mathbf{x}) + \partial_x^{l+1} \partial_y u(\mathbf{x}_0) \beta_{l+2}^{l+1}(\mathbf{x}).
\end{aligned} \tag{77}$$

Proof. Let $l \in \mathbb{N}$, $0 \le l \le n-2$. For $l_1 \in \mathbb{Z}$, $-1 \le l_1 \le l-1$ we define the function

$$\begin{aligned}
f(l_1) = &\sum_{p=0}^{l_1} \partial_x^p \partial_y^{l-p} u(\mathbf{x}_0) \alpha_l^p(\mathbf{x}) + \sum_{p=l_1+1}^{l} \partial_x^p \partial_y^{l-p} u(\mathbf{x}_0) T_l^p(\mathbf{x}) + \sum_{p=l_1+3}^{l+2} \partial_x^p \partial_y^{l+2-p} u(\mathbf{x}_0) \alpha_{l+2}^p(\mathbf{x}) \\
&+ \partial_x^{l_1+2} \partial_y^{l-l_1} u(\mathbf{x}_0) \beta_{l+2}^{l_1+2}(\mathbf{x}) + \partial_x^{l_1+1} \partial_y^{l+1-l_1} u(\mathbf{x}_0) \beta_{l+2}^{l_1+1}(\mathbf{x}),
\end{aligned} \tag{78}$$

where we use the convention $\sum_{p=a}^{b} = 0$ for $a, b \in \mathbb{Z}$ and $b < a$. First we show $f(l_1) = f(l_1+1)$ for $-1 \le l_1 \le l-1$. Because u is solution to the equation (77) one notices

$$\partial_x^{l_1+1} \partial_y^{l+1-l_1} u(\mathbf{x}_0) \beta_{l+2}^{l_1+1}(\mathbf{x}) = \Big(- \partial_x^{l_1+3} \partial_y^{l-l_1-1} + \omega \partial_x^{l_1+1} \partial_y^{l-l_1-1} \Big) u(\mathbf{x}_0) \beta_{l+2}^{l_1+1}(\mathbf{x}). \tag{79}$$

Now we consider the definition of the function f (78) and we study the difference $f(l_1+1) - f(l_1)$. After simplifications on the elements that appear in both $f(l_1)$ and $f(l_1+1)$ one finds

$$\begin{aligned}
f(l_1+1) - f(l_1) = &\partial_x^{l_1+1} \partial_y^{l-l_1-1} u(\mathbf{x}_0) \alpha_l^{l_1+1}(\mathbf{x}) - \partial_x^{l_1+1} \partial_y^{l-l_1-1} u(\mathbf{x}_0) T_l^{l_1+1}(\mathbf{x}) - \partial_x^{l_1+3} \partial_y^{l-l_1-1} u(\mathbf{x}_0) \alpha_{l+2}^{l_1+3}(\mathbf{x}) \\
&+ \partial_x^{l_1+3} \partial_y^{l-l_1-1} u(\mathbf{x}_0) \beta_{l+2}^{l_1+3}(\mathbf{x}) - \partial_x^{l_1+1} \partial_y^{l+1-l_1} u(\mathbf{x}_0) \beta_{l+2}^{l_1+1}(\mathbf{x}).
\end{aligned}$$

Using the equality (79) to reformulate the fifth term on the right hand side, one gets

$$\begin{aligned}
f(l_1+1) - f(l_1) = \quad &\partial_x^{l_1+1} \partial_y^{l-l_1-1} u(\mathbf{x}_0) \alpha_l^{l_1+1}(\mathbf{x}) - \partial_x^{l_1+1} \partial_y^{l-l_1-1} u(\mathbf{x}_0) T_l^{l_1+1}(\mathbf{x}) - \partial_x^{l_1+3} \partial_y^{l-l_1-1} u(\mathbf{x}_0) \alpha_{l+2}^{l_1+3}(\mathbf{x}) \\
&+ \partial_x^{l_1+3} \partial_y^{l-l_1-1} u(\mathbf{x}_0) \beta_{l+2}^{l_1+3}(\mathbf{x}) + \Big(\partial_x^{l_1+3} \partial_y^{l-l_1-1} - \omega \partial_x^{l_1+1} \partial_y^{l-l_1-1} \Big) u(\mathbf{x}_0) \beta_{l+2}^{l_1+1}(\mathbf{x}).
\end{aligned}$$

Ordering the terms with respect to the derivatives gives

$$f(l_1+1)-f(l_1) = \partial_x^{l_1+1}\partial_y^{l-l_1-1}u(\mathbf{x}_0)\Big(\alpha_l^{l_1+1}(\mathbf{x}) - T_l^{l_1+1}(\mathbf{x}) - \omega\beta_{l+2}^{l_1+1}(\mathbf{x})\Big) + \partial_x^{l_1+3}\partial_y^{l-l_1-1}u(\mathbf{x}_0)\Big(- \alpha_{l+2}^{l_1+3}(\mathbf{x}) + \beta_{l+2}^{l_1+1}(\mathbf{x}) + \beta_{l+2}^{l_1+3}(\mathbf{x})\Big).$$

Using the definitions (52) and (53) one finds $\alpha_l^{l_1+1}(\mathbf{x}) - T_l^{l_1+1}(\mathbf{x}) - \omega\beta_{l+2}^{l_1+1}(\mathbf{x}) = 0$ and $\beta_{l+2}^{l_1+3}(\mathbf{x}) - \alpha_{l+2}^{l_1+3}(\mathbf{x}) + \beta_{l+2}^{l_1+1}(\mathbf{x}) = 0$. Therefore one has $f(l_1+1)-f(l_1)=0$ for all $-1 \leq l_1 \leq l-1$. One deduces $f(-1)=f(l)$ which can be written

$$\sum_{p=0}^{l}\partial_x^p\partial_y^{l-p}u(\mathbf{x}_0)T_l^p(\mathbf{x}) + \sum_{p=2}^{l+2}\partial_x^p\partial_y^{l+2-p}u(\mathbf{x}_0)\alpha_{l+2}^p(\mathbf{x}) + \partial_y^{l+2}u(\mathbf{x}_0)\beta_{l+2}^0(\mathbf{x}) + \partial_x\partial_y^{l+1}u(\mathbf{x}_0)\beta_{l+2}^1(\mathbf{x}) = \sum_{p=0}^{l}\partial_x^p\partial_y^{l-p}u(\mathbf{x}_0)\alpha_l^p(\mathbf{x}) + \partial_x^{l+2}u(\mathbf{x}_0)\beta_{l+2}^{l+2}(\mathbf{x}) + \partial_x^{l+1}\partial_y u(\mathbf{x}_0)\beta_{l+2}^{l+1}(\mathbf{x}).$$

Noticing from (54) $\alpha_{l+2}^0(\mathbf{x}) = \beta_{l+2}^0(\mathbf{x})$ and $\alpha_{l+2}^1(\mathbf{x}) = \beta_{l+2}^1(\mathbf{x})$, one incorporates the two corresponding terms in the second sum so one finds the equality (77). This completes the proof. □

Proof of proposition 7.2. Start from the Taylor expansion (51). From definition (52) one has $\alpha_n^p(\mathbf{x}) = T_n^p(\mathbf{x})$ and $\alpha_{n-1}^p(\mathbf{x}) = T_{n-1}^p(\mathbf{x})$. Therefore

$$u(\mathbf{x}) = \sum_{k=0}^{n-2}\sum_{p=0}^{k}\partial_x^p\partial_y^{k-p}u(\mathbf{x}_0)T_k^p(\mathbf{x}) + \sum_{p=0}^{n-1}\partial_x^p\partial_y^{n-1-p}u(\mathbf{x}_0)\alpha_{n-1}^p(\mathbf{x}) + \sum_{p=0}^{n}\partial_x^p\partial_y^{n-p}u(\mathbf{x}_0)\alpha_n^p(\mathbf{x}) + \sum_{p=0}^{n+1}\partial_x^p\partial_y^{n+1-p}u(\mathbf{x}_s)T_{n+1}^p(\mathbf{x}).$$

One can recursively use the equality (77) from $l=n-2$ to $l=0$. More precisely, rearranging the first sum one has

$$u(\mathbf{x}) = \sum_{k=0}^{n-3}\sum_{p=0}^{k}\partial_x^p\partial_y^{k-p}u(\mathbf{x}_0)T_k^p(\mathbf{x}) + \sum_{p=0}^{n-1}\partial_x^p\partial_y^{n-1-p}u(\mathbf{x}_0)\alpha_{n-1}^p(\mathbf{x}) + \left(\sum_{p=0}^{n-2}\partial_x^p\partial_y^{n-2-p}u(\mathbf{x}_0)T_{n-2}^p(\mathbf{x}) + \sum_{p=0}^{n}\partial_x^p\partial_y^{n-p}u(\mathbf{x}_0)\alpha_n^p(\mathbf{x})\right) + \sum_{p=0}^{n+1}\partial_x^p\partial_y^{n+1-p}u(\mathbf{x}_s)T_{n+1}^p(\mathbf{x}).$$

One can reformulate the terms between brackets using (77) with the index correspondence $n-2=l$. One finds

$$u(\mathbf{x}) = \sum_{k=0}^{n-3}\sum_{p=0}^{k}\partial_x^p\partial_y^{k-p}u(\mathbf{x}_0)T_k^p(\mathbf{x}) + \sum_{p=0}^{n-1}\partial_x^p\partial_y^{n-1-p}u(\mathbf{x}_0)\alpha_{n-1}^p(\mathbf{x}) + \sum_{p=0}^{n-2}\partial_x^p\partial_y^{n-2-p}u(\mathbf{x}_0)\alpha_{n-2}^p(\mathbf{x}) \\ \left[+\partial_x^n u(\mathbf{x}_0)\beta_n^n(\mathbf{x}) + \partial_x^{n-1}\partial_y u(\mathbf{x}_0)\beta_n^{n-1}(\mathbf{x})\right] + \sum_{p=0}^{n+1}\partial_x^p\partial_y^{n+1-p}u(\mathbf{x}_s)T_{n+1}^p(\mathbf{x}). \tag{80}$$

One can now recursively repeat this simple operation using the equality (77) for $l=n-3, \ldots,$ to $l=0$. One finally gets the formula (80) where the first line is written for $n=2$, the term $[\cdot]$ becomes a sum and the last term remains unchanged

$$u(\mathbf{x}) = 0 + \sum_{p=0}^{1}\partial_x^p\partial_y^{1-p}u(\mathbf{x}_0)\alpha_1^p(\mathbf{x}) + u(\mathbf{x}_0)\alpha_0^0(\mathbf{x}) + \sum_{k=2}^{n}\left[\partial_x^k u(\mathbf{x}_0)\beta_k^k(\mathbf{x}) + \partial_x^{k-1}\partial_y u(\mathbf{x}_0)\beta_k^{k-1}(\mathbf{x})\right] + \sum_{p=0}^{n+1}\partial_x^p\partial_y^{n+1-p}u(\mathbf{x}_s)T_{n+1}^p(\mathbf{x}).$$

That is

$$u(\mathbf{x}) = u(\mathbf{x}_0)\alpha_0^0(\mathbf{x}) + \partial_x u(\mathbf{x}_0)\alpha_1^1(\mathbf{x}) + \partial_y u(\mathbf{x}_0)\alpha_1^0(\mathbf{x}) + \sum_{k=2}^{n}\left[\partial_x^k u(\mathbf{x}_0)\beta_k^k(\mathbf{x}) + \partial_x^{k-1}\partial_y u(\mathbf{x}_0)\beta_k^{k-1}(\mathbf{x})\right] + \sum_{p=0}^{n+1}\partial_x^p\partial_y^{n+1-p}u(\mathbf{x}_s)T_{n+1}^p(\mathbf{x}).$$

Noticing from (54) that $\alpha_0^0(\mathbf{x}) = \beta_0^0(\mathbf{x}), \alpha_1^0(\mathbf{x}) = \beta_1^0(\mathbf{x}), \alpha_1^1(\mathbf{x}) = \beta_1^1(\mathbf{x})$ one finds the expression (55). This completes the proof. □

C Interpretation of the one dimensional TDG method as a finite difference scheme

The goal of this section is to obtain the FD scheme (33) based on the Trefftz discontinuous Galerkin method (15) for the one dimensional hyperbolic heat equation

$$\begin{cases} \partial_t p + \frac{c}{\varepsilon}\partial_x v = 0, \\ \partial_t v + \frac{c}{\varepsilon}\partial_x p = -\frac{\sigma_s}{\varepsilon^2}v, \end{cases} \tag{81}$$

$\varepsilon \in \mathbb{R}_*^+$, $\sigma_s, c \in \mathbb{R}^+$. For the sake of simplicity we assume that σ_s is constant in the domain. This model can be written in the form of the Friedrichs system (2) with

$$A_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad A_1 = \frac{c}{\varepsilon}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad R = \begin{pmatrix} 0 & 0 \\ 0 & -\frac{\sigma_s}{\varepsilon^2} \end{pmatrix}.$$

We consider basis functions $\mathbf{e}_{i,l}$ where i is the global number of the cell and l the local number of the basis function in the cell i. We denote by $x_{i-\frac{1}{2}}$ and $x_{i+\frac{1}{2}}$ the edges of the spatial cell $\Omega_{S,i}$, i.e. $\Omega_{S,i} = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$ and x_i the midpoint. We use two stationary basis functions defined as

$$\mathbf{e}_{k,1}(t,x) = \begin{cases} \begin{pmatrix} 1 \\ 0 \end{pmatrix}, & \text{if } (t,x) \in \Omega_k, \\ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, & \text{else,} \end{cases} \qquad \mathbf{e}_{k,2}(t,x) = \begin{cases} \begin{pmatrix} -\frac{\sigma_s}{c\varepsilon}(x - x_k) \\ 1 \end{pmatrix}, & \text{if } (t,x) \in \Omega_k, \\ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, & \text{else.} \end{cases} \tag{82}$$

We use the notation $\mathbf{e}_{i,1}^n, \mathbf{e}_{i,2}^n$ when designing the basis functions from the spatial cell $\Omega_{S,i}$ at the time step n. Consider the bilinear and linear forms obtained from the decoupled formulation (17)

$$\begin{aligned} a_T^n(\mathbf{u},\mathbf{v}) = &-\sum_k \sum_{j<k} \int_{\Sigma_{k^n j^n}} (M^-_{k^n j^n}\mathbf{v}_k^n + M^+_{k^n j^n}\mathbf{v}_j^n)^T(\mathbf{u}_k^n - \mathbf{u}_j^n) - \sum_k \int_{\partial\Omega_S \cap \partial\Omega_{k^n}} (\mathbf{v}_k^n)^T M^-_{k^n}\mathbf{u}_k^n \\ &- \sum_k \int_{\Sigma_{k^n k^{n-1}}} (\mathbf{v}_k^n)^T M^-_{k^n k^{n-1}}\mathbf{u}_k^n, \quad \mathbf{u},\mathbf{v} \in V(\mathcal{T}_h), \\ l^n(\mathbf{v}) = &-\sum_k \int_{\partial\Omega_S \cap \partial\Omega_{k^n}} (\mathbf{v}_k^n)^T \mathbf{g}_S - \sum_k \int_{\Sigma_{k^n k^{n-1}}} (\mathbf{v}_k^n)^T M^-_{k^n k^{n-1}}\mathbf{u}_k^{n-1}, \quad \mathbf{v} \in V(\mathcal{T}_h). \end{aligned} \tag{83}$$

In the following we will write explicitly the equality

$$a_T^n(\mathbf{u}, \mathbf{e}_{l,i}^n) = l^n(\mathbf{e}_{l,i}^n), \quad l = 1, 2, \tag{84}$$

for any time step n and any spatial cell $\Omega_{S,i}$. For simplicity we will consider periodic boundary conditions, a uniform space step h and a uniform time step Δt. We define some notation.

Definition C.1. *Define* $C^n_{S,i,l}$, $C^{n-1}_{T,i,l}$ *and* $C^n_{T,i,l}$ *as*

$$C^n_{S,i,l} = -\sum_k \sum_j \int_{\Sigma_{k^n j^n}} (M^-_{k^n j^n} \mathbf{e}^n_{i,l})^T (\mathbf{u}^n_k - \mathbf{u}^n_j), \tag{85}$$

$$C^{n-1}_{T,i,l} = -\sum_k \int_{\Sigma_{k^n k^{n-1}}} (\mathbf{e}^n_{i,l})^T M^-_{k^n k^{n-1}} \mathbf{u}^{n-1}_k, \tag{86}$$

$$C^n_{T,i,l} = -\sum_k \int_{\Sigma_{k^n k^{n-1}}} (\mathbf{e}^n_{i,l})^T M^-_{k^n k^{n-1}} \mathbf{u}^n_k. \tag{87}$$

Since $\mathbf{u}_k$ is a combination of the basis functions in each cell, one can make the following assumption.

Assumption C.2. *Assume that* $\mathbf{u}_k$ *admits the following decomposition in each cell* Ω_k

$$\mathbf{u}_k = \alpha_k \mathbf{e}_{k,1} + \beta_k \mathbf{e}_{k,2}, \quad \alpha_k, \beta_k \in \mathbb{R},$$

or, in an identical way, when considering the time step n *and the spatial cell* $\Omega_{S,i}$

$$\mathbf{u}^n_i = \alpha^n_i \mathbf{e}^n_{i,1} + \beta^n_i \mathbf{e}^n_{i,2}, \quad \alpha^n_i, \beta^n_i \in \mathbb{R}. \tag{88}$$

We can now write the equality (84) using the Definition C.1.

Proposition C.3. *Consider the model* (81) *and the basis functions* (82). *The equality* (84) *with periodic boundary conditions at the time step* n *in any spatial cell* $\Omega_{S,i}$ *reads*

$$\begin{aligned} C^n_{T,i,1} - C^{n-1}_{T,i,1} + C^n_{S,i,1} = 0, \\ C^n_{T,i,2} - C^{n-1}_{T,i,2} + C^n_{S,i,2} = 0. \end{aligned} \tag{89}$$

Proof. Since we consider periodic boundary conditions, the term $\int_{\partial\Omega_S \cap \partial\Omega_{k^n}} (\mathbf{v}^n_k)^T M^-_{k^n} \mathbf{u}^n_k$ in the bilinear form and the term $\int_{\partial\Omega_S \cap \partial\Omega_{k^n}} (\mathbf{v}^n_k)^T \mathbf{g}_S$ in the linear form of (83) are equal to zero. One notices that

$$-\sum_k \sum_{j<k} \int_{\Sigma_{k^n j^n}} (M^-_{k^n j^n} \mathbf{v}^n_k - M^+_{k^n j^n} \mathbf{v}^n_j)^T (\mathbf{u}^n_k - \mathbf{u}^n_j) = -\sum_k \sum_j \int_{\Sigma_{k^n j^n}} (M^-_{k^n j^n} \mathbf{v}^n_k)^T (\mathbf{u}^n_k - \mathbf{u}^n_j). \tag{90}$$

Therefore one has $a_T(\mathbf{u}, \mathbf{e}^n_{l,i}) = C^n_{T,i,l} + C^n_{S,i,l}$ and $l(\mathbf{e}^n_{l,i}) = C^{n-1}_{T,i,l}$. The equality (84) gives for $l = 1$ and $l = 2$, respectively the first and second equations of (89). This completes the proof. □

Now we can study the values of the coefficients $C_{S,i,l}$ and $C_{T,i,l}$.

Proposition C.4. *One has*

$$C^n_{S,i,1} = \frac{c\Delta t}{2\varepsilon} \Big(-\alpha_{i-1} + 2\alpha_i - \alpha_{i+1} + (1 - \frac{\sigma_s h}{2c\varepsilon})(\beta_{i+1} - \beta_{i-1}) \Big)^n, \tag{91}$$

and

$$C^n_{S,i,2} = \frac{c\Delta t}{2\varepsilon} \Big((\frac{\sigma_s h}{2c\varepsilon})^2 (\beta_{i+1} + 2\beta_i + \beta_{i-1}) + \frac{\sigma_s h}{2c\varepsilon} \beta_i + (-\beta_{i-1} + 2\beta_i - \beta_{i+1}) + (1 + \frac{\sigma_s h}{2c\varepsilon})(\alpha_{i+1} - \alpha_{i-1}) \Big)^n. \tag{92}$$

Proof. For simplicity we will use the notation $M^-_{\pm 1} = M^-((0, \pm 1)^T)$, $M^+_{\pm 1} = M^+((0, \pm 1)^T)$ and $(\lambda^{m,l}_{k,j})^\pm = (M^+_{\pm 1} \mathbf{e}_{j,l})^T \mathbf{e}_{k,m}$. Since the function $\mathbf{e}_{i,l}$ is only non-zero in the cell Ω_i one can write $C_{S,i,l}$ from (85) as

$$C_{S,i,l} = \int_{t^{n-1}}^{t^n} \Big(-(M^-_{-1} \mathbf{e}_{i,l})^T (\mathbf{u}_i - \mathbf{u}_{i-1})(x_{i-\frac{1}{2}}) - (M^-_1 \mathbf{e}_{i,l})^T (\mathbf{u}_i - \mathbf{u}_{i+1})(x_{i+\frac{1}{2}}) \Big). \tag{93}$$

Using $M^-_{\pm 1} = -M^+_{\mp 1}$, the decomposition of $\mathbf{u}_i^n$ (88) and the fact that the basis (82) does not depend on time, the equality (93) reads

$$\begin{aligned} C^n_{S,i,l} =& \Delta t \Big(\alpha_i (\lambda^{1l}_{i,i})^+(x_{i-\frac{1}{2}}) + \beta_i (\lambda^{2l}_{i,i})^+(x_{i-\frac{1}{2}}) - \alpha_{i-1}(\lambda^{l1}_{i,i-1})^+(x_{i-\frac{1}{2}}) - \beta_{i-1}(\lambda^{l2}_{i,i-1})^+(x_{i-\frac{1}{2}}) \\ &+ \alpha_i (\lambda^{1l}_{i,i})^-(x_{i+\frac{1}{2}}) + \beta_i (\lambda^{2l}_{i,i})^+(x_{i+\frac{1}{2}}) - \alpha_{i+1}(\lambda^{l1}_{i,i+1})^-(x_{i+\frac{1}{2}}) - \beta_{i+1}(\lambda^{l2}_{i,i+1})^-(x_{i+\frac{1}{2}}) \Big)^n. \end{aligned} \tag{94}$$

For $n_t = 0$, one has

$$M(\mathbf{n}) = M(0, n_x) = \frac{c}{\varepsilon} \begin{pmatrix} 0 & n_x \\ n_x & 0 \end{pmatrix}, \quad M^+(0, n_x) = \frac{c}{2\varepsilon} \begin{pmatrix} 1 & n_x \\ n_x & 1 \end{pmatrix}, \quad M^-(0, n_x) = \frac{c}{2\varepsilon} \begin{pmatrix} -1 & n_x \\ n_x & -1 \end{pmatrix},$$

and one notices that

$$\begin{aligned} (\lambda^{11}_{ji})^\pm(x) &= \frac{c}{2\varepsilon}, \\ (\lambda^{12}_{ji})^\pm(x) &= \frac{c}{2\varepsilon} \Big(-\frac{\sigma_s}{c\varepsilon}(x - x_i) \pm 1 \Big), \\ (\lambda^{22}_{ji})^\pm(x) &= \frac{c}{2\varepsilon} \Big(1 \mp \frac{\sigma_s}{c\varepsilon} \big((x - x_i) + (x - x_j) \big) + (\frac{\sigma_s}{c\varepsilon})^2 (x - x_i)(x - x_j) \Big). \end{aligned} \tag{95}$$

Recalling that for simplicity $h = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$ for all i and inserting (95) in (94) one finds for $l = 1$

$$C^n_{S,i,1} = \frac{c\Delta t}{2\varepsilon} \Big(-\alpha_{i-1} + 2\alpha_i - \alpha_{i+1} + (1 - \frac{\sigma_s h}{2c\varepsilon})(\beta_{i+1} - \beta_{i-1}) \Big)^n,$$

and for $l = 2$

$$C^n_{S,i,2} = \frac{c\Delta t}{2\varepsilon} \Big((\frac{\sigma_s h}{2c\varepsilon})^2 (\beta_{i+1} + 2\beta_i + \beta_{i-1}) + \frac{\sigma_s h}{2c\varepsilon} \beta_i + (-\beta_{i-1} + 2\beta_i - \beta_{i+1}) + (1 + \frac{\sigma_s h}{2c\varepsilon})(\alpha_{i+1} - \alpha_{i-1}) \Big)^n.$$

This completes the proof. □

Proposition C.5. *One has*

$$C^n_{T,i,1} = h \tag{96}$$

$$C^n_{T,i,2} = h(1 + \frac{\sigma_s^2 h^2}{48 c^2 \varepsilon^2}). \tag{97}$$

Proof. Since $-M^-_{k^n k^{n-1}} = I_m$, $C^n_{T,i,l}$ reads

$$C^n_{T,i,l} = -\sum_k \int_{\Sigma_{k^n k^{n-1}}} (\mathbf{e}^n_{i,l})^T M^-_{k^n k^{n-1}} \mathbf{u}^n_k = \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} (\mathbf{e}^n_{i,l})^T \mathbf{u}^n_i.$$

One notices that $\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} (\mathbf{e}^n_{i,1})^T \mathbf{e}^n_{i,2} = 0$. Therefore using the decomposition of $\mathbf{u}_i^n$ (88) one finds

$$C^n_{T,i,1} = \alpha_i^n \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} (\mathbf{e}^n_{i,1})^T \mathbf{e}^n_{i,1} = h\alpha_i^n,$$

$$C^n_{T,i,2} = \beta_i^n \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} (\mathbf{e}^n_{i,2})^T \mathbf{e}^n_{i,2} = h(1 + \frac{\sigma_s^2 h^2}{48 c^2 \varepsilon^2}) \beta_i^n.$$

This completes the proof. □

Proposition C.6. *The scheme* (89) *reads*

$$\begin{cases} \dfrac{p_i^n - p_i^{n-1}}{\Delta t} + \dfrac{c}{2\varepsilon h}\Big[- p_{i+1} + 2p_i - p_{i-1} + (1-a)(v_{i+1} - v_{i-1})\Big]^n = 0, \\ (1 + \dfrac{a^2}{3})\dfrac{v_i^n - v_i^{n-1}}{\Delta t} + \dfrac{c}{2\varepsilon h}\Big[a^2(v_{i+1} + 2v_i + v_{i-1}) + (-v_{i+1} + 2v_i - v_{i-1}) \\ +(1+a)(p_{i+1} - p_{i-1})\Big]^n = -\dfrac{\sigma_s}{\varepsilon^2} v_i^n, \end{cases} \tag{98}$$

Proof. Starting from (89) one has

$$C_{T,i,1}^n - C_{T,i,1}^{n-1} + C_{S,i,1}^n = 0,$$
$$C_{T,i,2}^n - C_{T,i,2}^{n-1} + C_{S,i,2}^n = 0.$$

We recall the decomposition (88) $\mathbf{u}_i^n(x) = \alpha_i^n e_{i,1}^n(x) + \beta_i^n e_{i,2}(x) = (p_i^n, v_i^n)^T(x)$. In particular considering the center of the cell one finds $\alpha_i^n = p_i^n(x_i)$ and $\beta_i^n = v_i^n(x_i)$. Therefore using (91), (92), (96) and (97) in (89) and making the simplification $\alpha_i^n = p_i^n$ and $\beta_i^n = v_i^n$, one finally gets the scheme (98). This completes the proof. □

Acknowledgment

The authors thank the referees for their comments and remarks which deeply help to improve the final quality of this work.

References

[1] D. N. Arnold, *An interior penalty finite element method with discontinuous elements.*, SIAM J. Numer. Anal., 19 (1982), pp. 742–760.

[2] S. C. Brenner, *Poincaré–Friedrichs inequalities for piecewise H^1 functions.*, SIAM J. Numer. Anal., 41 (2003), pp. 306–324.

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

[4] C. Buet, B. Després, and E. Franck, *Asymptotic preserving schemes on distorted meshes for Friedrichs systems with stiff relaxation: application to angular models in linear transport.*, J. Sci. Comput., 62 (2015), pp. 371–398.

[5] C. Buet, B. Després, E. Franck, and T. Leroy, *Proof of uniform convergence for a cell-centered AP discretization of the hyperbolic heat equation on general meshes*, Mathematics of Computation, (2016).

[6] 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.

[7] O. Cessenat and B. Despres, *Application of an ultra weak variational formulation of elliptic pdes to the two-dimensional helmholtz problem*, SIAM J. Numer. Anal., 35 (1998), pp. 255–299.

[8] S. Chandrasekhar, *Radiative transfer.* (International Series of Monographs on Physics) Oxford: Clarendon Press; London: Oxford University Press. XIV, 394 p. (1950)., 1950.

[9] A. Ern and J.-L. Guermond, *Discontinuous Galerkin methods for Friedrichs' systems.*, in Numerical mathematics and advanced applications. Proceedings of ENUMATH 2005, the 6th European conference on numerical mathematics and advanced applications, Santiago de Compostela, Spain, July 18–22, 2005., Berlin: Springer, 2006, pp. 79–96.

[10] W. Fleming, *Functions of several variables. 2nd ed.* Undergraduate Texts in Mathematics. New York - Heidelberg - Berlin: Springer-Verlag. XI, 411 p. DM 41.00; $ 18.10 (1977)., 1977.

[11] K. O. Friedrichs, *Symmetric positive linear differential equations*, Communications on Pure and Applied Mathematics, 11 (1958), pp. 333–418.

[12] G. Gabard, *Discontinuous galerkin methods with plane waves for the displacement-based acoustic equation*, International Journal for Numerical Methods in Engineering, 66 (2006), pp. 549–569.

[13] G. Gabard, *Discontinuous galerkin methods with plane waves for time-harmonic problems*, Journal of Computational Physics, 225 (2007), pp. 1961–1984.

[14] 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.

[15] L. Gosse, *Computing qualitatively correct approximations of balance laws. Exponential-fit, well-balanced and asymptotic-preserving.*, Milano: Springer, 2013, https://doi.org/10.1007/978-88-470-2892-0.

[16] L. Gosse and G. Toscani, *An asymptotic-preserving well-balanced scheme for the hyperbolic heat equations.*, C. R., Math., Acad. Sci. Paris, 334 (2002), pp. 337–342.

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

[18] 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.

[19] 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.

[20] R. Hiptmair, A. Moiola, and I. Perugia, *A survey of trefftz methods for the helmholtz equation*, in Building Bridges: Connections and Challenges in Modern Approaches to Numerical Partial Differential Equations, vol. 114, Springer, 2016, pp. 237–278.

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

[22] L.-M. Imbert-Gérard, *Interpolation properties of generalized plane waves*, Numer. Math., 131 (2015), pp. 683–711, https://doi.org/10.1007/s00211-015-0704-y, http://dx.doi.org/10.1007/s00211-015-0704-y.

[23] L.-M. Imbert-Gérard, *Well-posedness and generalized plane waves simulations of a 2D mode conversion model*, J. Comput. Phys., 303 (2015), pp. 105–124, https://doi.org/10.1016/j.jcp.2015.09.033, http://dx.doi.org/10.1016/j.jcp.2015.09.033.

[24] 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, https://doi.org/10.1093/imanum/drt030, http://dx.doi.org/10.1093/imanum/drt030.

[25] S. Jin, *Asymptotic preserving (AP) schemes for multiscale kinetic and hyperbolic equations: a review.*, Riv. Mat. Univ. Parma (N.S.), 3 (2012), pp. 177–216.

[26] S. Jin and C. Levermore, *Numerical schemes for hyperbolic conservation laws with stiff relaxation terms.*, J. Comput. Phys., 126 (1996), pp. 449–467, art. no. 0149.

[27] S. Jin, M. Tang, and H. Han, *A uniformly second order numerical method for the one-dimensional discrete-ordinate transport equation and its diffusion limit with interface.*, Netw. Heterog. Media, 4 (2009), pp. 35–65.

[28] 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.

[29] Q. Li, J. Lu, and W. Sun, *Diffusion approximations and domain decomposition method of linear transport equations: asymptotics and numerics.*, J. Comput. Phys., 292 (2015), pp. 141–167, https://doi.org/10.1016/j.jcp.2015.03.014.

[30] E. A. Maunder, *Trefftz in translation.*, Comput. Assist. Mech. Eng. Sci., 10 (2003), pp. 545–563.

[31] J. Melenk and I. Babuška, *The partition of unity finite element method: Basic theory and applications.*, Comput. Methods Appl. Mech. Eng., 139 (1996), pp. 289–314.

[32] 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.

[33] J. Ragusa, J.-L. Guermond, and G. Kanschat, *A robust S_N-DG-approximation for radiation transport in optically thick and diffusive regimes.*, J. Comput. Phys., 231 (2012), pp. 1947–1962.

[34] M. Tang, *A uniform first-order method for the discrete ordinate transport equation with interfaces in X,Y-geometry.*, J. Comput. Math., 27 (2009), pp. 764–786.

[35] L. Wu and Y. Guo, *Geometric correction for diffusive expansion of steady neutron transport equation.*, Commun. Math. Phys., 336 (2015), pp. 1473–1553, https://doi.org/10.1007/s00220-015-2315-y.