Analysis of a combined spherical harmonics and discontinuous Galerkin discretization for the Boltzmann transport equation

Kenneth Assogba, Grégoire Allaire, Lahbib Bourhrara

Analysis of a combined spherical harmonics and discontinuous Galerkin discretization for the Boltzmann transport equation

Kenneth Assogba[1], Grégoire Allaire[2], Lahbib Bourhrara[1]
[1]Université Paris-Saclay, CEA, Service d'Études des Réacteurs
et de Mathématiques Appliquées
91191 Gif-sur-Yvette, France
[2]CMAP, École polytechnique, Institut Polytechnique de Paris
91120 Palaiseau, France

{kenneth.assogba, lahbib.bourhrara}@cea.fr, gregoire.allaire@polytechnique.fr

Abstract

In [11], a numerical scheme based on a combined spherical harmonics and discontinuous Galerkin finite element method for the resolution of the Boltzmann transport equation is proposed. One of its features is that a streamline weight is added to the test function to obtain the variational formulation. In this paper, we prove the convergence and provide error estimates of this numerical scheme. To this end, the original variational formulation is restated in a broken functional space. The use of broken functional spaces enables to build a conforming approximation, that is the finite element space is a subspace of the broken functional space. The setting of a conforming approximation simplifies the numerical analysis, in particular the error estimates, for which a Céa's type lemma and standard interpolation estimates are sufficient for our analysis. For our numerical scheme, based on $\mathbb{P}^k$ discontinuous Galerkin finite elements (in space) on a mesh of size h and a spherical harmonics approximation of order N (in the angular variable), the convergence rate is of order $\mathcal{O}\big(N^{-t} + h^k\big)$ for a smooth solution which admits partial derivatives of order $k+1$ and t with respect to the spatial and angular variables, respectively. For $k = 0$ (piecewise constant finite elements) we also obtain a convergence result of order $\mathcal{O}\big(N^{-t} + h^{1/2}\big)$. Numerical experiments in 1, 2 and 3 dimensions are provided, showing a better convergence behavior for the L^2-norm, typically of one more order, $\mathcal{O}\big(N^{-t} + h^{k+1}\big)$.

Keywords— Boltzmann transport equation · discontinuous Galerkin · spherical harmonics

1 Introduction

We are interested in the numerical resolution of the linear transport equation. This equation describes the streaming and collisions of neutral particles, for example neutrons, in matter. In the phase space $X = D \times \mathbf{S}^2$, where D is the space domain and $\mathbf{S}^2$ is the 3D unit sphere, the angular flux of particles u, subjected to sources q (collisions, possibly fission and external sources), at any point $(x, \omega) \in X$, is given as the solution of the linear transport equation

$$\begin{aligned} \omega \cdot \nabla u(x,\omega) + \sigma u(x,\omega) &= q(x,\omega) \quad \text{in } X, \\ u(x,\omega) &= f(x,\omega) \quad \text{on } \Gamma_-. \end{aligned}$$

The function f is a given incoming flux on the incoming boundary Γ_-. The most popular numerical methods to solve this equation, consist in removing the angular dependence by either approximating the streaming operator by a diffusion operator [2, 10] or by using a quadrature rule to approximate the angular flux [15]. A different approach is to approximate the angular flux by a truncated series of real spherical harmonics [31, Appendix A], it is the spherical harmonics method or $\mathrm{P_N}$ method [17].

The present work relies on the numerical scheme proposed by [11] for the resolution of the transport equation. This scheme combines the spherical harmonics method with the discontinuous Galerkin finite element method in space. It allows the treatment of 2D unstructured, non-conforming and curved meshes and 3D prismatic meshes.

Our previous studies carried out in [6, 7] show that the obtained accuracy is similar to that of high-fidelity solvers such as the method of characteristics [21, 33]. The purpose of this paper is to study the convergence and provide error estimates of this combined spherical harmonics – discontinuous Galerkin approximation.

Discontinuous Galerkin (DG) methods were introduced by Reed and Hill [37] to solve the neutron transport equation without imposing continuity at the interface between two cells. These schemes avoid spurious numerical oscillations when the solution of the transport problem is not smooth enough. In addition, DG methods provide the flexibility to handle unstructured and non-conforming meshes, that is meshes with hanging nodes. Shortly afterwards, Lesaint and Raviart [30, 29] proposed a first estimate of the numerical error of the DG method: for a triangular mesh of size h, an approximation by polynomials of degree k yields convergence of order $\mathcal{O}\big(h^k\big)$ in the L^2 norm. This estimate was further improved in [27, 25], where the error is proved to be actually $\mathcal{O}\Big(h^{k+\frac{1}{2}}\Big)$. Finally [40] establish a converges rate of $\mathcal{O}\big(h^{k+1}\big)$ on semi-uniform triangulations [40, §3].

In addition, [12] introduced the idea of writing the upwind DG as a jump penalty term on the mesh interfaces. However all these authors treated only the linear advection problem, i.e. the ω direction is fixed. For a more comprehensive overview of DG methods one refers to [20] for Friedrichs' systems, [3] for elliptic problems and [19] for a big picture.

Concerning the angular dependence of the problem, [17] proposed to approximate the angular flux $u(x,\omega)$ by a truncated series on a real spherical harmonics basis $\{y_n^m;\ 0 \le n \le N,\ -n \le m \le n\}$

$$u(x,\omega) \approx P_N u(x,\omega) = \sum_{n=0}^{N} \sum_{m=-n}^{n} u_n^m(x) y_n^m(\omega).$$

Like for Fourier series expansion, the error resulting from the approximation of u by the truncated expansion $P_N u$ is related to the smoothness of u regarding the angular variable. A result for Lipschitz functions is given by [23, Th. 1], and a general result for t-times continuously differentiable functions is provided by [36, Th. 3.3] and summarized in [8, Th. 7.5.10]. There exists a positive constant c such that, for any $v \in C^{t,\gamma}(\mathbf{S}^2)$ a γ-Hölder function t-times continuously differentiable, with $0 < \gamma \le 1$, and for every integer $N \ge 1$,

$$\|v - P_N v\|_{L^2(\mathbf{S}^2)} \le \frac{c}{N^{t+\gamma}} \|v\|_{C^{t,\gamma}(\mathbf{S}^2)}.$$

More recently, an alternative result has been proposed by [22, Lemma 3.3] for functions in Sobolev space $H^t(\mathbf{S}^2)$, see [8, §7.5.5]. There exists a positive constant c such that, for any $v \in H^t(\mathbf{S}^2)$, the truncation error satisfies

$$\|v - P_N v\|_{L^2(\mathbf{S}^2)} \le \frac{c}{N^t} \|v\|_{H^t(\mathbf{S}^2)}.$$

In both cases, the constant c is independent of N. A result for the eigenvalue neutron transport problem in 1D with isotropic scattering is given by [18]: for a layered medium, solving exactly with respect to the space variable, the error on the eigenvalue due to the truncation of the spherical harmonics expansion is found to be $O(\frac{1}{N^2})$.

On the analysis of the coupled angle-space problem, we refer to [35, 28, 4], where the authors use the discrete ordinates method and limit themselvesed to estimate for the scalar flux error. In [32, 38], a scaled least-squares finite element method for the neutron transport problem is proposed to well capture the diffusion limit. The $\mathrm{P_N}$ method is used for the angular discretization and the discretization error is found to be $\mathcal{O}\big(\frac{1}{N} + h^k\big)$ for an approximation by polynomials of degree k. More recent works [43, 22], in the context of radiative transfer, treat the angular flux but do not use the discontinuous Galerkin discretization.

In this paper we prove that the combined spherical harmonics – discontinuous Galerkin method converges as the discretization parameters tend to zero and we provide an error estimate on the angular flux. For this purpose, in section 2, the linear transport problem and its variational formulation (as proposed in [9]) are recalled. Section 3 is devoted to a new variational formulation, so-called broken, where no continuity between subdomains (or mesh cells) is imposed. This broken problem is shown to be equivalent to the original problem. It is then discretized in section 4 by $\mathbb{P}^k$ discontinuous Galerkin finite elements (in space) and a spherical harmonics approximation of order N (in the angular variable), and we prove that the discretized problem admits a unique solution. Finally, in sections 5 and 6, we prove that the numerical scheme is convergent and give its convergence rate (see our main results, Theorems 5.2 and 5.4). Lastly, section 7 provides numerical experiments which support and improve our error estimates.

2 Analysis of the original problem

In this section we introduce the linear Boltzmann transport problem used in the context of neutron transport [39, 16, 24]. Then the variational formulation, introduced in [9] and further studied in [11], is recalled. This weak formulation is the starting point of the broken weak formulation presented in section 3.

2.1 The neutron transport equation

Let us denote by x the spatial variable, which belongs to an open bounded set D in $\mathbb{R}^d$ ($d = 1$, 2 or 3), not necessarily convex with piecewise C^1 boundary ∂D. The direction variable ω belongs to $\mathbf{S}^2$, the unit sphere of $\mathbb{R}^3$. We consider the neutron transport equation in the phase space $X = D \times \mathbf{S}^2$, that is we look for the angular flux $u(x, \omega)$ which is the solution of:

$$\omega \cdot \nabla u + \sigma u = q \quad \text{in } X, \tag{1a}$$
$$u = f \quad \text{on } \Gamma_-. \tag{1b}$$

The outward unit normal of D is denoted by n, and the incoming Γ_- and outgoing Γ_+ boundaries of X are defined as:

$$\Gamma_\pm = \Gamma_\pm(D) = \{(x, \omega) \in \partial D \times \mathbf{S}^2;\ \pm\omega \cdot n(x) > 0\}.$$

The data of the problem are a given incoming flux f, an external source q and the total macroscopic cross section σ, assumed to be strictly positive. These functions are assumed to be given as:

$$f(x, \omega) \in L^2_- = L^2(\Gamma_-, |\omega \cdot n| \, \mathrm{d}s \, \mathrm{d}\omega), \tag{2a}$$
$$q(x, \omega) \in L^2(X), \tag{2b}$$
$$\sigma(x) \in L^\infty(D) \quad \text{and} \quad 0 < \sigma_0 \leq \sigma(x) \leq \sigma_\infty. \tag{2c}$$

We also define the space $L^2_+ = L^2(\Gamma_+, (\omega \cdot n) \, \mathrm{d}s \, \mathrm{d}\omega)$. We introduce the spaces V and W defined by:

$$V = V(D) = \left\{v \in L^2(X);\ \omega \cdot \nabla v \in L^2(X)\right\},$$
$$W = W(D) = \left\{v \in V;\ v|_{\Gamma_+} \in L^2_+\right\}, \tag{3}$$

with $v|_{\Gamma_+}$ denoting the restriction of v to Γ_+.

Let us denote $(u, v)_{L^2(X)}$ the standard scalar product in $L^2(X)$ and $(u, v)_{L^2_+} = \int_{\Gamma_+} uv(\omega \cdot n) \, \mathrm{d}s \, \mathrm{d}\omega$ the weighted scalar product of L^2_+, where $\mathrm{d}s$ is the surface measure on ∂D. The space W is then equipped with the scalar product $(u, v)_W$ defined by

$$(u, v)_W = (u, v)_{L^2(X)} + (\omega \cdot \nabla u, \omega \cdot \nabla v)_{L^2(X)} + (u, v)_{L^2_+}.$$

Equipped with the scalar product $(u, v)_W$, W is a Hilbert space. Let us denote by $\|\cdot\|_W$ the associated norm,

$$\|v\|^2_W = \|v\|^2_{L^2(X)} + \|\omega \cdot \nabla v\|^2_{L^2(X)} + \|v\|^2_{L^2_+}. \tag{4}$$

Remark 2.1. *The space W is required since the trace $v|_{\Gamma_+}$ on Γ_+ of functions in V do not satisfy in general $\int_{\Gamma_+} |v|^2(\omega \cdot n) \, \mathrm{d}s \, \mathrm{d}\omega < \infty$ (the same applies to the traces on Γ_-) [16, p. 219]. Thanks to [14, Proposition 1] it is known that the trace application $\gamma_- : W \to L^2_-$, such that $\gamma_-(u) = u|_{\Gamma_-}$, is continuous. As a consequence, definition* (3) *of the space W is equivalent to $W = \left\{v \in V;\ v|_{\Gamma_-} \in L^2_-\right\}$.*

Finally recall the Green formula for functions $(u, v) \in W \times W$ [16, p. 225]:

$$\int_X ((\omega \cdot \nabla u)v + u(\omega \cdot \nabla v)) \, \mathrm{d}x \, \mathrm{d}\omega = \int_\Gamma uv(\omega \cdot n) \, \mathrm{d}s \, \mathrm{d}\omega, \tag{5}$$

where $\Gamma = \partial D \times \mathbf{S}^2$. The above Green formula will play an important role in what follows. In the sequel, the volume mesure $\mathrm{d}x \, \mathrm{d}\omega$ and surface measure $\mathrm{d}s \, \mathrm{d}\omega$ will sometimes be ommitted in order to lighten the writing.

To conclude this section, let us recall a classical result of existence and uniqueness [16].

Lemma 2.2. *Under assumptions* (2) *the transport problem* (1) *admits a unique solution $u \in W$.*

Remark 2.3. *We adopt the assumptions of [13, 14] for the existence and uniqueness of solutions to transport problem* (1)*, where the convexity of D is not required. However, in Theorem 5.2, more regularity is required for the solutions, and this is usually obtained by making the assumption that D is convex, on top of smoothness of the data.*

2.2 Original variational formulation

Let $v \in W$ be a test function. Since it is assumed that σ is stritly positive, (1a) can be multiplied by $(v + \frac{1}{\sigma}\omega \cdot \nabla v)$ and integrated over the phase space $X = D \times \mathbf{S}^2$,

$$\int_X \left(\frac{1}{\sigma}(\omega \cdot \nabla u)(\omega \cdot \nabla v) + \sigma uv \right) + \int_X \left(u(\omega \cdot \nabla v) + (\omega \cdot \nabla u)v \right) = \int_X q \left(v + \frac{1}{\sigma}(\omega \cdot \nabla v) \right).$$

After using Green's formula (5) and boundary condition (1b), the resulting variational problem is written:

$$\text{find } u \in W \text{ such that } a(u, v) = L(v) \quad \forall v \in W, \tag{6}$$

with

$$a(u, v) = \int_X \left(\frac{1}{\sigma}(\omega \cdot \nabla u)(\omega \cdot \nabla v) + \sigma uv \right) + \int_{\Gamma_+} uv(\omega \cdot n), \tag{7}$$

$$L(v) = \int_X q \left(v + \frac{1}{\sigma}(\omega \cdot \nabla v) \right) - \int_{\Gamma_-} fv(\omega \cdot n). \tag{8}$$

Proposition 2.4 ([9]). *The variational formulation* (6) *admits a unique solution $u \in W$. Furthermore,* (6) *is equivalent to the original transport problem* (1)*, that is, if one admits a solution $u \in W$, u is also solution of the other.*

The proof of Proposition 2.4 can be found in [9]. It is based on the Lax-Milgram theorem, since the bilinear form a and the linear form L are continuous in W and on the other hand the form a is coercive.

3 A broken formulation

The goal of this section is to restate the original variational problem (6) on a broken functional space. The new formulation so obtained is called broken weak formulation. The main result is Theorem 3.8, stating that the broken problem is well-posed and its solution coincides with the solution to the original problem. In addition, we proved a weaker notion of continuity of $\tilde{a}(u, v)$ by choosing different norms for u and v, see Proposition 3.9. The use of a broken functional space enables to build a conforming approximation in section 4.

3.1 Functional setting

Let us introduce D_h a partition or mesh of the domain D into disjoint regions or mesh elements D_r

$$\overline{D} = \bigcup_{D_r \in D_h} \overline{D}_r. \tag{9}$$

Here, h denotes the meshsize, and is defined as the maximum of the diameter h_r of the mesh elements. It should be noted that the mesh D_h in what follows is not necessarily conformal in the sense of classical finite element [1]. We denote by $\mathcal{F}_h^i$ the set of interior faces (or interfaces), that is $F \in \mathcal{F}_h^i$ if there exist two distinct regions D_{r_1} and D_{r_2} such that $F = \partial D_{r_1} \cap \partial D_{r_2}$. Let $\mathcal{F}_h^b$ be the set of boundary faces, i.e $F \in \mathcal{F}_h^b$ if there exists a region D_r such that $F = \partial D_r \cap \partial D$ and $\mathcal{F}_h = \mathcal{F}_h^i \cup \mathcal{F}_h^b$, is the set of all faces. We assume that all faces $F \in \mathcal{F}_h$ are $(d-1)$-dimensional with non-vanishing Lebesgue measure. We define the local phase space $X_r = D_r \times \mathbf{S}^2$, with incoming and outgoing boundaries $\Gamma_\pm^r = \Gamma_\pm(D_r) = \{(x, \omega) \in \partial D_r \times \mathbf{S}^2,\ \pm\omega \cdot n_r(x) > 0\}$, where n_r is the outward unit normal of D_r.

With each element $D_r \in D_h$ we associate the space W_r of functions defined on X_r

$$W_r = W(D_r) = \left\{ v \in L^2(X_r);\ \omega \cdot \nabla v \in L^2(X_r),\ v|_{\Gamma_+^r} \in L^2_{r,+} = L^2(\Gamma_+^r, (\omega \cdot n_r)\, \mathrm{d}s\, \mathrm{d}\omega) \right\}.$$

We then introduce the space $\widetilde{W}$ as the collection of independent spaces W_r over the regions D_r

$$\widetilde{W} = \left\{ v \in L^2(X) \text{ such that } \forall D_r \in D_h,\ v|_{D_r} \in W_r \right\}, \tag{10}$$

with $v|_{D_r}$ denoting the restriction of v to D_r. The natural norm on $\widetilde{W}$ (making it a Hilbert space) is

$$\|v\|_{\widetilde{W}}^2 = \sum_r \left(\|v\|_{L^2(X_r)}^2 + \|\omega \cdot \nabla v\|_{L^2(X_r)}^2 + \|v\|_{L^2_{r,+}}^2 \right). \tag{11}$$

Remark 3.1. *Following Remark 2.1, the trace operators $\gamma^r_\pm : W_r \to L^2(\Gamma^r_\pm, |\omega \cdot n_r| \mathrm{d}s\, \mathrm{d}\omega)$ are continuous. However, for two regions D_{r_1} and D_{r_2} sharing the same face F, by denoting $v_r \in W_r$ the restriction of $v \in \widetilde{W}$ to D_r, in full generality $v_{r_1}|_F \neq v_{r_2}|_F$. In other words, the functions of $\widetilde{W}$ are not necessarily continuous through interfaces F.*

Definition 3.1. *To each internal face $F \in \mathcal{F}^i_h$ we associate a triple (D_{r_1}, D_{r_2}, n_F), where D_{r_1} and D_{r_2} are the two regions located on either side of the face F and n_F is the unit normal vector to the face F oriented from D_{r_1} to D_{r_2}, where by convention the labels are chosen such that $r_1 < r_2$. The vector n_F is called the face normal vector, which is outgoing from the region D_{r_1} and incoming into the region D_{r_2}. The region D_{r_1} is called the first region of face F and D_{r_2} is called the second region of face F. Furthermore, D_{r_1} and D_{r_2} are said to be adjacent to each other by the face F. Note that the normal vector n_F can be non-constant when the face F is not flat.*

Since functions $v \in \widetilde{W}$ have two traces on an inner face $F \in \mathcal{F}^i_h$, we define their mean and jump over F:

$$\{\!\{v\}\!\}_F(x,\omega) = \frac{1}{2}\left(v|_{D_{r_1}}(x,\omega) + v|_{D_{r_2}}(x,\omega)\right), \tag{12}$$

$$[\![v]\!]_F(x,\omega) = v|_{D_{r_1}}(x,\omega) - v|_{D_{r_2}}(x,\omega), \tag{13}$$

where D_{r_1} and D_{r_2} are respectively the first and the second region of the interface F in the sense of Definition 3.1. To lighten the notation we omit both the subscript F and variabes (x,ω), and simply write $\{\!\{v\}\!\}$ and $[\![v]\!]$. It is worth noticing that

$$[\![uv]\!] = \{\!\{u\}\!\}[\![v]\!] + [\![u]\!]\{\!\{v\}\!\}. \tag{14}$$

Green's formula (5) is not valid in general in the full domain D for $\widetilde{W}$-functions, but it is in each region D_r. Therefore we write a variant of (5) which is valid for $\widetilde{W}$-functions and will be of great use in what follows:

Lemma 3.2. *(Broken Green's formula) For any $(u,v) \in \widetilde{W} \times \widetilde{W}$ we have:*

$$\sum_r \int_{X_r} ((\omega \cdot \nabla u)v + u(\omega \cdot \nabla v))\, \mathrm{d}x\, \mathrm{d}\omega = \int_\Gamma uv(\omega \cdot n)\, \mathrm{d}s\, \mathrm{d}\omega + \sum_{F \in \mathcal{F}^i_h} \int_{F \times \mathbf{S}^2} [\![uv]\!](\omega \cdot n_F)\, \mathrm{d}s\, \mathrm{d}\omega\,. \tag{15}$$

Proof. Let $(u,v) \in \widetilde{W} \times \widetilde{W}$ and use Green's formula (5) for each phase space X_r.

$$\begin{aligned}\sum_r \int_{X_r} ((\omega \cdot \nabla u)v + u(\omega \cdot \nabla v))\, \mathrm{d}x\, \mathrm{d}\omega &= \sum_r \int_{\Gamma^r} uv(\omega \cdot n_r)\, \mathrm{d}s\, \mathrm{d}\omega \\ &= \sum_{F \in \mathcal{F}^b_h} \int_{F \times \mathbf{S}^2} uv(\omega \cdot n_F)\, \mathrm{d}s\, \mathrm{d}\omega + \sum_{F \in \mathcal{F}^i_h} \int_{F \times \mathbf{S}^2} [\![uv]\!](\omega \cdot n_F)\, \mathrm{d}s\, \mathrm{d}\omega\,,\end{aligned}$$

where $\Gamma^r = \partial D_r \times \mathbf{S}^2$. Formula (15) is obtained by recognizing that

$$\sum_{F \in \mathcal{F}^b_h} \int_{F \times \mathbf{S}^2} uv(\omega \cdot n_F)\, \mathrm{d}s\, \mathrm{d}\omega = \int_\Gamma uv(\omega \cdot n)\, \mathrm{d}s\, \mathrm{d}\omega\,.$$

□

The next lemma gives a sufficient condition for a discontinuous function in $\widetilde{W}$ to belong to W.

Lemma 3.3. *A function $u \in \widetilde{W}$ belongs to W if it satisfies*

$$\forall F \in \mathcal{F}^i_h \quad (\omega \cdot n_F)[\![u]\!]_F(x,\omega) = 0 \quad \text{for a.e. } (x,\omega) \in F \times \mathbf{S}^2. \tag{16}$$

Proof. This prof use the ideas developed in [19, §1.2.5] for the broken gradient. We start by defining a broken advection operator acting on the space $\widetilde{W}$, $A_h : \widetilde{W} \to L^2(D \times \mathbf{S}^2)$ is defined by:

$$\text{for all } D_r \in D_h, \quad (A_h v)|_{D_r} := \omega \cdot \nabla(v|_{D_r}).$$

Let $v \in \widetilde{W}$ and $\varphi \in C_c^\infty(D \times \mathbf{S}^2)$, integrating by part element-wise we observe that

$$\begin{aligned}\int_{D\times\mathbf{S}^2} v(\omega\cdot\nabla\varphi) &= \sum_{D_r}\int_{D_r\times\mathbf{S}^2} v(\omega\cdot\nabla\varphi)\\ &= -\sum_{D_r}\int_{D_r\times\mathbf{S}^2} \omega\cdot\nabla(v|_{D_r})\varphi + \sum_{D_r}\int_{\partial D_r\times\mathbf{S}^2} v\varphi(\omega\cdot n)\\ &= -\int_{D\times\mathbf{S}^2}(A_h v)\varphi + \sum_{F\in\mathcal{F}_h^i}\int_{F\times\mathbf{S}^2}[\![v\varphi]\!](\omega\cdot n_F) + \sum_{F\in\mathcal{F}_h^b}\int_{F\times\mathbf{S}^2} v\varphi(\omega\cdot n)\end{aligned}$$

Since φ is continuous over interfaces and vanishes on ∂D, we obtain

$$\int_{D\times\mathbf{S}^2} v(\omega\cdot\nabla\varphi) = -\int_{D\times\mathbf{S}^2}(A_h v)\varphi + \sum_{F\in\mathcal{F}_h^i}\int_{F\times\mathbf{S}^2}[\![v]\!]\varphi(\omega\cdot n_F).$$

If for all interfaces $[\![v]\!]_F(\omega\cdot n_F) = 0$, we obtain

$$\int_{D\times\mathbf{S}^2} v(\omega\cdot\nabla\varphi) = -\int_{D\times\mathbf{S}^2}(A_h v)\varphi \quad \forall\varphi \in C_c^\infty(D\times\mathbf{S}^2).$$

Meaning that the weak advective derivative of v, denoted $\omega\cdot\nabla v$ exist and is equal to $A_h v$ in $L^2(D\times\mathbf{S}^2)$. Noting that $\Gamma_+ \subset \bigcup_{D_r\in D_h}\Gamma_+^r$, we deduce $v|_{\Gamma_+} \in L^2(\Gamma_+, |\omega\cdot n|\,\mathrm{d}s\,\mathrm{d}\omega)$ and conclude that $v \in W$. □

Remark that equality (16) says nothing about the jump of a function $u \in \widetilde{W}$ in the directions ω which are tangent to the face F. One may wonder if there is a reciprocal to Lemma 3.3, that is, under what conditions a function $u \in W$ belongs to $\widetilde{W}$. The only obstacle is that such a function u should have traces in $L^2(\Gamma_+^r, (\omega\cdot n_r)\,\mathrm{d}s\,\mathrm{d}\omega)$, for any r, which is not obvious. Of course, as soon as u is bounded, it is true.

3.2 Local variational formulation

For each region D_r let us define the set of all its faces $\mathcal{F}_r$, its boundary faces $\mathcal{F}_r^b$ and its internal faces $\mathcal{F}_r^i$ by

$$\begin{aligned}\mathcal{F}_r &= \{F \in \mathcal{F}_h;\ F \subset \partial D_r\},\\ \mathcal{F}_r^b &= \{F \in \mathcal{F}_h^b;\ F \subset \partial D_r\},\\ \mathcal{F}_r^i &= \{F \in \mathcal{F}_h^i;\ F \subset \partial D_r\}.\end{aligned}$$

The subset $\mathcal{F}_r^b$ is empty if the region D_r does not intersect the boundary of the domain ∂D. The subset $\mathcal{F}_r^i$ is empty when there are no inner faces, in other words in the case where the mesh is reduced to a single region. Finaly note that $\partial D_r = (\cup_{F\in\mathcal{F}_r} F) = (\cup_{F\in\mathcal{F}_r^b} F) \cup (\cup_{F\in\mathcal{F}_r^i} F)$. For each face $F \in \mathcal{F}_r$, define the subsets $\Gamma_-^r(F)$ and $\Gamma_+^r(F)$ of Γ_-^r and Γ_+^r, respectively by:

$$\Gamma_\pm^r(F) = (F\times\mathbf{S}^2) \cap \Gamma_\pm^r.$$

The approach proposed in [11] consists in applying the variational formulation (6) for each region D_r by imposing as boundary conditions on ∂D_r:

$$u|_{\Gamma_-^r} = \begin{cases} f & \text{on } \Gamma_-^r(F) \quad \forall F \in \mathcal{F}_r^b,\\ u_r^F & \text{on } \Gamma_-^r(F) \quad \forall F \in \mathcal{F}_r^i,\end{cases} \tag{17}$$

where u_r^F is the trace on F of the flux in the adjacent region to D_r by the face F (in the sense of Definition 3.1). In other words, the incoming flux f, given by the boundary condition (1b), is imposed on the boundary faces $F \in \mathcal{F}_r^b$ of D_r and the outgoing flux from the adjacent region to D_r through the face F is imposed on the internal faces $F \in \mathcal{F}_r^i$. The condition $u|_{\Gamma_-^r} = u_r^F$ on the internal faces is called upwind condition in the literature [12].

As suggested in [11], a local variational formulation for each region D_r is then:

$$\text{find } u \in W_r, \quad \text{such that } \ a_r(u,v) = L_r(v) \quad \forall v \in W_r, \tag{18}$$

with

$$a_r(u,v) = \int_{X_r} \left(\frac{1}{\sigma}(\omega\cdot\nabla u)(\omega\cdot\nabla v) + \sigma uv \right) + \sum_{F\in\mathcal{F}_r^b} \int_{\Gamma_+^r(F)} uv(\omega\cdot n) + \sum_{F\in\mathcal{F}_r^i} \int_{\Gamma_+^r(F)} uv(\omega\cdot n_r), \tag{19}$$

$$L_r(v) = \int_{X_r} q\left(v + \frac{1}{\sigma}(\omega\cdot\nabla v) \right) - \sum_{F\in\mathcal{F}_r^b} \int_{\Gamma_-^r(F)} fv(\omega\cdot n) - \sum_{F\in\mathcal{F}_r^i} \int_{\Gamma_-^r(F)} u_r^F v(\omega\cdot n_r). \tag{20}$$

This local variational formulation is nothing but the previous variational formulation (6) applied in the region D_r, where the role of the internal or boundary faces is highlighted because of the different type of boundary condition in (17).

3.3 Global variational formulation

In order to pass from a local variational formulation to a global one (which will be our new broken variational formulation), we simply sum the local problems (18) over all regions D_r of the mesh and remark that $\widetilde{W}$ is just the collection of spaces W_r. Before performing this summation, let us give the resulting global or broken variational formulation:

$$\text{find } u\in\widetilde{W}, \quad \text{such that } \ \tilde{a}(u,v) = \tilde{L}(v) \quad \forall v\in\widetilde{W}, \tag{21}$$

where

$$\tilde{a}(u,v) = \sum_r \int_{X_r} \left(\frac{1}{\sigma}(\omega\cdot\nabla u)(\omega\cdot\nabla v) + \sigma uv \right) + \int_{\Gamma_+} uv(\omega\cdot n) + \tilde{a}^i(u,v), \tag{22}$$

$$\tilde{a}^i(u,v) = \sum_{F\in\mathcal{F}_h^i} \int_{F\times\mathbf{S}^2} \left(\{\!\{u\}\!\}(\omega\cdot n_F) + \frac{1}{2}[\![u]\!]|\omega\cdot n_F| \right) [\![v]\!], \tag{23}$$

$$\tilde{L}(v) = \sum_r \int_{X_r} q\left(v + \frac{1}{\sigma}(\omega\cdot\nabla v) \right) - \int_{\Gamma_-} fv(\omega\cdot n), \tag{24}$$

and $\{\!\{u\}\!\}, [\![u]\!]$ are defined by (12) and (13), respectively.

Proposition 3.4. *The collection of local variational formulations* (18) *is equivalent to the global (or broken) variational formulation* (21)*, in the sense that if u is a solution of one of them then u is a solution of the other.*

Proof. The proof is constructive in the sense that we shall obtain (21) by summing the local variational formulations (18). Doing so leads indeed to (21) but with a different formula for $\tilde{a}^i$ which is

$$\tilde{a}^i(u,v) = \sum_r \sum_{F\in\mathcal{F}_r^i} \left(\int_{\Gamma_+^r(F)} u_r v_r(\omega\cdot n_r) + \int_{\Gamma_-^r(F)} u_r^F v_r(\omega\cdot n_r) \right), \tag{25}$$

where u_r, v_r are the restrictions to X_r of u, v. Note that $\tilde{a}$, defined by (22), is not the sum of the a_r's and similarly $\tilde{L}$, defined by (24), is not the sum of the L_r's. Formulas (22) and (24) are obtained by observing that

$$\sum_r \sum_{F\in\mathcal{F}_r^b} \int_{\Gamma_+^r(F)} uv(\omega\cdot n) = \int_{\Gamma_+} uv(\omega\cdot n),$$

$$\sum_r \sum_{F\in\mathcal{F}_r^b} \int_{\Gamma_-^r(F)} fv(\omega\cdot n) = \int_{\Gamma_-} fv(\omega\cdot n),$$

and passing the integrals on the internal faces in L_r to the bilinear form $\tilde{a}$. The bilinear form $\tilde{a}^i$ gathers all contributions from the internal faces. The proof that (25) is equivalent (23) is given by Lemma 3.5 below. In other words we have just proved that, if u is a solution of the collection of local variational formulations (18) for each region D_r, then it is also a solution of (21) by the very construction of this global formulation. The converse is obtained by using in the global formulation (21) a test function v that vanishes everywhere except in a region D_r and using the expression (25) for the bilinear form $\tilde{a}^i(u,v)$. □

Recall the notations for the positive $x^\oplus$ and negative $x^\ominus$ parts of a real number x:

$$x^\oplus = \frac{1}{2}(|x|+x), \quad x^\ominus = \frac{1}{2}(|x|-x),$$

which satifies $(-x)^\oplus = x^\ominus$ and $(-x)^\ominus = x^\oplus$.

Lemma 3.5. *The bilinear form $\tilde{a}^i(u,v)$, defined by* (25)*, can be rewritten:*

$$\tilde{a}^i(u,v) = \sum_{F\in\mathcal{F}_h^i} \int_{F\times\mathbf{S}^2} \left(u_{r_1}(\omega\cdot n_F)^{\oplus} - u_{r_2}(\omega\cdot n_F)^{\ominus}\right)[\![v]\!], \tag{26}$$

where u_{r_1} and u_{r_2} are the fluxes in the first and second regions D_{r_1} and D_{r_2} of the face F in the sense of Definition 3.1. Finally, (26)*, and thus* (25)*, is equivalent to* (23)*.*

Proof. In (25) write the double sum $\sum_r \sum_{F\in\mathcal{F}_r^i}$ as a simple sum over all internal faces $F\in\mathcal{F}_h^i$, using for each face $F\in\mathcal{F}_h^i$ its associated triple (D_{r_1}, D_{r_2}, n_F):

$$\tilde{a}^i(u,v) = \sum_{F\in\mathcal{F}_h^i} \left(\int_{\Gamma_+^{r_1}(F)} u_{r_1}v_{r_1}(\omega\cdot n_{r_1}) + \int_{\Gamma_+^{r_2}(F)} u_{r_2}v_{r_2}(\omega\cdot n_{r_2}) + \int_{\Gamma_-^{r_1}(F)} u_{r_1}^F v_{r_1}(\omega\cdot n_{r_1}) + \int_{\Gamma_-^{r_2}(F)} u_{r_2}^F v_{r_2}(\omega\cdot n_{r_2}) \right),$$

where n_{r_1} is the unit normal vector, outgoing from D_{r_1} and similarly for n_{r_2}. Next we rely on the notations for the positive and negative parts to rewrite all integrals on $F\times\mathbf{S}^2$:

$$\tilde{a}^i(u,v) = \sum_{F\in\mathcal{F}_h^i} \int_{F\times\mathbf{S}^2} \left(u_{r_1}v_{r_1}(\omega\cdot n_{r_1})^{\oplus} + u_{r_2}v_{r_2}(\omega\cdot n_{r_2})^{\oplus} - u_{r_1}^F v_{r_1}(\omega\cdot n_{r_1})^{\ominus} - u_{r_2}^F v_{r_2}(\omega\cdot n_{r_2})^{\ominus}\right),$$

Replacing (n_{r_1}, n_{r_2}) by $(n_F, -n_F)$ and $(u_{r_1}^F, u_{r_2}^F)$ by (u_{r_2}, u_{r_1}), we deduce

$$\tilde{a}^i(u,v) = \sum_{F\in\mathcal{F}_h^i} \int_{F\times\mathbf{S}^2} \left(u_{r_1}v_{r_1}(\omega\cdot n_F)^{\oplus} + u_{r_2}v_{r_2}(\omega\cdot n_F)^{\ominus} - u_{r_2}v_{r_1}(\omega\cdot n_F)^{\ominus} - u_{r_1}v_{r_2}(\omega\cdot n_F)^{\oplus}\right),$$

which yields (26) after regrouping terms in u_{r_1} and in u_{r_2}. In order to conclude we use

$$\begin{aligned} u_1(\omega\cdot n_F)^{\oplus} - u_2(\omega\cdot n_F)^{\ominus} &= \frac{1}{2}u_1\left(|\omega\cdot n_F| + (\omega\cdot n_F)\right) - \frac{1}{2}u_2\left(|\omega\cdot n_F| - (\omega\cdot n_F)\right) \\ &= \frac{1}{2}[\![u]\!]\,|\omega\cdot n_F| + \{\!\{u\}\!\}(\omega\cdot n_F), \end{aligned}$$

which, applied to (26), leads to (23). □

3.4 Existence and uniqueness

Proposition 3.4 does not say anything on the existence of solutions for the broken variational formulation (21) (it is just an equivalence result). The goal of this subsection is to provide an existence and uniqueness result for (21) in a quite indirect way. Indeed, we are not able to apply standard results like the Lax-Milgram theorem. This will be clear because the norm (11) of $\widetilde{W}$ is too strong to prove coercivity of the bilinear form. Rather, we introduce a new weaker norm on $\widetilde{W}$ which is not equivalent to (11) and for which $\widetilde{W}$ is not closed.

Let us define a new norm $\|v\|_{\widetilde{W}^\star}$ on $\widetilde{W}$ by

$$\|v\|_{\widetilde{W}^\star}^2 = \sum_r \|v\|_{L^2(X_r)}^2 + \sum_r \|\omega\cdot\nabla v\|_{L^2(X_r)}^2 + \|v\|_{L_+^2}^2 + \sum_{F\in\mathcal{F}_h^i} \int_{F\times\mathbf{S}^2} [\![v]\!]^2 |\omega\cdot n_F|, \tag{27}$$

where $L_+^2 = L^2(\Gamma_+, (\omega\cdot n)\,\mathrm{d}s\,\mathrm{d}\omega)$. It is clear that there exist $C>0$ such that $\|v\|_{\widetilde{W}^\star} \le C\|v\|_{\widetilde{W}}$ for any $v\in\widetilde{W}$ but the reciprocal inequality does not hold true.

Proposition 3.6. *The bilinear form $\tilde{a}$, defined by* (22)*, is coercive on $\widetilde{W}$ for the norm* (27)*. Namely, for $\alpha = \frac{1}{2}\min(\sigma_0, \sigma_\infty^{-1}, 1)$ (independent of the choice of the mesh), we have*

$$\tilde{a}(v,v) \ge \alpha\|v\|_{\widetilde{W}^\star}^2 \quad \forall v\in\widetilde{W}. \tag{28}$$

Proof. For any $v\in\widetilde{W}$, using Lemma 3.5, we have

$$\tilde{a}(v,v) = \sum_r \int_{X_r} \left(\frac{1}{\sigma}(\omega\cdot\nabla v)^2 + \sigma v^2\right) + \int_{\Gamma_+} v^2(\omega\cdot n) + \sum_{F\in\mathcal{F}_h^i} \int_{F\times\mathbf{S}^2} \{\!\{v\}\!\}[\![v]\!](\omega\cdot n_F) + \frac{1}{2}[\![v]\!]^2|\omega\cdot n_F|. \tag{29}$$

The term $\int_{F\times\mathbf{S}^2}\{\!\{v\}\!\}[\![v]\!](\omega\cdot n_F)$, a priori signless, is eliminated by using Green's formula. For this purpose, first observe that

$$\int_{X_r}\left(\frac{1}{\sigma}(\omega\cdot\nabla v)^2+\sigma v^2\right)=\int_{X_r}\left(\frac{1}{2\sigma}\left((\omega\cdot\nabla v)^2+\sigma^2v^2+(\omega\cdot\nabla v+\sigma v)^2\right)-v(\omega\cdot\nabla v)\right),$$

thus

$$\sum_r\int_{X_r}\left(\frac{1}{\sigma}(\omega\cdot\nabla v)^2+\sigma v^2\right)=\sum_r a_r^\star(v)-\sum_r\int_{X_r}v(\omega\cdot\nabla v),$$

where

$$a_r^\star(v)=\int_{X_r}\frac{1}{2\sigma}\left((\omega\cdot\nabla v)^2+\sigma^2v^2+(\omega\cdot\nabla v+\sigma v)^2\right).$$

On the other hand, owing to Green's formula (15)

$$\sum_r\int_{X_r}v(\omega\cdot\nabla v)=\frac{1}{2}\int_\Gamma v^2(\omega\cdot n)+\frac{1}{2}\sum_{F\in\mathcal{F}_h^i}\int_{F\times\mathbf{S}^2}[\![v^2]\!](\omega\cdot n_F).$$

Next using $[\![v^2]\!]=2\{\!\{v\}\!\}[\![v]\!]$, we obtain

$$\sum_r\int_{X_r}\left(\frac{1}{\sigma}(\omega\cdot\nabla v)^2+\sigma v^2\right)=\sum_r a_r^\star(v)-\frac{1}{2}\int_\Gamma v^2(\omega\cdot n)-\sum_{F\in\mathcal{F}_h^i}\int_{F\times\mathbf{S}^2}\{\!\{v\}\!\}[\![v]\!](\omega\cdot n_F),$$

thus

$$\tilde a(v,v)=\sum_r a_r^\star(v)-\frac{1}{2}\int_\Gamma v^2(\omega\cdot n)+\int_{\Gamma_+}v^2(\omega\cdot n)+\frac{1}{2}\sum_{F\in\mathcal{F}_h^i}\int_{F\times\mathbf{S}^2}[\![v]\!]^2|\omega\cdot n_F|.$$

On the other hand

$$-\frac{1}{2}\int_\Gamma v^2(\omega\cdot n)+\int_{\Gamma_+}v^2(\omega\cdot n)=\int_\Gamma v^2\left(-\frac{1}{2}(\omega\cdot n)+(\omega\cdot n)^\oplus\right)=\frac{1}{2}\int_\Gamma v^2|\omega\cdot n|.$$

Thereby

$$\tilde a(v,v)=\sum_r\int_{X_r}\frac{1}{2\sigma}\left((\omega\cdot\nabla v)^2+\sigma^2v^2+(\omega\cdot\nabla v+\sigma v)^2\right)+\frac{1}{2}\int_\Gamma v^2|\omega\cdot n|+\frac{1}{2}\sum_{F\in\mathcal{F}_h^i}\int_{F\times\mathbf{S}^2}[\![v]\!]^2|\omega\cdot n_F|.$$

Since $0<\sigma_0\le\sigma\le\sigma_\infty$, we deduce

$$\tilde a(v,v)\ge\frac{1}{2}\sum_r\int_{X_r}\left(\sigma_\infty^{-1}(\omega\cdot\nabla v)^2+\sigma_0v^2\right)+\frac{1}{2}\int_{\Gamma_+}v^2|\omega\cdot n|+\frac{1}{2}\sum_{F\in\mathcal{F}_h^i}\int_{F\times\mathbf{S}^2}[\![v]\!]^2|\omega\cdot n_F|,$$

which implies the desired result (28). □

Proposition 3.7. *Let $u\in W$ be the unique solution to problem* (1)*. If $u\in\widetilde{W}$, then it solves the broken variational formulation* (21) *too.*

Proof. This is immediate by the very construction of the broken variational formulation (21). Indeed, let $u\in W$ be the solution of problem (1). If $u\in\widetilde{W}$, then its restriction to D_r is obviously a solution in W_r of the local variational formulation (18) by virtue of Proposition 2.4, applied to the region D_r. Then, by Proposition 3.4, u is also a solution of (21). The condition $u\in\widetilde{W}$ can be seen as a kind of boundedness assumption since it requires that u has suitable traces in $L^2(\Gamma_+^r,(\omega\cdot n_r)\,\mathrm{d}s\,\mathrm{d}\omega)$, for any subdomain r. □

Theorem 3.8. *(well-posedness). Assume that the solution $u\in W$ of* (1) *belongs to $\widetilde{W}$. Then, under assumptions* (2) *on the data f, q and σ, problem* (21) *admits a unique solution in $\widetilde{W}$, which coincides with the solution to* (1)*.*

Proof. The uniqueness of the solution follows from the coercivity of the bilinear form $\tilde a$ established in Proposition 3.6. The existence of a solution was proved in Propostion 3.7. □

3.5 Upper bound on the bilinear form

We obtain an upper bound on $\tilde{a}(u,v)$ in terms of the norm $\|v\|_{\widetilde{W}^\star}$, defined by (27) (the one used for coercivity), and a new stronger norm $\|u\|_{\widetilde{W}^+}$ defined by

$$\|u\|^2_{\widetilde{W}^+} = \|u\|^2_{\widetilde{W}^\star} + \sum_{F\in\mathcal{F}_h^i} \|\{\!\{u\}\!\}\|^2_{L^2_F}. \tag{30}$$

Here $L^2_F \coloneqq L^2(F\times\mathbf{S}^2, |\omega\cdot n_F|\,\mathrm{d}s\,\mathrm{d}\omega)$ and

$$\|\{\!\{v\}\!\}\|^2_{L^2_F} = \int_{F\times\mathbf{S}^2} \{\!\{v\}\!\}^2 |\omega\cdot n_F|.$$

By definition $\|v\|_{\widetilde{W}^\star} \le \|v\|_{\widetilde{W}^+}$, $\forall v\in\widetilde{W}$.

Proposition 3.9. *The bilinear form $\tilde{a}$, defined by (22), is bounded in the sense that there is $M>0$ such that*

$$|\tilde{a}(u,v)| \le M\|u\|_{\widetilde{W}^+}\|v\|_{\widetilde{W}^\star} \quad \textit{for all } (u,v)\in\widetilde{W}\times\widetilde{W}, \tag{31}$$

with M independent of the choice of the mesh.

Proof. Using $0<\sigma_0\le\sigma\le\sigma_\infty$ and Cauchy-Schwarz inequality, we can bound from above $\tilde{a}(u,v)$, defined by (22),

$$\begin{aligned}|\tilde{a}(u,v)| \le{}& \frac{1}{\sigma_0}\sum_r \|\omega\cdot\nabla u\|_{L^2_r}\,\|\omega\cdot\nabla v\|_{L^2_r} + \sigma_\infty\sum_r \|u\|_{L^2_r}\,\|v\|_{L^2_r} + \|u\|_{L^2_+}\|v\|_{L^2_+}\\ &+ \sum_{F\in\mathcal{F}_h^i}\|\{\!\{u\}\!\}\|_{L^2_F}\|[\![v]\!]\|_{L^2_F} + \frac{1}{2}\sum_{F\in\mathcal{F}_h^i}\|[\![u]\!]\|_{L^2_F}\|[\![v]\!]\|_{L^2_F},\end{aligned} \tag{32}$$

where $L^2_r = L^2(X_r)$. We now consider the entire right hand side of (32) as a single scalar product $\sum_{i=1}^I a_ib_i$ where the vector a collects all norms in u and b all norms in v. Using the inequality

$$\sum_{i=1}^I a_ib_i \le \left(\sum_{i=1}^I (a_i)^2\right)^{1/2}\left(\sum_{i=1}^I (b_i)^2\right)^{1/2},$$

we obtain

$$|\tilde{a}(u,v)| \le \max\left((\sigma_0)^{-1},\sigma_\infty,1\right)\alpha(u)\beta(v),$$

with

$$\begin{aligned}\alpha(u)^2 &= \sum_r\|\omega\cdot\nabla u\|^2_{L^2_r} + \sum_r\|u\|^2_{L^2_r} + \|u\|^2_{L^2_+} + \sum_{F\in\mathcal{F}_h^i}\|\{\!\{u\}\!\}\|^2_{L^2_F} + \sum_{F\in\mathcal{F}_h^i}\|[\![u]\!]\|^2_{L^2_F},\\ \beta(v)^2 &= \sum_r\|\omega\cdot\nabla v\|^2_{L^2_r} + \sum_r\|v\|^2_{L^2_r} + \|v\|^2_{L^2_+} + 2\sum_{F\in\mathcal{F}_h^i}\|[\![v]\!]\|^2_{L^2_F}.\end{aligned}$$

Recalling (27) and (30), we check that $\alpha(u)^2 = \|u\|^2_{\widetilde{W}^+}$ and $\beta(v)^2 \le 2\|v\|^2_{\widetilde{W}^\star}$, which yields the desired upper bound (31). □

4 The discrete problem

The goal of this section is to construct a discrete approximation of the problem (21). In order to do this, we build a finite dimension subspace $\widetilde{W}_{N,k}$ of $\widetilde{W}$ and consider the approximate problem:

$$\text{find } u\in\widetilde{W}_{N,k} \text{ such that } \tilde{a}(u,v) = \tilde{L}(v),\ \ \forall v\in\widetilde{W}_{N,k}$$

with the same bilinear form $\tilde{a}$ and linear form $\tilde{L}$ defined respectively at (22) and (24). The N index is related to the spherical harmonics method $\mathrm{P_N}$ used for the angular discretization and is described in section 4.1. In section 4.2 we present the Discontinuous Galerkin (DG) method, related to the index k, the maximum degree of the polynomials used in the spatial approximation.

4.1 Spherical harmonics method

The unit sphere $\mathbf{S}^2$ of $\mathbb{R}^3$ is parametrized by two angles θ and φ such that $\omega(\theta,\varphi) \in \mathbf{S}^2$ is defined by its components $\omega_x = \sin\theta\cos\varphi$, $\omega_y = \sin\theta\sin\varphi$ and $\omega_z = \cos\theta$, with $\theta \in [0,\pi]$ the colatitude (polar angle), and $\varphi \in [0,2\pi]$ the longitude. A complete orthonormal basis of $L^2(\mathbf{S}^2)$ is given by the real spherical harmonics $y_n^m(\omega)$ (see e.g [8, §7.5.5], [11]) and leads to the expansion formula of u:

$$u(x,\omega) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} u_n^m(x) y_n^m(\omega). \tag{33}$$

The components $u_n^m(x) = (u, y_n^m)_{L^2(\mathbf{S}^2)}$ of the flux u are called angular flux moments. The spherical harmonics or $\mathrm{P_N}$ method consist in the truncation of (33) to term of degree at most N:

$$u^N(x,\omega) = \sum_{n=0}^{N} \sum_{m=-n}^{n} u_n^m(x) y_n^m(\omega). \tag{34}$$

The operator $P_N : u \mapsto u^N$ define the orthogonal projection of $L^2(\mathbf{S}^2)$ to $\mathcal{H}_N$ the set of all finite linear combinations of spherical harmonics of degrees $n \le N$, $\mathcal{H}_N = \mathrm{span}\{y_n^m;\ 0 \le n \le N,\ -n \le m \le n\}$. Finally, we introduce the subspace $\widetilde{W}_N$ of $\widetilde{W}$-functions that can be written as linear combination of spherical harmonics of degree not exceeding N:

$$\widetilde{W}_N = \left\{ u \in \widetilde{W};\ u = \sum_{n=0}^{N} \sum_{m=-n}^{n} u_n^m(x) y_n^m(\omega) \right\}. \tag{35}$$

The finite dimension space $\widetilde{W}_{N,k}$ will be build as a subspace of $\widetilde{W}_N$.

4.2 Spatial approximation

The space discretization amounts to approximate the angular flux moments u_n^m by piecewise polynomial over the mesh D_h. To any $D_r \in D_h$ we associate a finite-dimensional space $\mathbb{P}^k(D_r)$ of d-variate polynomials of total degree at most k on D_r. We introduce the space of piecewise polynomials over D_h

$$\mathbb{P}_h^k = \left\{ v \in L^2(D);\ \forall D_r \in D_h,\ v|_{D_r} \in \mathbb{P}^k(D_r) \right\}. \tag{36}$$

Then we consider the finite-dimensional space $\widetilde{W}_{N,k}$ of functions in $\widetilde{W}_N$ that have piecewise polynomials angular moments u_n^m on the mesh D_h:

$$\widetilde{W}_{N,k} = \left\{ u(x,\omega) = \sum_{n=0}^{N} \sum_{m=-n}^{n} u_n^m(x) y_n^m(\omega) \text{ such that } u_n^m \in \mathbb{P}_h^k \right\}. \tag{37}$$

By construction $\widetilde{W}_{N,k} \subset \widetilde{W}_N \subset \widetilde{W}$.

4.3 The full discretization

The fully discretized problem that we consider is then stated as follows:

$$\text{find } u_h^N \in \widetilde{W}_{N,k} \text{ such that } \tilde{a}(u_h^N, v) = \tilde{L}(v),\ \ \forall v \in \widetilde{W}_{N,k}, \tag{38}$$

where $\tilde{a}$ and $\tilde{L}$ have been defined in (22) and (24) respectively.

Proposition 4.1. *The discrete problem* (38) *has a unique solution* $u_h^N \in \widetilde{W}_{N,k}$.

Proof. Since $\widetilde{W}_{N,k}$ is of finite dimension, the coercivity of $\tilde{a}$ obtained from Proposition 3.6 is sufficient to prove existence and uniqueness of a solution to problem (38). □

The $\mathrm{P_N}$ method comes in several variants depending on the variational formulation used and the spatial approximation adopted. Some of these variants present problems at the interfaces which are still not completely elucidated, the literature is rich in articles on this subject, see [41] and all the references cited therein. It is in part for this reason that some solvers based on the $\mathrm{P_N}$ method only deal with odd orders N. Proposition 4.1 shows that the numerical scheme studied in this work is not affected by these interface difficulties. The `NYMO` solver [11] based on this numerical scheme works for odd and even orders N.

Remark 4.2. *Once a basis of $\mathbb{P}_h^k$ is chosen, let us call it $(\varphi_j(x))_{1\le j\le J}$, the coefficients $u_{n,j}^m \in \mathbb{R}$ fully determine the discrete flux*

$$u_h^N(x,\omega) = \sum_{n=0}^{N} \sum_{m=-n}^{n} \sum_{j=1}^{J} u_{n,j}^m \varphi_j(x) y_n^m(\omega).$$

The dimension of the discrete space is

$$\dim \widetilde{W}_{N,k} = \dim(\mathbb{P}_h^k)\dim(\mathcal{H}_N) = \mathrm{card}(D_h)\binom{k+d}{d}(N+1)(\frac{d-1}{2}N+1).$$

The writing of the completely discretized problem (38) in matrix form reveals integrals over space and angle in the entries of the matrices. These integrals can be calculated exactly for a large class of meshes having plane or cylindrical faces, which can take into account the complexity of the geometries encountered in real applications in nuclear reactor calculations, see [11].

5 Error analysis

The goal of this section is to prove the convergence of the above numerical scheme and exhibit a precise order of convergence. If the broken variational formulation (21) were to fall in the scope of Lax-Milgram theorem, the error analysis would be fairly standard. However it is not the case. Indeed, the bilinear form $\tilde{a}$ is coercive for a weaker norm than that of $\widetilde{W}$, as shown in Proposition 3.6, and unfortunately, we are unable to prove the continuity of the bilinear form $\tilde{a}$ with the same norm, see Proposition 3.9. As a first step, using the coercivity result of Proposition 3.6 and upper bound on $\tilde{a}$ in Proposition 3.9, we obtain a Céa's Lemma 5.1. In a second step, using this Céa's lemma, we obtain an upper bound on the numerical error.

5.1 Céa's lemma

The goal is to derive an upper bound for the approximation error $u - u_h^N$, where u solves the original problem (1) and u_h^N is the numerical solution of (38).

Lemma 5.1. *Let $u \in W$ be the unique solution of (1) and assume that $u \in \widetilde{W}$. Let $u_h^N \in \widetilde{W}_{N,k}$ be the discrete solution of (38). There exists a constant C independent of k, h and N such that*

$$\big\|u - u_h^N\big\|_{\widetilde{W}^\star} \le C \inf_{v_h^N \in \widetilde{W}_{N,k}} \big\|u - v_h^N\big\|_{\widetilde{W}^+}. \tag{39}$$

Proof. Let $v_h^N \in \widetilde{W}_{N,k}$. Since $u \in \widetilde{W}$, from the continuous and discrete variational formulations (21) and (38), we infer $\tilde{a}(u_h^N, w_h^N) = \tilde{a}(u, w_h^N)$ for any $w_h^N \in \widetilde{W}_{N,k}$. So it is easy to verify $\tilde{a}(u_h^N - v_h^N, u_h^N - v_h^N) = \tilde{a}(u - v_h^N, u_h^N - v_h^N)$. Thus from the coercivity result of Proposition 3.6 and upper bound on $\tilde{a}$ in Proposition 3.9, we obtain

$$\alpha\big\|u_h^N - v_h^N\big\|_{\widetilde{W}^\star}^2 \le \tilde{a}(u_h^N - v_h^N, u_h^N - v_h^N) = \tilde{a}(u - v_h^N, u_h^N - v_h^N) \le M\big\|u - v_h^N\big\|_{\widetilde{W}^+}\big\|u_h^N - v_h^N\big\|_{\widetilde{W}^\star},$$

thus, we have $\big\|u_h^N - v_h^N\big\|_{\widetilde{W}^\star} \le \alpha^{-1}M\big\|u - v_h^N\big\|_{\widetilde{W}^+}$. To conclude, we write

$$\big\|u - u_h^N\big\|_{\widetilde{W}^\star} \le \big\|u - v_h^N\big\|_{\widetilde{W}^\star} + \big\|v_h^N - u_h^N\big\|_{\widetilde{W}^\star} \le \big(1 + \alpha^{-1}M\big)\big\|u - v_h^N\big\|_{\widetilde{W}^+},$$

from which we deduce (39), with $C = 1 + \alpha^{-1}M$, since v_h^N is arbitrary in $\widetilde{W}_{N,k}$. □

5.2 Error estimate

From now on it is assumed that $(D_h)_{h>0}$ is a sequence of admissible meshes [19, Definition 1.57] with mesh-width h going to zero. As is usual for obtaining error estimates of a numerical scheme, the solution u of the transport equation (1) is assumed to be smooth (and therefore obviously belongs to $\widetilde{W}$), more precisely to belong to some Sobolev space. Following the notations of [22], we introduce Sobolev spaces with mixed smoothness order. Let $s \in \mathbb{N}$ and $t \in \mathbb{N}$, two positive integers, $\alpha \in \mathbb{N}^d$ and $\beta \in \mathbb{N}^{d-1}$ two multi-indices, with $|\alpha| = \alpha_1 + \cdots + \alpha_d$. Let us denote by ∂_x^α the α-th weak derivative with respect to x and ∂_ω^β the β-th weak derivative with respect to ω. We then define

$$H^{s,t}(D \times \mathbf{S}^2) = \left\{u \in L^2(D \times \mathbf{S}^2);\ \partial_x^\alpha \partial_\omega^\beta u \in L^2(D \times \mathbf{S}^2),\ |\alpha| \le s,\ |\beta| \le t\right\}.$$

The space $H^{s,t}(D\times\mathbf{S}^2)$ is equipped with the norm

$$\|u\|^2_{H^{s,t}(D\times\mathbf{S}^2)} = \sum_{\substack{|\alpha|\le s\\ |\beta|\le t}} \left\|\partial_x^\alpha\partial_\omega^\beta u\right\|^2_{L^2(D\times\mathbf{S}^2)}.$$

We shall also used the semi-norm

$$|u|^2_{H^{s,t}(D\times\mathbf{S}^2)} = \sum_{\substack{|\alpha|= s\\ |\beta|= t}} \left\|\partial_x^\alpha\partial_\omega^\beta u\right\|^2_{L^2(D\times\mathbf{S}^2)}.$$

Our first main result is the following convergence theorem.

Theorem 5.2. *Assume that the unique solution u of the transport equation (1) belongs to $H^{k+1,t}(D\times\mathbf{S}^2)$. For $N\ge 1$ and $k\ge 1$, let u_h^N be the discrete solution of (38). The following error estimate holds true*

$$\left\|u-u_h^N\right\|_{\widetilde{W}^\star} \le \frac{c}{N^t}\|u\|_{H^{1,t}(D\times\mathbf{S}^2)} + ch^k\,|u|_{H^{k+1,0}(D\times\mathbf{S}^2)}\,.$$

As a consequence

$$\left\|u-u_h^N\right\|_{\widetilde{W}^\star} \le c\|u\|_{H^{k+1,t}(D\times\mathbf{S}^2)}\left(\frac{1}{N^t}+h^k\right).$$

Remark 5.3. *Theorem 5.2 does not prove convergence for $k=0$. Furthermore, for $k\ge 1$ it is suboptimal since we expect an order of convergence of the type $\mathcal{O}\big(N^{-t}+h^{k+1}\big)$, at least for the $L^2(D\times\mathbf{S}^2)$-norm. The numerical experiments of section 7 confirm that a better rate of convergence holds true.*

In the case $k=0$ of a piecewise constant approximation (which corresponds to the finite volume method), Theorem 5.2 can be improved because the streamline derivative term $\int_{X_r}(\omega\cdot\nabla u)(\omega\cdot\nabla v)$ vanishes in the bilinear form $\tilde a$ when one of its two arguments, u or v is piecewise constant in each element. Let us introduce a weaker norm $\|\cdot\|_{\widetilde{W}_0^\star}$ than $\|\cdot\|_{\widetilde{W}^\star}$, which is precisely defined by (27) but without the streamline derivative term, namely

$$\|u\|^2_{\widetilde{W}_0^\star} = \sum_r \|u\|^2_{L^2(X_r)} + \|u\|^2_{L^2_+} + \sum_{F\in\mathcal{F}_h^i}\int_{F\times\mathbf{S}^2}[\![u]\!]^2|\omega\cdot n_F|. \tag{40}$$

Our second main result is the following convergence theorem.

Theorem 5.4. *For $k=0$, assume that the unique solution u of the transport equation (1) belongs to $H^{1,t}(D\times\mathbf{S}^2)$. For $N\ge 1$, let $u_h^N\in\widetilde{W}_{N,0}$ be the discrete solution of (38). The following error estimate holds true*

$$\left\|u-u_h^N\right\|_{\widetilde{W}_0^\star} \le c\|u\|_{H^{1,t}(D\times\mathbf{S}^2)}\left(\frac{1}{N^t}+h^{\frac12}\right).$$

The proofs of both theorems are given in section 6.

6 Proof of the error estimates

6.1 Strategy of the proof of Theorem 5.2

The proof of Theorem 5.2 starts from inequality (39) in Céa's lemma with the special choice $u_h^N=\pi_h^k P_N u$

$$\left\|u-u_h^N\right\|_{\widetilde{W}^\star} \le C\left\|u-\pi_h^k P_N u\right\|_{\widetilde{W}^+},$$

where P_N is the angular projection operator, defined by $P_N u = u^N$ in (34) and π_h^k is the spatial projection operator, defined as the L^2-orthogonal projection of $L^2(D)$ onto $\mathbb{P}_h^k$ defined by (36),

$$\pi_h^k P_N u = \sum_{n,m}^{N}(\pi_h^k u_n^m)y_n^m. \tag{41}$$

Using the triangle inequality leads to

$$\left\|u-u_h^N\right\|_{\widetilde{W}^\star} \le C\left(\|e_N\|_{\widetilde{W}^+}+\|e_k\|_{\widetilde{W}^+}\right), \tag{42}$$

where $e_N = u-P_N u$ is the angular truncation error and $e_k = P_N u-\pi_h^k P_N u$ is the spatial approximation error inside the space $\widetilde{W}_N$. The aim is therefore to obtain separate estimates of $\|e_N\|_{\widetilde{W}^+}$ and $\|e_k\|_{\widetilde{W}^+}$.

6.2 Interpolation estimates

In this subsection we recall polynomial interpolation inequalities in Lemma 6.1 and a spherical harmonics approximation error in Lemma 6.2, as well as a bound on a partial sum of $|u_n^m|^2_{H^{k+1}(D_r)}$ in Lemma 6.3.

Lemma 6.1. *([19, §1.4.4] and [26, §4.3]) Let D_r be a cell of the mesh and F a face of the mesh. There exists a positive constant c, independent of D_r, F and h, such that the L^2-orthogonal projection operator π_h^k satisfies*

$$\left\|v - \pi_h^k v\right\|_{L^2(D_r)} \leq c h^{k+1} \left|v\right|_{H^{k+1}(D_r)} \qquad \forall v \in H^{k+1}(D_r), \tag{43}$$

$$\left|v - \pi_h^k v\right|_{H^1(D_r)} \leq c h^{k} \left|v\right|_{H^{k+1}(D_r)} \qquad \forall v \in H^{k+1}(D_r), \tag{44}$$

$$\left\|v - \pi_h^k v\right\|_{L^2(F)} \leq c h^{k+\frac{1}{2}} \left|v\right|_{H^{k+1}(D_r)} \qquad \forall v \in H^{k+1}(D_r). \tag{45}$$

Lemma 6.2. *([22, Lemma 3.2]) There exists a constant c, independent of N, such that, for any $g \in H^t(\mathbf{S}^2)$, the Soblove space on the unit sphere of order $t \in \mathbb{N}$, the truncation error satisfies*

$$\|g - P_N g\|_{L^2(\mathbf{S}^2)} \leq \frac{c}{N^t} \|g\|_{H^t(\mathbf{S}^2)}.$$

Finally we recall a classical result. In the sequel, we use the notation $\sum_{n,m}^N = \sum_{n=0}^N \sum_{m=-n}^{m=n}$.

Lemma 6.3. *Under the assumption that the exact solution u belongs to $H^{k+1,0}(D \times \mathbf{S}^2)$, the following inequality holds*

$$\sum_r \sum_{n,m}^N |u_n^m|^2_{H^{k+1}(D_r)} \leq |u|^2_{H^{k+1,0}(D\times\mathbf{S}^2)},$$

where u_n^m is the spherical harmonic decomposition of u, defined by (33).

Proof. Let $\alpha \in \mathbb{N}^d$ a multi-index with $|\alpha| = \sum_{i=1}^d \alpha_i$. By applying the derivation operator ∂_x^α to (33), it is easy to see

$$\int_{D_r\times\mathbf{S}^2} (\partial_x^\alpha u)^2 = \int_{D_r\times\mathbf{S}^2} \left(\sum_{n,m}^\infty (\partial_x^\alpha u_n^m) y_n^m \right)^2,$$

as the family $(y_n^m)_{n,m}$ form an orthonormal basis of $L^2(\mathbf{S}^2)$, we obtain

$$\|\partial_x^\alpha u\|^2_{L^2(D_r\times\mathbf{S}^2)} = \sum_{n,m}^\infty \|\partial_x^\alpha u_n^m\|^2_{L^2(D_r)},$$

which gives by summation over all α such that $|\alpha| = k+1$

$$|u|^2_{H^{k+1,0}(D_r\times\mathbf{S}^2)} = \sum_{n,m}^\infty |u_n^m|^2_{H^{k+1}(D_r)}.$$

The result announced in the lemma is obtained by summation over all regions of the mesh. □

6.3 End of the proof of Theorem 5.2

From inequality (42) it remains to provide upper bounds for $\|e_N\|_{\widetilde{W}^+}$ and $\|e_k\|_{\widetilde{W}^+}$. First, we rewrite

$$e_k = P_N u - \pi_h^k P_N u = \sum_{n,m}^N (u_n^m - \pi_h^k u_n^m) y_n^m,$$

and use the fact that $(y_n^m)_{n,m}$ is an orthonormal basis of $L^2(\mathbf{S}^2)$. We separate each term in the definition of the norm $\|e_k\|_{\widetilde{W}^+}$.

(i) Upper bound on $\|e_k\|_{L^2(X_r)}$

By othonormality of $(y_n^m)_{n,m}$ we obtain

$$\|e_k\|^2_{L^2(X_r)} = \sum_{n,m}^{N} \left\|u_n^m - \pi_h^k u_n^m\right\|^2_{L^2(D_r)}.$$

Using the interpolation inequality (43) leads to

$$\left\|u_n^m - \pi_h^k u_n^m\right\|_{L^2(D_r)} \leq ch^{k+1} \left|u_n^m\right|_{H^{k+1}(D_r)}.$$

It follows that,

$$\sum_r \|e_k\|^2_{L^2(X_r)} \leq (ch^{k+1})^2 \sum_r \sum_{n,m}^{N} |u_n^m|^2_{H^{k+1}(D_r)}.$$

Using Lemma 6.3 we conclude that

$$\sum_r \|e_k\|^2_{L^2(X_r)} \leq c(h^{k+1})^2 \left|u\right|^2_{H^{k+1,0}(D\times\mathbf{S}^2)}.$$

The estimates of the remaining terms follow the same kind of proof.

(ii) Upper bound on $\|\omega \cdot \nabla e_k\|_{L^2(X_r)}$

For the convective term, since $|\omega| = 1$, we have

$$\|\omega \cdot \nabla e_k\|^2_{L^2(X_r)} \leq \int_{D_r\times\mathbf{S}^2} \sum_{i=1}^{d} |\partial_{x_i} e_k|^2 \leq \sum_{n,m}^{N} \sum_{i=1}^{d} \int_{D_r} \left|\partial_{x_i}(u_n^m - \pi_h^k u_n^m)\right|^2 \leq \sum_{n,m}^{N} \left|u_n^m - \pi_h^k u_n^m\right|^2_{H^1(D_r)}.$$

Using the interpolation inequality (44) yields

$$\left|u_n^m - \pi_h^k u_n^m\right|_{H^1(D_r)} \leq ch^k \left|u_n^m\right|_{H^{k+1}(D_r)},$$

which, after summation, leads to

$$\sum_r \|\omega \cdot \nabla e_k\|^2_{L^2(X_r)} \leq ch^{2k} \left|u\right|^2_{H^{k+1,0}(D\times\mathbf{S}^2)}.$$

(iii) Upper bound on $\|e_k\|_{L^2_+}$

For the boundary term on ∂D, we rely on the interpolation inequality (45)

$$\|e_k\|^2_{L^2_+} = \sum_{F\in\mathcal{F}_h^b} \int_{F\times\mathbf{S}^2} e_k^2(\omega\cdot n)^{\oplus} \leq \sum_{F\in\mathcal{F}_h^b} \int_{F\times\mathbf{S}^2} e_k^2 \leq \sum_{n,m}^{N} \sum_{F\in\mathcal{F}_h^b} \int_F (u_n^m - \pi_h^k u_n^m)^2.$$

Inequality (45) implies

$$\left\|u_n^m - \pi_h^k u_n^m\right\|_{L^2(F)} \leq ch^{k+\frac{1}{2}} \left|u_n^m\right|_{H^{k+1}(D_r)},$$

thus

$$\|e_k\|^2_{L^2_+} \leq c(h^{k+\frac{1}{2}})^2 \left|u\right|^2_{H^{k+1,0}(D\times\mathbf{S}^2)}.$$

(iv) Upper bound on the jump $\|[\![e_k]\!]\|$ and mean $\|\{\!\{e_k\}\!\}\|$

A completely similar calulation leads to the following upper bound for the terms on the inner faces

$$\sum_{F\in\mathcal{F}_h^i} \int_{F\times\mathbf{S}^2} [\![e_k]\!]^2 |\omega\cdot n_F| + \int_{F\times\mathbf{S}^2} \{\!\{e_k\}\!\}^2 |\omega\cdot n_F| \leq c(h^{k+\frac{1}{2}})^2 \left|u\right|^2_{H^{k+1,0}(D\times\mathbf{S}^2)}.$$

(v) Upper bound on $\|e_k\|_{\widetilde{W}^+}$

Therefore, summing up all terms, we obtain

$$\|e_k\|_{\widetilde{W}^+} \leq ch^k \, |u|_{H^{k+1,0}(D\times\mathbf{S}^2)} \,. \tag{46}$$

(vi) Upper bound on $\|e_N\|_{\widetilde{W}^+}$

We now turn to the estimate of the angular approximation $e_N = u - P_N u$. By virtue of Lemma 6.2 we obtain

$$\sum_r \int_{D_r\times\mathbf{S}^2} |u - P_N u|^2 = \sum_r \int_{D_r} \|u - P_N u\|^2_{L^2(\mathbf{S}^2)} \leq \frac{c}{N^{2t}} \sum_r \int_{D_r} \|u\|^2_{H^t(\mathbf{S}^2)} = \frac{c}{N^{2t}} \|u\|^2_{H^{0,t}(D\times\mathbf{S}^2)}.$$

We have $|\omega\cdot\nabla(e_N)| \leq |\nabla e_N|$, since $|\omega| \leq 1$. Then,

$$\begin{aligned}\sum_r \int_{D_r\times\mathbf{S}^2} |\omega\cdot\nabla(u - P_N u)|^2 &\leq \sum_r \int_{D_r\times\mathbf{S}^2} |\nabla(u - P_N u)|^2 \\ &\leq \sum_r \int_{\mathbf{S}^2} |u - P_N u|^2_{H^1(D_r)} \leq \frac{c}{N^{2t}} \|u\|^2_{H^{1,t}(D\times\mathbf{S}^2)}.\end{aligned}$$

The other boundary terms are bounded from above in a similar fashion, leading to

$$\|e_N\|_{\widetilde{W}^+} \leq \frac{c}{N^t} \|u\|_{H^{1,t}(D\times\mathbf{S}^2)}. \tag{47}$$

Finally, summing inequalities (46) and (47) leads to the desired result.

6.4 Proof of Theorem 5.4

In this subsection the polynomial order is $k = 0$. In a first step, we revisit the proof of Céa's Lemma 5.1, using the fact that the streamline derivative of piecewise constant functions is zero. Recall that, by its very definition (40), the norm $\|\cdot\|_{\widetilde{W}_0^\star}$ satisfy $\|v\|_{\widetilde{W}_0^\star} \leq \|v\|_{\widetilde{W}^\star}$ for any $v \in \widetilde{W}$. Similarly, one can define a new norm $\|\cdot\|_{\widetilde{W}_0^+}$ from $\|\cdot\|_{\widetilde{W}^+}$ (see (30)) by removing the streamline derivative term, namely

$$\|u\|^2_{\widetilde{W}_0^+} = \|u\|^2_{\widetilde{W}_0^\star} + \sum_{F\in\mathcal{F}_h^i} \|\{\!\{u\}\!\}\|^2_{L^2_F}.$$

When $k = 0$ the result of Céa's lemma (with the same proof) reads

$$\|u - u_h^N\|_{\widetilde{W}_0^\star} \leq C \inf_{v_h^N\in\widetilde{W}_{N,k}} \|u - v_h^N\|_{\widetilde{W}_0^+}. \tag{48}$$

Indeed, coercivity of the bilinear form $\tilde{a}$ holds for the weaker norm $\|\cdot\|_{\widetilde{W}_0^\star}$. Furthermore, since there are no streamline derivatives in the formula for $\tilde{a}(u,v)$ if one of the functions u or v is piecewise constant, the upper bound on $\tilde{a}$ (coming from Proposition 3.9) is valid for the two norms $\|\cdot\|_{\widetilde{W}_0^\star}$ and $\|\cdot\|_{\widetilde{W}_0^+}$.

Eventually, the desired estimate is deduced from (48) by using the interpolation errors (43) and (45).

7 Numerical experiments

This section is devoted to some numerical tests performed with the `NYMO` software [11, 6, 7, 5], which is a P_N-transport solver of the CEA[1] reactor physics platform `APOLLO3`® [42, 34]. The goal is to compare the actual numerical errors with the theoretical error estimates obtained in section 5. To this end, we design numerical experiments of escalating complexity.

We are interested in the transport source problem, i.e. the incoming flux f is considered to be zero in (1). We prescribe a manufactured solution u, from which we deduce the corresponding source q such that equation (1) is satisfied. The manufactured solutions are constructed with a finite number of angular moments. The angular approximation orders are then chosen to be sufficiently high, so that the exact solution is in the approximation space. There is therefore no error in angle, and only the error in space is studied. The solution is chosen such

[1] Commissariat à l'Énergie Atomique et aux Énergies Alternatives

that it vanishes on the domain boundary. The chosen flux moments are polynomial, and when calculating the source q, their derivatives are calculated exactly.

From now on, the numerical solution is simply noted u_h. Then, we compute the L^2-error on the angular flux $\|u(x,\omega) - u_h(x,\omega)\|_{L^2(X)}$ and the L^2-error on the derivative in the streamline direction $\|\omega \cdot \nabla(u - u_h)\|_{L^2(X)}$. In practice, we merely compute the error projected in the space $\widetilde{W}_{N,k}$ [1, §6.2.4]. For two successive calculations where the mesh-width is divided by n, the order of convergence is calculated as $p = \log_n(\frac{e_h}{e_{\frac{h}{n}}})$.

For all experiments below, the calculation are performed in double-precision and linear systems are solved by GMRES with a 10^{-10} tolerance and a Jacobi preconditioner is used. Coordinates in 2 and 3 dimensions are denoted by $x = (x_1, x_2)$ and $x = (x_1, x_2, x_3)$.

7.1 1D homogeneous

The spatial domain is the interval $D = [0, 1]$. It is homogeneous in the sense that the total cross-section is constant on the domain, given by $\sigma = 0.48$. The transport source problem in 1D is written

$$\mu \frac{\partial u}{\partial x} + \sigma u = q \quad \text{in }]0,1[\times[-1,1] \ni (x,\mu), \qquad u = 0 \quad \text{on } \Gamma_-.$$

Recall that $y_n(\mu)$ are the normalized Legendre polynomials. The manufactured solution is chosen of order 2 in μ

$$u(x,\mu) = \sum_{n=0}^{N=2} u_n(x) y_n(\mu), \tag{49}$$

with $u_n(x) = x(1-x)$ for all n. It is necessary to use at least a P_3 approximation in angle in order to represent the source q. The mesh of D is uniformly refined and the errors on the angular flux $\|u(x,\mu) - u_h(x,\mu)\|_{L^2(X)}$ and on the derivative of the angular flux $\|\mu\partial_x(u(x,\mu) - u_h(x,\mu))\|_{L^2(X)}$ are displayed in Table 1.

Nbr. of elements	$\|u - u_h\|_{L^2(X)}$	Order (angular flux)	$\|\mu\partial_x(u - u_h)\|_{L^2(X)}$	Order (derivative)
$k = 0$ spatial approximation, $N = 3$ angular approximation				
2	1.910e-01	-	7.164e-01	-
20	2.423e-02	0.90	8.262e-01	-
200	2.463e-03	0.99	8.273e-01	-
2000	2.466e-04	1.00	8.273e-01	-
$k = 1$ spatial approximation, $N = 3$ angular approximation				
2	4.797e-02	-	4.137e-01	-
20	5.411e-04	1.95	4.136e-02	1.00
200	5.484e-06	1.99	4.136e-03	1.00
2000	5.492e-08	2.00	4.136e-04	1.00
$k = 2$ spatial approximation, $N = 3$ angular approximation				
2	2.213e-09	-	7.796e-09	-
20	2.721e-10	-	8.145e-09	-
200	3.191e-11	-	8.179e-09	-
2000	1.088e-11	-	8.182e-09	-

Table 1: Convergence orders for the scalar flux and its derivative for the 1D problem.

The numerical convergence orders are better than those given by Theorems 5.2 and 5.4. More precisely, we obtain a convergence rate of order $\mathcal{O}(k+1)$ for the angular flux in the L^2-norm (thus better by one order for $k \geq 1$ and by one half order for $k = 0$) and a convergence rate of order $\mathcal{O}(k)$ for the derivative in the L^2-norm (exactly as predicted by our theoretical analysis). Moreover the error is zero (up to the GMRES tolerance) when the approximation order reaches the one of the manufactured solution. Furthermore, in the $k = 1$ spatial approximation, $N = 3$ angular approximation case, we note that the error on the streamline derivative is divided

almost exactly by 10, resulting in a perfect slope of 1. The exact solution being continuous and polynomial, a quick calculation shows that in this case, the error depends solely on an integral on its derivative. This integral over a polynomial is calculated exactly.

7.2 2D homogeneous

Consider a square domain $D = [0,1]^2$. It is homogeneous in the sense that the total cross-section is constant on the domain, given by $\sigma = 0.48$. The manufactured solution is chosen as

$$u(x,\omega) = u_0^0(x)y_0^0(\omega) + u_1^{-1}(x)y_1^{-1}(\omega) + u_1^1(x)y_1^1(\omega),$$

with $u_0^0(x) = u_1^{-1}(x) = u_1^1(x) = x_1(1-x_1)x_2(1-x_2)$, and $y_n^m(\omega)$ being the real spherical harmonics. Here, the solution u is of order $N = 1$ regarding the angular variable, hence it is necessary to go at least to order $N = 2$ in angle in order to represent the source q. The mesh of D is uniformly refined as displayed on Figure 1. The error on the angular flux $\|u - u_h\|_{L^2(X)}$, the error on the derivative in the streamline direction $\|\omega \cdot \nabla(u - u_h)\|_{L^2(X)}$ and the error on the full gradient $\|\nabla(u - u_h)\|_{[L^2(X)]^2}$ are reported in Table 2.

The same remark on the convergence orders, as in the 1D case, are in order. The numerical errors match the theoretical error for the L^2-norm of the derivative, but is 1 less (for $k \geq 1$) or 1/2 less (for $k = 0$) for the L^2-norm of the angular flux. In addition, the rates of convergence for the streamline derivative and for the full gradient are the same. Consequently, only the error on the streamline derivative will be presented in the rest of this section. Finally, for $N = 2$ in angle and $k = 4$, the exact solution is included in the the approximation space, therefore the errors reaches the requested GMRES tolerance.

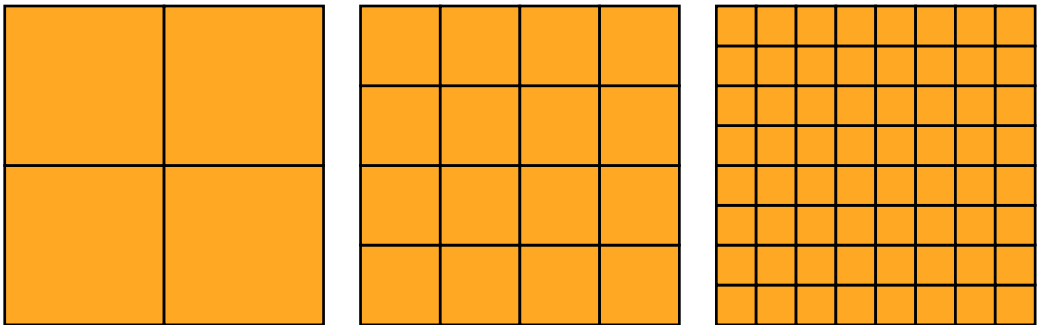

Figure 1: Uniform mesh refinement for the 2D problem.

7.3 3D homogeneous and heterogeneous

Consider a cubic domain $D = [0,1]^3$. The mesh is uniformly refined in each direction as shown in Figure 2. The manufactured solution is chosen as

$$u(x,\omega) = \sum_{n=0}^{N=1} \sum_{m=-n}^{m=n} u_n^m(x)y_n^m(\omega),$$

with

$$u_n^m(x) = \frac{m+2}{n+1} x_1^{n+1}(1-x_1)x_2^{m+2}(1-x_2)x_3(1-x_3),$$

for $n \in \{0,1\}$ and $m \in \{-n,\ldots,n\}$. The solution u is P_1 in angle and it is necessary to use at least order 2 in angle in order to represent the source q.

Two different experiments are performed. In Table 3 the medium is homogeneous, with a constant total cross-section $\sigma = 0.48$. In Table 4 the medium is heterogeneous and the total cross-sections are piecewise constant, given on Figure 2. As in 1D and 2D, the numerical convergence order for the derivative is in accordance with the theoretical one provided by Theorem 5.2, while it is one degree higher for the angular flux.

8 Conclusion

We proved an error estimate for the fully discrete linear Boltzmann transport equation using a discontinuous Galerkin – spherical harmonics method. More precisely, for $H^{k+1,t}(D \times \mathbf{S}^2)$ smooth solutions, we proved that an

Nbr. of elements	$\|u-u_h\|_{L^2(X)}$	Order	$\|\omega\cdot\nabla(u-u_h)\|_{L^2(X)}$	Order	$\|\nabla(u-u_h)\|_{[L^2(X)]^2}$	Order
$k=1$ spatial approximation, $N=2$ angular approximation						
4	1.039e-02	-	9.062e-02	-	1.511e-01	-
16	2.790e-03	1.90	5.147e-02	0.82	8.529e-02	0.83
64	6.878e-04	2.02	2.657e-02	0.95	4.373e-02	0.96
256	1.674e-04	2.04	1.339e-02	0.99	2.195e-02	0.99
$k=2$ spatial approximation, $N=2$ angular approximation						
4	1.348e-03	-	3.441e-02	-	5.699e-02	-
16	2.008e-04	2.75	9.608e-03	1.84	1.609e-02	1.82
64	2.700e-05	2.90	2.460e-03	1.97	4.154e-03	1.95
256	3.472e-06	2.96	6.187e-04	1.99	1.049e-03	1.99
$k=4$ spatial approximation, $N=2$ angular approximation						
4	3.306e-10	-	1.392e-09	-	2.113e-09	-
16	2.841e-10	-	1.433e-09	-	5.272e-09	-
64	3.232e-10	-	1.449e-09	-	1.364e-08	-
256	5.329e-11	-	1.446e-09	-	1.961e-09	-

Table 2: Convergence orders for the angular flux and its derivatives for the 2D problem.

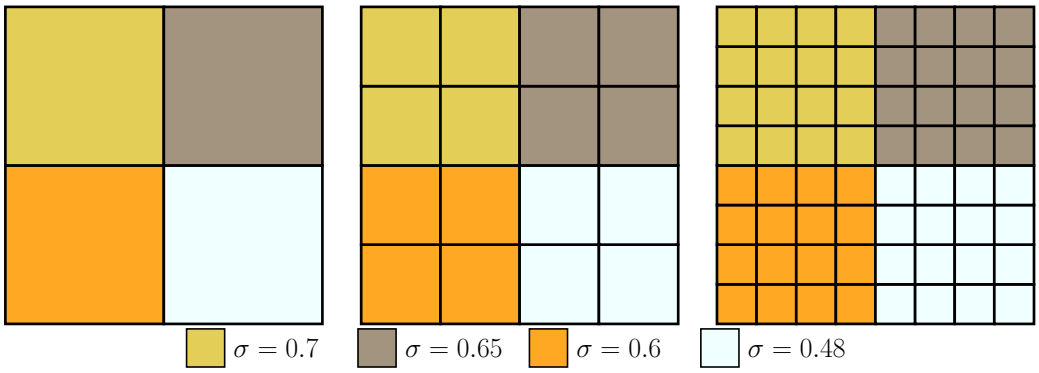

Figure 2: Radial section and uniform mesh refinement for the heterogeneous 3D problem.

approximation of order $k \geq 1$ in space and N in angle converges at rate $\mathcal{O}\big(N^{-t} + h^k\big)$. For $k = 0$ we obtained that the convergence rate is $\mathcal{O}\big(N^{-t} + h^{1/2}\big)$. Numerical experiments in 1, 2 and 3 dimensions show that our theoretical error estimate is optimal for the derivative but is pessimistic for the angular flux. Actually, the numerical error estimate is $\mathcal{O}\big(N^{-t} + h^{k+1}\big)$ for the L^2-norm of the angular flux.

Therefore an obvious perspective and future work is to improve our theoretical error estimate. So far, our efforts have not been successful although we tried some classical tricks like using weighted norms in terms of h as in [25], [19, §2]. Concerning other aspects not addressed by this study, the analysis of the eigenvalue problem (criticity) would be very relevant, as well as taking into account explicitly scattering and fission in the transport model, especially anisotropic scattering.

Acknowledgments

K. Assogba's PhD research work is supported by the CEA NUMERICS program, which has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 800945.

Nbr. of elements	$\|u - u_h\|_{L^2(X)}$	Order (angular flux)	$\|\omega \cdot \nabla(u - u_h)\|_{L^2(X)}$	Order (derivative)
$k = 3$ spatial approximation, $N = 2$ angular approximation				
8	1.651e-04	-	3.866e-03	-
64	1.051e-05	3.97	6.520e-04	2.57
512	6.554e-07	4.00	8.771e-05	2.89
4096	4.117e-08	3.99	1.116e-05	2.97
$k = 4$ spatial approximation, $N = 2$ angular approximation				
8	4.019e-05	-	1.294e-03	-
64	1.489e-06	4.75	1.043e-04	3.63
512	4.896e-08	4.93	6.923e-06	3.91
4096	1.555e-09	4.98	4.391e-07	3.97
$k = 5$ spatial approximation, $N = 2$ angular approximation				
8	9.888e-06	-	3.308e-04	-
64	1.867e-07	5.73	1.226e-05	4.75
512	3.085e-09	5.92	3.992e-07	4.94
4096	4.962e-11	5.96	1.260e-08	4.99

Table 3: Convergence orders for the angular flux and its derivative for the homogeneous 3D problem.

Nbr. of elements	$\|u - u_h\|_{L^2(X)}$	Order (angular flux)	$\|\omega \cdot \nabla(u - u_h)\|_{L^2(X)}$	Order (derivative)
$k = 3$ spatial approximation, $N = 2$ angular approximation				
8	1.619e-04	-	3.869e-03	-
64	1.041e-05	3.96	6.521e-04	2.57
512	6.496e-07	4.00	8.772e-05	2.89
4096	4.073e-08	4.00	1.116e-05	2.97
$k = 4$ spatial approximation, $N = 2$ angular approximation				
8	4.028e-05	-	1.295e-03	-
64	1.485e-06	4.76	1.043e-04	3.63
512	4.880e-08	4.93	6.924e-06	3.91
4096	1.551e-09	4.97	4.391e-07	3.98
$k = 5$ spatial approximation, $N = 2$ angular approximation				
8	9.808e-06	-	3.309e-04	-
64	1.853e-07	5.73	1.227e-05	4.75
512	3.066e-09	5.92	3.992e-07	4.94
4096	4.911e-11	5.96	1.260e-08	4.99

Table 4: Convergence orders for the angular flux and its derivative for the heterogeneous 3D problem.

References

[1] G. Allaire. *Numerical Analysis and Optimization: An Introduction to Mathematical Modelling and Numerical Simulation.* Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford, 2007.

[2] G. Allaire, X. Blanc, B. Després, and F. Golse. *Transport et diffusion.* Editions de l'École polytechnique, Palaiseau, 2018.

[3] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini. Unified Analysis of Discontinuous Galerkin Methods for Elliptic Problems. *SIAM Journal on Numerical Analysis*, 39(5):1749–1779, January 2002.

[4] M. Asadzadeh. Analysis of a Fully Discrete Scheme for Neutron Transport in Two-Dimensional Geometry. *SIAM Journal on Numerical Analysis*, 23(3):543–561, June 1986.

[5] K. Assogba and L. Bourhrara. The PN form of the Neutron Transport Problem Achieves Linear Scalability Through Domain Decomposition. In *To Appear in Proceedings of International Conference on Mathematics and Computational Methods Applied to Nuclear Science and Engineering (M&C 2023)*, Niagara Falls, Ontario, Canada, August 2023.

[6] K. Assogba, L. Bourhrara, I. Zmijarevic, and G. Allaire. Precise 3D Reactor Core Calculation Using Spherical Harmonics and Discontinuous Galerkin Finite Element Methods. In *Proceedings of International Conference on Physics of Reactors 2022 (PHYSOR 2022)*, pages 1224–1233, Pittsburgh, PA, United States, May 2022. American Nuclear Society.

[7] K. Assogba, L. Bourhrara, I. Zmijarevic, G. Allaire, and A. Galia. Spherical Harmonics and Discontinuous Galerkin Finite Element Methods for the Three-Dimensional Neutron Transport Equation: Application to Core and Lattice Calculation. *Nuclear Science and Engineering*, 197(8):1584–1599, August 2023.

[8] K. E. Atkinson and W. Han. *Theoretical Numerical Analysis: A Functional Analysis Framework*. Number 39 in Texts in Applied Mathematics. Springer, New York, 2nd ed edition, 2005.

[9] L. Bourhrara. New Variational Formulations for the Neutron Transport Equation. *Transport Theory and Statistical Physics*, 33(2):93–124, January 2004.

[10] L. Bourhrara. H1 Approximations of the Neutron Transport Equation and Associated Diffusion Equations. *Transport Theory and Statistical Physics*, 35(3-4):89–108, August 2006.

[11] L. Bourhrara. A new numerical method for solving the Boltzmann transport equation using the PN method and the discontinuous finite elements on unstructured and curved meshes. *Journal of Computational Physics*, 397, July 2019.

[12] F. Brezzi, L. D. Marini, and E. Süli. Discontinuous galerkin methods for first-order hyperbolic problems. *Mathematical Models and Methods in Applied Sciences*, 14(12):1893–1903, December 2004.

[13] M. Cessenat. Théorèmes de trace Lp pour des espaces de fonctions de la neutronique. *C. R. Acad. Sci. Paris*, 299(16):831–834, 1984.

[14] M. Cessenat. Théorèmes de trace pour des espaces de fonctions de la neutronique. *C. R. Acad. Sci. Paris*, 300(1):89–92, 1985.

[15] S. Chandrasekhar. On the Radiative Equilibrium of a Stellar Atmosphere. II. *The Astrophysical Journal*, 100:76, 1944.

[16] R. Dautray and J.-L. Lions. *Mathematical Analysis and Numerical Methods for Science and Technology*, volume 6. Springer Berlin Heidelberg, Berlin, Heidelberg, 1988.

[17] B. Davison. Spherical-harmonics method for neutron transport theory problems with incomplete symmetry. *Canadian Journal of Physics*, 36(4):462–475, April 1958.

[18] B. Davison. On the rate of convergence of the spherical harmonics method: (for the plane case, isotropic scattering). *Canadian Journal of Physics*, 38(11):1526–1545, November 1960.

[19] D. A. Di Pietro and A. Ern. *Mathematical Aspects of Discontinuous Galerkin Methods*. Number 69 in Mathématiques et Applications. Springer, Berlin ; New York, 2012.

[20] A. Ern and J. L. Guermond. Discontinuous Galerkin Methods for Friedrichs' Systems. I. General theory. *SIAM Journal on Numerical Analysis*, 44(2):753–778, January 2006.

[21] A. Gammicchia, S. Santandrea, and S. Dulla. Cross sections polynomial axial expansion within the APOLLO3® 3D characteristics method. *Annals of Nuclear Energy*, 165:108673, January 2022.

[22] K. Grella and Ch. Schwab. Sparse tensor spherical harmonics approximation in radiative transfer. *Journal of Computational Physics*, 230(23):8452–8473, September 2011.

[23] T. H. Gronwall. On the Degree of Convergence of Laplace's Series. *Transactions of the American Mathematical Society*, 15(1):1–30, 1914.

[24] A. Hébert. *Applied Reactor Physics*. Presses internationales Polytechnique, 2016.

[25] C. Johnson and J. Pitkäranta. An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation. *Mathematics of Computation*, 46(173):1–26, 1986.

[26] C. Johnson. *Numerical Solution of Partial Differential Equations by the Finite Element Method*. Cambridge University Press, Cambridge, cambridge university press edition, 1987.

[27] C. Johnson, U. Nävert, and J. Pitkäranta. Finite element methods for linear hyperbolic problems. *Computer Methods in Applied Mechanics and Engineering*, 45(1):285–312, September 1984.

[28] C. Johnson and J. Pitkäranta. Convergence of a Fully Discrete Scheme for Two-Dimensional Neutron Transport. *SIAM Journal on Numerical Analysis*, 20(5):951–966, October 1983.

[29] P. Lesaint. *Sur la résolution des systèmes hyperboliques du premier ordre par des méthodes d'éléments finis*. PhD thesis, Université Paris VI, 1975.

[30] P. Lesaint and P. A. Raviart. On a Finite Element Method for Solving the Neutron Transport Equation. In C. de Boor, editor, *Mathematical Aspects of Finite Elements in Partial Differential Equations*, pages 89–123. Academic Press, January 1974.

[31] E. E. Lewis and W. F. Miller. *Computational Methods of Neutron Transport*. Wiley, New York, 1984.

[32] T. A. Manteuffel and K. J. Ressel. Least-Squares Finite-Element Solution of the Neutron Transport Equation in Diffusive Regimes. *SIAM Journal on Numerical Analysis*, 35(2):806–835, January 1998.

[33] E. Masiello, R. Sanchez, and I. Zmijarevic. New Numerical Solution with the Method of Short Characteristics for 2-D Heterogeneous Cartesian Cells in the APOLLO2 Code: Numerical Analysis and Tests. *Nuclear Science and Engineering*, 161(3):257–278, March 2009.

[34] P. Mosca, L. Bourhrara, A. Calloo, A. Gammicchia, F. Goubioud, L. Mao, F. Madiot, F. Malouch, E. Masiello, F. Moreau, S. Santandrea, D. Sciannandrone, I. Zmijarevic, E. Y. Garcias-Cervantes, G. Valocchi, J. Vidal, F. Damian, P. Laurent, A. Willien, A. Brighenti, L. Graziano, and B. Vezzoni. APOLLO3®: Overview of the new code capabilities for reactor physics analysis. In *To Appear in Proceedings of International Conference on Mathematics and Computational Methods Applied to Nuclear Science and Engineering (M&C 2023)*, 2023.

[35] J. Pitkäranta and L. R. Scott. Error Estimates for the Combined Spatial and Angular Approximations of the Transport Equation for Slab Geometry. *SIAM Journal on Numerical Analysis*, 20(5):922–950, October 1983.

[36] D. L. Ragozin. Constructive polynomial approximation on spheres and projective spaces. *Transactions of the American Mathematical Society*, 162:157–170, 1971.

[37] W. Reed and T. Hill. Triangular mesh methods for the neutron transport equation. Technical Report LA-UR-73-479, Los Alamos Scientific Laboratory, United States, 1973.

[38] K. J. Ressel. *Least-Squares Finite-Element Solution of the Neutron Transport Equation in Diffusive Regimes*. PhD thesis, University of Colorado at Denver, 1994.

[39] P. Reuss and J. Bussac. *Traité de neutronique*. Hermann, Paris, 1978.

[40] G. R. Richter. An optimal-order error estimate for the discontinuous Galerkin method. *Mathematics of Computation*, 50(181):75–88, 1988.

[41] R. Sanchez. On PN Interface and Boundary Conditions. *Nuclear Science and Engineering*, 177(1):19–34, May 2014.

[42] D. Schneider, F. Dolci, F. Gabriel, J.-M. Palau, M. Guillo, and B. Pothet. APOLLO3® CEA/DEN deterministic multi-purpose code for reactor physics analysis. In *PHYSOR 2016 – Unifying Theory and Experiments in the 21st Century*, Sun Valley, United States, May 2016.

[43] G. Widmer, R. Hiptmair, and Ch. Schwab. Sparse adaptive finite elements for radiative transfer. *Journal of Computational Physics*, 227(12):6071–6105, June 2008.