IMPROVED FULLY-IMPLICIT SPHERICAL HARMONICS METHODS FOR FIRST AND SECOND ORDER FORMS OF THE TRANSPORT EQUATION USING GALERKIN FINITE ELEMENT

A Dissertation

by

VINCENT MATTHIEU LABOURE

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Ryan G. McClarren
Committee Members,	Marvin L. Adams
	Jean-Luc Guermond
	Jim E. Morel
Head of Department,	Yassin A. Hassan

August 2016

Major Subject: Nuclear Engineering

ABSTRACT

In this dissertation, we focus on solving the linear Boltzmann equation – or transport equation – using spherical harmonics (P$_N$) expansions with fully-implicit time-integration schemes and Galerkin Finite Element spatial discretizations within the Multiphysics Object Oriented Simulation Environment (MOOSE) framework. The presentation is composed of two main ensembles.

On one hand, we study the first-order form of the transport equation in the context of Thermal Radiation Transport (TRT). This nonlinear application physically necessitates to maintain a positive material temperature while the P$_N$ approximation tends to create oscillations and negativity in the solution. To mitigate these flaws, we provide a fully-implicit implementation of the Filtered P$_N$ (FP$_N$) method and investigate local filtering strategies. After analyzing its effect on the conditioning of the system and showing that it improves the convergence properties of the iterative solver, we numerically investigate the error estimates derived in the linear setting and observe that they hold in the non-linear case. Then, we illustrate the benefits of the method on a standard test problem and compare it with implicit Monte Carlo (IMC) simulations.

On the other hand, we focus on second-order forms of the transport equation for neutronics applications. We mostly consider the Self-Adjoint Angular Flux (SAAF) and Least-Squares (LS) formulations, the former being globally conservative but void incompatible and the latter having – in all generality – the opposite properties. We study the relationship between these two methods based on the weakly-imposed LS boundary conditions. Equivalences between various parity-based P$_N$ methods are also established, in particular showing that second-order filters are not an appropriate fix to retrieve void compatibility. The importance of global conservation is highlighted on a heterogeneous multigroup k-eigenvalue test problem.

Based on these considerations, we propose a new method that is both globally conservative and compatible with voids. The main idea is to solve the LS form in the void regions and the SAAF form elsewhere. For the LS form to be conservative in void, a non-symmetric fix is required, yielding the Conservative LS (CLS) formulation. From there, an hybrid SAAF–CLS method can be derived, having the desired properties. We also show how to extend it to near-void regions and time-dependent problems. While such a second-order form already existed for discrete-ordinates (S_N) discretizations (Wang et al. 2014), we believe that this method is the first of its kind, being well-suited to both S_N and P_N discretizations.

DEDICATION

A mes parents, qui m'ont toujours laissé libre de mes décisions

If you can dream - and not make dreams your master;

If you can think - and not make thoughts your aim;

If you can meet with Triumph and Disaster

And treat those two impostors just the same;

[...]

Yours is the Earth and everything that's in it

If – Rudyard Kipling

ACKNOWLEDGEMENTS

This work has been very enjoyable and I am so grateful to my advisor, Ryan McClarren, for giving me that opportunity as well as his continuous support, ideas and friendliness. I am also indebted to the other members of my committee, Marvin Adams, Jim Morel and Jean-Luc Guermond for being inspiring and devoted professors. I do not forget my undergraduate professors Michel Colin and Jean-Loup Baudy who significantly contributed to my love for mathematics and physics.

My deepest gratitude also goes out to my mentors at the Idaho and Oak Ridge National Laboratories (INL, ORNL), respectively Yaqi Wang and Cory Hauck, from whom I have learned tremendously and who deeply influenced this work. The positive influence from Matthieu Guillo, my mentor at Electricité de France (EDF), is much valued, too. I truly appreciate the support from Phillip Finck who has motivated and helped me achieve my goal of pursuing this degree in the United States.

I will never be able to thank my family enough, especially my parents, for always being supporting and great examples. The Piepers and the Gormans hold a very dear place in my heart for being like family to me, in Idaho and Tennessee, respectively. I am very thankful to Oliver Hatfield and his family, for their continuous friendship.

I am grateful for many other great people in my life, such as my fellow graduate students Weixiong Zheng, Marco Delchini, Simon Bolding and many more. Special thanks also go to Tarek Ghaddar, Carolyn McGraw, Anthony Barbu and Pablo Vaquer for providing very valuable results to compare my own code with.

Lastly, I appreciate the National Science Foundation, INL and ORNL for supporting my research with financial and computational resources. In particular, I very much value the help from the MOOSE and Rattlesnake team at the INL.

NOMENCLATURE

Abbreviations

AFE	Angular Flux Equation
AMG	Algebraic Multigrid
BDF-2	Implicit time discretization method based on the Backward Differentiation Formula (BDF) method
C5G7	A reactor physics benchmark problem with 7 energy groups and 5 different configurations
CG	Conjugate Gradient
CGFEM	Continuous Galerkin Finite Element Method
CFL	Courant–Friedrichs–Lewy
CLS	Conservative Least-Squares
DGFEM	Discontinuous Galerkin Finite Element Method
gmsh	A multi-dimensional finite element mesh generator
GMRES	Generalized Minimum Residual
IMC	Implicit Monte-Carlo, a class of stochastic methods for Thermal Radiation Transport
INL	Idaho National Laboratory
libMesh	A C++ framework designed for the numerical resolution of partial differential equations
LS	Least-Squares
MMS	Method of Manufactured Solutions
MOOSE	Multiphysics Object Oriented Simulation Environment, a finite element framework developed by the Idaho National Laboratory

NDA	Nonlinear Diffusion Acceleration
ORNL	Oak Ridge National Laboratory
pcm	per cent mille (one one-thousandth of a percent)
PETSc	Portable, Extensible Toolkit for Scientific Computation, a parallel software library for numerical calculations, developed by the Argonne National Laboratory
PJNFK	Preconditioned Jacobian-Free Newton Krylov
P$_N$	Spherical Harmonics
Rattlesnake	Radiation transport solver based on MOOSE developed by the Idaho National Laboratory
SAAF	Self-Adjoint Angular Flux
SAAF–LS	Hybrid method combining the Self-Adjoint Angular Flux and Least-Squares methods
SAAF–CLS	Hybrid method combining the Self-Adjoint Angular Flux and Conservative Least-Squares methods
SAAF–VT	Self-Adjoint Angular Flux with a Void Treatment
S$_N$	Discrete Ordinates
SOR	Successive Over-relaxation
SPD	Symmetric Positive Definite
SSpline	Spherical Spline, type of filter defined by Eq. 3.14
VisIt	Open source, interactive, scalable, visualization, animation and analysis tool , developed by the Lawrence Livermore National Laboratory

Symbols (with default units, as appropriate)

A	Global matrix
a	Bilinear form

a	Radiation constant ($a = (8\pi^5 k^4)/(15h^3c^3) = 7.56573164 \times 10^{-16}$ J–m^{-3}–K^{-4})
B	Frequency-integrated Planckian blackbody source (in J–m^{-2}–s^{-1})
C	Fully-discretized collision operator
C_ℓ^m	Spherical harmonics normalization constant
C_v	Material heat capacity (in J–m^{-3}–K^{-1})
c	Speed of light in vacuum ($c = 299792458$ m/s)
c	Scaling parameter to weakly impose the Least-Squares boundary conditions (in m^{-1}), further assumed to be a strictly positive constant in Chapter 5
$\mathcal{D}$	Spatial domain
$\mathcal{D}_0$	Subdomain composed of all the void (or near-void) regions in the spatial domain
$\mathcal{D}_1$	Subdomain composed of all the non-void regions in the domain
d	Number of spatial dimensions the solution depends on
$\mathrm{dist}(\cdot,\cdot)$	Minimum distance between the elements of two sets
E	Energy of the particle (in J)
E	L^2-error of the angular flux, defined by Eq. 3.51
e	Material internal energy (in J)
$\mathrm{eig}(\cdot)$	Set of the eigenvalues of a matrix
eV	Electronvolt, fundamentally an energy unit but used here as a temperature unit, by abuse of notations (1 eV $= q\tilde{v}_0/k$ K, $\tilde{v}_0 = 1$ V)
$\vec{e}_\chi$	Unit vector defining the χ-axis ($\chi \in \{x, y, z\}$); may also be referred to as $\vec{e}_u$, $u \in \{1, 2, 3\}$ (see Section 2.3)
$\mathbf{F}$	Filtering operator

F	Fully-discretized filtering operator
$\mathcal{F}$	Numerical flux
f	Filtering function
H	Scattering plus fission operator (in m^{-1})
h	Planck constant ($h = 6.626070040(81) \times 10^{-34}$ J–s)
h	Size of the largest disk that can be inscribed inside any cell of the mesh (in m)
$\mathbb{I}$	Identity matrix, with its size implied by the context
$\mathbb{I}_p$	Identity matrix of size p, $p \in \mathbb{N}$
k	Boltzmann constant ($k = 1.38064852(79) \times 10^{-23}$ J/K)
keV	A thousand electronvolts
L	Linear form
L	Streaming plus collision operator (in m^{-1})
$\mathbf{M}_u$	Dissipation matrix corresponding to $\vec{e}_u$
max	Maximum element of a set (in magnitude)
min	Minimum element of a set (in magnitude)
N	Spherical harmonics expansion truncation degree
N_0	Spherical harmonics expansion truncation degree for which an unfiltered calculation is *satisfying*
N_s	Spherical harmonics expansion truncation degree of the scattering source
$\mathcal{N}$	Set containing all the $(\ell, m) \in \mathbb{N}^2$ such that $0 \leq \|m\| \leq \ell \leq N$, see Eq. 2.14
$\vec{n}$	Unit vector normal to an interior facet (arbitrary orientation)
$\vec{n}_0$	Outward unit normal vector on a cell boundary

$\vec{n}_b$	Outward unit normal vector on the domain boundary
P	Number of non-redundant moments needed to describe the solution
q	Elementary charge ($q = 1.6021766208(98) \times 10^{-19}$ C)
R	Spherical harmonics
$\mathbf{R}$	Vector of the spherical harmonics up to degree N
$\vec{r}$	Spatial coordinate vector (in m)
r	Convergence rate, defined by Eq. 3.56
S	Energy-integrated volumetric source (in m^{-3}–s^{-1} or J–m^{-3}–s^{-1})
$\mathbf{S}$	Vector of the volumetric source moments (in m^{-3}–s^{-1} or J–m^{-3}–s^{-1})
$\mathbb{S}^2$	Unit sphere (in 1-D, 2-D or 3-D depending on the dimension of the problem)
T	Material temperature (in K)
t	Time (in s)
V	Finite-Element space
V_0	Finite-Element space restricted to a void or near-void region
V_1	Complement of V_0
v	Particle velocity (in m/s)
w	Solid angle of the unit sphere, equal to 2, 2π and 4π in 1-D, 2-D or 3-D, respectively (i.e. if d is equal to 1, 2 or 3, respectively)
x, y, z	Spatial coordinates (in m)
α	Order of the filter, defined by Eq. 3.15
Γ	Interface between the void/near-void regions and the non-void regions
Γ_{even}	Boundary terms in the variational form of even-parity based methods, see Eq. 4.67
Γ_{int}	Set of all the interior facets

$\delta_{i,j}$	Kronecker delta (equal to one if $i = j$, to zero otherwise)
$\boldsymbol{\eta}$	Scattering plus fission P_N operator (in m^{-1})
κ	Radius of the eigenspectrum relative to its distance to the origin, as defined by Eq. 3.48
λ	Eigenvalue
μ	Cosine of the polar angle
ν	Frequency (in s^{-1})
ν	Average number of neutrons emitted per fission
σ	Constant to scale the SAAF and LS terms (in m^{-1})
σ_t	Total macroscopic cross-section (in m^{-1})
$\sigma_\mathrm{t}^\star$	Quasi-steady total macroscopic cross-section (in m^{-1}), defined by Eq. 3.41
σ_a	Absorption macroscopic cross-section (in m^{-1})
σ_s	Scattering macroscopic cross-section (in m^{-1})
$\boldsymbol{\sigma_s}$	Scattering P_N operator (in m^{-1})
$\sigma_{s,\ell}$	ℓ-th moment of the scattering macroscopic cross-section (in m^{-1})
σ_f	Fission macroscopic cross-section (in m^{-1})
σ_f	Filtering strength (in m^{-1})
τ	Stabilization parameter (in m)
Φ	Energy-integrated scalar flux (in m^{-2}–s^{-1} or J–m^{-2}–s^{-1})
$\boldsymbol{\Phi}$	Solution vector of the energy-integrated angular flux moments (in m^{-2}–s^{-1} or J–m^{-2}–s^{-1})
φ	Azimuthal angle
Ψ	Energy-integrated angular flux (in m^{-2}–s^{-1} or J–m^{-2}–s^{-1})
$\vec{\Omega}$	Angular coordinate vector

$\vec{\nabla}$	Gradient operator
$\|\| \cdot \|\|_{L^1}$	L^1-norm
$\|\| \cdot \|\|_{L^2}$	L^2-norm
$\{\cdot\}$	Average operator, as defined by Eq. 3.21
$[\![\cdot]\!]$	Jump operator, as defined by Eq. 3.21
$(\cdot,\cdot)_{\mathcal{V}}$	Scalar product over a volume $\mathcal{V}$ and the unit sphere, defined by Eq. 2.18
$\langle\cdot,\cdot\rangle_{\partial\mathcal{V}}$	Scalar product over a surface $\partial\mathcal{V}$ and the unit sphere, defined by Eq. 2.19
$\otimes$	Tensor product
$\equiv$	Equality used as a definition

Subscripts

d	Relative to an incoming Dirichlet boundary condition
e	Even
g	Energy group
init	Initial (i.e. at $t = 0$)
ℓ	Spherical harmonics degree (same as Legendre polynomial degree)
l	Left value
max	Maximum
min	Minimum
o	Odd
r	Reflecting
r	Right value
u	Relative to the u-th spatial dimension ($u \in \{1, ..., d\}$)

Superscripts

$*$	Test function

$\dagger$	Adjoint operator
$+$	Interior side of a facet, defined by the orientation of the normal vector as shown in Fig. 3.1
$+$	Angular integration only defined for the positive half-range ($\vec{\Omega}\cdot\vec{n} > 0$), see Eq. 2.18
$-$	Exterior side of a facet, defined by the orientation of the normal vector as shown in Fig. 3.1
$-$	Angular integration only defined for the negative half-range ($\vec{\Omega}\cdot\vec{n} < 0$), see Eq. 2.19
0	Unit normal vector chosen locally pointing towards the outside of $\mathcal{D}_0$, see Section 5.2.2
1	Unit normal vector chosen locally pointing towards the outside of $\mathcal{D}_1$, see Section 5.2.2
$\oplus$	Relative to the positive half-range integral (for $\vec{\Omega} \cdot \vec{n}_b > 0$)
$\ominus$	Relative to the negative half-range integral (for $\vec{\Omega} \cdot \vec{n}_b < 0$)
inc	Incoming
m	Spherical harmonics order
n	Time step index (the absence thereof implying an evaluation at t^{n+1})
T	Transposed operator

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

The propagation of particles such as neutrons or photons in a spatial domain $\mathcal{D}$ and their interaction with the surrounding medium can be accurately described by the linear Boltzmann equation, or *transport equation*. It is however particularly challenging to solve because it involves seven independent variables (three for the position, two for the angle of propagation, one for time and one for energy) while allowing, in all generality, for discontinuities in the solution. As a matter of fact, exact solutions can rarely be derived analytically, except in simplistic – yet useful – configurations. With the advent of the Computer Age, numerical methods present themselves as an ever more attractive, complementary alternative to experiments. Since scientific computations only deal with finite quantities while trying to describe continuous[1] functions, some approximations need be somehow performed so as to reduce the dimension of that seven-dimensional phase-space.

It can become even more arduous when it is coupled to other potentially nonlinear physics. In the present work, we also consider the case of Thermal Radiation Transport (TRT) which describes the propagation of photons as well as their absorption by and re-emission from the material, thereby modifying the temperature T of the surrounding material. The transport equation can in particular be seen as the limiting case as the material heat capacity becomes infinite with a zero initial temperature.

1.1 Angular Discretization

A number of studies has been focusing on developing various numerical methods to solve this nonlinear multiphysics problem as efficiently and accurately as possible.

[1] In the sense that it can take on infinitely many different values.

Techniques include stochastic approaches, such as Implicit Monte Carlo [3, 4], as well as a variety of deterministic methods, among which are discrete ordinate (S_N) methods [5], spectral approximations [6], finite element discretizations [7], and nonlinear moments methods [8, 9, 10].

Although some of the work presented in this thesis also applies to S_N discretizations, we mostly focus on the spherical harmonics – or P_N – method which is a spectral Galerkin Finite Element method in angle. It essentially consists of expanding the angular-dependent variables using a truncated spherical harmonics expansion up to degree N. The P_N equations are then obtained by testing the equation against these same spherical harmonics.[2] This results in a method that preserves the rotational invariance of the transport equation while offering spectral convergence, i.e. exponential convergence with respect to N for infinitely smooth solutions [11]. Nevertheless, this method can give solutions that are oscillatory or even negative, especially in regions where the material properties are either small or rapidly varying. Negativity in the angular flux is not only non-physical (as it means a negative particle density) but can in turn lead to a negative material temperature, in which case the Planckian re-emission term is not well-defined.

1.2 Filtering

Several techniques have been studied to cope with this serious deficiency, one of which being the Filtered P_N (FP_N) method. Originally introduced by McClarren and Hauck [12], it was designed to reduce the anisotropy of the solution by smoothing out its angular dependency. Compared to other methods addressing the same issue, such as positive P_N (PP_N) closures [13, 14], entropy-based closures [9, 10] or moment closures based on residual minimization [15], the filtering approach is a lot

[2] The spherical harmonics thus serve as both the basis and test functions, hence the qualifying adjective "Galerkin".

less computationally expensive but is not robustly positive, meaning that it does not ensure strict positivity of the solution. Furthermore, FP$_N$ methods usually have a free parameter characterizing the filter strength that needs to be chosen by the user. Several works have already presented promising results [12, 16, 17, 18, 19]. In particular, the work in [19] shows how to incorporate the filtering directly into the nonlinear system, essentially recasting it as an additional collision term in the transport equation. Nevertheless, the FP$_N$ method has yet to be studied on implicit time-integration schemes, which are often preferred due to the extremely fast scales in the transport equation. Indeed, in a lot of applications, the stability condition imposed by explicit schemes on the time step Δt is too restrictive. It is typically bounded by $\Delta x/c$ where Δx characterizes the spatial mesh size and c is the speed of the particles (the speed of light in the TRT case).

1.3 Spatial Finite Element Discretization

Finite Element Methods have been widely used and studied, in particular because it was suggested that methods with high order of accuracy tend to perform better than low-order methods on comparatively finer meshes [20]. Nevertheless, due to the hyperbolic nature of the transport equation, most of these methods – if not all – require some sort of stabilization. In this work, we are considering two main approaches[3]: Discontinuous Galerkin Finite Element Methods (DGFEM) for TRT applications and Continuous Galerkin Finite Element Methods (CGFEM) for reactor physics applications.

DGFEM were introduced for transport problems in Ref. [26] where it was observed that the discontinuous basis, while more expensive than its standard continuous

[3] In no way are these the only existing methods. Alternative approaches include streamlined-upwind Petrov-Galerkin methods [21, 22], parity-based formulations [23, 24] and artificial viscosity based schemes – such as the entropy viscosity method [25].

counterpart, give better approximations to problems with non-smooth solutions. In addition to being robust in streaming regimes, where discontinuous solutions may occur, DGFEM (with a sufficiently rich basis set) also perform well in the diffusion limit [27, 28, 29].[4] A semi-implicit discretization of the P_N equations with DGFEM, which treats the flux terms explicitly, can be found in Ref. [32].

For reactor problems, particularly steady-state neutron transport calculations, CGFEM are typically preferred because of their relatively cheaper computational cost, while offering a good accuracy. Stabilization of the method can then be achieved by recasting the standard transport equation as one of the so-called *second-order* forms.

1.4 Second-Order Forms

The terminology *first-order* and *second-order* forms refer to the order of the spatial derivative. While the original – or first-order – transport equation only contains a first-order spatial derivative, there exists several ways to rewrite it with second-order ones. Such partial differential equations (PDE) are then referred to as *second-order forms* and have the advantage of being intrinsically similar to diffusion equations. As such, they can be well-suited to solution techniques initially designed for elliptic problems, such as CGFEM. Furthermore, they may result in a symmetric positive definite (SPD) matrix, thus ideal for the use of conjugate-gradient (CG) Krylov solvers [33, 34].

Nevertheless, these attractive features do not come without its share of challenges. Most of them, such as the Self-Adjoint Angular Flux (SAAF) [33, 35, 36, 37, 38] or

[4] Roughly speaking, this limit occurs when particle interactions with the surrounding medium isotropize the radiation field and the angular average of the photon distribution satisfies a much simpler diffusion equation [30, 31]. More specifically, whenever the absorption mean free path $\lambda = \sigma_{\mathrm{a}}^{-1}$ of the problem becomes very small compared to the mesh size, an important property of the scheme is to preserve the equilibrium diffusion limit. This ensures that the solution will be accurate in that limit despite being underresolved with respect to λ [27].

the even-parity [39] formulations, formally break down in void regions because it requires the evaluation of σ_{t}^{-1}, the inverse of the total macroscopic cross-section. Approximating a zero cross-section with an arbitrary low number is not a viable solution as it drastically degrades the solver performance [40]. An alternative is the SAAF formulation with a Void Treatment (SAAF–VT) [38], one of the downsides of this method being that it no longer results in an SPD system. In addition, it is not well suited to a P_N expansion because the bilinear form in the void regions goes – in steady state and as the mesh is refined – to the first-order form in void, which is ill-conditioned for P_N.

As the available computational resources keep growing, numerical methods allowing for more and more spatially and angularly refined calculations will become increasingly attractive. Yet, a natural way of improving the spatial resolution in the context of reactor physics is to give up on homogenization techniques. This often leads to strong discontinuities in the material properties, let alone the necessity of dealing with void or near-void regions. If there is hope in such second-order forms, this latter challenge need necessarily be overcome.

Another class of methods based on weighted residual minimizations, such as Least-Squares (LS) formulations [41, 42, 43], may not have this formal problem when σ_{t} goes to zero but are in general not globally conservative. This is in particular the case of the LS method compatible with voids [44]. Acceleration schemes such as Nonlinear Diffusion Acceleration (NDA) can help recover global conservation [38] but such schemes have yet to be developed for P_N methods. Thus, to the best our knowledge, there did not exist – prior to the present work – any second-order form for P_N that would result in a globally conservative and void compatible scheme.

1.5 Outline

The work in this thesis can be decomposed into two main ensembles: first, the study of the fully-implicit Filtered P_N method for the first-order form of the transport equation in the TRT case, with the hope to reduce and mitigate the negativity and unphysical oscillations inherent to the P_N approximation; second, the study of second-order forms for reactor physics applications to invent, as a result, a P_N method that would be both globally conservative and void compatible.

The remainder of this dissertation is organized as follows. In Chapter 2, we present the neutronics and TRT problems with the notation to be used throughout this work. We also briefly describe the spherical harmonics expansion – with more details available in Appendix A.

In Chapter 3, we focus on the FP_N method for TRT applications. We present the angular and spatial discretizations and discuss the choice of numerical flux and boundary conditions. We propose a general filtering strategy and study both the impact of the filter on the iterative solver convergence properties and on the error estimates derived in the linear setting [11]. Finally, we show results on a challenging benchmark problem, known as the Crooked Pipe [45] and compare the code with IMC calculations.

In Chapter 4, we study the relationship between the SAAF, SAAF–VT and LS second-order forms and show that equivalence or consistency can be achieved depending on the boundary conditions chosen and some assumptions pertaining to the problem. Further analysis is conducted to better understand the connection between several parity-based second-order forms and to show why a second-order filter fails to create an efficient, void compatible method. Numerical results on a heterogeneous multigroup k-eigenvalue problem are presented to illustrate how critical global conservation is, for numerical schemes to yield accurate answers.

Chapter 5 takes the considerations from Chapter 4 one step further: after having identified the reason why LS is not – in general – globally conservative, a new method with that very property in a void region, named Conservative LS (CLS), is proposed. It is consistent with the transport equation but sacrifices the symmetry of the bilinear form. From there, an hybrid method, coined SAAF–CLS, is derived by combining the CLS and SAAF formulations in the void and non-void regions, respectively. Generalization to near-void regions is demonstrated. Numerical results show that this method indeed constitutes the first void-compatible, globally conservative second-order form for P_N.

1.6 Novelty

The novelty of this research lies in the following aspects:

✓ Fully-implicit time-integration implementation and study of the FP_N method for TRT with a Lax-Friedrich numerical flux to reduce the computational cost.

✓ Investigation of local filtering strategies and extension to energy-dependent filters.

✓ Analysis of the effect of the filter on the conditioning of the system in the streaming limit, showing in particular that the filter improves the convergence properties of the iterative solver.

✓ Study of FP_N error estimates – previously derived for linear transport – in the TRT case with non-linear material properties.

✓ Formal derivation of the conditional equivalence between the SAAF and LS formulations, depending on the weakly-imposed LS boundary conditions, independently of the angular discretization.

✓ Formal derivation of SAAF–VT consistent boundary conditions for LS, independently of the angular discretization.

✓ Formal derivation of the equivalence between the SAAF–P$_N$ and SAAF–VT–P$_N$ methods while solving only for the even-parity moments.

✓ Formal derivation of the equivalence of the previous two methods and the Even-Parity P$_N$ formulation.

✓ Derivation of a fix consistent with the transport equation to make LS globally conservative in void regions.

✓ Invention of the SAAF–CLS second-order form both compatible with void and globally conservative. This method, unlike the SAAF–VT method, shows good results for both P$_N$ and S$_N$. Extension to near-void regions.

2. PROBLEM AND NOTATION

In this chapter, we first present the two applications that we are interested in: neutronics and Thermal Radiation Transport (TRT) which both give prominence to the transport equation. We briefly describe the P_N approximation (with more details available in Appendix A) and introduce notation and assumptions.

2.1 Physics

In this thesis, we focus our interest on the so-called *transport equation*, which here designates the linear Boltzmann equation. In Chapter 3, we more specifically consider Thermal Radiation Transport (TRT) which is the nonlinear coupling of the transport equation with the material temperature equation. As Chapters 4 and 5 are oriented towards neutronics applications, we do not account for that latter equation therein, or, equivalently, we set the material heat capacity to infinity and the initial temperature to zero.

In all generality, the transport equation depends on seven independent variables: three to describe the spatial position, noted $\vec{r} = (x, y, z)^T$; two to indicate the direction of propagation through the angular variable $\vec{\Omega} = \sqrt{1-\mu^2}\cos\varphi\,\vec{e}_x + \sqrt{1-\mu^2}\sin\varphi\,\vec{e}_y + \mu\,\vec{e}_z$ where μ is the cosine of the polar angle and φ is the azimuthal angle; one to specify the energy E of the particles and one for the time, noted t. The problem is defined for $\vec{r} \in \mathcal{D}$, $\vec{\Omega} \in \mathbb{S}^2$, $t > 0$ and $E > 0$, where $\mathcal{D}$ and $\mathbb{S}^2$ respectively designate the spatial domain and the unit sphere.

None of the following theory is restricted to energy-independent problems. As a matter of fact, some of results presented below correspond to multigroup calculations. Nevertheless, except in Section 3.4.4, the treatment of the energy variable is not the point of emphasis of this work. Therefore, for simplicity in the notation – and without loss of generality – we omit the dependency on E and study the energy-integrated transport equation, unless otherwise specified.

We describe in Section 2.1.1 the linear Boltzmann equation, along with its corresponding boundary and initial conditions. In Section 2.1.2, we detail the TRT system of partial differential equations (PDE) to be studied in Chapter 3 and point out the subtle difference in physical meaning the variable Ψ then has.

2.1.1 Neutronics

The energy-integrated linear Boltzmann equation reads:

$$\frac{1}{v}\frac{\partial \Psi}{\partial t} + \vec{\Omega}\cdot\vec{\nabla}\Psi + \sigma_{\mathrm{t}}(\vec{r})\Psi(\vec{r},\vec{\Omega}) = \int_{\mathbb{S}^2} \sigma_s(\vec{r},\vec{\Omega}'\cdot\vec{\Omega})\Psi(\vec{r},\vec{\Omega}')\,\mathrm{d}\Omega' + \nu\sigma_f(\vec{r})\Phi(\vec{r}) + S(\vec{r},\vec{\Omega}), \tag{2.1}$$

where Ψ and Φ represent respectively the angular and scalar fluxes, with the following relation:

$$\Phi(t,\vec{r}) \equiv \int_{\mathbb{S}^2} \Psi(t,\vec{r},\vec{\Omega})\,\mathrm{d}\Omega. \tag{2.2}$$

In addition, σ_{t}, σ_s and σ_f respectively denote the total, scattering and fission macroscopic cross-sections. Besides, v and ν are the neutron velocity and the average number of neutrons emitted per fission, respectively and S is the (known) volumetric source. The boundary conditions are applied at the boundary of the domain $\partial\mathcal{D}$. It is decomposed into two non-overlapping surfaces $\partial\mathcal{D} = \partial\mathcal{D}^{\mathrm{d}} \cup \partial\mathcal{D}^{\mathrm{r}}$, incoming Dirichlet and reflecting boundary conditions being respectively imposed on $\partial\mathcal{D}^{\mathrm{d}}$ and $\partial\mathcal{D}^{\mathrm{r}}$:

$$\Psi(\vec{r}_b,\vec{\Omega}) \equiv \Psi^{\mathrm{inc}}(\vec{r}_b,\vec{\Omega}) = \begin{cases} \Psi^{\mathrm{d}}(\vec{r}_b,\vec{\Omega}), & \vec{r}_b \in \partial\mathcal{D}^{\mathrm{d}}, \\ \Psi(\vec{r}_b,\vec{\Omega}_{\mathrm{r}}), & \vec{r}_b \in \partial\mathcal{D}^{\mathrm{r}}, \end{cases} \tag{2.3}$$

$$\text{for } \vec{\Omega}\cdot\vec{n}_b < 0,$$

where Ψ^{d} is assumed to be given and the reflecting angle is defined as:

$$\vec{\Omega}_{\mathrm{r}} = \vec{\Omega} - 2(\vec{\Omega} \cdot \vec{n}_b)\vec{n}_b, \tag{2.4}$$

$\vec{n}_b$ being the outward unit vector at a point $\vec{r}_b$ on the boundary. An initial condition must also be provided:

$$\Psi(t = 0, \vec{r}, \vec{\Omega}) = \Psi_{\mathrm{init}}(\vec{r}, \vec{\Omega}). \tag{2.5}$$

2.1.2 Thermal Radiation Tranport (TRT)

We consider the grey (frequency integrated) form of the thermal radiation transport equations, given by [46]:

$$\frac{1}{c}\frac{\partial \Psi}{\partial t} + \vec{\Omega} \cdot \vec{\nabla}\Psi + \sigma_{\mathrm{t}}(\vec{r}, T)\Psi = \sigma_a(\vec{r}, T)B(T) + \int_{\mathbb{S}^2} \sigma_s(\vec{r}, \vec{\Omega}' \cdot \vec{\Omega})\Psi(\vec{r}, \vec{\Omega}')\,\mathrm{d}\Omega' + S, \tag{2.6}$$

$$\frac{\partial e}{\partial t} = \sigma_a(\vec{r}, T)\big(\Phi - wB(T)\big). \tag{2.7}$$

These equations describe the energy balance in the radiation field and in the material. The angular flux of the photon radiation is governed by Eq. 2.6 while Eq. 2.7 governs the evolution of the material energy $e(T)$, where $T(t, \vec{r})$ is the material temperature. The constant c is the speed of light and $w = \int_{\mathbb{S}^2} \mathrm{d}\Omega$ is the total angular weight The energy-integrated Planckian blackbody source is defined as:

$$B(T) \equiv \frac{4\pi}{w}\int_0^\infty \frac{2h\nu^3}{c^2}\frac{1}{\exp(\frac{h\nu}{kT}) - 1}\mathrm{d}\nu = \frac{acT^4}{w}, \tag{2.8}$$

where h, k and $a = (8\pi^5 k^4)/(15h^3c^3)$ are the Planck, Boltzmann and radiation constants, respectively. This integral is not defined for $T \leq 0$, which is why maintaining a positive material temperature is crucial for numerical codes.

For simplicity, we assume that the derivative $C_v = e'(T)$, referred to as the material heat capacity, is independent of T, although it is not required. Eq. 2.7 can then be rewritten:

$$C_v \frac{\partial T}{\partial t} = \sigma_a(\vec{r}, T)\big(\Phi - wB(T)\big). \tag{2.9}$$

The boundary and initial conditions for Ψ are still given by Eqs. 2.3 and 2.5. An initial condition for T need also be provided:

$$T(t = 0, \vec{r}) = T_{\text{init}}(\vec{r}). \tag{2.10}$$

At this point, it seems appropriate to point out that – although we have named it identically – the angular flux Ψ does not have the same units in Eqs. 2.1 and 2.6. In the former and latter equations, it has the units of neutrons per area per time and of energy per area per time, respectively. This fundamentally reflects that, while Eq. 2.1 conserves the number of neutrons, Eq. 2.6 does not conserve the number of photons, but rather their energy. In the TRT literature, Ψ is commonly named angular intensity and referred to as I (e.g. in [47]). We have estimated however that the difference is too subtle to be worth maintaining throughout this thesis. The same goes for Φ and S.

Occasionally, we may consider Eqs. 2.6 and 2.7 in the context of *pure transport*, which is to say that we are in the limit $C_v \longrightarrow \infty$ and a zero initial temperature (or equivalently to replace Eq. 2.7 with $T = 0$).

2.2 Spherical Harmonics Expansion

The P_N equations are derived by applying a spectral Galerkin method to Eqs. 2.1 or 2.6 using the spherical harmonics of degree less than or equal to N as both the basis and test functions. Roughly speaking, Ψ is then represented as a truncated linear combination of the spherical harmonics and the residual of our system of equations is minimized in that same space. The expansion reads:

$$\Psi(\vec{r},\vec{\Omega}) \approx \mathbf{R}^T(\vec{\Omega})\,\mathbf{\Phi}(\vec{r}) = \sum_{\ell=0}^{N}\sum_{m=-\ell}^{\ell} \Phi_\ell^m(\vec{r})\,R_\ell^m(\vec{\Omega}), \tag{2.11}$$

where the real-form spherical harmonics are defined as:

$$R_\ell^m(\vec{\Omega}) = \begin{cases} \sqrt{2}\,C_\ell^m\,P_\ell^m(\mu)\cos(m\varphi), & 0 < m \leq \ell \leq N \\ C_\ell^0\,P_\ell^0(\mu), & 0 \leq \ell \leq N \\ \sqrt{2}\,C_\ell^{|m|}\,P_\ell^{|m|}(\mu)\sin(|m|\varphi), & 0 < -m \leq \ell \leq N \end{cases}, \tag{2.12}$$

P_ℓ^m designating the associated Legendre polynomial of degree ℓ and order m and $C_\ell^m = \sqrt{\frac{(2\ell+1)}{w}\frac{(\ell-m)!}{(\ell+m)!}}$ being a normalization constant chosen such that the spherical harmonics are orthonormal to each other, that is:

$$\int_{\mathbb{S}^2} R_\ell^m(\vec{\Omega})\,R_{\ell'}^{m'}(\vec{\Omega})\mathrm{d}\Omega = \delta_{\ell,\ell'}\,\delta_{m,m'} \quad , \quad \forall\ (\ell,m),(\ell',m') \in \mathcal{N}, \tag{2.13}$$

where δ is the Kronecker delta and

$$\mathcal{N} \equiv \{(\ell,m) \in \mathbb{N}^2 : 0 \leq |m| \leq \ell \leq N\}. \tag{2.14}$$

Besides, we have defined:

$$\mathbf{R} \equiv \{R_\ell^m, m=-\ell,\cdots,\ell; \ell=0,\cdots,N\}\ ,\ \mathbf{\Phi} \equiv \{\Phi_\ell^m, m=-\ell,\cdots,\ell; \ell=0,\cdots,N\}\,. \tag{2.15}$$

The Φ_ℓ^m coefficients are called *moments* of the angular flux Ψ and satisfy the following relationship:

$$\Phi_\ell^m(\vec{r}) \equiv \int_{\mathbb{S}^2} R_\ell^m(\vec{\Omega})\Psi(\vec{r},\vec{\Omega})\mathrm{d}\Omega, \quad \text{or, equivalently,} \quad \mathbf{\Phi} = \int_{\mathbb{S}^2} \Psi\,\mathbf{R}\,\mathrm{d}\Omega. \tag{2.16}$$

Once this expansion has been performed for all angular dependent quantities, the P_N equations are then obtained by testing Eqs. 2.1 or 2.6 against each individual R_ℓ^m, that is by multiplying it with $\mathbf{R}$ and integrating over $\mathbb{S}^2$.

2.3 More Notation and Conventions

Let V be the finite-dimensional trial space of functions in which we look for our solution. In this context, the superscript $*$ is used to designate a test function. We further define the vector of moments of Ψ^* and of the volumetric source S as:

$$\mathbf{\Phi}^* = \int_{\mathbb{S}^2} \Psi^*\,\mathbf{R}\,\mathrm{d}\Omega \quad , \quad \mathbf{S} = \int_{\mathbb{S}^2} S\,\mathbf{R}\,\mathrm{d}\Omega. \tag{2.17}$$

Considering a spatial domain $\mathcal{V}$ with boundary $\partial\mathcal{V}$, we also define the following operators, for any function f, g:

$$(f,g)_{\mathcal{V}} \equiv \int_{\mathcal{V}}\int_{\mathbb{S}^2} f\,g\,\mathrm{d}\Omega\mathrm{d}r\ ,\ \langle f,g\rangle_{\partial\mathcal{V}}^{+} \equiv \int_{\partial\mathcal{V}}\int_{\vec{\Omega}\cdot\vec{n}(\vec{r})>0} f\,g\,|\vec{\Omega}\cdot\vec{n}|\mathrm{d}\Omega\mathrm{d}r, \tag{2.18}$$

$$\langle f,g\rangle_{\partial\mathcal{V}} \equiv \int_{\partial\mathcal{V}}\int_{\mathbb{S}^2} f\,g\,\vec{\Omega}\cdot\vec{n}\,\mathrm{d}\Omega\mathrm{d}r\ ,\ \langle f,g\rangle_{\partial\mathcal{V}}^{-} \equiv \int_{\partial\mathcal{V}}\int_{\vec{\Omega}\cdot\vec{n}(\vec{r})<0} f\,g\,|\vec{\Omega}\cdot\vec{n}|\mathrm{d}\Omega\mathrm{d}r. \tag{2.19}$$

In particular, we have:

$$\langle f,g\rangle_{\partial\mathcal{V}} = \langle f,g\rangle_{\partial\mathcal{V}}^{+} - \langle f,g\rangle_{\partial\mathcal{V}}^{-}. \tag{2.20}$$

In addition, the superscript n is used to indicate the discrete approximation of a time-dependent quantity at time t^n. In the absence thereof, it is implied that such approximations are evaluated at t^{n+1}. Likewise, the subscript g refers to the energy group. As mentioned above, this superscript is omitted whenever the generalization to a multigroup theory does not present any difficulty.

We fundamentally consider three-dimensional problems. However, in a lot of cases, the solution may depend on only one or two spatial dimensions. Occasionally, we may refer to such problems respectively as 1-D or 2-D. This does not mean that the problem is restricted to a line or a surface but rather that the problem is infinite in the other directions. Let d be the number of spatial variables the solution depends on and P the number of non-redundant moments needed to describe the solution. It can be shown that the value of P is equal to $(N+1)$, $(N+1)(N+2)/2$ and $(N+1)^2$ if d is one, two and three, respectively [48] (see also Appendix A.2.2). For convenience, we also choose to rename (and reorder, if need be) the basis unit vectors as $(\vec{e}_1, ..., \vec{e}_d)$.

3. FULLY-IMPLICIT FILTERED P$_N$ METHOD FOR THERMAL RADIATION TRANSPORT USING DISCONTINUOUS GALERKIN FINITE ELEMENTS*

In this chapter, we discuss the Filtered P$_N$ (FP$_N$) method using a fully-implicit time discretization and Discontinuous Galerkin Finite Element Method (DGFEM) for the spatial discretization, in the context of Thermal Radiation Transport (TRT).*

The fully-implicit TRT FP$_N$ equations are presented in Section 3.1. The spatial discretization of the FP$_N$ equations is derived in Section 3.2. In Section 3.3, we investigate the filter in terms of convergence properties of the iterative solver and then consider the error estimates derived in Ref. [11] for the linear setting. Finally, in Section 3.4, we test the method with different filtering strategies on the challenging benchmark problem known as the Crooked Pipe [45]. Because this problem is particularly hard to converge, we first show good agreement between our code and implicit Monte-Carlo calculations on a simplified version. We then show for the harder problem that the filter mitigates deficiencies in the P$_N$ solutions, especially for smaller values of N.

3.1 Implicit Filtered P$_N$

We focus here on a Backward Euler time discretization without restricting the work to this particular implicit time-integration scheme. The TRT equations (2.6)

*Part of this chapter is reprinted from "Implicit Filtered P$_N$ for High-Energy Density Thermal Radiation Transport using Discontinuous Galerkin Finite Elements" by Vincent M. Laboure, Ryan G. McClarren and Cory D. Hauck, 2016. *Journal of Computational Physics* [49] Copyright 2016 Elsevier. The author exercises his right granted by the copyright agreement to use the published journal article for personal use (specifically the inclusion in a dissertation).

and (2.9) then read:

$$\frac{1}{c}\frac{\Psi-\Psi^n}{\Delta t}+\vec{\Omega}\cdot\vec{\nabla}\Psi+\sigma_{\mathrm{t}}\Psi=\sigma_a B+\int_{\mathbb{S}^2}\sigma_s(\vec{r},\vec{\Omega}'\cdot\vec{\Omega})\Psi(t,\vec{r},\vec{\Omega}')\,\mathrm{d}\Omega+S, \tag{3.1}$$

$$C_v(T)\frac{T-T^n}{\Delta t}=\sigma_a(\vec{r},T)\left(\Phi(t,\vec{r})-wB(T)\right), \tag{3.2}$$

with Δt being the time step size. Recall (now and throughout this thesis) that the rest of the notation was defined in Chapter 2.

3.1.1 $\mathbf{P}_N$ Expansion

As explained in Section 2.2, the P_N equations are obtained by approximating $\Psi \approx \mathbf{R}^T\,\mathbf{\Phi}$, $\Psi^n \approx \mathbf{R}^T\,\mathbf{\Phi}^n$ and $S \approx \mathbf{R}^T\,\mathbf{S}$, multiplying Eq. 3.1 with $\mathbf{R}$ and integrating over the unit sphere $\mathbb{S}^2$. The spherical harmonics having been normalized so as to be orthonormal, it reads:

$$\frac{1}{c}\frac{\mathbf{\Phi}-\mathbf{\Phi}^n}{\Delta t}+\vec{\mathbf{D}}\cdot\vec{\nabla}\mathbf{\Phi}+\sigma_{\mathrm{t}}\mathbf{\Phi}=\boldsymbol{\sigma}_a\mathbf{B}(T)+\boldsymbol{\sigma}_s\mathbf{\Phi}+\mathbf{S}, \tag{3.3}$$

where:

$$\vec{\mathbf{D}}=\sum_{u=1}^{d}\mathbf{D}_u\vec{e}_u \quad , \quad \mathbf{D}_u=\int_{\mathbb{S}^2}\Omega_u\,\mathbf{R}(\vec{\Omega})\mathbf{R}^T(\vec{\Omega})\,\mathrm{d}\Omega, \tag{3.4}$$

$$\mathbf{B}=\int_{\mathbb{S}^2}\mathbf{R}\,B(T)\mathrm{d}\Omega, \tag{3.5}$$

$$\boldsymbol{\sigma}_{\mathrm{s}}=\mathrm{diag}\Big\{\sigma_{\mathrm{s},\ell}\,,\,m=-\ell,...,\ell\,;\,\ell=0,...,N\Big\}, \tag{3.6}$$

$$\boldsymbol{\sigma}_{\mathrm{a}}=\mathrm{diag}\Big\{\sigma_a\,\delta_{\ell,0}\,,\,m=-\ell,...,\ell\,;\,\ell=0,...,N\Big\}, \tag{3.7}$$

$$\sigma_{\mathrm{s},\ell}=\int_{\mathbb{S}^2}\sigma_s\,P_\ell^0\,\mathrm{d}\Omega. \tag{3.8}$$

To avoid having zero-speed waves in the problem, it is common practice to impose N to be odd [6]. Therefore, in this entire chapter, we will limit ourselves to that particular situation.

3.1.2 Filtering

The filtering idea – first introduced by McClarren and Hauck [12] – originally consisted in modifying the moments after each time step as follows:

$$\Phi_\ell^m \longleftarrow \frac{\Phi_\ell^m}{1+\alpha_0 \ell^2(\ell+1)^2}, \tag{3.9}$$

where:

$$\alpha_0 = \frac{\omega}{N^2(\sigma_\mathrm{t} L + N)^2}, \tag{3.10}$$

and ω and L are free parameters representing respectively the filter strength and a characteristic length. The coefficient by which each moment is divided had several desirable features: the preservation of the energy (or particle) balance, the rotational invariance of the solution, the formal convergence as N goes to infinity and the equilibrium diffusion limit were all preserved. This is respectively because the filter coefficient is zero for $\ell = m = 0$, does not depend on m, and vanishes as N goes to infinity and if σ_t goes to infinity. One of its most concerning flaws was that the filter strength ω had to be adjusted if the simulation was refined spatially and temporally. Alternatively, another formulation of the filter is possible [19] and eliminates that very flaw. It is applied to Eq. 3.3 by adding an extra collision term:

$$\frac{1}{c}\frac{\mathbf{\Phi}-\mathbf{\Phi}^n}{\Delta t} + \vec{\mathbf{D}}\cdot\vec{\nabla}\mathbf{\Phi} + \sigma_\mathrm{t}(\vec{r})\mathbf{\Phi} + \sigma_\mathrm{f}(\vec{r})\mathbf{F}\mathbf{\Phi} = \boldsymbol{\sigma}_\mathrm{a}\mathbf{B}(T) + \boldsymbol{\sigma}_\mathrm{s}\mathbf{\Phi} + \mathbf{S}. \tag{3.11}$$

The expression of $\mathbf{F}$ is given by:

$$\mathbf{F} = \mathrm{diag}\Big\{f(\ell,N)\,,\, m=-\ell,...,\ell\,;\,\ell=0,...,N\Big\}, \tag{3.12}$$

with the *filter function* f being:

$$f(\ell, N) \equiv -\log \rho_{\text{filterType}} \left(\frac{\ell}{N+1} \right). \tag{3.13}$$

Several filter types are considered here:

$$\rho_{\text{Lanczos}}(\zeta) \equiv \frac{\sin\zeta}{\zeta}\,; \quad \rho_{\text{SSpline}}(\zeta) \equiv \frac{1}{1+\zeta^4}\,; \quad \rho_{\exp}(\zeta) \equiv \exp(c_0\,\zeta^{\alpha}). \tag{3.14}$$

In the exponential filter, $\alpha \in \mathbb{N}$ and $c_0 = \log(\epsilon_M)$ where ϵ_M is the machine accuracy. The main difference between the filters is their order which by definition [11] is equal to α if and only if α satisfies the three following conditions:

$$\rho_{\text{filterType}}(0) = 1 \quad , \quad \rho^{(a)}_{\text{filterType}}(1) = 0 \text{ for } a = 1, ..., \alpha - 1, \quad \text{and} \quad \rho^{(\alpha)}_{\text{filterType}}(1) \neq 0. \tag{3.15}$$

The Lanczos and spherical spline filters are respectively 2 and 4. The integer α, defined by Eq. 3.14 happens to also be the order of that filter[1].

The variable σ_{f} in Eq. 3.11 is a tuning parameter – henceforth called *filter strength* – that may be spatially (and energy) dependent. Strategies for determining a good local value of σ_{f} are discussed in Section 3.3.1. In this context, one of the strengths of the reformulation in [19] is that — unlike the original implementation in [12] — the filter strength is independent of the size of the time step and the spatial mesh [19]. Thus the value of σ_{f} needs to be tuned only once, and this can be done using relatively cheap simulations on coarse meshes.

[1] This is why we use the same notation for these two *a priori* different variables.

3.2 Spatial Discretization

This section is dedicated to the Discontinuous Galerkin Finite Element Method (DGFEM) discretization of Eq. 3.11 in space, along with Eq. 3.2 for the material temperature. Each component of $\mathbf{\Phi}$ as well as T are approximated as piecewise polynomial functions. Stability is achieved by specifying the numerical flux at cells interfaces. Although this method is fairly standard and more details can be found in [50, 51], we show how to derive the P$_N$ streaming term in Section 3.2.1. We then discuss how to choose a good numerical flux in Section 3.2.2, describe how to impose reflecting and incoming Dirichlet boundary conditions in Section 3.2.3 and summarize the variational formulation in Section 3.2.5.

Let $\mathcal{K}_h$ be a collection of open convex, polyhedral cells $K \subset \mathcal{D}$ such that $\cup\overline{K} = \mathcal{D}$ with $h > 0$ being the size of the largest disk that can be inscribed inside any cell K. Let Γ_{int} be the set of interior facets:

$$\Gamma_{\text{int}} = \{e : e = \overline{K}_1 \cap \overline{K}_2 \text{ for any } K_1, K_2 \in \mathcal{K}_h,\ K_1 \neq K_2\}\,. \tag{3.16}$$

3.2.1 Streaming Term

The weak formulation is obtained by multiplying Eqs. 3.11 and 3.2 respectively with $\mathbf{\Phi}^* \in V^P$ and $T^* \in V$ and integrating over $\mathcal{D}$. The difficulty of this method lies in the treatment of the streaming term. An integration by parts is performed thereon to make boundary terms appear:

$$\begin{aligned}\left(\mathbf{\Phi}^*, \vec{\mathbf{D}} \cdot \vec{\nabla}\mathbf{\Phi}\right)_{\mathcal{D}} &= \sum_{K\in\mathcal{K}_h} \left(\mathbf{\Phi}^*, \vec{\mathbf{D}} \cdot \vec{\nabla}\mathbf{\Phi}\right)_K \\ &= \sum_{K\in\mathcal{K}_h} \left(-\left(\vec{\nabla}\mathbf{\Phi}^*, \vec{\mathbf{D}}\mathbf{\Phi}\right)_K + \left\langle \mathbf{\Phi}^*, \vec{n}_0 \cdot \vec{\mathbf{D}}\,\mathbf{\Phi}\right\rangle_{\partial K}\right),\end{aligned} \tag{3.17}$$

where $\vec{n}_0$ is the unit normal vector directed towards the outside of an element $K \in \mathcal{K}_h$. Reordering the summation to sum over the interior and exterior facets[2] rather than over the elements, it yields:

$$\Big(\mathbf{\Phi}^*, \vec{\mathbf{D}} \cdot \vec{\nabla}\mathbf{\Phi}\Big)_{\mathcal{D}} = -\Big(\vec{\nabla}\mathbf{\Phi}^*, \vec{\mathbf{D}}\mathbf{\Phi}\Big)_{\mathcal{D}} + \sum_{e\in\Gamma_{\text{int}}} \Big(\big\langle \mathbf{\Phi}^{*,+}, \vec{n} \cdot \vec{\mathbf{D}}\,\mathbf{\Phi}^+\big\rangle_e - \big\langle \mathbf{\Phi}^{*,-}, \vec{n} \cdot \vec{\mathbf{D}}\,\mathbf{\Phi}^-\big\rangle_e\Big) + \big\langle \mathbf{\Phi}^*, \vec{n}_b \cdot \vec{\mathbf{D}}\,\mathbf{\Phi}\big\rangle_{\partial\mathcal{D}}, \tag{3.18}$$

where $\vec{n}$ is the unit normal vector[3] and $\vec{n}_b$ represents the outward normal unit vector[4] on the boundary $\partial\mathcal{D}$. For any discontinuous variable ψ, we define, on any facet, ψ^+ and ψ^- depending on the orientation of $\vec{n}$, as shown in Fig. 3.1. Using the following identity [51]:

$$\eta\xi - \zeta\sigma = \frac{1}{2}(\eta+\zeta)(\xi-\sigma) - \frac{1}{2}(\eta-\zeta)(\xi+\sigma), \tag{3.19}$$

we have:

$$\big\langle \mathbf{\Phi}^{*,+}, \vec{n} \cdot \vec{\mathbf{D}}\,\mathbf{\Phi}^+\big\rangle_\Gamma - \big\langle \mathbf{\Phi}^{*,-}, \vec{n} \cdot \vec{\mathbf{D}}\,\mathbf{\Phi}^-\big\rangle_\Gamma = \big\langle [\![\mathbf{\Phi}^*]\!], \{\vec{n} \cdot \vec{\mathbf{D}}\,\mathbf{\Phi}\}\big\rangle_\Gamma + \big\langle \{\mathbf{\Phi}^*\}, [\![\vec{n} \cdot \vec{\mathbf{D}}\,\mathbf{\Phi}]\!]\big\rangle_\Gamma, \tag{3.20}$$

where the following operators are given, for any variable ψ, by:

$$[\![\psi]\!] \equiv (\psi^+ - \psi^-) \quad , \quad \{\psi\} \equiv \frac{\psi^+ + \psi^-}{2} \quad , \tag{3.21}$$

[2] An exterior facet is defined to be a facet which belongs to the boundary $\partial\mathcal{D}$.

[3] Note that the orientation of $\vec{n}$ does not matter since the expression is still valid if we replace it with $-\vec{n}$.

[4] For $\vec{n}_b$, the orientation matters.

at any point on a facet. Imposing to have continuity of the numerical flux, i.e. $[\![\vec{n} \cdot \vec{\mathbf{D}}\,\mathbf{\Phi}]\!] = 0$, we end up with:

$$\left(\mathbf{\Phi}^*, \vec{\mathbf{D}} \cdot \vec{\nabla}\mathbf{\Phi}\right)_{\mathcal{D}} = -\left(\vec{\nabla}\mathbf{\Phi}^*, \vec{\mathbf{D}}\mathbf{\Phi}\right)_{\mathcal{D}} + \left\langle [\![\mathbf{\Phi}^*]\!], \{\vec{n} \cdot \vec{\mathbf{D}}\,\mathbf{\Phi}\}\right\rangle_{\Gamma} + \left\langle \mathbf{\Phi}^*, \vec{n}_b \cdot \vec{\mathbf{D}}\,\mathbf{\Phi}\right\rangle_{\partial\mathcal{D}}. \quad (3.22)$$

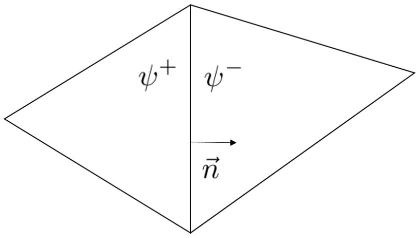


Figure 3.1: Notation for discontinuous variables, given a unit normal vector $\vec{n}$.

3.2.2 Numerical Flux

In Eq. 3.22, we need to specify what the numerical flux, defined as $\vec{n} \cdot \vec{\mathbf{D}}\,\mathbf{\Phi}$ and $\vec{n}_b \cdot \vec{\mathbf{D}}\,\mathbf{\Phi}$ respectively are on Γ and $\partial\mathcal{D}$, given that $\mathbf{\Phi}$ is potentially discontinuous thereon. It is typically done through a penalty term whose effect is to stabilize the scheme [51]:

$$\{\vec{n} \cdot \vec{\mathbf{D}}\,\mathbf{\Phi}\} = \sum_{u=1}^{d} \left(\vec{e}_u \cdot \vec{n}\,\mathbf{D}_u\{\mathbf{\Phi}\} + \frac{|\vec{e}_u \cdot \vec{n}\,|}{2}\mathbf{M}_u[\![\mathbf{\Phi}]\!] \right), \quad (3.23)$$

for all $\vec{r} \in \Gamma$. Let us now step back a little to justify the choice of the dissipation matrices $\mathbf{M}_u$.

Defining the numerical flux in the case of the P_N equations presents a challenge because each moment does not have a given direction of flow. If it were the case, an upwinded numerical flux would be easy to compute. Previously, it has been common practice to rely on Riemann solvers [48, 52]. It uses the fact that the $\mathbf{D}_u$ matrices are diagonalizable and therefore can be expressed as $\mathbf{D}_u \equiv \mathcal{R}_u \mathbf{\Lambda}_u \mathcal{L}_u$ where $\mathbf{\Lambda}_u$ is a diagonal matrix. The dissipation matrices are then defined as:

$$\mathbf{M}_u \equiv \mathcal{R}_u |\mathbf{\Lambda}_u| \mathcal{L}_u, \tag{3.24}$$

where $|\mathbf{\Lambda}_u|$ is obtained by taking the absolute value of each component of $\mathbf{\Lambda}_u$. There are two main issues with doing so. First, while $\mathbf{D}_u$ is very sparse, $\mathbf{M}_u$ – with that definition – is not, which can induce a large computational cost, especially for large values of N. Second, unless N is odd and the problem only depends on one spatial dimension, $\mathbf{\Lambda}_u$ has at least one zero eigenvalue, the effect of which is to have zero-speed waves in the problem [6]. A way around that is to replace all zero eigenvalues by some small value (e.g. the smallest nonzero eigenvalue). Although this fix does not change the order of convergence of the method [6], it is no longer a strict upwinding. It is therefore not perfectly rigorous to use this numerical flux on the boundary domain as it would couple the incoming data (imposed by the boundary conditions) and outgoing data.

It is then proposed to use a different type of numerical flux, namely a Lax-Friedrich flux which consists of choosing:

$$\mathbf{M}_u = \lambda \mathbb{I}, \tag{3.25}$$

where λ is the maximum eigenvalue of the $\mathbf{D}_u$ matrices and $\mathbb{I}$ denotes the identity matrix. As this value approaches one from below as $N \to \infty$; we set $\lambda = 1$. The advantage is to drastically increase the sparsity of $\mathbf{M}_u$. A special treatment of the numerical flux at the boundary has to be applied to make sure the boundary conditions are only used to set incoming data. As mentioned above, this treatment would – rigorously speaking – be required also in the case of an upwinded numerical flux if the zero eigenvalues are modified. Thus, this does not constitute an additional inconvenience *per se* and can be done using matrices whose elements consists of integrals over half of the angular domain. We are going to expand on this in Section 3.2.3.

In the meantime, we have:

$$\{\vec{n} \cdot \vec{\mathbf{D}}\, \mathbf{\Phi}\} = \vec{n} \cdot \vec{\mathbf{D}}\{\mathbf{\Phi}\} + \frac{||\vec{n}||_{L^1}}{2} [\![\mathbf{\Phi}]\!], \tag{3.26}$$

for all $\vec{r} \in \Gamma$. The streaming term can then be expressed as:

$$\begin{aligned} \left(\mathbf{\Phi}^*, \vec{\mathbf{D}} \cdot \vec{\nabla} \mathbf{\Phi}\right)_{\mathcal{D}} = &- \left(\vec{\nabla} \mathbf{\Phi}^*, \vec{\mathbf{D}} \mathbf{\Phi}\right)_{\mathcal{D}} \\ &+ \left\langle [\![\mathbf{\Phi}^*]\!], \vec{n} \cdot \vec{\mathbf{D}}\, \{\mathbf{\Phi}\} \right\rangle_{\Gamma} + \left\langle [\![\mathbf{\Phi}^*]\!], \frac{||\vec{n}||_{L^1}}{2} [\![\mathbf{\Phi}]\!] \right\rangle_{\Gamma} \\ &+ \left\langle \mathbf{\Phi}^*, \mathcal{F}(\mathbf{\Phi}, \Psi^{\text{inc}}) \right\rangle_{\partial \mathcal{D}}, \end{aligned} \tag{3.27}$$

where $\mathcal{F}(\mathbf{\Phi}, \Psi^{\text{inc}})$ is the numerical flux at the boundary. Besides the L^1 norm is defined for any vector $\vec{v}$ as:

$$||\vec{v}||_{L^1} \equiv \sum_{u=1}^{d} |\vec{v} \cdot \vec{e}_u|. \tag{3.28}$$

The treatment for the numerical flux on the boundaries depends on the type of boundaries.

3.2.3 Boundary Conditions

We consider two types of boundary conditions: the reflecting and the incoming Dirichlet boundary conditions, as described by Eq. 2.3.

3.2.3.1 Reflecting Boundary

In the case of a reflecting boundary condition, the numerical flux is very similar to the one described by Eq. 3.26 except that the value of Ψ^+ designates the value inside the exterior facet (i.e. Ψ) and the value of Ψ^- corresponds to the value outside the exterior facet. The latter is determined by the boundary condition and is nothing but $\Psi(\vec{\Omega}_{\rm r})$. Therefore, the 'reflecting' numerical flux is given by:

$$\mathcal{F}(\mathbf{\Phi}, \Psi^{\rm inc}) = \mathcal{F}_{\rm r}(\mathbf{\Phi}) \equiv \vec{n}_b \cdot \vec{\mathbf{D}}\,\mathbf{\Phi} = \frac{1}{2}\vec{n}_b \cdot \vec{\mathbf{D}}\mathbf{\Phi} + \frac{||\vec{n}_b||_{L^1}}{2}\mathbf{\Phi} + \frac{1}{2}\mathbf{L}_{\rm r}\mathbf{\Phi} - \frac{||\vec{n}_b||_{L^1}}{2}\mathbf{Q}_{\rm r}\mathbf{\Phi}, \quad (3.29)$$

for all $\vec{r} \in \partial\mathcal{D}_{\rm r}$, where:

$$\mathbf{L}_{\rm r} \equiv \int_{\mathbb{S}^2} \vec{\Omega} \cdot \vec{n}_b \, \mathbf{R}(\vec{\Omega})\, \mathbf{R}^T(\vec{\Omega}_{\rm r})\, {\rm d}\Omega \quad , \quad \mathbf{Q}_{\rm r} \equiv \int_{\mathbb{S}^2} \mathbf{R}(\vec{\Omega})\, \mathbf{R}^T(\vec{\Omega}_{\rm r})\, {\rm d}\Omega. \quad (3.30)$$

As a reminder, $\vec{\Omega}_{\rm r}$ designates the reflected direction corresponding to $\vec{\Omega}$ and depends on $\vec{n}_b$ (see Eq. 2.4).

The first two terms correspond to the inside of the exterior facet while the last two correspond to the outside of the exterior facet. We then further define:

$$\mathcal{F}_{\rm r}(\mathbf{\Phi}) = \mathcal{F}_{\rm ext}(\mathbf{\Phi}) + \mathcal{F}_{\rm r}^{\rm BC}(\mathbf{\Phi}), \quad (3.31)$$

where:

$$\mathcal{F}_{\rm ext}(\mathbf{\Phi}) \equiv \frac{1}{2}\, \vec{n}_b \cdot \vec{\mathbf{D}}\, \mathbf{\Phi} + \frac{||\vec{n}_b||_{L^1}}{2}\, \mathbf{\Phi}, \quad (3.32)$$

and

$$\mathcal{F}_{\mathrm{r}}^{\mathrm{BC}}(\boldsymbol{\Phi}) \equiv \frac{1}{2}\,\mathbf{L}_{\mathrm{r}}\boldsymbol{\Phi} - \frac{||\vec{n}_b||_{L^1}}{2}\mathbf{Q}_{\mathrm{r}}\boldsymbol{\Phi}. \tag{3.33}$$

We dropped the subscript r for $\mathcal{F}_{\mathrm{ext}}$ because we will see that it is identical for an incoming Dirichlet boundary.

3.2.3.2 Incoming Dirichlet Boundary

In the case of an incoming Dirichlet boundary, the numerical flux will again be determined again by Eq. 3.26 with the Ψ^+ being the value inside the exterior facet and Ψ^- being the value outside the exterior facet, i.e. imposed by the boundary condition. The problem is that we can only impose the incoming data of Ψ because that is all the boundary conditions give us. For the outgoing data, we reconstruct them using the outgoing data from the inside of the exterior facet. In other words, for the outside of the exterior facet, we impose the incoming data through the boundary conditions and the outgoing data imposing continuity with the inside of the exterior facet.

We start by taking what could be seen as the *half-range* moments of Ψ^{inc}:

$$\boldsymbol{\Phi}^{\mathrm{inc}}(\vec{r}) \equiv \int_{\vec{\Omega}\cdot\vec{n}_b<0} \mathbf{R}\,\Psi^{\mathrm{inc}}\,\mathrm{d}\Omega \quad , \quad \mathbf{J}^{\mathrm{inc}}(\vec{r}) \equiv \int_{\vec{\Omega}\cdot\vec{n}_b<0} |\vec{\Omega}\cdot\vec{n}_b|\,\mathbf{R}\,\Psi^{\mathrm{inc}}\,\mathrm{d}\Omega. \tag{3.34}$$

Then, we construct the 'Dirichlet' numerical flux as follows:

$$\mathcal{F}_{\mathrm{d}}(\boldsymbol{\Phi},\Psi^{\mathrm{inc}}) = \frac{1}{2}\left(\vec{n}_b\cdot\vec{\mathbf{D}}\,\boldsymbol{\Phi} + \mathbf{L}^{\oplus}\,\boldsymbol{\Phi} - \mathbf{J}^{\mathrm{inc}}\right) + \frac{||\vec{n}_b||_{L^1}}{2}\left(\boldsymbol{\Phi} - \mathbf{Q}^{\oplus}\,\boldsymbol{\Phi} - \boldsymbol{\Phi}^{\mathrm{inc}}\right), \tag{3.35}$$

for all $\vec{r}\in\partial\mathcal{D}_{\mathrm{d}}$, where we have defined:

$$\mathbf{L}^{\oplus} = \int_{\vec{\Omega}\cdot\vec{n}_b>0} |\vec{\Omega}\cdot\vec{n}_b|\,\mathbf{R}\mathbf{R}^T\,\mathrm{d}\Omega \quad , \quad \mathbf{Q}^{\oplus} \equiv \int_{\vec{\Omega}\cdot\vec{n}_b>0} \mathbf{R}\mathbf{R}^T\,\mathrm{d}\Omega. \tag{3.36}$$

Again, the terms $\left(\vec{n}_b \cdot \vec{\mathbf{D}}\,\mathbf{\Phi} + \mathbf{L}^{\oplus}\,\mathbf{\Phi} - \mathbf{J}^{\text{inc}}\right)$ are nothing but $\vec{n} \cdot \vec{\mathbf{D}}\{\mathbf{\Phi}\}$ where $\mathbf{\Phi}^{+} = \mathbf{\Phi}$ and where $\mathbf{\Phi}^{-}$ is chosen to be $\mathbf{\Phi}$ if it corresponds to the outgoing data and to be $\mathbf{\Phi}^{\text{inc}}$ if it corresponds to the incoming data. Similarly, the terms $\left(\mathbf{\Phi} - \mathbf{Q}^{\oplus}\,\mathbf{\Phi} - \mathbf{\Phi}^{\text{inc}}\right)$ correspond to $[\![\mathbf{\Phi}]\!]$.

We can rearrange to get an expression similar to the numerical flux on reflecting boundaries (see Eq. 3.31):

$$\mathcal{F}_{\text{d}}(\mathbf{\Phi}, \Psi^{\text{inc}}) = \mathcal{F}_{\text{ext}}(\mathbf{\Phi}) + \mathcal{F}_{\text{d}}^{\text{BC}}(\mathbf{\Phi}, \Psi^{\text{inc}}), \tag{3.37}$$

where $\mathcal{F}_{\text{ext}}$ was defined in Eq. 3.32 and:

$$\mathcal{F}_{\text{d}}^{\text{BC}}(\mathbf{\Phi}, \Psi^{\text{inc}}) \equiv \frac{1}{2}\,\left(\mathbf{L}^{\oplus}\,\mathbf{\Phi} - \mathbf{J}^{\text{inc}}\right) - \frac{||\vec{n}_b||_{L^1}}{2}\left(\mathbf{Q}^{\oplus}\,\mathbf{\Phi} + \mathbf{\Phi}^{\text{inc}}\right). \tag{3.38}$$

3.2.4 Mass Matrix Lumping

For robustness in optically thick regions, it may be necessary to lump the matrices corresponding to the collision terms. This was demonstrated in [29] in the context of discontinuous Galerkin discretizations of discrete ordinate equations. In practice, lumping a matrix is done by replacing it by a diagonal matrix whose i-th term is the sum of the elements on the i-th row of the original matrix. For the Crooked Pipe test problem (see Section 3.4) this lumping proved to be necessary to avoid non-physical instabilities in the solution.

3.2.5 Variational Formulation

Our weak formulation[5] is given by: Find $(\mathbf{\Phi}, T) \in V^P \times V$ such that:

$$\forall\, (\mathbf{\Phi}^*, T^*) \in V^P \times V,\ a((\mathbf{\Phi}^*, T^*), (\mathbf{\Phi}, T)) = \mathrm{L}((\mathbf{\Phi}^*, T^*)), \tag{3.39}$$

where the bilinear form a is defined for all $(\mathbf{\Phi}, T), (\mathbf{\Phi}^*, T^*) \in V^P \times V$ by[6]:

$$\begin{aligned} a((\mathbf{\Phi}^*, T^*), (\mathbf{\Phi}, T)) = \\ &\big(\mathbf{\Phi}^*, \sigma_{\mathrm{t}}^{\star}\, \mathbf{\Phi}\big)_{\mathcal{D}} + \big(\mathbf{\Phi}^*, \sigma_{\mathrm{f}}\, \mathbf{F}\, \mathbf{\Phi}\big)_{\mathcal{D}} - \big(\mathbf{\Phi}^*, \boldsymbol{\sigma}_s\, \mathbf{\Phi}\big)_{\mathcal{D}} - \big(\mathbf{\Phi}^*, \boldsymbol{\sigma}_a\, \mathbf{B}(T)\big)_{\mathcal{D}} \\ &- \Big(\vec{\nabla}\mathbf{\Phi}^*, \vec{\mathbf{D}}\mathbf{\Phi}\Big)_{\mathcal{D}} + \Big\langle \llbracket \mathbf{\Phi}^* \rrbracket, \vec{n} \cdot \vec{\mathbf{D}} \left\{\mathbf{\Phi}\right\} \Big\rangle_{\Gamma_{\mathrm{int}}} + \Big\langle \llbracket \mathbf{\Phi}^* \rrbracket, \frac{||\vec{n}||_{L^1}}{2} \llbracket \mathbf{\Phi} \rrbracket \Big\rangle_{\Gamma_{\mathrm{int}}} \\ &+ \big\langle \mathbf{\Phi}^*, \mathcal{F}_{\mathrm{ext}}(\mathbf{\Phi}) \big\rangle_{\partial\mathcal{D}} + \big\langle \mathbf{\Phi}^*, \mathcal{F}_{\mathrm{r}}^{\mathrm{BC}}(\mathbf{\Phi}) \big\rangle_{\partial\mathcal{D}^{\mathrm{r}}} + \big\langle \frac{1}{2}\mathbf{\Phi}^*, \mathbf{L}^{\oplus}\, \mathbf{\Phi} - ||\vec{n}_b||_{L^1} \mathbf{Q}^{\oplus}\, \mathbf{\Phi} \big\rangle_{\partial\mathcal{D}^{\mathrm{d}}} \\ &+ \Bigg(T^*, \frac{1}{\Delta t} T - \sigma_a \left(\sqrt{w}\, \Phi_0^0 - w\, B(T)\right)\Bigg)_{\mathcal{D}}, \end{aligned} \tag{3.40}$$

where:

$$\sigma_{\mathrm{t}}^{\star} \equiv \sigma_{\mathrm{t}} + \frac{1}{c\Delta t}, \tag{3.41}$$

and the linear form L is defined for all $(\mathbf{\Phi}^*, T^*) \in V^P \times V$ by:

$$\begin{aligned} \mathrm{L}((\mathbf{\Phi}^*, T^*)) = &\big(\mathbf{\Phi}^*, \frac{1}{c\Delta t}\mathbf{\Phi}^n\big)_{\mathcal{D}} + \big(\mathbf{\Phi}^*, \mathbf{S}\big)_{\mathcal{D}} + \big\langle \frac{1}{2}\mathbf{\Phi}^*, \mathbf{J}^{\mathrm{inc}} + ||\vec{n}_b||_{L^1} \mathbf{\Phi}^{\mathrm{inc}} \big\rangle_{\partial\mathcal{D}^{\mathrm{d}}} \\ &+ \Big(T^*, \frac{1}{\Delta t} T^n\Big)_{\mathcal{D}}. \end{aligned} \tag{3.42}$$

[5] We consider here the case of Backward Euler for the time discretization but our implementation allows for other time discretizations.

[6] The term $\sqrt{w}\Phi_0^0$ comes from the fact that $\Phi = \int_{\mathbb{S}^2} \Psi \, \mathrm{d}\Omega = w^{1/2} \int_{\mathbb{S}^2} R_0^0\, \Psi \, \mathrm{d}\Omega = \sqrt{w}\, \Phi_0^0$. As a reminder $w = 2, 2\pi, 4\pi$ for $d = 1, 2, 3$, respectively.

3.2.6 Implementation

The algorithm is implemented in two codes: a simple code that is used to explore the eigenstructure of the linearized system and a production code to generate numerical solutions.

3.2.6.1 Simple Code

A pure transport code[7] has been written in C++ using cubic elements and piecewise constant material properties. The spatial discretization was performed using a DGFEM with the space V built from the tensor product of linear polynomials on each cell.[8]

From the variational formulation 3.39, one can generate a linear system of the form $AU = b$ where U is the solution vector. If $v^{(k)}$ and $v^{(l)}$ are, respectively, the k-th and l-th basis functions ($v^{(k)}, v^{(l)} \in V^{P+1}$), then the components of the global matrix A are

$$A_{kl} = a(v^{(l)}, v^{(k)}) \quad \text{and} \quad b_k = \mathrm{L}(v^{(k)}). \tag{3.43}$$

This matrix was assembled in order to study its eigenspectrum as a function of the filter strength σ_{f}. This was done using MATLAB [53].

3.2.6.2 Production Code

To generate numerical solutions for Eqs. 3.1 and 3.2, a code has been implemented in Rattlesnake, the transport solver of the Idaho National Laboratory (INL), based on the Multiphysics Object Oriented Simulation Environment (MOOSE) framework [54]. Nonlinear solves are performed using the Jacobian Free Newton Krylov (JFNK) method, and the PETSc [55] restarted generalized minimal residual (GMRES) solver

[7] As a reminder, by pure transport, we mean that the re-emission term in Eq. 3.1 is neglected. This can be done, for instance, by setting the temperature T uniformly to zero.

[8] The number of basis functions is thus equal to eight times the number of cells.

for the linear solves. In this method, the Jacobian is never explicitly formed but its action is computed with two nonlinear residual evaluations. All the results from this code are obtained using the first order LAGRANGE elements from libMesh [56]. The meshes are generated using `gmsh` [57] and the results are visualized with VisIt [58]. Several convergence tests were performed to verify the spatial and temporal accuracy of the code.

The linear system for $\mathbf{\Phi}$ in Eq. 3.11 can be ill-conditioned in streaming regimes. Specifically, $\sigma_{\mathrm{t}}^{\star} \to \frac{1}{c\Delta t}$ when $\sigma_{\mathrm{t}} \to 0$. Hence when σ_{t} is small and Δt is large, the system is dominated by the streaming operator $\vec{\mathbf{D}} \cdot \vec{\nabla}$, which is singular and not diagonally dominant. The loss of diagonal dominance makes most iterative schemes (Jacobi, Gauss-Seidel, SOR, etc.) unstable. To our knowledge, there does not exist a universally effective preconditioner for the P_N equations in the streaming limit, though some multigrid in angle preconditioners have been studied in the past for the even-parity form of the P_N equations [59]. For the results in this chapter, we have used the built-in algebraic multigrid (AMG) preconditioners in PETSc.

3.3 Study of the Filter

In this section, we discuss the selection of filter parameters. We then investigate how the filter affects (i) the convergence of the iterative solver for the fully discretized system and (ii) the convergence of the angular discretization as $N \to \infty$.

3.3.1 Filtering Strategy

In this subsection, we discuss the strategy for selecting the location, type, and strength of the filter. Fig. 3.2 shows the dependence of different filter functions f (cf. Eq. 3.13) on ℓ when $N = 7$. These results illustrate several general trends. First, for fixed N, f is a monotonically increasing function of ℓ. Second, the value of σ_{f} must be adjusted according to the type of filter used. Third, as the order of the

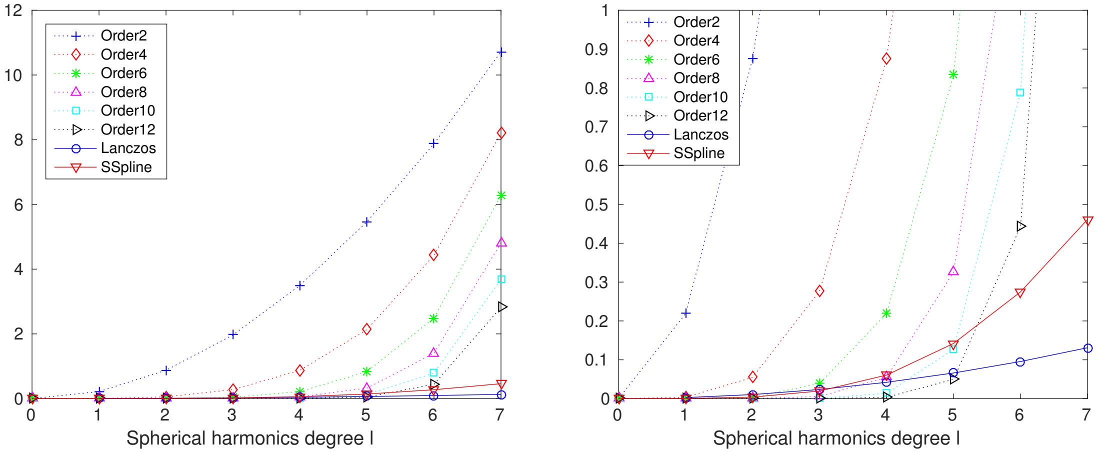


Figure 3.2: Filter function f vs. degree ℓ for exponential filters of various orders, the Lanczos filter, and the spherical spline filter. The results are shown for $N = 7$. The right plot is the same as the left one with a different scale to see the latter two curves more clearly.

filter increases, the higher-order moments are filtered more, comparatively to the low-order ones. In particular, an unfiltered calculation can be seen as a filtered one with an infinite order. It is therefore intuitive – and empirically found – that, for difficult problems, the higher-order filters do not perform as well as low-order ones, even if the value of σ_f is adjusted. For this reason, we use the Lanczos filter for the rest of the paper, unless specified otherwise.

The major drawback of the filter is that σ_f must be tuned by the user for each individual problem. Unfortunately, the numerical solution can be very sensitive to the value of σ_f, especially for small values of N. The choice of filter strength is a trade-off between removing unphysical oscillations and excessive damping of the solution. Since the appropriate balance may be different in different parts of the spatial domain, it is usually advantageous to allow σ_f to vary in space. Often a basic understanding of radiation transport can help guide the strategy for setting σ_f without the need for extensive knowledge of the solution beforehand. When more information is needed, a relatively coarse simulation (in space and time) may be used as a proxy. This is one of the main benefits of using the consistent formulation in Eq. 3.11: the value σ_{f} does not need to be recomputed when the space-time mesh is refined.

In our experience, we have found the following to be good practices for setting the filter strength.

- **Location.** Run a calculation with no filter and find local regions where Φ_0^0 becomes negative. Activate the filter in these 'negative' regions as well as in upstream regions of comparable sizes. For the other parts of the problem, the filter can typically be set to zero or to a much smaller value. If the problem is uniform, then activate the filter everywhere.

- **Filter type.** Set the order of the filter to match the expected regularity (with respect to angle) of the transport solution.[9] If unsure, it is better to underestimate the regularity. Lower order filters are typically more robust; for the most difficult problems, we have found that the second-order Lanczos filter works well.

- **Filter strength.** Using a coarse mesh, determine N_0 which yields an acceptable[10] unfiltered calculation. A good scaling is usually obtained by setting $\sigma_{\mathrm{f}}(\vec{r}) \approx \sigma_{\mathrm{t}}(\vec{r})/f(1, N_0)$ in the previously determined regions. Another option is to tune the filter strength empirically (still on a coarse mesh).

These guidelines are quite broad but they usually are precise enough to determine a suitable σ_{f}. The relative freedom that is left to the user is also an advantage since

[9] See Section 3.3.3 for a more precise statement of the regularity.

[10] As $N \to \infty$, the numerical solution converges to the analytical solution so there exists an integer N_0 such that the numerical solution is good enough. This notion is of course subjective and is up to the user. In practice, N_0 can for instance be chosen such that the unfiltered Φ_0^0 is non-negative or that the unfiltered T does not reach unphysical values.

the extent to which the negativity and oscillations should be reduced can vary from one application to another.

3.3.2 Effects of the Filter on the Iterative Solver

In this section, we study how the filter affects the convergence properties of the GMRES solver for the fully discretized system. We first study the pure transport case and propose empirical and theoretical predictions regarding the behavior of the spectrum of the global matrix as a function of σ_{f} and N. After predicting how the solver should be affected by these parameters, we present the practical number of linear iterations for a complicated TRT benchmark problem.

3.3.2.1 Linear Setting: Pure Transport

We investigate first the spectrum of the linear system corresponding to the pure transport problem. The matrix elements for this system are computed using the simple code described in Section 3.2.6.1. We consider a cubic domain that is 1 m on each side, with periodic boundary conditions in z and open boundaries elsewhere. We set $\sigma_{\mathrm{t}} = 0$ because we are mostly interested in the effect of the filter when the unfiltered P_N equations are ill-conditioned. We set $N = 3$, use a $10 \times 10 \times 1$ mesh ($\Delta x = \Delta y = 0.1$ m, $\Delta z = 1$ m), and set $c\Delta t = 0.1$ m (therefore, $\sigma_{\mathrm{t}}^{\star} = 10\ \mathrm{m}^{-1}$). Though relatively small, this problem already has $12,800$ unknowns.[11]

We write the global matrix A from Eq. 3.43 as $A = R + C + F$, where C and F are the matrices corresponding to the collision and filtering operators, respectively. More specifically, $C = \sigma_{\mathrm{t}}^{\star} M_0$ and $F = \sigma_{\mathrm{f}}\, \mathbf{F} \otimes M$ where[12]

$$M_0 = \mathbb{I}_P \otimes \int_{\mathcal{D}} \mathcal{V}\,\mathcal{V}^T\,\mathrm{dx} \quad \text{and} \quad M = \int_{\mathcal{D}} \mathcal{V}\,\mathcal{V}^T\,\mathrm{dx}. \tag{3.44}$$

[11] There are sixteen moments, 100 cells, and eight basis functions per cell.

[12] Note that this block-diagonal form of the mass matrix M_0 assumes that the basis functions have been grouped according to their corresponding moments.

Here $\mathbb{I}_k$ denotes the identity matrix of size k for any $k \in \mathbb{N}$; $\otimes$ designates the tensor product; and $\mathcal{V}$ is a vector whose components form an orthogonal basis of V. We then compute the spectrum of the operator[13]

$$\mathcal{A} \equiv C^{-1}A = C^{-1}R + \mathbb{I}_{SP} + \frac{\sigma_{\mathrm{f}}}{\sigma_{\mathrm{t}}^{\star}}\mathbf{F} \otimes \mathbb{I}_S, \tag{3.45}$$

where S is the dimension of V.

Unfiltered spectrum. In the unfiltered case ($\sigma_{\mathrm{f}} = 0$), we observe (i) that all eigenvalues have a real part greater than one; (ii) that the eigenvalue of smallest magnitude is $\lambda_{\min} = 1$; and (iii) that the eigenvalue of largest magnitude has the form

$$\lambda_{\max} = 1 + \eta \frac{1}{\Delta x\, \sigma_{\mathrm{t}}^{\star}}. \tag{3.46}$$

This formula is based on a linear fit of the numerical results (see Fig. 3.3). The constant η, which depends on the other physical and numerical parameters, is determined by the fit. In the streaming limit ($\sigma_{\mathrm{t}} = 0$), the ratio $\lambda_{\max}/\lambda_{\min}$, which gives a lower bound on the condition number of $\mathcal{A}$, is given by

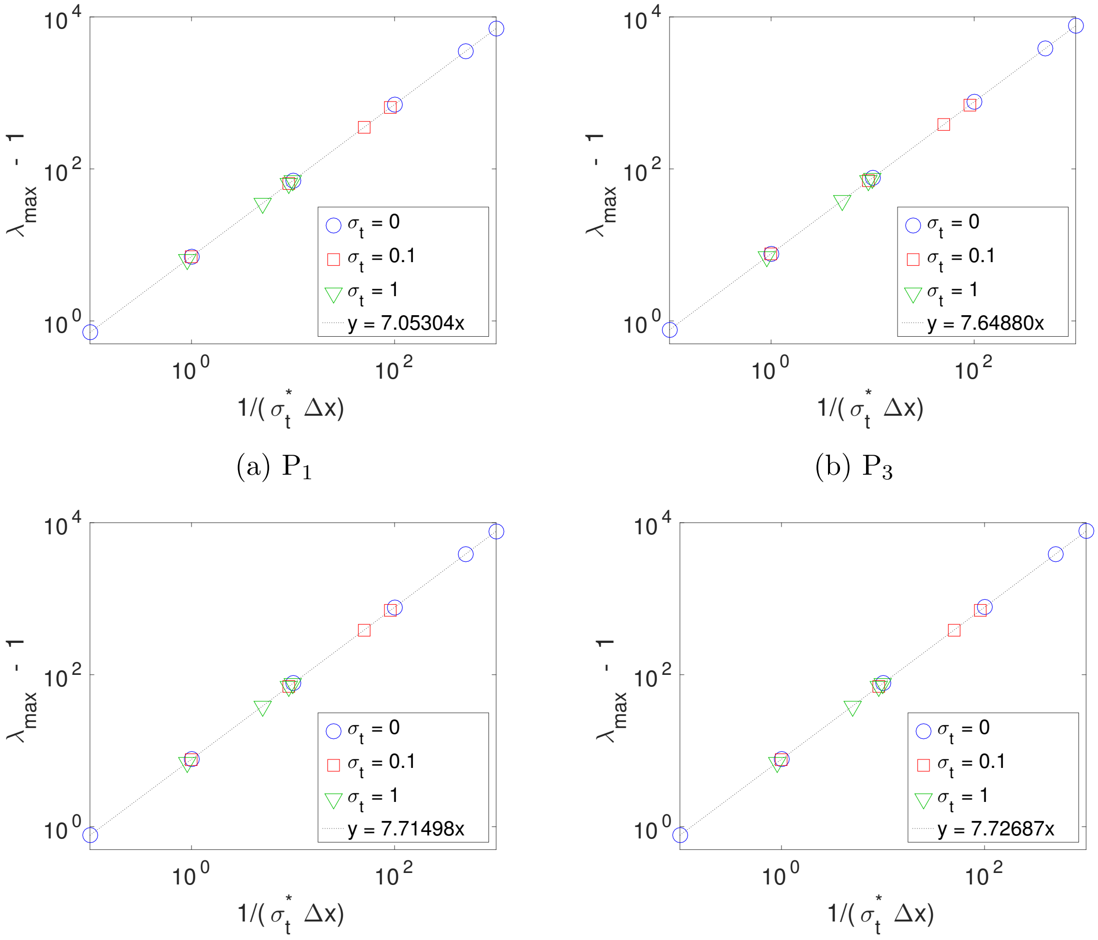
$$\frac{\lambda_{\max}}{\lambda_{\min}} = 1 + \eta \frac{1}{\Delta x\, \sigma_{\mathrm{t}}^{\star}} = 1 + \eta \frac{c\Delta t}{\Delta x}. \tag{3.47}$$

Thus while $\mathcal{A}$ is never singular, it does become ill-conditioned as the CFL number increases.

Filtered spectrum. We consider again the same system, but now with the Lanczos filter. Fig. 3.4 shows the eigenspectrum for different values of σ_{f}.

For small values of σ_{f}, all the eigenvalues are shifted to the right, away from the

[13] The choice of $\mathcal{A}$ is for convenience. Because M_0 is diagonal with minimum and maximum elements $1/27$ and 1 respectively, the asymptotic behavior of $\mathcal{A}$ is identical to that of A.



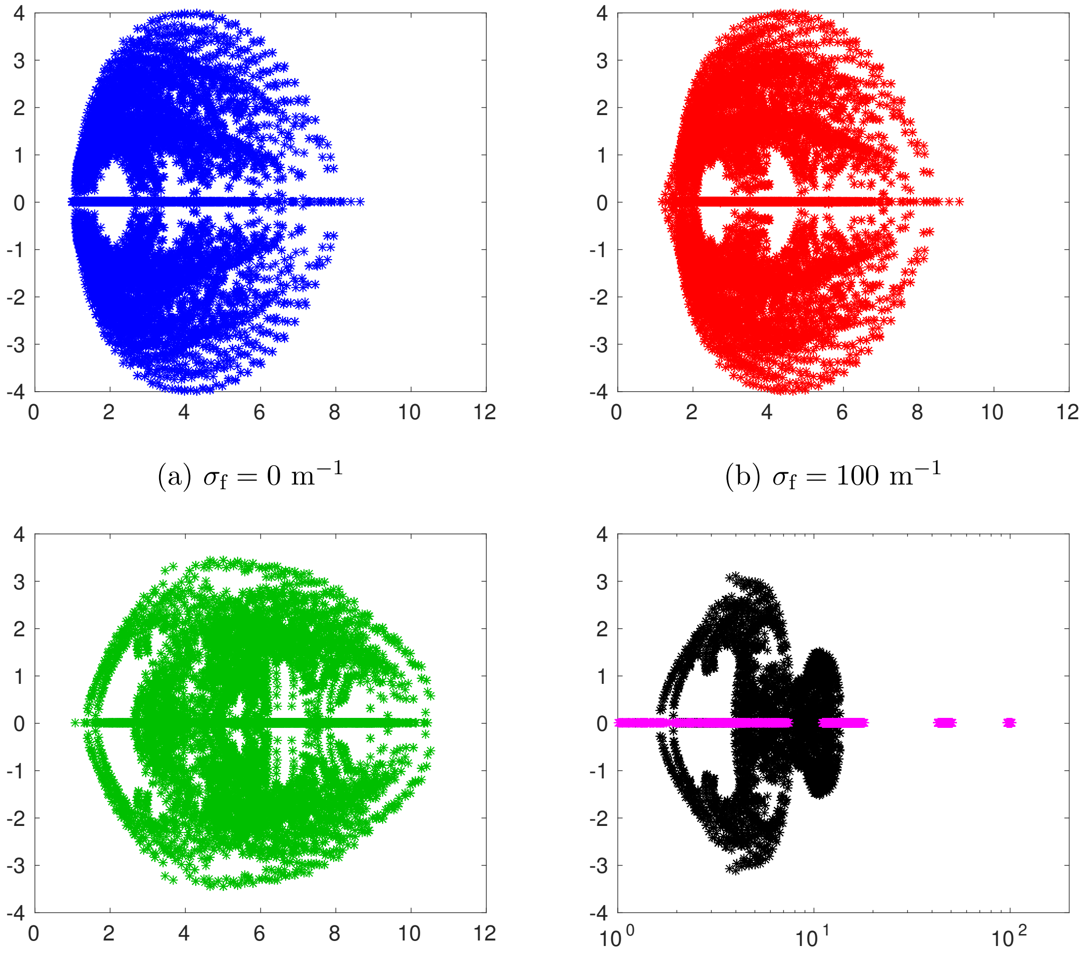
(a) P_1 (b) P_3

(c) P_5 (d) P_7

Figure 3.3: The value $\lambda_{\max} - 1$ as a function of $1/(\sigma_{\mathrm{t}}^{\star}\,\Delta x)$ for different values of σ_t, $c\Delta t/\Delta x$ and N. For a given value of N, the computed values of $\lambda_{\max} - 1$ are all exactly predicted by the linear fit to the precision of the calculation (6 digits). The slope changes slightly depending on the other parameters of the simulation.

origin (see Figs. 3.4b and 3.4c). Previous observations [5] and theoretical results [60] show that the solver converges faster when eigenvalues are clustered away from the origin. However, since there are no distinct clusters, we quantify this notion using the ratio:

$$\kappa(\sigma_{\mathrm{f}}) \equiv \left|\frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\min}}\right|. \tag{3.48}$$



(a) $\sigma_{\mathrm{f}} = 0\ \mathrm{m}^{-1}$

(b) $\sigma_{\mathrm{f}} = 100\ \mathrm{m}^{-1}$

(c) $\sigma_{\mathrm{f}} = 500\ \mathrm{m}^{-1}$

(d) $\sigma_{\mathrm{f}} = 10^3$, $10^4\ \mathrm{m}^{-1}$ (black, magenta)

Figure 3.4: Spectrum of $\mathcal{A}$ in the complex plane for different values of σ_{f} for a P_3 calculation. Note the log scale for the real part in the last subfigure.

These values, which are given in Table 3.1, suggest that the filter will improve the solver convergence, at least for small values of σ_{f}.

As σ_{f} continues to increase (see Fig. 3.4d), the spectrum continues to stretch along the real axis. At this point, it is more complicated to predict the behavior of the solver.

For very large values of σ_{f}, the spectrum appears in $N+1$ clusters very near the

$\dfrac{c\Delta t}{\Delta x}$	σ_{f}	$\lambda_{\max}$	$\lambda_{\min}$	$\kappa(\sigma_{\mathrm{f}})/\kappa(0)$
1	0	1.0000	8.6488	100.00%
1	50	1.1392	8.8633	88.65%
1	100	1.1737	9.0737	88.00%
1	200	1.1494	9.4792	94.75%
10	0	1.0000	77.488	100.00%
10	50	2.3920	79.633	42.22%
10	100	2.7366	81.737	37.74%
10	200	2.4940	85.792	43.67%

Table 3.1: Behavior for small values of σ_{f} of the maximum and minimum eigenvalues in the P_1 case for different values of σ_{f} and different CFL numbers. The ratio $\kappa(\sigma_{\mathrm{f}})/\kappa(0)$, which characterizes the radius of the eigenspectrum relative to its distance to the origin, compared to the unfiltered case, highlights that the spectrum is more clustered for "small" values of σ_{f}.

real-axis (see Fig. 3.4d for the P_3 case). Indeed, it follows from the decomposition in Eq. 3.45 and Gerschgorin's Circle Theorem that, because $\mathbf{F}$ is diagonal,

$$\max_{\lambda(\sigma_{\mathrm{f}})\in\mathrm{eig}(\mathcal{A})} \min_{\mu\in\mathrm{eig}(\mathbf{F})} \left|\lambda(\sigma_{\mathrm{f}}) - \frac{\sigma_{\mathrm{f}}\mu}{\sigma_{\mathrm{t}}^{\star}}\right| \leq \varrho, \tag{3.49}$$

where ϱ is independent of σ_{f}. Hence as $\sigma_{\mathrm{f}} \to \infty$, the spectrum separates into $N+1$ distinct clusters $\mathcal{C}_\mu$ contained in discs with centers $\sigma_{\mathrm{f}}\mu/\sigma_{\mathrm{t}}^{\star}$, $\mu \in \mathrm{eig}(\mathbf{F})$, and radius ϱ. Moreover for $\mu \neq 0$, the relative size of the closure $\mathcal{C}_\mu$ converges to zero:

$$\frac{\mathrm{dist}(\sigma_{\mathrm{f}}\mu/\sigma_{\mathrm{t}}^{\star}, \mathcal{C}_\mu)}{\sigma_{\mathrm{f}}\mu/\sigma_{\mathrm{t}}^{\star}} \to 0 \quad \text{as} \quad \sigma_{\mathrm{f}} \longrightarrow 0. \tag{3.50}$$

In has been shown in [60] that for a matrix with d outlying eigenvalues, the convergence properties of the GMRES solver are no longer affected by those outlying eigenvalues after d iterations. Based on Eq. 3.50, we conjecture that for large σ_{f}, the solver treats the clusters $\mathcal{C}_\mu$, $\mu \neq 0$, as isolated points and that the majority of iterations is due to the spectrum in the cluster $\mathcal{C}_0$.

This would imply in particular that, for a uniform filtering strategy, the number of iterations required should not depend on N as $\sigma_{\mathrm{f}} \longrightarrow \infty$. For a local filtering strategy, the behavior is more complicated to predict as only the eigenvalues that correspond to filtered regions would be shifted as $\sigma_{\mathrm{f}} \longrightarrow \infty$ while the eigenvalues corresponding to unfiltered regions would be unchanged. It is then conceivable that the number of iterations might depend on N in that limit because the size of the cluster staying close to the origin increases as N gets larger.

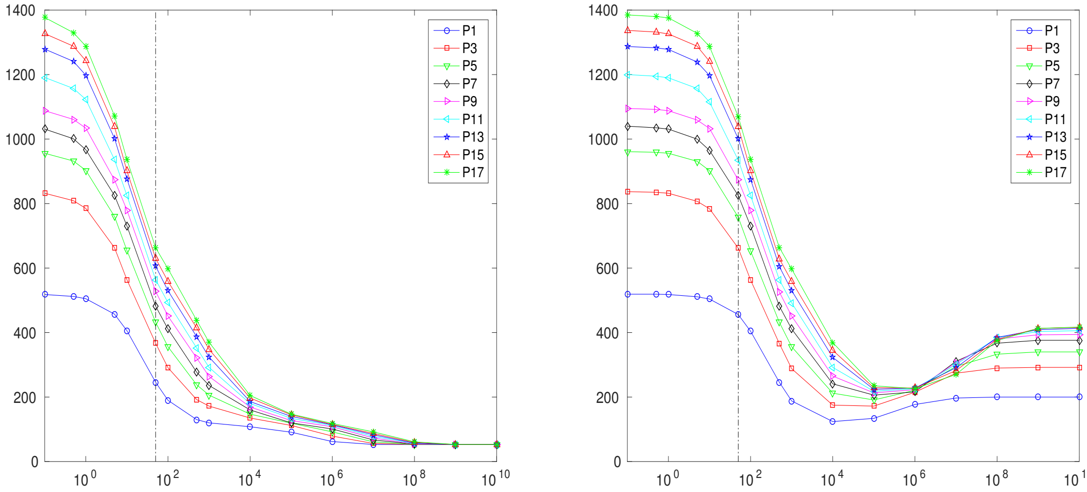
Although further analysis is required to confirm this, it is believed that this considerations on pure transport are useful to understand the nonlinear TRT case because each nonlinear solve consists of multiple linear solves with the additional terms mostly contributing to the diagonal terms on the global matrix. The above reasoning relying on Gerschgorin discs would most likely translate to the general case, though the derivation would be more tedious.

Let us now see if these results are indeed seen in practice.

3.3.2.2 Solver Efficiency for the Crooked Pipe

We next consider the effect of the filter on the iteration count for the full nonlinear system when solving the Crooked Pipe problem using the MOOSE implementation (cf. Section 3.2.6.2). A full description of this problem and numerical solutions can be found in Section 3.4 (see Fig. 3.6 for the layout).

In Fig. 3.5, the total number of GMRES iterations are displayed for the first time step, which is typically the most expensive. The numbers are generally consistent with the analysis above in the linear case. For the uniform filter, the number of iterations decreases monotonically as σ_{f} increases to a fixed number that is independent of N. For the local filter, the iterations decrease initially and increase to a fixed value that is different for each N. (Note however, that this increase occurs well beyond any practical value of σ_{f} and that the number of iterations as $\sigma_{\mathrm{f}}, N \longrightarrow \infty$ seems to converge.) The difference in performance between the uniform and local strategies is due to the fact that the local filter introduces an artificial discontinuity in the effective material cross-section.



(a) Uniform filtering

(b) Local filtering

Figure 3.5: Iteration count for the first time step as a function of N and the filter strength σ_{f} (in cm^{-1}), using the Lanczos filter. As a reference, the value of σ_{f} for this test problem was in practice chosen to be 50 cm^{-1} (vertical line). The value of σ_{f} for the local filter designates the maximum value; see Fig. 3.9 for a complete description.

In Table 3.2, we show the gain that we obtain in the number of iterations for the practical value of σ_{f} compared to the unfiltered case. In both strategies, the improvement in performance is noteworthy. Indeed, the number of iterations for the practical value of σ_{f} decreases by more than one-half when compared to the unfiltered case for uniform filtering and by more than 20% for the locally filtered P_N with

$N > 1$.

Filtering	P_1	P_3	P_5	P_7	P_9	P_{11}	P_{13}	P_{15}	P_{17}
Uniform	0.473	0.439	0.449	0.463	0.482	0.470	0.471	0.471	0.479
Local	0.877	0.791	0.788	0.793	0.797	0.779	0.777	0.776	0.773

Table 3.2: Factor by which the number of iterations is multiplied for $\sigma_{\mathrm{f}} = 50$ cm^{-1} compared to the unfiltered case ($\sigma_{\mathrm{f}} = 0$).

3.3.3 Comparison to Error Estimates

Frank, Hauck and Kuepper [11] have derived error estimates for the convergence of Filtered P_N for the case of pure transport. Here we compare these estimates to numerical results for smooth and non-smooth solutions of thermal radiative transfer with non-linear material properties. Define the angular error

$$E_N = ||\Psi_N - \Psi||_{L^2} = \left(\sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} \int_{\mathcal{D}} \left((\Psi_N)_\ell^m - \Phi_\ell^m \right)^2 \mathrm{d}x \right)^{1/2}, \tag{3.51}$$

where the expansion coefficients of Ψ_N solve the (time continuous) FP_N equations and we have added the subscript N to emphasize the dependence on N. Based on [11], we expect

$$E_N = \mathcal{O}\left(N^{-\min\{k,\alpha\}} \right), \tag{3.52}$$

where k is the order of convergence in the unfiltered case, α is the order of the filter and the implied constant in Eq. 3.52 depends on Ψ and the time t. As a reminder, the Lanczos and Spherical Spline filter orders are two and four, respectively. The order of the exponential filter of order α is precisely α (see Eq. 3.14).

As a test problem, we use the smooth Marshak Wave [61]. This problem is defined on a slab geometry, which implies that $\mathbf{\Phi}$ only depends on $x \in [0,1]$ and t. It assumes a purely absorbing medium with cross-section $\sigma_{\mathrm{t}} = \sigma_a = (ac)^{-3/4}\, T^{-3}$. The material heat capacity C_v is set to $a^{1/4}c^{-3/4}$.[14] The initial conditions are

$$\Phi_\ell^m(x,0) = \frac{\delta_{\ell,0}}{\sqrt{w}}\Big(\phi_l + (\phi_r - \phi_l)\frac{1+\tanh\big(50(x-0.25)\big)}{2}\Big), \tag{3.53}$$

$$T(x,0) = \Big(\frac{\sqrt{w}}{ac}\Phi_0^0(x,0)\Big)^{1/4}, \tag{3.54}$$

where $\phi_l = 4$, $\phi_r = 0.004$. We use Dirichlet boundary conditions (see Sec. 3.2.3.2) at both boundaries: $\Psi^{\mathrm{inc}} = \phi_l/w$ at $x = 0$ and $\Psi^{\mathrm{inc}} = \phi_r/w$ at $x = 1$. We use 200 uniform cells of width $\Delta x = 0.005$. The final time is $t_{\max} = \Delta t = 0.005/c$ and the filter strength is $\sigma_{\mathrm{f}} = 100$.

To test both aspects of (3.52), we consider two problems. In the first one, $S = 0$; in the second, we add a non-smooth, volumetric source that is constant in x and t and a hat function in μ.

$$S(x,\mu,t) = \begin{cases} 20\big|(|\mu| - 0.5)\big|, & 0 \le |\mu| \le 0.5 \\ 0, & 0.5 \le |\mu| \le 1 \end{cases}. \tag{3.55}$$

Thus the angular derivative is not continuous. To estimate the error E_N, we use Ψ_{99} and Ψ_{199}, respectively, to approximate Ψ in the smooth and non-smooth cases.[15]

[14] In the original paper, the equations solved can be obtained by setting $a = c = C_v = 1$. Here we prefer to keep the physical constants unchanged and use a different scaling, which leads to slightly different expressions for the cross-sections, the heat capacity, the time step and the temperature. We are however solving the same equations.

[15] In the non-smooth case, the reference solution must be more refined in order to see a more saturated convergence rate. Similarly, the exponential filters being stronger than the Lanczos and SSpline filters for a given σ_{f} (see Fig. 3.2), σ_{f} is set to 1 for the exponential filters in the non-smooth case.

In Tables 3.3–3.8 we show numerical values of E_{N_i} and the convergence rate

$$r_i = -\frac{\log(E_{N_i}/E_{N_{i+1}})}{\log(N_i/N_{i+1})}, \tag{3.56}$$

for several different filters in the smooth case. As expected, the order of convergence is close to the order of the filter.

N	E_N	r
1	1.29E-06	4.35
3	1.08E-08	7.59
7	1.74E-11	5.04
15	3.75E-13	3.78
29	3.10E-14	0.68
49	2.17E-14	NA
99	Reference	NA

Table 3.3: Unfiltered (smooth)

N	E_N	r
1	1.59E-06	2.01
3	1.75E-07	1.64
7	4.36E-08	1.83
15	1.08E-08	1.97
29	2.95E-09	2.15
49	9.56E-10	NA
99	Reference	NA

Table 3.4: Lanczos (smooth)

N	E_N	r
1	1.78E-06	2.84
3	7.89E-08	3.25
7	5.00E-09	3.64
15	3.13E-10	3.82
29	2.53E-11	3.91
49	3.26E-12	NA
99	Reference	NA

Table 3.5: SSpline (smooth)

N	E_N	r
1	2.01E-05	0.65
3	9.77E-06	1.30
7	3.25E-06	1.70
15	8.90E-07	1.87
29	2.59E-07	1.94
49	9.39E-08	NA
99	Reference	NA

Table 3.6: Exponential filter, order 2 (smooth)

N	E_N	r
1	9.87E-06	2.06
3	1.02E-06	3.17
7	6.98E-08	3.63
15	4.38E-09	3.81
29	3.54E-10	3.90
49	4.58E-11	NA
99	Reference	NA

Table 3.7: Exponential filter, order 4 (smooth)

N	E_N	r
1	1.72E-06	3.51
3	3.64E-08	6.86
7	1.09E-10	6.77
15	6.23E-13	2.24
29	1.42E-13	1.83
49	5.43E-14	NA
99	Reference	NA

Table 3.8: Exponential filter, order 8 (smooth)

In Tables 3.9–3.14 we show the results in the non-smooth case. We observe that the order of convergence is not affected by the order of the filter. This is as expected, since $k < \alpha$. Because lower-order filters are more robust, it is generally best to choose α no less than k, but as close to k as possible.

N	E_N	r
1	1.60E-02	0.23
3	1.23E-02	1.40
7	3.76E-03	1.43
15	1.27E-03	1.50
29	4.73E-04	1.63
49	2.01E-04	1.57
69	1.18E-04	1.11
89	8.87E-05	1.65
109	6.35E-05	NA
199	Reference	NA

Table 3.9: Unfiltered (non-smooth)

N	E_N	r
1	1.60E-02	0.23
3	1.23E-02	1.40
7	3.78E-03	1.43
15	1.27E-03	1.49
29	4.75E-04	1.63
49	2.02E-04	1.56
69	1.19E-04	1.12
89	8.93E-05	1.64
109	6.40E-05	NA
199	Reference	NA

Table 3.10: Lanczos (non-smooth)

N	E_N	r
1	1.60E-02	0.23
3	1.23E-02	1.38
7	3.84E-03	1.45
15	1.27E-03	1.48
29	4.79E-04	1.61
49	2.06E-04	1.55
69	1.21E-04	1.14
89	9.06E-05	1.63
109	6.51E-05	NA
199	Reference	NA

Table 3.11: SSpline (non-smooth)

N	E_N	r
1	1.60E-02	0.23
3	1.23E-02	1.40
7	3.77E-03	1.43
15	1.27E-03	1.49
29	4.74E-04	1.63
49	2.02E-04	1.56
69	1.18E-04	1.11
89	8.91E-05	1.64
109	6.39E-05	NA
199	Reference	NA

Table 3.12: Exponential filter, order 2 (nonsmooth)

N	E_N	r
1	1.60E-02	0.23
3	1.23E-02	1.40
7	3.77E-03	1.43
15	1.27E-03	1.50
29	4.73E-04	1.63
49	2.01E-04	1.56
69	1.18E-04	1.11
89	8.88E-05	1.65
109	6.36E-05	NA
199	Reference	NA

Table 3.13: Exponential filter, order 4 (nonsmooth)

N	E_N	r
1	1.60E-02	0.23
3	1.23E-02	1.40
7	3.76E-03	1.43
15	1.27E-03	1.50
29	4.73E-04	1.63
49	2.01E-04	1.56
69	1.18E-04	1.11
89	8.88E-05	1.65
109	6.36E-05	NA
199	Reference	NA

Table 3.14: Exponential filter, order 8 (nonsmooth)

3.3.4 Spatial and Temporal Convergence Studies

We check in this section that our code gives the expected orders of convergence in time and in space. In time, we expect an order of convergence for the global error to be one since we use the backward Euler discretization. In space, we expect to get a second order convergence using Linear DGFEM. In both cases, we use the solution on a very refined mesh as the reference solution and we compute the error with respect to that reference, leaving all the other parameters unchanged. We consider the same test problem studied and described in Section 3.3.3. Tables 3.15 and 3.16 show that the error in the angular flux indeed behaves as expected. The filter type does not change the order of convergence within three digits of accuracy.[16]

[16] Note however that the reference solution depends on the filter type since it is obtained by running the same calculation on a finer mesh.

Δt	E_N	r
1/10	5.99×10^{-2}	0.711
1/20	3.66×10^{-2}	0.771
1/40	2.15×10^{-2}	0.850
1/80	1.19×10^{-2}	0.958
1/160	6.13×10^{-3}	1.128
1/320	2.80×10^{-3}	1.524
1/640	9.75×10^{-4}	NA
1/1280	reference	NA

Table 3.15: Convergence in time for a P_9 calculation and 80 spatial cells. The final time is $t_{\max} = 1$.

Cells	E_N	r
10	5.14×10^{-2}	0.252
20	4.31×10^{-2}	1.845
40	1.20×10^{-2}	2.582
80	2.01×10^{-3}	2.265
160	4.18×10^{-4}	2.071
320	9.94×10^{-5}	1.977
640	2.52×10^{-5}	NA
1280	reference	NA

Table 3.16: Convergence in space for a P_9 calculation. The final time is $t_{\max} = 10\Delta t = 0.01$.

It was also checked that the same orders of convergence are obtained for a 2-D problem with an isotropic initial condition and a source given by $S = 1+\sin(4\pi x)\sin(4\pi y)$ in a unit square with uniform material properties and reflective boundary conditions at every boundary. We have also checked that our code converges to the correct semi-analytical solution for the P_1 Su-Olson test problem [62]. Using the BDF-2 time discretization yields second-order convergence in time.

3.4 Numerical Results

In this section, we study a variation[17] of the Crooked Pipe benchmark [45]. In this problem, there are two purely absorbing materials in a two-dimensional, Cartesian domain that is 7 cm × 2 cm, respectively, in the x and y directions (as shown in Fig. 3.6), with the origin located at the bottom left corner. There is no z-dependence.

[17] The original Crooked Pipe problem has a cylindrical geometry; here we use Cartesian coordinates.

The location of the two materials is shown in Fig. 3.6. In the thin one, $\sigma_a = 20$ m^{-1} and $C_v = 4.3 \times 10^4$ J/m^3/K; in the thick one, $\sigma_a = 2 \times 10^4$ m^{-1} and $C_v = 4.3 \times 10^7$ J/m^3/K.[18]

On the left boundary, we apply an isotropic incoming source (see Eq. 2.3):

$$\Psi^{\mathrm{d}} = \frac{ac}{w} T_L^4, \quad T_L = 0.3 \text{ keV}, \tag{3.57}$$

at $x = 0$ for $0 \leq y \leq 0.5$ cm – that is, only along the thin region of the left boundary. A reflective boundary condition is imposed on the bottom boundary and open boundaries everywhere else. The initial temperature is set to $T_0 = 0.05$ keV, and the expansion coefficients of the initial scalar flux are

$$\Phi_\ell^m(x, 0) = \frac{acT_0^4}{\sqrt{w}} \delta_{\ell,0}. \tag{3.58}$$

As explained in Section 3.2.4, we lump the mass matrix for the collision terms in order to increase robustness. The time step is set to 0.05 ns using a BDF-2 time-discretization scheme.[19]

3.4.1 Comparison with IMC: Simplified Problem

The sharp material interfaces and the absence of scattering in the Crooked Pipe make it very difficult to solve. Furthermore, because σ_a in the thick region is very large, fully converging the solution requires a significant amount of computational resources. Thus, for verification purposes, we begin with a simpler test problem and compare it to a solution obtained from an IMC calculation. In this problem, $\sigma_a = 20$ m^{-1} everywhere and the source on the left is applied along the entire left boundary.

[18] Note that 4.3×10^7 J/m^4/K ≈ 0.5 GJ/cm^4/keV, which are the units that were actually used.

[19] The difference with the Backward-Euler scheme was barely noticeable, suggesting that the temporal error is not dominant with this time step. Increasing the time step to 0.1 ns also had a negligible impact.

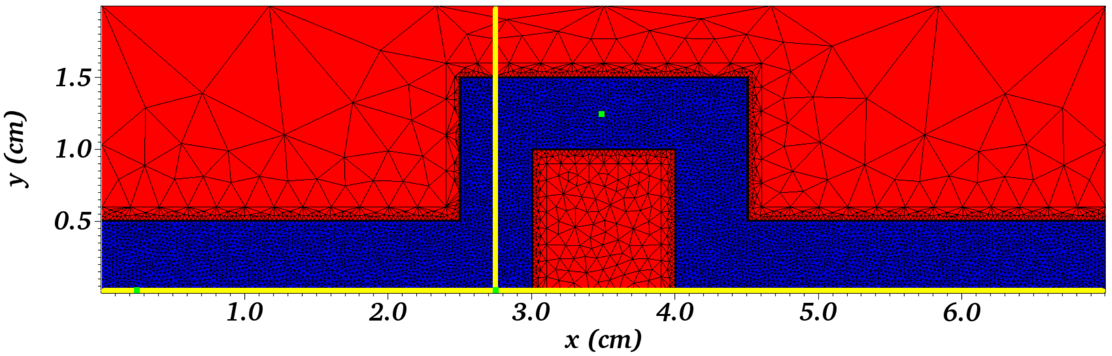


Figure 3.6: Mesh for the Crooked Pipe test problem. In the thin regions (shown in blue), $\sigma_\mathrm{t} = \sigma_a = 20$ m^{-1} and $C_v = 4.3 \times 10^4$ J/m^4/K. In the thick regions (shown in red), each of these constants is factor of 1000 greater. The two straight lines (in yellow) are $y = 0$ and $x = 2.75$ cm; the three points (in green) are $(x_1, y_1) = (0.25\,\mathrm{cm}, 0)$, $(x_2, y_2) = (2.75\,\mathrm{cm}, 0)$ and $(x_3, y_3) = (3.5\,\mathrm{cm}, 1.25\,\mathrm{cm})$. The interface between thick and thin regions is refined so that there are several cells per mean free path. (The first layer of cells has a width of 0.005 cm.) The entire mesh contains 20,106 triangular elements.

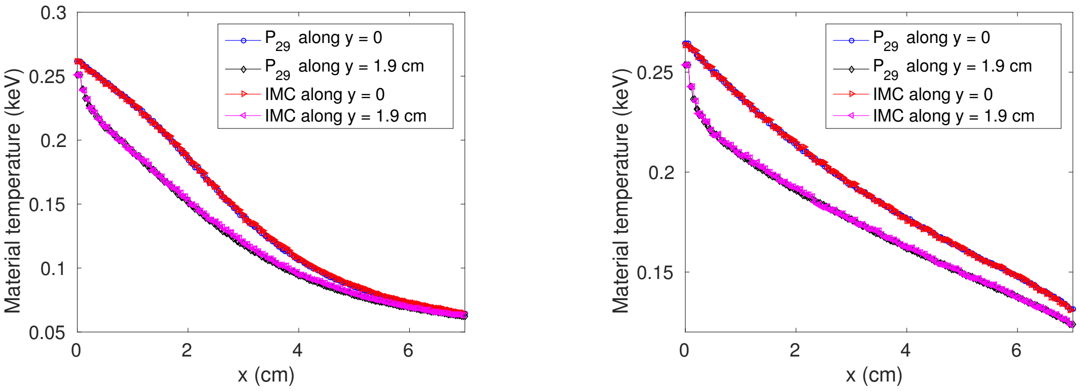
We verify that a P$_{29}$ solution agrees well with the IMC one; see Fig. **??**. With this fact in mind, we use a P$_{39}$ solution with the spatial mesh shown in Fig. 3.6 as the reference solution below.

3.4.2 Filtering Strategy

For robustness, we use the Lanczos filter in all of the filtered calculations. Based on the guidelines detailed in the previous section, we consider three filtering strategies.

- **Unfiltered.** This is the original P$_N$ method, obtained by setting $\sigma_\mathrm{f} = 0$.

- **Uniformly filtered.** Here σ_f is a fixed constant across the domain. Based on the discussion in Section 3.3.1 and given that the material temperature T is virtually always above the initial temperature for $N = 7$ (see Fig. 3.10), we choose a value such that $\sigma_\mathrm{f}\, f(1, N_0 = 7)$ is comparable to the cross-section in



(a) T at $t = 1$ ns.

(b) T at $t = 10$ ns.

Figure 3.7: Temperature profile along $y = 0$ and $y = 1.9$ cm as a function of x. For convergence purposes, these results are obtained on the same geometry as Fig. 3.6 except that the material properties are set to the thin region everywhere and that the source is applied on the entire left boundary. The mesh however was a uniform rectangular grid (100×50 for the IMC, 112×32 for the P_{29}).

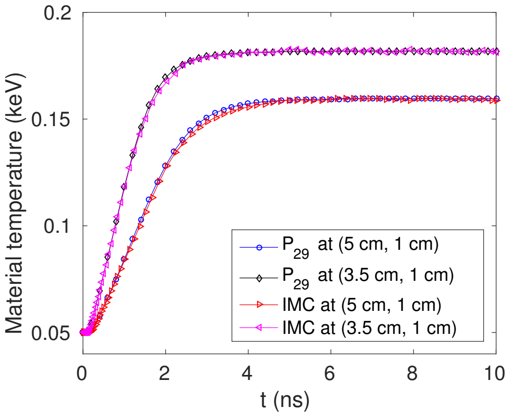


Figure 3.8: Temperature profile $(x, y) = (3.5$ cm, 1 cm$)$ and $(x, y) = (5$ cm, 1 cm$)$ as a function of time. For convergence purposes, these results are obtained on the same geometry as Fig. 3.6 except that the material properties are set to the thin region everywhere and that the source is applied on the entire left boundary. The mesh however was a uniform rectangular grid (100×50 for the IMC, 112×32 for the P_{29}).

the thin part of the problem. Setting $\sigma_f = 5 \times 10^3$ m^{-1} gives $\sigma_f f(1,7) \approx 13$ m^{-1} for the Lanczos filter. (Recall that $\sigma_t = 20$ m^{-1} in the thin region.)

- **Locally filtered.** The spatial profile of σ_f in this case is provided by Fig. 3.9. Following the guidelines of Section 3.3.1, we set it to 5×10^3 m^{-1} after the first elbow of the pipe (where the radiation tends to become negative) as well as in an upstream region of comparable size. It is set to a value ten times smaller in the rest of the pipe because it appeared to slightly improve the profile therein.

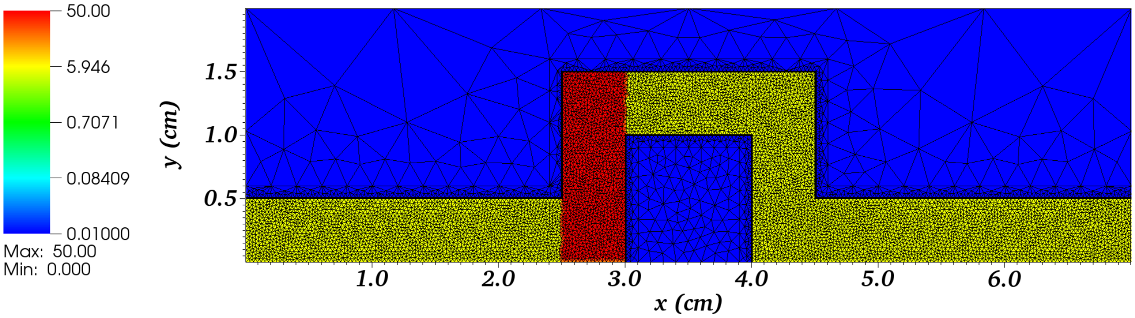


Figure 3.9: Value of σ_f (in cm^{-1}) for the locally filtered calculations.

3.4.3 Results

In all simulations, radiation flows rapidly from the left boundary to the first elbow of the pipe. It is then absorbed and re-emitted by the material. Isotropic re-emission allows for some of the radiation to change direction and propagate further down the pipe.

In the following subsections, we present 2-D maps of the different solutions at a fixed time. We then examine these solutions in more detail: first along specified lines in space with time fixed and then at fixed points in space over a given time interval. As expected, the locally filtered strategy generally produces the best solutions: it maintains a positive scalar flux without damping its profile too strongly.

3.4.3.1 Scalar Flux 2-D Maps

In Fig. 3.10, we plot the heat map of the material temperature T for the unfiltered case as $t = 0.35$ sh, the approximate time that gives the minimum values. It is based on that figure that we chose the value for σ_{f} in Section 3.4.2.

In Figs. 3.11-3.13, we plot heat maps of the scalar flux Φ_0^0 for the unfiltered, uniformly filtered, and locally filtered spherical harmonic calculations, respectively, at time $t = 0.05$ sh. It is around this time that the value of Φ_0^0 in the unfiltered solution reaches its minimum. Each figure contains solutions for $N = 1$, 3, 5 and 7. The filtered P_{39} solution with uniform filtering is included for reference.

Fig. 3.11 shows the defects of the P_N closures. P_1 allows energy to flow through the thin region around the bend in the pipe. Meanwhile, the P_3, P_5 and P_7 calculations have regions – the edge of shadows – where the scalar flux becomes negative. If a low enough initial temperature is chosen, the temperature will actually become negative, then yielding nonsensical results. Fig. 3.12 shows that uniform filtering efficiently removes regions of negativity, but also over-damps the scalar flux profile for low values of N.

3.4.3.2 Lineouts

In this section and the following, we provide L2-error tables to quantify the filter performances. It is generally defined as $(\int_{u_{\min}}^{u_{\max}} (\Phi_0^0 - \Phi_{\mathrm{ref}})^2 \, \mathrm{d}u)^{1/2}$, the reference being the P_{39} curve. For the lineouts, u represents the corresponding spatial variable ($u = x$ for Fig. 3.14, $u = y$ for Fig. 3.15). For the time histories, it represents the time t.

Figs. 3.14 and 3.15 show lineouts of the scalar flux profile at time $t = 0.05$ sh along the lines $y = 0$ cm and $x = 2.75$ cm, respectively. Except for P_1, all of the unfiltered P_N solutions (Fig. 3.14a) along $y = 0$ are very similar and agree with the reference solution to within 12%. In the uniformly filtered case (Fig. 3.14b), over damping has slowed the effective flow of radiation down the pipe, causing solutions to be much less accurate. Meanwhile, the locally filtered results (Fig. 3.14c) are slightly better than the unfiltered ones.

Along the line $x = 2.75$ cm, nonphysical oscillations cause the scalar flux profile for the unfiltered equations (Fig. 3.15a) to reach negative values. The filter helps significantly in this region, with the local filter (Fig. 3.15c) again outperforming the uniform one, especially for small values of N. Even so, the filtered solutions do over-predict the scalar flux compared to the P_{39} solution after the first elbow.

3.4.3.3 Time Histories

As suggested in [45], we also monitor the evolution of Φ_0^0 as a function of time at 3 different points in space: $(x_1, y_1) = (0.25\,\mathrm{cm}, 0)$, $(x_2, y_2) = (2.75\,\mathrm{cm}, 0)$, and $(x_3, y_3) = (3.5\,\mathrm{cm}, 1.25\,\mathrm{cm})$. These results are given in Figs. 3.16 - 3.18.

At (x_1, y_1) (Fig. 3.16), all the filtering approaches give reasonable results. The values of Φ_0^0 for uniform filtering in Fig. 3.16b are slightly higher than with the other two types because the radiation propagates more slowly and is therefore more concentrated at the entrance of the pipe. For the same reason, the unfiltered calculations tend to underestimate the temperature at that point for small values of N.

At (x_2, y_2), in Fig. 3.17a, the unfiltered solutions are all reasonably close to the P_{39} solution at early times (see Table 3.17d), except for the P_3 solution, which is affected by the time history at this point. Similar behavior for P_5 or P_7 can be observed at different points in space. The uniformly filtered solutions (Fig. 3.17b) again suffer from over damping, while the locally filtered results (Fig. 3.17c) agree well with the reference solution. Only P_1 does not capture the shape accurately.

At (x_3, y_3) in Fig. 3.18a, the unfiltered scalar intensities are too high. The filtering improves this, with the uniform filter giving the best results for $N = 1$ and $N = 3$. For $N = 5$ and $N = 7$, the local and uniform filters have similar errors.

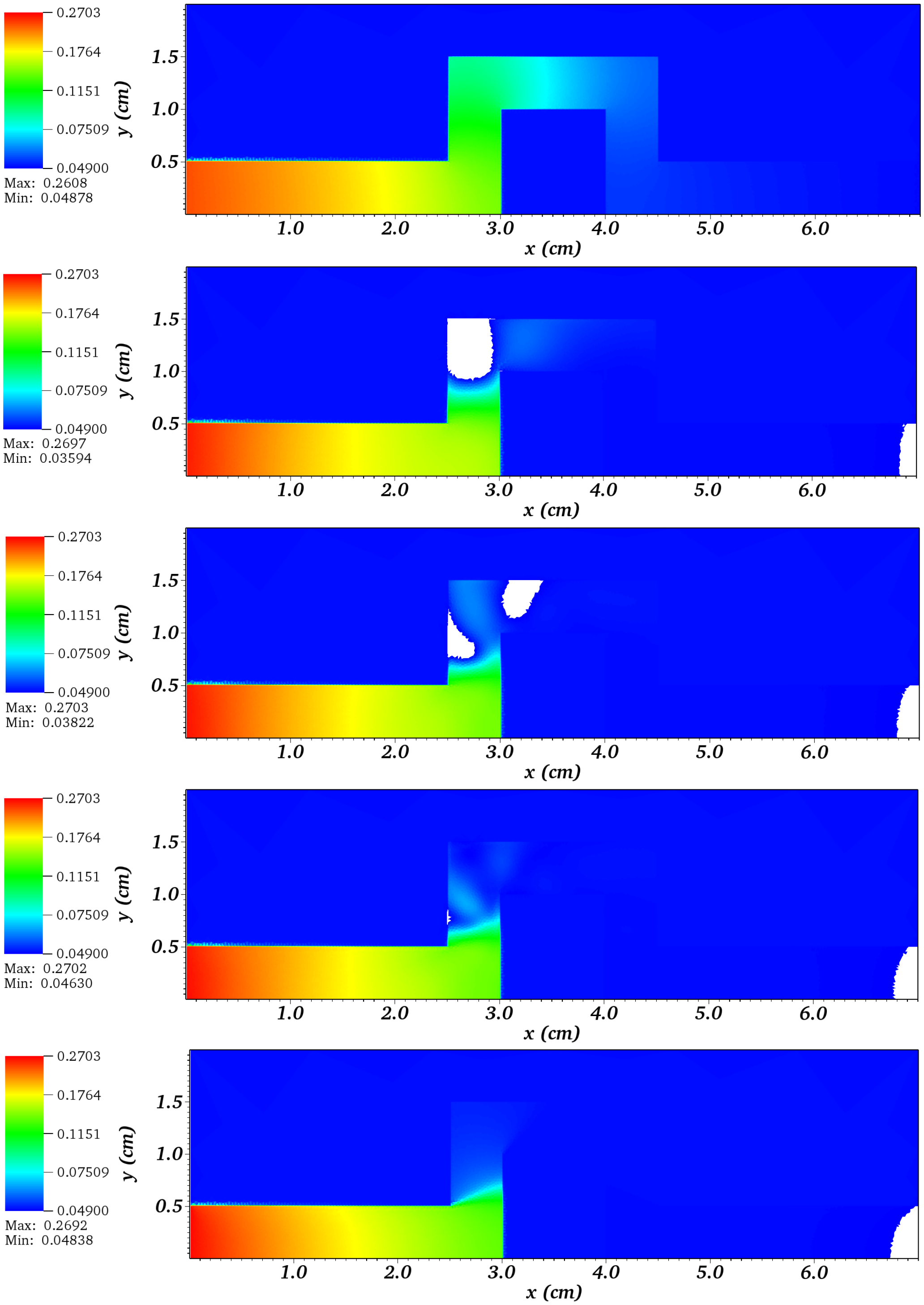


Figure 3.10: Material temperature T (in keV) at $t = 0.35$ sh for unfiltered P_1, P_3, P_5, and P_7 calculations (from top to bottom). The last plot is a uniformly filtered P_{39} calculation for reference. The white regions show where T is less than 98% of the initial temperature. Only the piecewise constant component of the solution is shown.

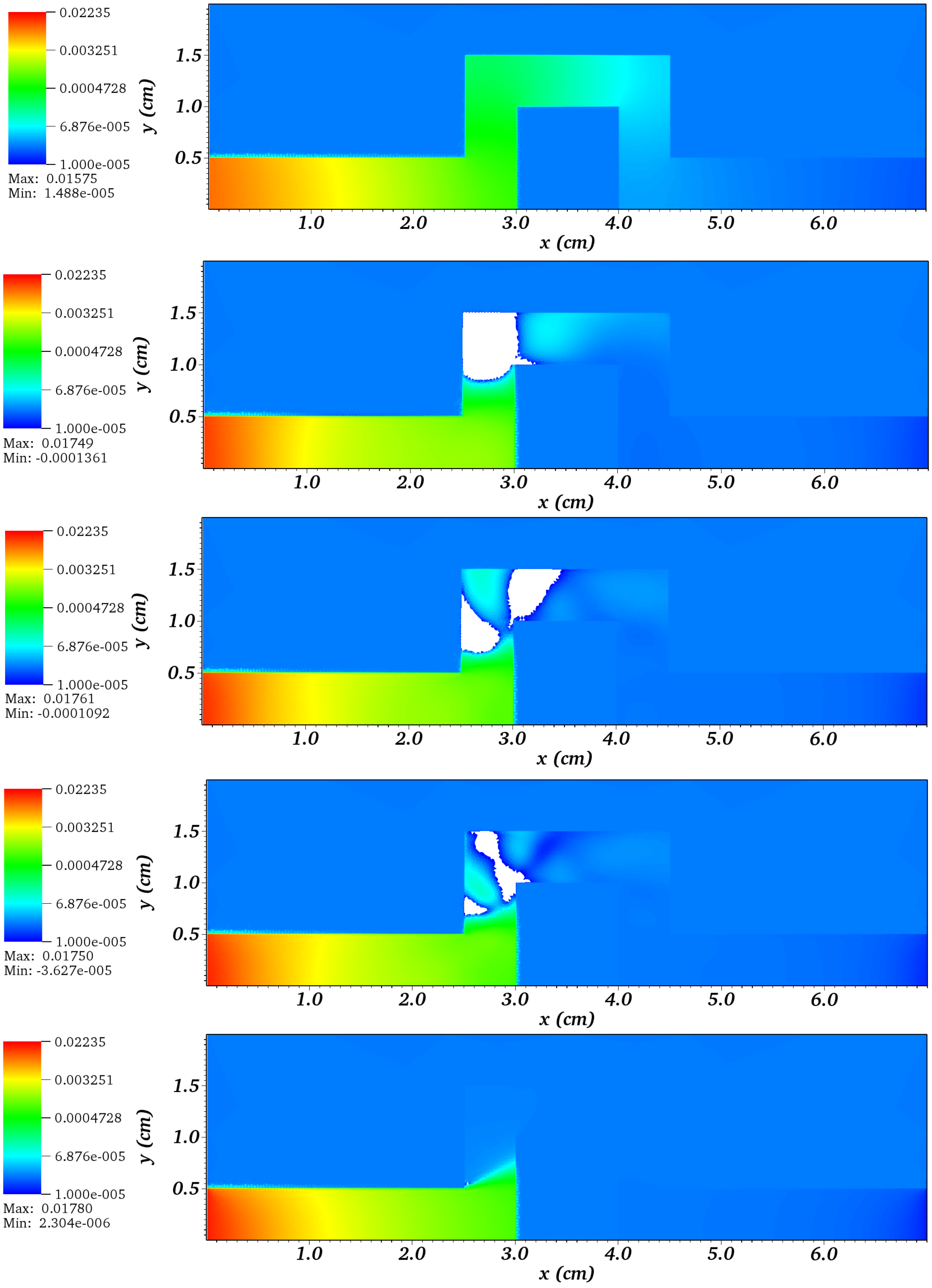


Figure 3.11: Scalar flux Φ_0^0 (in GJ/cm^2/sh) at $t = 0.05$ sh for unfiltered P$_1$, P$_3$, P$_5$, and P$_7$ calculations (from top to bottom). The last plot is a uniformly filtered P$_{39}$ calculation for reference. The white regions show where Φ_0^0 is less than 10^{-5} (i.e. essentially negative with such a log scale). Only the piecewise constant component of the solution is shown.

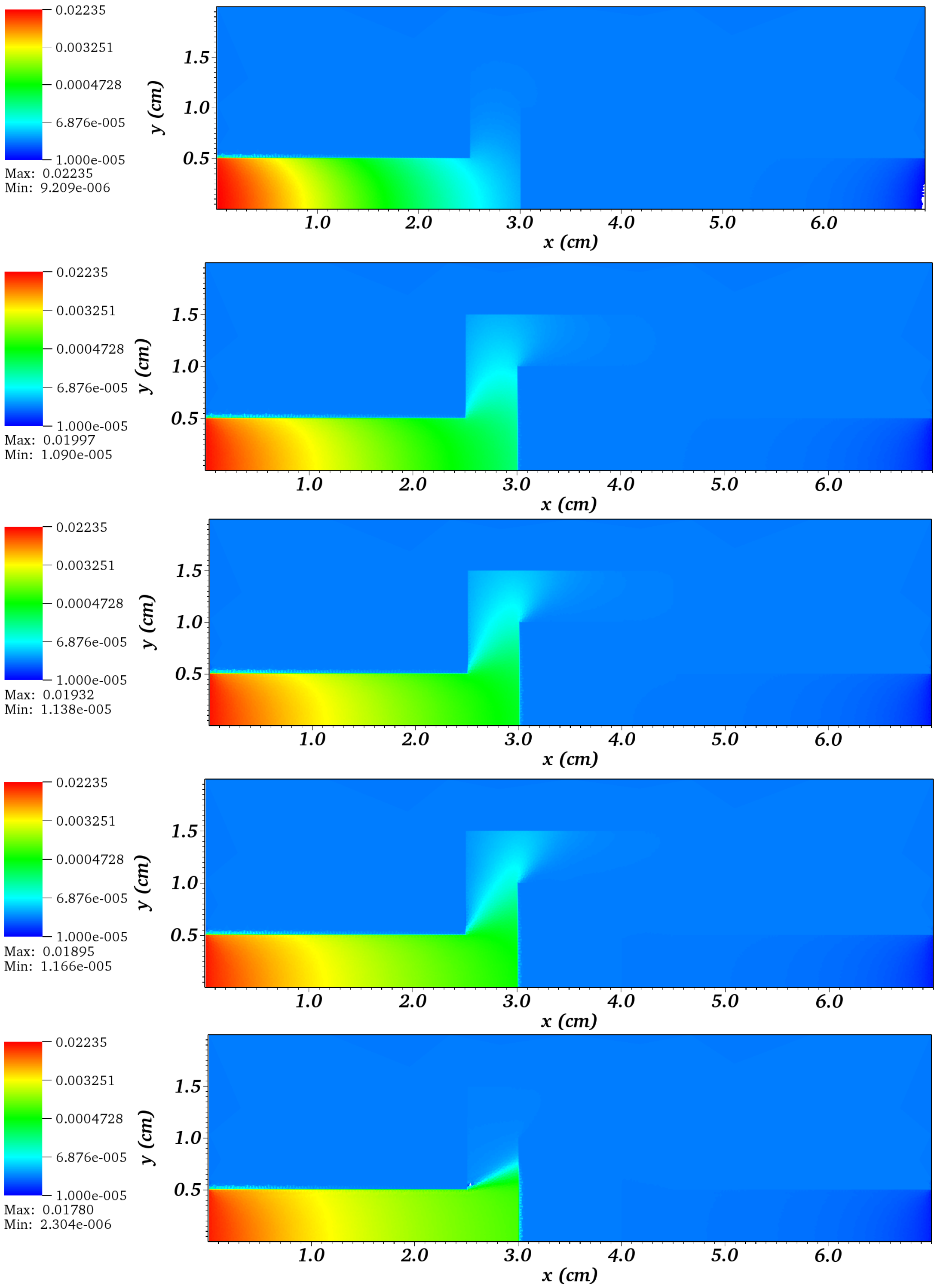


Figure 3.12: Scalar flux Φ_0^0 (in GJ/cm^2/sh) at $t = 0.05$ sh for uniformly filtered P$_1$, P$_3$, P$_5$, and P$_7$ calculations (from top to bottom). The last plot is a uniformly filtered P$_{39}$ calculation for reference. Only the piecewise constant component of the solution is shown.

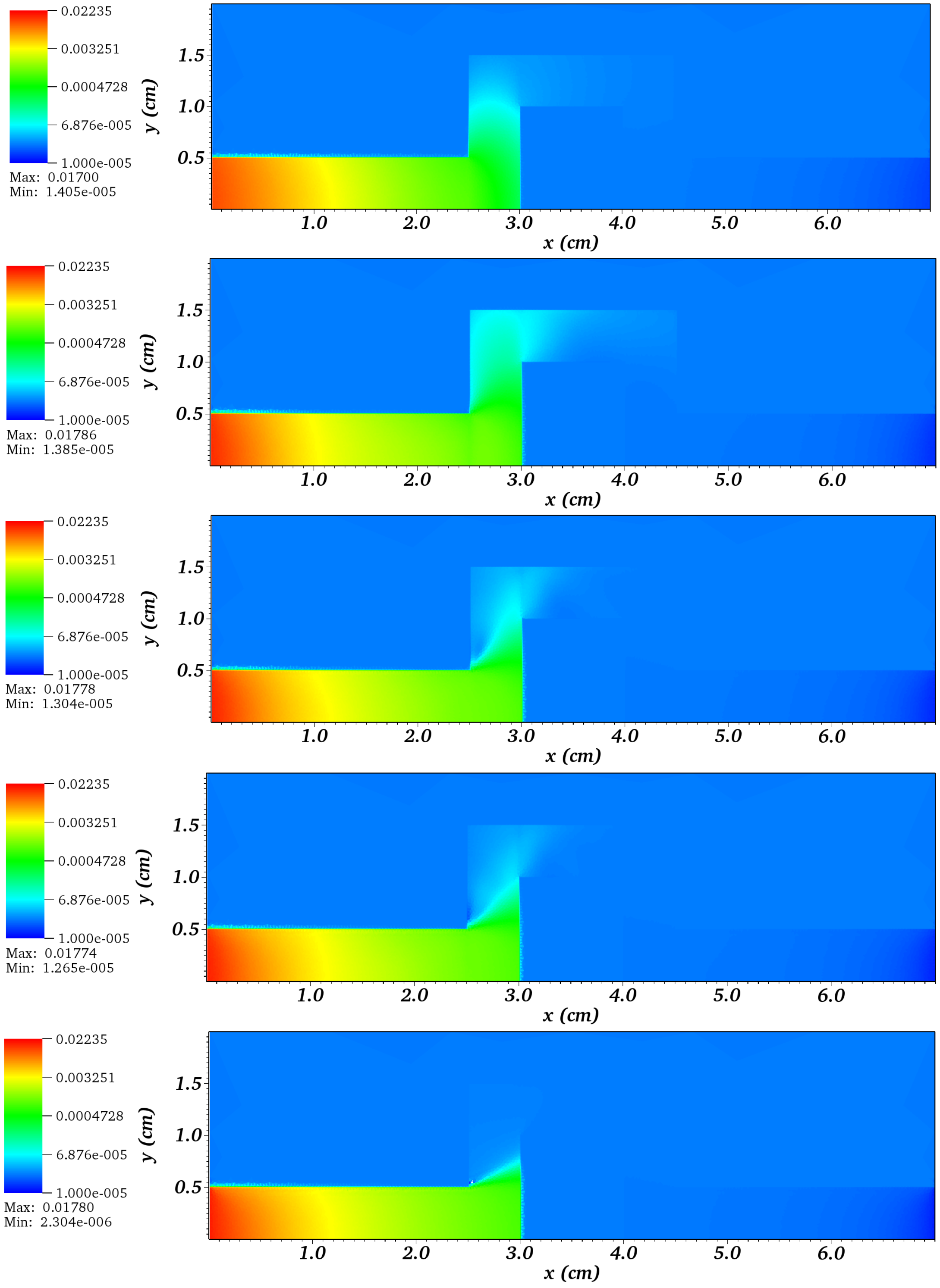
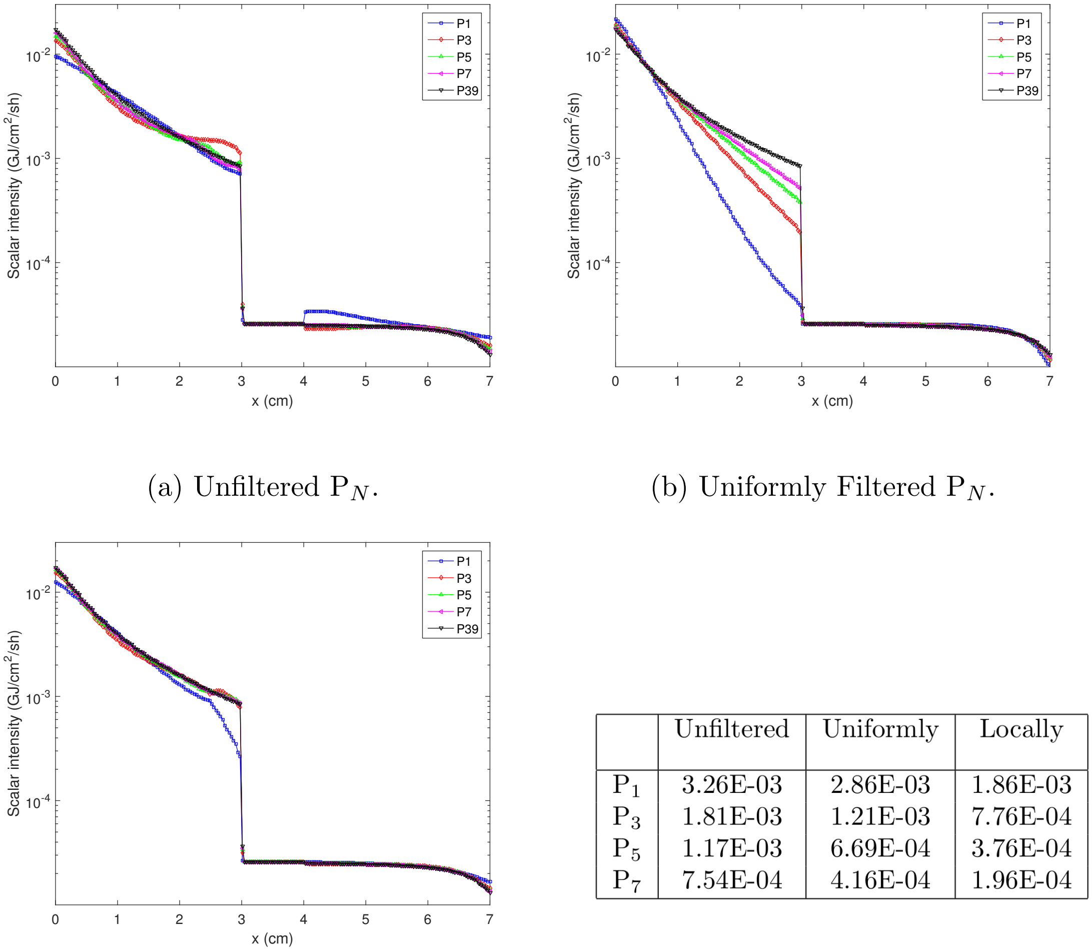


Figure 3.13: Scalar flux Φ_0^0 (in GJ/cm^2/sh) at $t = 0.05$ sh for locally filtered P$_1$, P$_3$, P$_5$, and P$_7$ calculations (from top to bottom). The last plot is a uniformly filtered P$_{39}$ calculation for reference. Only the piecewise constant component of the solution is shown.



(a) Unfiltered P_N.

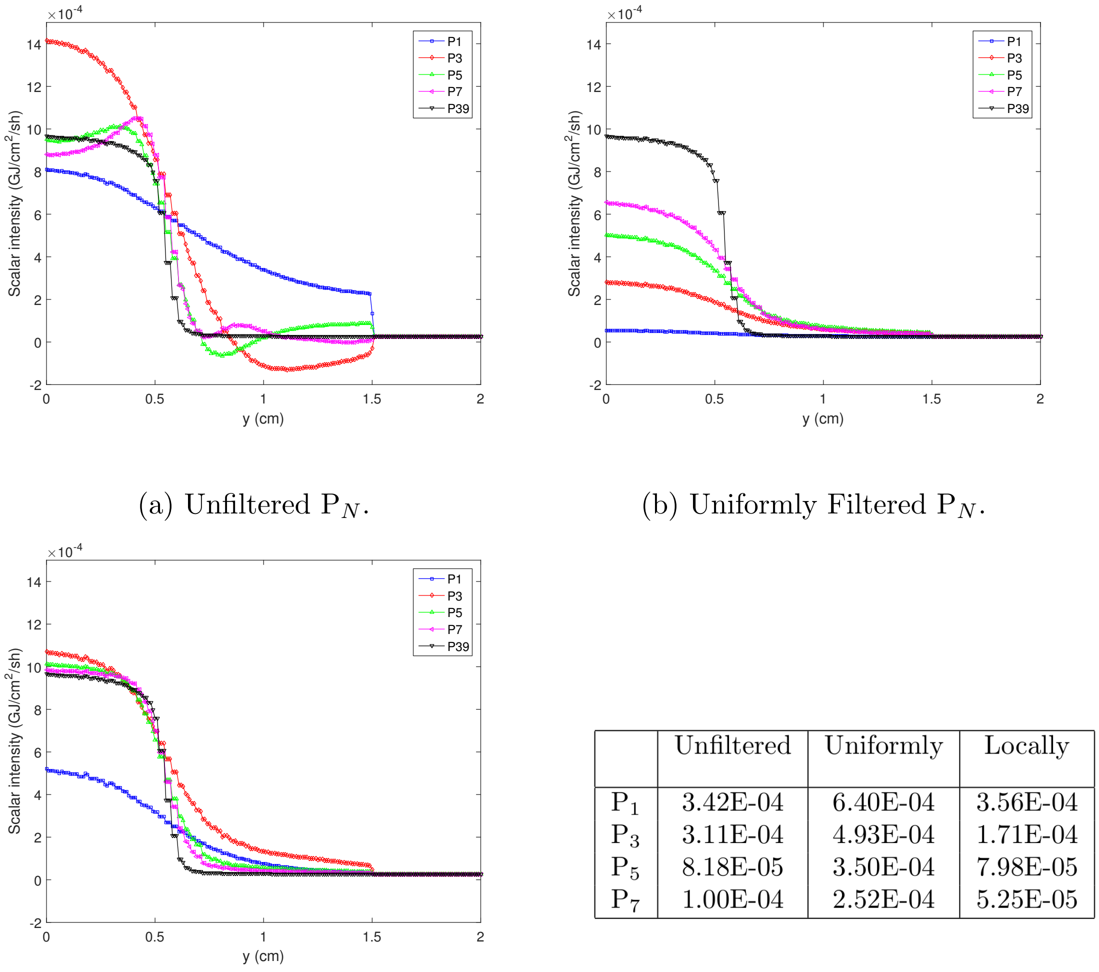
(b) Uniformly Filtered P_N.

(c) Locally Filtered P_N.

	Unfiltered	Uniformly	Locally
P_1	3.26E-03	2.86E-03	1.86E-03
P_3	1.81E-03	1.21E-03	7.76E-04
P_5	1.17E-03	6.69E-04	3.76E-04
P_7	7.54E-04	4.16E-04	1.96E-04

(d) L2-error in GJ-cm$^{-3/2}$-sh^{-1}.

Figure 3.14: Scalar flux profile along the straight line $y = 0$ at $t = 0.05$ sh (refer to Fig. 3.6 to see where the straight line is with respect to the geometry). The stair-casing is an artifact of the visualization software, which plots piece-wise constants.



(a) Unfiltered P_N.

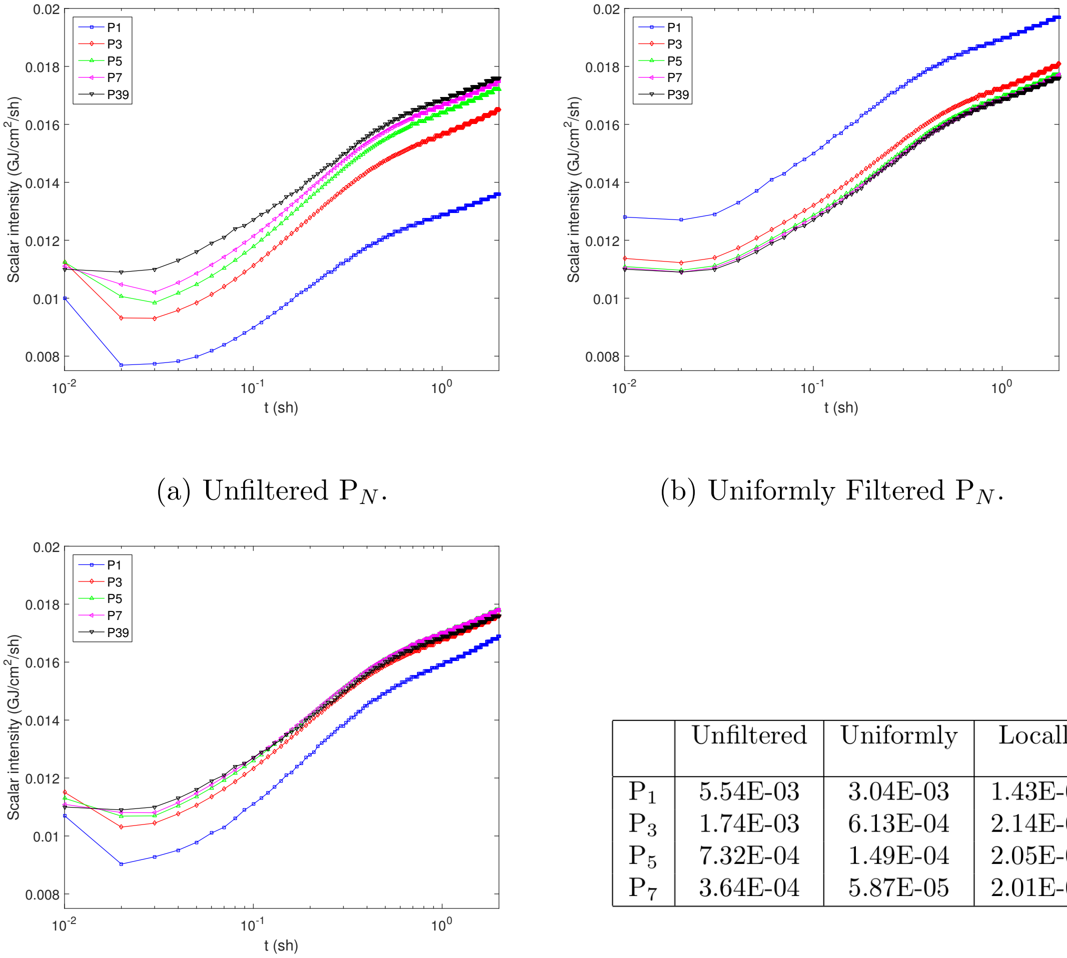
(b) Uniformly Filtered P_N.

(c) Locally Filtered P_N.

	Unfiltered	Uniformly	Locally
P_1	3.42E-04	6.40E-04	3.56E-04
P_3	3.11E-04	4.93E-04	1.71E-04
P_5	8.18E-05	3.50E-04	7.98E-05
P_7	1.00E-04	2.52E-04	5.25E-05

(d) L2-error in GJ-cm$^{-3/2}$-sh^{-1}.

Figure 3.15: Scalar flux profile along the straight line $x = 2.75$ cm at $t = 0.05$ sh (refer to Fig. 3.6 to see where the straight line is with respect to the geometry). The stair-casing is an artifact of the visualization software, which plots piecewise constants.



(a) Unfiltered P$_N$.

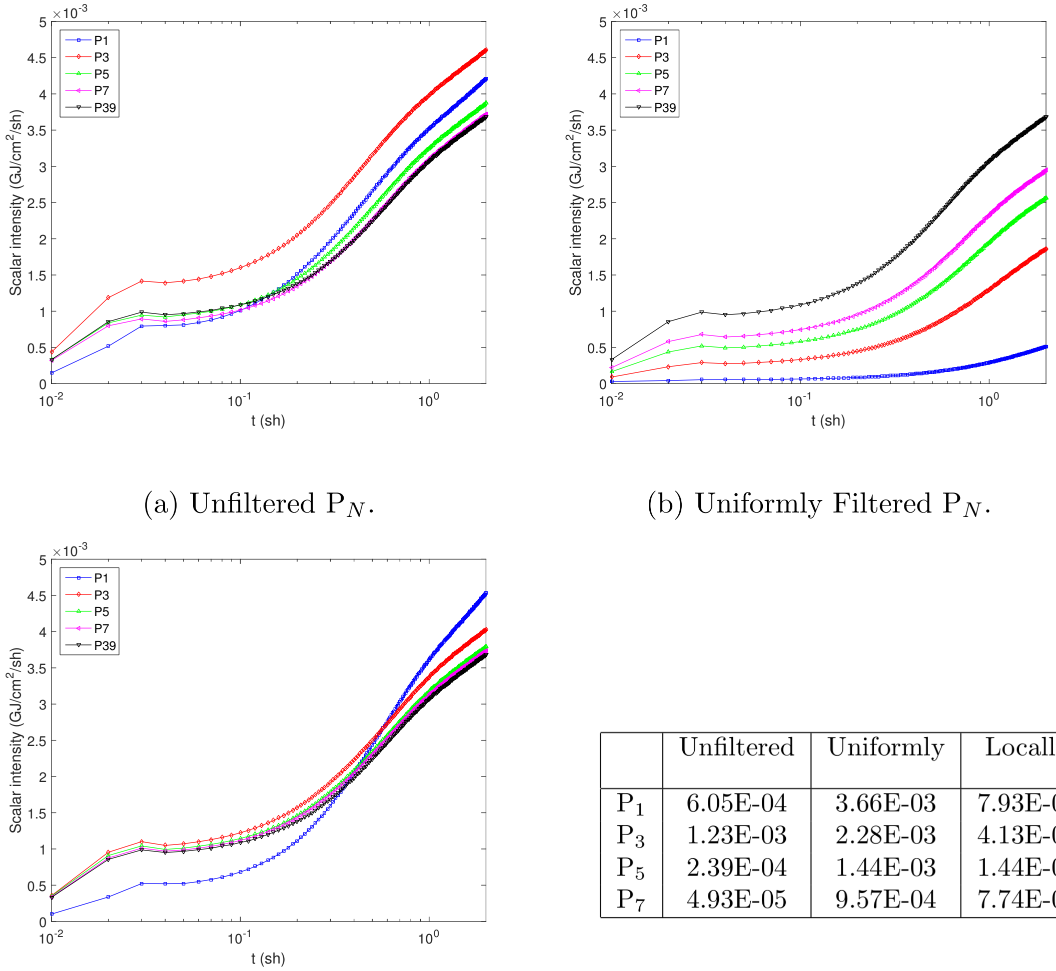
(b) Uniformly Filtered P$_N$.

(c) Locally Filtered P$_N$.

	Unfiltered	Uniformly	Locally
P$_1$	5.54E-03	3.04E-03	1.43E-03
P$_3$	1.74E-03	6.13E-04	2.14E-04
P$_5$	7.32E-04	1.49E-04	2.05E-04
P$_7$	3.64E-04	5.87E-05	2.01E-04

(d) L2-error in GJ-cm^{-2}-sh$^{-1/2}$.

Figure 3.16: Scalar flux profile at the point $(x_1, y_1) = (0.25\,\text{cm}, 0)$. Refer to Fig. 3.6 to see where this point lies with respect to the geometry.



(a) Unfiltered P_N.

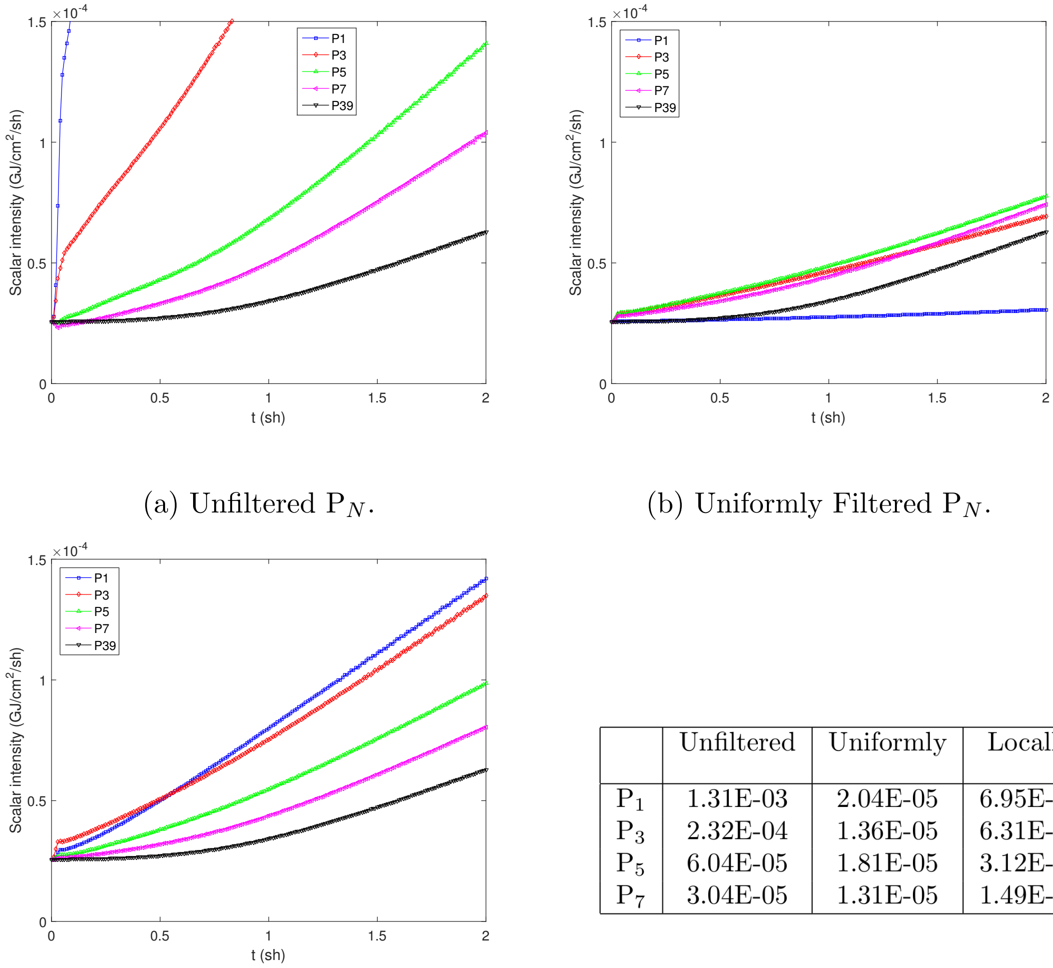
(b) Uniformly Filtered P_N.

(c) Locally Filtered P_N.

	Unfiltered	Uniformly	Locally
P_1	6.05E-04	3.66E-03	7.93E-04
P_3	1.23E-03	2.28E-03	4.13E-04
P_5	2.39E-04	1.44E-03	1.44E-04
P_7	4.93E-05	9.57E-04	7.74E-05

(d) L2-error in GJ-cm^{-2}-sh$^{-1/2}$.

Figure 3.17: Scalar flux profile at the point $(x_2, y_2) = (2.75\,\mathrm{cm}, 0)$. Refer to Fig. 3.6 to see where this point lies with respect to the geometry.



(a) Unfiltered P_N.

(b) Uniformly Filtered P_N.

(c) Locally Filtered P_N.

	Unfiltered	Uniformly	Locally
P_1	1.31E-03	2.04E-05	6.95E-05
P_3	2.32E-04	1.36E-05	6.31E-05
P_5	6.04E-05	1.81E-05	3.12E-05
P_7	3.04E-05	1.31E-05	1.49E-05

(d) L2-error in GJ-cm^{-2}-sh$^{-1/2}$.

Figure 3.18: Scalar flux profile at the point $(x_3, y_3) = (3.5\,\text{cm}, 1.25\,\text{cm})$. Refer to Fig. 3.6 to see where this point lies with respect to the geometry.

3.4.4 Energy-Dependent Filtering

3.4.4.1 Specifications

Although we have only considered one-group calculations so far, it turns out that extending the filtering approach to multigroup cases is a minor complication, the filtering strength σ_{f} being then potentially group-dependent. In this section, we illustrate it on the previous test problem with energy- and temperature-dependent cross-sections. In Appendix B, we derive – based on the model opacity given in [63] – the following two-group cross-sections:

$$\sigma_{a,\,\mathrm{thick},\,g=0}(T) = 41.00\,T^{-0.5163}, \tag{3.59}$$

$$\sigma_{a,\,\mathrm{thick},\,g=1}(T) = 0.01702\,T^{-2.564}, \tag{3.60}$$

where T is expressed in keV and σ_a in cm^{-1}. The frequency bounds for the energy groups are $(h\nu_{\min}, h\nu_1, h\nu_{\max}) \equiv (0.001\,\mathrm{keV},\ 0.3\,\mathrm{keV},\ 10\,\mathrm{keV})$. The cross-sections in the thin region are 1000 times less (see Fig. 3.6).

Besides, the initial condition is given by:

$$\Psi_g(t=0) = B_g(T_0) = \int_{\nu_{g-1}}^{\nu_g} B(\nu, T_0)\mathrm{d}\nu, \tag{3.61}$$

while the boundary condition at the left entrance of the pipe reads:

$$\Psi_g^{\mathrm{d}} = B_g(T_L) = \int_{\nu_{g-1}}^{\nu_g} B(\nu, T_L)\mathrm{d}\nu, \tag{3.62}$$

for $g \in \{0, 1\}$.

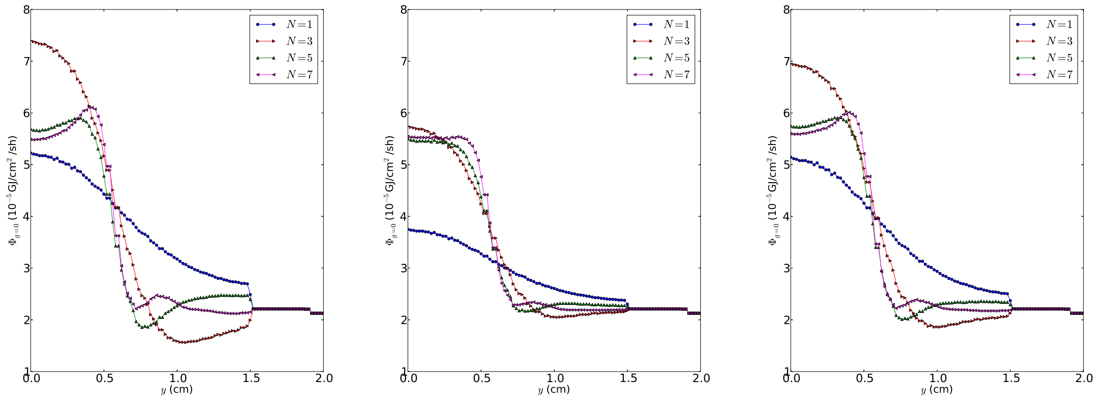
3.4.4.2 Energy-Dependent Filter Strength

Since most of the energy at early times corresponds to the fast group ($g = 1$) – see for instance Fig. B.1 in Appendix B, it can be expected that the filter strength in that group need be larger than in the thermal group ($g = 0$).

After testing several values for $\sigma_{\mathrm{f},g}$ on coarse meshes, it was found that $\sigma_{\mathrm{f},1} = 50$ cm^{-1} in the fast group gives good results along $x = 2.75$ cm,[20] as shown on Fig. 3.20: the effect of filtering for the fast scalar flux is very much similar to that of the gray case with an unfiltered approach, yielding negative values, while a uniformly filtered calculation removes the negativity but converges more slowly than a locally filtered one.

For the thermal scalar flux, the magnitude of the solution is almost 50 times less, which implies than we do not need to activate the filter as strongly. Fig. 3.19 shows the results for $\sigma_{\mathrm{f},0} = 5$ cm^{-1}. The unfiltered solution has a shape very similar to that of the fast group. However, the uniformly filtered solution seems to perform better than the locally filtered one because the filter strength is fairly small compared to the fast group. It would then be possible to increase $\sigma_{\mathrm{f},0}$ but it does not seem crucial, as the temperature profile on Fig. 3.21 tends to indicate: although the oscillations and negativity from the thermal scalar flux have not been completely removed, the locally filtered temperature is much more satisfying than the unfiltered and uniformly filtered calculations for small values of N.

[20] Because the profiles are fairly close to those from the one-group calculations, we only show results along $x = 2.75$ cm, for compactness.

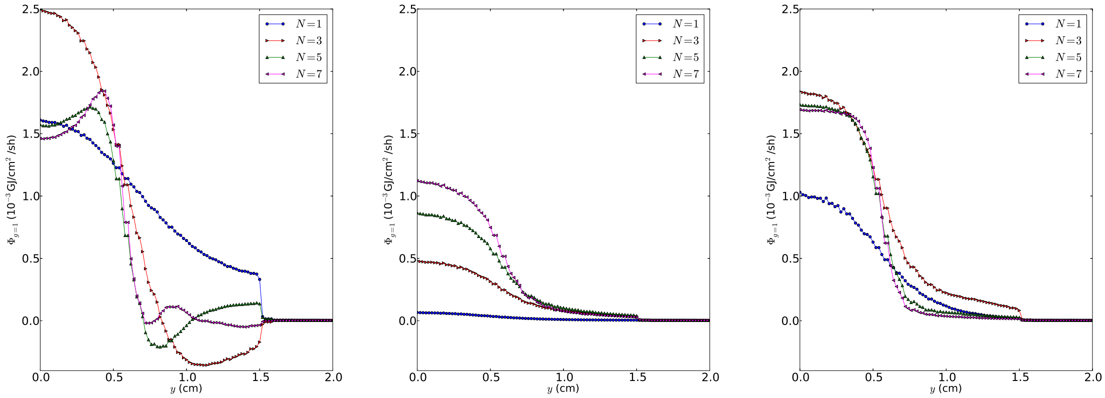


(a) Unfiltered.

(b) Uniformly Filtered.

(c) Locally Filtered.

Figure 3.19: Scalar flux for the thermal group along the straight line $x = 2.75$ cm at $t = 0.03$ sh. The filter strength when filtering is activated is chosen to be $(\sigma_{\text{f},0}, \sigma_{\text{f},1}) = (5, 50)$ cm^{-1}.

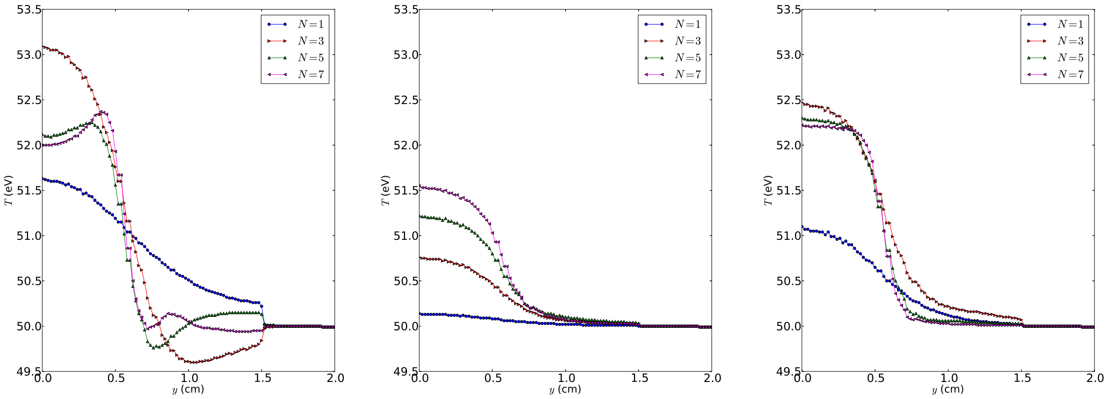


(a) Unfiltered.

(b) Uniformly Filtered.

(c) Locally Filtered.

Figure 3.20: Scalar flux for the fast group along the straight line $x = 2.75$ cm at $t = 0.03$ sh. The filter strength when filtering is activated is chosen to be $(\sigma_{\text{f},0}, \sigma_{\text{f},1}) = (5, 50)$ cm^{-1}.



(a) Unfiltered.

(b) Uniformly Filtered.

(c) Locally Filtered.

Figure 3.21: Temperature along the straight line $x = 2.75$ cm at $t = 0.03$ sh. The filter strength when filtering is activated is chosen to be $(\sigma_{\mathrm{f},0}, \sigma_{\mathrm{f},1}) = (5, 50)$ cm^{-1}.

4. CONSIDERATIONS ON SECOND-ORDER FORMS*

After studying the first-order form of the transport equation in Chapter 3, we shift our focus to second-order forms for the remainder of this thesis.*

As explained in Section 1.4, they can be preferred over first-order forms whenever the solution is smooth enough because they allow for the use of CGFEM, as opposed to the relatively more costly DGFEM. However, second-order forms are not exempt from any flaws, for they often are either incompatible with void or fail to be globally conservative. As both of these properties are highly desired[1], this chapter aims at better understanding these concepts before introducing, in Chapter 5, a new P_N method having the desired features.

The remainder of this chapter is structured as follows. In Section 4.1, we introduce and recall how to derive the main two second-order forms of interest in this work: the SAAF method, globally conservative but void incompatible, and the LS method, void compatible but not globally conservative. Conditional equivalences based on the choice of the weakly-imposed LS boundary conditions are derived in Section 4.2. Some insight is presented on the concept of particle conservation in Section 4.3, emphasizing where the lack thereof in the LS form generally comes from. Up to that point all the derivations are angular discretization agnostic.

From Section 4.4 on, the more specific case of P_N discretizations is discussed, first introducing the P_N weak formulations of the previously described methods. In

*Part of this chapter is reprinted from "Least-Squares P_N Formulation of the Transport Equation Using Self-Adjoint-Angular-Flux Consistent Boundary Conditions." by Vincent M. Laboure, Yaqi Wang and Mark D. DeHart, 2016. *PHYSOR 2016 Conference* [64]. Copyright 2016 American Nuclear Society. The author exercises his right granted by the copyright agreement to use the work for personal use (specifically the inclusion in a dissertation).

[1] This is especially crucial in the context of giving up on homogenization techniques in reactor physics applications.

Section 4.5, we focus on parity-based methods and show equivalences with the even-parity P$_N$ second-order form. We also explain why the idea of a second-order filter, a seemingly attractive extension of the work in Chapter 3 is unfruitful.

Finally, numerical results highlighting why global conservation can be so crucial are presented in Section 4.6 and conclusions are drawn in Section 4.7.

In this chapter and the following, we consider the steady-state version of Eq. 2.1 and rewrite it as [65]:

$$\vec{\Omega} \cdot \vec{\nabla}\Psi + \sigma_{\mathrm{t}}(\vec{r})\Psi(\vec{r},\vec{\Omega}) = \sum_{\ell=0}^{N_s} \sigma_{s,\ell}(\vec{r}) \sum_{m=-\ell}^{\ell} \Phi_\ell^m(\vec{r}) R_\ell^m(\vec{\Omega}) + \nu\sigma_f(\vec{r})\Phi(\vec{r}) + S(\vec{r},\vec{\Omega}), \quad (4.1)$$

where N_s designates the degree of anisotropy of the scattering source. For any function $f = f(\vec{r},\vec{\Omega})$, define the streaming plus collision operator as:

$$Lf \equiv \vec{\Omega} \cdot \vec{\nabla} f + \sigma_{\mathrm{t}}\, f, \quad (4.2)$$

and the scattering plus fission operator as:

$$Hf \equiv \int_{\mathbb{S}^2} \sigma_s(\vec{r},\vec{\Omega}' \cdot \vec{\Omega}) f(\vec{r},\vec{\Omega}')\,\mathrm{d}\Omega' + \nu\sigma_f(\vec{r}) \int_{\mathbb{S}^2} f(\vec{r},\vec{\Omega}')\,\mathrm{d}\Omega'. \quad (4.3)$$

Eq. 4.1 can then be expressed in operator form:

$$L\Psi = H\Psi + S. \quad (4.4)$$

4.1 Existing Methods

In this section, we show some well-known second-order forms and remind the reader how to derive their respective variational formulations. First we consider the

Self-Adjoint Angular Flux (SAAF) method. Second, we focus on Least-Squares (LS) formulations.

4.1.1 Self-Adjoint Angular Flux Formulations

This method has been studied in numerous publications [33, 35, 36, 37, 38] and relies on stabilizing it by transforming the first-order streaming term into a second-order term using a so-called Angular Flux Equation (AFE).

4.1.1.1 Derivation

We start by testing Eq. 4.4 with $\Psi^* \in V$:

$$(\Psi^*, L\Psi)_{\mathcal{D}} = (\Psi^*, H\Psi + S)_{\mathcal{D}} \,. \tag{4.5}$$

An integration by parts on the streaming term yields:

$$-\left(\vec{\Omega} \cdot \vec{\nabla}\Psi^*, \Psi\right)_{\mathcal{D}} + \langle \Psi^*, \Psi \rangle_{\partial\mathcal{D}} + (\Psi^*, \sigma_{\mathrm{t}}\Psi)_{\mathcal{D}} = (\Psi^*, H\Psi + S)_{\mathcal{D}} \,. \tag{4.6}$$

Going back to Eq. 4.4, we can also express Ψ as:

$$\Psi = \frac{1}{\sigma_{\mathrm{t}}}\left(H\Psi + S - \vec{\Omega} \cdot \vec{\nabla}\Psi\right), \tag{4.7}$$

which constitutes an AFE. The choice of the AFE is not unique – as we could for instance choose $\Psi = (\sigma_{\mathrm{t}} - H)^{-1}\left(S - \vec{\Omega} \cdot \vec{\nabla}\Psi\right)$ – but this former choice is recommended in [38], in particular because no difficulty arises in purely scattering regions.

A substitution for Ψ from Eq. 4.7 in the streaming term of Eq. 4.6 yields the final weak formulation: find $\Psi \in V$ such that for all $\Psi^* \in V$,

$$\left(\vec{\Omega} \cdot \vec{\nabla}\Psi^*, \frac{1}{\sigma_\mathrm{t}}\vec{\Omega} \cdot \vec{\nabla}\Psi\right)_\mathcal{D} + \langle \Psi^*, \Psi \rangle_{\partial\mathcal{D}} + (\Psi^*, \sigma_\mathrm{t}\Psi)_\mathcal{D} = \left(\frac{1}{\sigma_\mathrm{t}}\vec{\Omega} \cdot \vec{\nabla}\Psi^* + \Psi^*, H\Psi + S\right)_\mathcal{D}. \tag{4.8}$$

The boundary conditions are imposed through the surface term on $\partial\mathcal{D}$:

$$\langle \Psi^*, \Psi \rangle_{\partial\mathcal{D}} = \langle \Psi^*, \Psi \rangle^+_{\partial\mathcal{D}} - \langle \Psi^*, \Psi^\mathrm{inc} \rangle^-_{\partial\mathcal{D}}. \tag{4.9}$$

This formulation is clearly not compatible with void, because of the σ_t^{-1} term. Furthermore, the bilinear form may be SPD if H and S have no anisotropic contribution and if no reflecting boundary condition is used.

4.1.1.2 Void Treatment for S_N

This previous formulation can be modified so as to make it compatible with void, as shown in [38]. For convenience, we recall the corresponding variational formulation and henceforth refer to it as the Self-Adjoint Angular Flux with a Void Treatment (SAAF–VT) formulation: find $\Psi \in V$ such that for all $\Psi^* \in V$,

$$\begin{aligned}\left(\vec{\Omega} \cdot \vec{\nabla}\Psi^\star, \tau\,\vec{\Omega} \cdot \vec{\nabla}\Psi\right)_\mathcal{D} - \left(\vec{\Omega} \cdot \vec{\nabla}\Psi^\star, (1 - \tau\sigma_\mathrm{t})\Psi\right)_\mathcal{D} + (\sigma_\mathrm{t}\Psi^\star, \Psi)_\mathcal{D} \\ + \langle \Psi^\star, \Psi \rangle^+_{\partial\mathcal{D}} - \langle \Psi^\star, \Psi^\mathrm{inc} \rangle^-_{\partial\mathcal{D}} = \left(\tau\,\vec{\Omega} \cdot \vec{\nabla}\Psi^\star + \Psi^\star, H\Psi + S\right)_\mathcal{D},\end{aligned} \tag{4.10}$$

with τ being defined as:

$$\tau \equiv \min\left(\frac{1}{\sigma_\mathrm{t}}, \frac{h}{\varsigma}\right), \tag{4.11}$$

where h characterizes the mesh size and ς is a constant, typically chosen to be 2.

While nothing in this formulation *a priori* suggests that it only works with S_N discretizations, numerical experiments with a P_N approximation show that the solver does not converge in that latter case. If we take a closer look at Eq. 4.10, it appears that in void, as $h \to 0$, the weak formulation tends to become:

$$-\left(\vec{\Omega} \cdot \vec{\nabla}\Psi^\star, \Psi\right)_{\mathcal{D}} + \langle \Psi^\star, \Psi \rangle^+_{\partial\mathcal{D}} - \langle \Psi^\star, \Psi^{\text{inc}} \rangle^-_{\partial\mathcal{D}} = 0, \tag{4.12}$$

which is precisely the first-order form in void. Yet, as we saw in Section 3.3.2, the first-order P_N formulation is ill-conditioned in void.[2] This fundamentally lies in the singularity of the $\vec{\mathbf{D}}$ matrices. It is therefore not surprising to see the SAAF–VT method have conditioning problems with P_N. Numerical results highlighting this behavior will be shown in Section 5.3.5.1 (in particular, see Table 5.1).

4.1.2 Least-Squares Formulation

Several LS formulations have been developed, but we consider here the LS transport equation compatible with voids from [44]. Besides being able to handle zero cross-sections, some of its advantages include the preservation of the intermediate and thick diffusion limit as well as its compatibility with source iteration algorithms.

One of its most concerning flaws – like most LS methods – is the loss of global conservation, as we shall study in detail in Section 5.2

4.1.2.1 Derivation

We now present one way of deriving this LS formulation, which can be obtained using the adjoint operator of L in infinite medium. As a reminder, $L^\dagger$ is the adjoint operator[3] corresponding to L if and only if, for all f, g,

$$(f, Lg)_{\mathcal{D}} = \left(L^\dagger f, g\right)_{\mathcal{D}}. \tag{4.13}$$

[2] Eq. 3.47 gives a lower bound to the condition number, which is therefore infinite in steady-state ($\Delta t \to \infty$).

[3] The definition of the adjoint operator presupposes to have defined a scalar product. Here, we use $(\cdot, \cdot)_{\mathcal{D}}$ as our scalar product.

It can be shown – through an integration by parts – that, in an infinite medium,

$$L^{\dagger} = -\vec{\Omega}\cdot\vec{\nabla} + \sigma_{\mathrm{t}}. \tag{4.14}$$

Applying $L^{\dagger}$ to Eq. 4.1:

$$L^{\dagger}L\Psi = L^{\dagger}\left(H\Psi + S\right), \tag{4.15}$$

it reads, after integrating against a test function Ψ^*:

$$\left(\Psi^*, L^{\dagger}L\Psi\right)_{\mathcal{D}} = \left(\Psi^*, L^{\dagger}\left(H\Psi + S\right)\right)_{\mathcal{D}}. \tag{4.16}$$

Using the definitions of L and $L^{\dagger}$, it becomes

$$\begin{aligned} &-\left(\Psi^*,\vec{\Omega}\cdot\vec{\nabla}\,\vec{\Omega}\cdot\vec{\nabla}\Psi\right)_{\mathcal{D}} - \left(\Psi^*,\vec{\Omega}\cdot\vec{\nabla}\left(\sigma_{\mathrm{t}}\Psi\right)\right)_{\mathcal{D}} + \left(\sigma_{\mathrm{t}}\Psi^*,\vec{\Omega}\cdot\vec{\nabla}\Psi\right)_{\mathcal{D}} + \left(\sigma_{\mathrm{t}}\Psi^*,\sigma_{\mathrm{t}}\Psi\right)_{\mathcal{D}}\\ &= -\left(\Psi^*,\vec{\Omega}\cdot\vec{\nabla}\left(H\Psi + S\right)\right)_{\mathcal{D}} + \left(\sigma_{\mathrm{t}}\Psi^*, H\Psi + S\right)_{\mathcal{D}}, \end{aligned} \tag{4.17}$$

and integrations by parts on the first two terms of the left-hand side and the first term of the right-hand side, we get:

$$\begin{aligned} &\left(\vec{\Omega}\cdot\vec{\nabla}\Psi^*,\vec{\Omega}\cdot\vec{\nabla}\Psi\right)_{\mathcal{D}} - \langle\Psi^{\star},\vec{\Omega}\cdot\vec{\nabla}\Psi\rangle_{\partial\mathcal{D}} + \left(\vec{\Omega}\cdot\vec{\nabla}\Psi^*,\sigma_{\mathrm{t}}\Psi\right)_{\mathcal{D}} - \langle\Psi^{\star},\sigma_{\mathrm{t}}\Psi\rangle_{\partial\mathcal{D}}\\ &+\left(\sigma_{\mathrm{t}}\Psi^*,\vec{\Omega}\cdot\vec{\nabla}\Psi\right)_{\mathcal{D}} + \left(\sigma_{\mathrm{t}}\Psi^*,\sigma_{\mathrm{t}}\Psi\right)_{\mathcal{D}}\\ &\quad= \left(\vec{\Omega}\cdot\vec{\nabla}\Psi^*, H\Psi + S\right)_{\mathcal{D}} - \langle\Psi^{\star}, H\Psi + S\rangle_{\partial\mathcal{D}} + \left(\sigma_{\mathrm{t}}\Psi^*, H\Psi + S\right)_{\mathcal{D}}. \end{aligned} \tag{4.18}$$

Grouping the terms yields

$$\begin{aligned} &\left(L\Psi^*,\vec{\Omega}\cdot\vec{\nabla}\Psi\right)_{\mathcal{D}} + \left(L\Psi^*,\sigma_{\mathrm{t}}\Psi\right)_{\mathcal{D}} - \langle\Psi^{\star},\vec{\Omega}\cdot\vec{\nabla}\Psi\rangle_{\partial\mathcal{D}} - \langle\Psi^{\star},\sigma_{\mathrm{t}}\Psi\rangle_{\partial\mathcal{D}}\\ &\quad= \left(L\Psi^*, H\Psi + S\right)_{\mathcal{D}} - \langle\Psi^{\star}, H\Psi + S\rangle_{\partial\mathcal{D}}, \end{aligned} \tag{4.19}$$

i.e.:

$$\begin{aligned}&(L\Psi^*, L\Psi)_{\mathcal{D}} - \langle \Psi^\star, \vec{\Omega} \cdot \vec{\nabla}\Psi + \sigma_{\mathrm{t}}\Psi - H\Psi + S \rangle_{\partial\mathcal{D}} \\ &\quad = (L\Psi^*, H\Psi + S)_{\mathcal{D}} \,.\end{aligned} \tag{4.20}$$

Enforcing the transport equation to hold on $\partial\mathcal{D}$, we end up with the following weak formulation: find $\Psi \in V$ such that for all $\Psi^* \in V$,

$$(L\Psi^*, L\Psi)_{\mathcal{D}} = (L\Psi^*, H\Psi + S)_{\mathcal{D}} \,. \tag{4.21}$$

Several remarks can be made at this point:

- Eq. 4.21 could be obtained directly from Eq. 4.16 using the mathematical definition of the adjoint in an infinite medium. In general however (i.e. if not in an infinite medium), we can no longer rigorously define $L^\dagger$ as the adjoint operator corresponding to L – because boundary terms would be needed. Rather, we directly use Eq. 4.14 as a definition.

- It can also be shown that the solution of the variational formulation (4.21) satisfies [66]:

  $$\Psi = \arg\min_{f \in V} \int_{\mathcal{D}} \int_{\mathbb{S}^2} \left| Lf - Hf - S \right|^2 \mathrm{d}\Omega \mathrm{d}r, \tag{4.22}$$

  i.e. minimizes the transport equation residual, in an L^2 sense. This in particular explains the name of such methods, since this constitutes a Least-Squares minimization problem.

4.1.2.2 Boundary Conditions

The boundary conditions can be imposed weakly by adding an additional term, yielding the final variational formulation [44, 67]: find $\Psi \in V$ such that for all

$\Psi^* \in V$,

$$\left(L\Psi^*, L\Psi\right)_{\mathcal{D}} + \langle c\Psi^*, (\Psi - \Psi^{\text{inc}})\rangle^-_{\partial\mathcal{D}} = \left(L\Psi^*, H\psi + S\right)_{\mathcal{D}}, \tag{4.23}$$

where c is a parameter, with units of a macroscopic cross-section. The choice of the value for c is an important question that is the focus of the following section.

4.2 Conditional Equivalences

This section aims at studying the relationships between the various aforementioned second-order forms, based in particular on the value of c. We say that, for a given problem, two variational formulations are *equivalent* if they are identical. We say that they are *consistent* if their difference only lies in the discretization error, or in other words, if the linear system to solve is identical upon angular and spatial convergence.

4.2.1 SAAF/LS Conditional Equivalence

Let us first point out the following property, deriving from the divergence theorem:

$$\begin{aligned}\left(\Psi^*, \vec{\Omega}\cdot\vec{\nabla}\Psi\right)_{\mathcal{D}} + \left(\vec{\Omega}\cdot\vec{\nabla}\Psi^*, \Psi\right)_{\mathcal{D}} &= \int_{\mathcal{D}}\int_{\mathbb{S}^2} \vec{\Omega}\cdot\vec{\nabla}(\Psi^*\Psi)\,\mathrm{d}\Omega\mathrm{d}\vec{r},\\ &= \int_{\partial\mathcal{D}}\int_{\mathbb{S}^2} \Psi^*\Psi\,\vec{\Omega}\cdot\vec{n}\,\mathrm{d}\Omega\mathrm{d}\vec{r},\\ &= \langle\Psi^*, \Psi\rangle^+_{\partial\mathcal{D}} - \langle\Psi^*, \Psi\rangle^-_{\partial\mathcal{D}}.\end{aligned} \tag{4.24}$$

Thus, assuming that σ_t is constant and non-zero across $\mathcal{D}$, the LS formulation divided by σ_t gives:

$$\begin{aligned}\left(\vec{\Omega}\cdot\vec{\nabla}\Psi^*, \frac{1}{\sigma_\mathrm{t}}\vec{\Omega}\cdot\vec{\nabla}\Psi\right)_\mathcal{D} + \langle\Psi^*,\Psi\rangle^+_{\partial\mathcal{D}} - \langle\Psi^*,\Psi\rangle^-_{\partial\mathcal{D}} + (\Psi^*,\sigma_\mathrm{t}\Psi)_\mathcal{D} \\ + \langle\frac{c}{\sigma_\mathrm{t}}\Psi^*,(\Psi-\Psi^\mathrm{inc})\rangle^-_{\partial\mathcal{D}} = \left(\frac{1}{\sigma_\mathrm{t}}\vec{\Omega}\cdot\vec{\nabla}\Psi^* + \Psi^*, H\Psi + S\right)_\mathcal{D},\end{aligned} \tag{4.25}$$

that is:

$$\begin{aligned}\left(\vec{\Omega}\cdot\vec{\nabla}\Psi^*, \frac{1}{\sigma_\mathrm{t}}\vec{\Omega}\cdot\vec{\nabla}\Psi\right)_\mathcal{D} + \langle\Psi^*,\Psi\rangle^+_{\partial\mathcal{D}} - \langle(1-\frac{c}{\sigma_\mathrm{t}})\Psi^*,\Psi\rangle^-_{\partial\mathcal{D}} - \langle\frac{c}{\sigma_\mathrm{t}}\Psi^*,\Psi^\mathrm{inc}\rangle^-_{\partial\mathcal{D}} \\ + (\Psi^*,\sigma_\mathrm{t}\Psi)_\mathcal{D} = \left(\frac{1}{\sigma_\mathrm{t}}\vec{\Omega}\cdot\vec{\nabla}\Psi^* + \Psi^*, H\Psi + S\right)_\mathcal{D}.\end{aligned} \tag{4.26}$$

Therefore, in the case of a constant and strictly positive σ_t, the SAAF and LS formulations, respectively given by Eqs. 4.8 and 4.23 are equivalent if and only if:

$$c = \sigma_\mathrm{t}. \tag{4.27}$$

Nevertheless, it is clear that this choice cannot always be pertinent because, for problems surrounded by void ($\sigma_\mathrm{t} = 0$ on $\partial\mathcal{D}$), the boundary terms would vanish.

4.2.2 SAAF–VT/LS Conditional Consistency

In this subsection, we assume that τ is constant (see Eq. 4.11) and we show that there is only one value for c such that Eqs. 4.23 and 4.8 are consistent. We first multiply Eq. 4.23 with τ and use the divergence theorem to get:

$$\begin{aligned}\left(\vec{\Omega}\cdot\vec{\nabla}\Psi^\star, \tau\vec{\Omega}\cdot\vec{\nabla}\Psi\right)_{\mathcal{D}} + \langle\tau\sigma_{\mathrm{t}}\Psi^\star, \Psi\rangle^+_{\partial\mathcal{D}} - \langle\tau\sigma_{\mathrm{t}}\Psi^\star, \Psi\rangle^-_{\partial\mathcal{D}} + (\tau\sigma_{\mathrm{t}}\Psi^\star, \sigma_{\mathrm{t}}\Psi)_{\mathcal{D}} \\ + c\,\langle\Psi^\star, \tau(\Psi - \Psi^{\mathrm{inc}})\rangle^-_{\partial\mathcal{D}} = \left(\tau\vec{\Omega}\cdot\vec{\nabla}\Psi^\star + \tau\sigma_{\mathrm{t}}\Psi^\star, H\Psi + S\right)_{\mathcal{D}}.\end{aligned} \tag{4.28}$$

Subtracting (4.8) from (4.28) and looking for the value of c such that this equation is satisfied for all $\Psi^\star \in V$, it yields:

$$\begin{aligned}(1-\tau\sigma_{\mathrm{t}})\left[-\langle\Psi^\star, \Psi\rangle^-_{\partial\mathcal{D}} - \left(\Psi^\star, \vec{\Omega}\cdot\vec{\nabla}\Psi\right)_{\mathcal{D}}\right] + (\tau\sigma_{\mathrm{t}} - 1)\,(\Psi^\star, \sigma_{\mathrm{t}}\Psi)_{\mathcal{D}} \\ + (c\tau - \tau\sigma_{\mathrm{t}})\,\langle\Psi^\star, \Psi\rangle^-_{\partial\mathcal{D}} + (1-c\tau)\langle\Psi^\star, \Psi^{\mathrm{inc}}\rangle^-_{\partial\mathcal{D}} = (\tau\sigma_{\mathrm{t}} - 1)\,(\Psi^\star, H\Psi + S)_{\mathcal{D}}.\end{aligned} \tag{4.29}$$

Besides, since Ψ is the solution of the transport equation in a weak sense we have, neglecting the discretization error[4]:

$$\left(\Psi^\star, \vec{\Omega}\cdot\vec{\nabla}\Psi + \sigma_{\mathrm{t}}\Psi\right)_{\mathcal{D}} \approx (\Psi^\star, H\Psi + S)_{\mathcal{D}}. \tag{4.30}$$

The previous terms then reduce to:

$$(c\tau - 1)\,\langle\Psi^\star, (\Psi - \Psi^{\mathrm{inc}})\rangle^-_{\partial\mathcal{D}} = 0. \tag{4.31}$$

Therefore, under the assumption that τ is constant and that the discretization error is negligible, the LS and SAAF formulations are equivalent if and only if:

$$c = \frac{1}{\tau} = \max\left(\sigma_{\mathrm{t}}, \frac{\varsigma}{h}\right). \tag{4.32}$$

[4] Note however that – as discussed in Section 5.2 – the discretization error may actually not be negligible at all. The point here is to find a boundary term making LS potentially equivalent to SAAF–VT in the limiting case.

This value for c presents the advantage of not vanishing even when $\sigma_\mathrm{t} = 0$ on $\partial\mathcal{D}$ and can therefore be used for a wide variety of problems.

4.3 Conservation

From the variational formulation of a method, a lot can be said regarding its conservation properties. In this section, we explain how the SAAF and SAAF–VT can be shown to be globally conservative whereas LS is not. While these properties are already known [44], the reasoning is worth being detailed as it will be useful in Chapter 5. We distinguish two types of conservation: global and local. The former means that the number of particles added to the domain $\mathcal{D}$ is rigorously equal to the number of particles removed. The latter is more restrictive as it requires this same balance equation to be satisfied locally on each cell.

4.3.1 Global Conservation

The beauty of variational formulations is that the numerical solution Ψ satisfies the equation for any test function $\Psi^* \in V$. In particular, choosing $\Psi^* = 1 \in V$ and plugging it into Eq. 4.8 gives:

$$\underbrace{\big(1, \sigma_\mathrm{t}\Psi\big)_\mathcal{D}}_{\text{Collision rate}} + \underbrace{\big\langle 1, \Psi\big\rangle^+_{\partial\mathcal{D}} - \big\langle 1, \Psi^\mathrm{inc}\big\rangle^-_{\partial\mathcal{D}}}_{\text{Net leakage rate}} = \underbrace{\big(1, H\Psi + S\big)_\mathcal{D}}_{\text{Production rate}}, \tag{4.33}$$

which is precisely a conservation statement over $\mathcal{D}$hat will be satisfied by the numerical solution. This proves that the SAAF formulation is globally conservative.

It turns out that the SAAF–VT variational formulation, given by Eq. 4.10 reduces to the exact same equation for $\Psi^* = 1$, which implies the same global conservation property.

Yet, Eq. 4.23 can also be written (this time, without assuming that σ_t is constant):

$$\begin{aligned}&\left(\vec{\Omega}\cdot\vec{\nabla}\Psi^*,\vec{\Omega}\cdot\vec{\nabla}\Psi\right)_{\mathcal{D}}+\langle\Psi^*,\sigma_\mathrm{t}\Psi\rangle^+_{\partial\mathcal{D}}-\langle(\sigma_\mathrm{t}-c)\Psi^*,\Psi\rangle^-_{\partial\mathcal{D}}-\langle c\Psi^*,\Psi^{\mathrm{inc}}\rangle^-_{\partial\mathcal{D}}\\&\quad-\left(\vec{\Omega}\cdot\vec{\nabla}(\sigma_\mathrm{t})\Psi^*,\Psi\right)_{\mathcal{D}}+(\sigma_\mathrm{t}\Psi^*,\sigma_\mathrm{t}\Psi)_{\mathcal{D}}=\left(\vec{\Omega}\cdot\vec{\nabla}\Psi^*+\sigma_\mathrm{t}\Psi^*,H\Psi+S\right)_{\mathcal{D}},\end{aligned}\tag{4.34}$$

which gives, choosing $\Psi^*=1$:

$$\begin{aligned}&(\sigma_\mathrm{t},\sigma_\mathrm{t}\Psi)_{\mathcal{D}}-\left(\vec{\Omega}\cdot\vec{\nabla}(\sigma_\mathrm{t}),\Psi\right)_{\mathcal{D}}\\&\qquad+\langle\sigma_\mathrm{t},\Psi\rangle^+_{\partial\mathcal{D}}-\langle\sigma_\mathrm{t}-c,\Psi\rangle^-_{\partial\mathcal{D}}-\langle c,\Psi^{\mathrm{inc}}\rangle^-_{\partial\mathcal{D}}=(\sigma_\mathrm{t},H\Psi+S)_{\mathcal{D}},\end{aligned}\tag{4.35}$$

or, equivalently:

$$\begin{aligned}&(\sigma_\mathrm{t},\sigma_\mathrm{t}\Psi)_{\mathcal{D}}-\left(\vec{\Omega}\cdot\vec{\nabla}c,\Psi\right)_{\mathcal{D}}+\left(\sigma_\mathrm{t}-c,\vec{\Omega}\cdot\vec{\nabla}\Psi\right)_{\mathcal{D}}\\&\qquad+\langle c,\Psi\rangle^+_{\partial\mathcal{D}}-\langle c,\Psi^{\mathrm{inc}}\rangle^-_{\partial\mathcal{D}}=(\sigma_\mathrm{t},H\Psi+S)_{\mathcal{D}}.\end{aligned}\tag{4.36}$$

These are not conservation statements because most terms are weighted either with σ_t or c unless we have $c=\sigma_\mathrm{t}$ and σ_t constant. Of course, this does not rigorously prove that LS is not conservative since there could be another $\Psi^*\in V$ such that Eq. 4.23 reduces to Eq. 4.33. One can show that there is no such Ψ^* though. Besides, some valuable insight can be gained from this equation:

- Eq. 4.35 actually would be a conservation statement under three conditions: (i) σ_t is constant; (ii) $c=\sigma_\mathrm{t}$ and (iii) $\sigma_\mathrm{t}\neq 0$, which are precisely the conditions under which SAAF and LS are equivalent (see Section 4.2.1).

- Although we saw in Section 4.2.2 that SAAF–VT and LS are consistent if: (i) τ is constant and (ii) $c=\tau$ and we just showed that SAAF–VT is globally conservative, it does not mean that LS has that same property if these two conditions are met. This is because the discretization error does not make these two methods equivalent.

We will actually show in Section 5.2 what term is missing for LS to be globally conservative in a uniform region.

4.3.2 Corollary

So far we have established that LS is globally conservative if and only if: (i) σ_{t} is constant; (ii) $c = \sigma_{\mathrm{t}}$ and (iii) $\sigma_{\mathrm{t}} \neq 0$. This in particular implies that if there is void anywhere in the domain, LS cannot be globally conservative.[5]

This important corollary will be the motivation behind the CLS method introduced in Section 5.2.

4.3.3 Local Conservation

Admittedly, it is not possible with CGFEM to find a test function $\Psi^* \in V$ constant across any given cell and zero elsewhere – which would otherwise yield a conservation statement over each individual cells. This is because any linear combination of continuous functions is also continuous. Nevertheless, although a common belief is that – unlike DGFEM – CGFEM must therefore lack local conservation, it was shown in [68] that element nodal fluxes are actually conserved within this class of methods. The consequence is then that CGFEM are locally conservative in that sense.

4.4 P$_N$ Expansion

The purpose of the present section is to introduce some notation specific to P$_N$ discretizations for second-order and/or parity-based formulations. It thus sets the stage for more complicated and tedious derivations in Section 4.5.

[5] Indeed, either the domain is pure void and we cannot choose $c = \sigma_{\mathrm{t}}$ or only part of the domain is void and σ_{t} is not constant.

Up to now, the considerations detailed in this chapter did not require any particular angular discretization. From now on, we use the P_N approximation to expand the angular-dependent variables, just like we did in Chapter 3. Most of the notation relative to spherical harmonics expansions can be found in Section 2.2.

We first recall and define some matrix notation, most of which being very similar to those used in Section 3.1.1. Then, we detail the SAAF–P_N, SAAF–VT–P_N and LS–P_N variational formulations.

4.4.1 Notation

Recall some definitions from Section 3.1.1:

$$\vec{\mathbf{D}} = \sum_{u=1}^{d} \mathbf{D}_u \vec{e}_u \quad , \quad \mathbf{D}_u = \int_{\mathbb{S}^2} \Omega_u \, \mathbf{R}(\vec{\Omega})\mathbf{R}^T(\vec{\Omega}) \, \mathrm{d}\Omega, \tag{4.37}$$

$$\boldsymbol{\eta} = \mathrm{diag}\Big\{\sigma_{\mathrm{s},\ell} + \nu\sigma_f \delta_{\ell,0} \, , \, m = -\ell, ..., \ell \, ; \, \ell = 0, ..., N\Big\}. \tag{4.38}$$

We further define:

$$\underline{\mathbf{H}} \equiv \int_{\mathbb{S}^2} \left(\vec{\Omega}\vec{\Omega}^T\right) \otimes \mathbf{R}\mathbf{R}^T \mathrm{d}\Omega, \tag{4.39}$$

where $\otimes$ is the tensor product. In particular, for all $u, v \in \{1, ..., d\}$, we have:

$$\mathbf{H}_{u,v} \equiv \int_{\mathbb{S}^2} \Omega_u \Omega_v \mathbf{R}\mathbf{R}^T \mathrm{d}\Omega. \tag{4.40}$$

4.4.2 SAAF–P$_N$

The SAAF–P_N formulation is obtained by expanding Ψ as $\mathbf{R}^T \, \boldsymbol{\Phi}$ in Eq. 4.8 and testing the equation against a test function $\Psi^* = \mathbf{R}^T \, \boldsymbol{\Phi}^*$, yielding: find $\boldsymbol{\Phi} \in V^P$ such that for all $\mathbf{\Phi}^* \in V^P$,

$$\begin{aligned}\left(\vec{\nabla}\mathbf{\Phi}^*, \frac{1}{\sigma_\mathrm{t}}\underline{\mathbf{H}} \cdot \vec{\nabla}\mathbf{\Phi}\right)_\mathcal{D} &+ \langle \mathbf{\Phi}^*, \mathbf{L}^\oplus(\vec{n}_b)\mathbf{\Phi}\rangle_{\partial\mathcal{D}} - \langle \mathbf{\Phi}^*, \mathbf{N}(\vec{n}_b)\mathbf{\Phi}\rangle_{\partial\mathcal{D}^\mathrm{r}} - \langle \mathbf{\Phi}^*, \mathbf{J}^\mathrm{inc}\rangle_{\partial\mathcal{D}^\mathrm{d}} \\ &+ (\mathbf{\Phi}^*, \sigma_\mathrm{t}\mathbf{\Phi})_\mathcal{D} = \left(\frac{1}{\sigma_\mathrm{t}}\vec{\mathbf{D}} \cdot \vec{\nabla}\mathbf{\Phi}^* + \mathbf{\Phi}^*, \boldsymbol{\eta}\mathbf{\Phi} + \mathbf{S}\right)_\mathcal{D}.\end{aligned} \tag{4.41}$$

where:

$$\mathbf{N}(\vec{n}) \equiv \int_{\vec{\Omega}\cdot\vec{n}_b<0} \left|\vec{\Omega} \cdot \vec{n}_b\right| \mathbf{R}(\vec{\Omega})\, \mathbf{R}^T(\vec{\Omega}_\mathrm{r})\, \mathrm{d}\Omega. \tag{4.42}$$

The definitions of $\mathbf{L}^\oplus$ and $\mathbf{J}^\mathrm{inc}$ are introduced in Eqs. 3.36 and 3.34.

Note that we are still using the operators $(\cdot,\cdot)_D$ and $\langle\cdot,\cdot\rangle_{\partial\mathcal{D}}$, which involve an angular integration, to avoid having to define new operators that would only integrate over space. The variational formulation does not fundamentally change since the angular integration over $\mathbb{S}^2$ of angular-independent quantities only multiplies all the terms with w.

4.4.3 SAAF–VT–P$_N$

The SAAF–VT–P$_N$ formulation is similarly obtained: find $\mathbf{\Phi} \in V^P$ such that for all $\mathbf{\Phi}^* \in V^P$,

$$\begin{aligned}&\left(\vec{\nabla}\mathbf{\Phi}^*, \tau\underline{\mathbf{H}} \cdot \vec{\nabla}\mathbf{\Phi}\right)_\mathcal{D} + \langle \mathbf{\Phi}^*, \mathbf{L}^\oplus(\vec{n}_b)\mathbf{\Phi}\rangle_{\partial\mathcal{D}} - \langle \mathbf{\Phi}^*, \mathbf{N}(\vec{n}_b)\mathbf{\Phi}\rangle_{\partial\mathcal{D}^\mathrm{r}} - \langle \mathbf{\Phi}^*, \mathbf{J}^\mathrm{inc}\rangle_{\partial\mathcal{D}^\mathrm{d}} \\ &\quad - \left(\vec{\nabla}\mathbf{\Phi}^*, (1-\tau\sigma_\mathrm{t})\vec{\mathbf{D}} \cdot \vec{\nabla}\mathbf{\Phi}\right)_\mathcal{D} + (\mathbf{\Phi}^*, \sigma_\mathrm{t}\mathbf{\Phi})_\mathcal{D} = \left(\frac{1}{\sigma_\mathrm{t}}\vec{\mathbf{D}} \cdot \vec{\nabla}\mathbf{\Phi}^* + \mathbf{\Phi}^*, \boldsymbol{\eta}\mathbf{\Phi} + \mathbf{S}\right)_\mathcal{D}.\end{aligned} \tag{4.43}$$

This formulation is mostly for formal considerations as we already mentioned that it suffers from conditioning issues (see Section 4.1.1.2). This is further demonstrated in Table 5.1.

4.4.4 LS–P_N

The LS–P$_N$ formulation is similarly obtained: find $\mathbf{\Phi} \in V^P$ such that for all $\mathbf{\Phi}^* \in V^P$,

$$\left(\vec{\nabla}\mathbf{\Phi}^*, \underline{\mathbf{H}} \cdot \vec{\nabla}\mathbf{\Phi}\right)_{\mathcal{D}} + \langle \sigma_{\mathrm{t}}\mathbf{\Phi}^*, \mathbf{L}^{\ominus}(\vec{n}_b)\mathbf{\Phi}\rangle_{\partial\mathcal{D}} - \langle \sigma_{\mathrm{t}}\mathbf{\Phi}^*, \mathbf{N}(\vec{n}_b)\mathbf{\Phi}\rangle_{\partial\mathcal{D}^{\mathrm{r}}} - \langle \sigma_{\mathrm{t}}\mathbf{\Phi}^*, \mathbf{J}^{\mathrm{inc}}\rangle_{\partial\mathcal{D}^{\mathrm{d}}} + (\sigma_{\mathrm{t}}\mathbf{\Phi}^*, \sigma_{\mathrm{t}}\mathbf{\Phi})_{\mathcal{D}} = \left(\vec{\mathbf{D}} \cdot \vec{\nabla}\mathbf{\Phi}^* + \sigma_{\mathrm{t}}\mathbf{\Phi}^*, \boldsymbol{\eta}\mathbf{\Phi} + \mathbf{S}\right)_{\mathcal{D}}, \tag{4.44}$$

where:

$$\mathbf{L}^{\ominus} = \int_{\vec{\Omega}\cdot\vec{n}_b<0} |\vec{\Omega} \cdot \vec{n}_b|\, \mathbf{R}\mathbf{R}^T \,\mathrm{d}\Omega. \tag{4.45}$$

4.5 Study of Parity-Based P$_N$ Methods

In this section, we discuss the even-parity form as well as the potential of the SAAF and LS methods to be solved solely for the even component of Ψ. The fundamental idea is to define the even- and odd-parity components of Ψ and Ψ^*:

$$\Psi_e(\vec{\Omega}) \equiv \frac{\Psi(\vec{\Omega}) + \Psi(-\vec{\Omega})}{2} \quad , \quad \Psi^*_e(\vec{\Omega}) \equiv \frac{\Psi^*(\vec{\Omega}) + \Psi^*(-\vec{\Omega})}{2}, \tag{4.46}$$

$$\Psi_o(\vec{\Omega}) \equiv \frac{\Psi(\vec{\Omega}) - \Psi(-\vec{\Omega})}{2} \quad , \quad \Psi^*_o(\vec{\Omega}) \equiv \frac{\Psi^*(\vec{\Omega}) - \Psi^*(-\vec{\Omega})}{2}, \tag{4.47}$$

which also implies:

$$\Psi = \Psi_e + \Psi_o \quad , \quad \Psi^* = \Psi^*_e + \Psi^*_o. \tag{4.48}$$

In this section, we introduce some notation and then describe some existing methods. Then, we prove the equivalences between some of these in the P$_N$ case. We give some interpretation and explain why the idea of a second-order filter for void compatibility is not successful.

4.5.1 Notation

In addition to the definitions introduced in Section 4.4.1, we introduce some consistent notation to specify the angular parity through a subscript e or o, respectively referring to the 'even' and 'odd' parity.[6]

In particular, using the fact that the parity of the spherical harmonics is equal to that of ℓ (see Appendix A.2.1), we define the even spherical harmonics and solution vectors:

$$\mathbf{R}_e \equiv \left\{ R_\ell^m, m = -\ell, \cdots, \ell; \ell = 0, \cdots, N,\ \ell \text{ even} \right\}, \tag{4.49}$$

$$\mathbf{\Phi}_e \equiv \left\{ \Phi_\ell^m, m = -\ell, \cdots, \ell; \ell = 0, \cdots, N;\ \ell \text{ even} \right\}, \tag{4.50}$$

and their odd counterparts as:

$$\mathbf{R}_o \equiv \left\{ R_\ell^m, m = -\ell, \cdots, \ell; \ell = 0, \cdots, N,\ \ell \text{ odd} \right\}, \tag{4.51}$$

$$\mathbf{\Phi}_o \equiv \left\{ \Phi_\ell^m, m = -\ell, \cdots, \ell; \ell = 0, \cdots, N;\ \ell \text{ odd} \right\}. \tag{4.52}$$

It follows:

$$\Psi \approx \mathbf{R}^T\, \mathbf{\Phi} = \underbrace{\mathbf{R}_e^T\, \mathbf{\Phi}_e}_{\approx \Psi_e} + \underbrace{\mathbf{R}_o^T\, \mathbf{\Phi}_o}_{\approx \Psi_o}, \tag{4.53}$$

and we then define[7]:

$$\vec{\mathbf{D}}_e \equiv \sum_{u=1}^{d} \mathbf{D}_{e,u} \vec{e}_u \quad , \quad \mathbf{D}_{e,u} \equiv \int_{\mathbb{S}^2} \Omega_u\, \mathbf{R}_e(\vec{\Omega}) \mathbf{R}_o^T(\vec{\Omega})\, \mathrm{d}\Omega \quad , \quad \vec{\mathbf{D}}_e^T \equiv \sum_{u=1}^{d} \mathbf{D}_{e,u}^T \vec{e}_u, \tag{4.54}$$

[6] This means that a quantity with a subscript e is implied to be an even function of $\vec{\Omega}$.

[7] The notation $\vec{\mathbf{D}}_e^T$ is ambiguous, which is why we explicitly detail its definition: it refers to the vector of the $\mathbf{D}_{e,u}^T$ matrices and not to the transposed vector of the $\mathbf{D}_{e,u}$ matrices.

as well as:

$$\underline{\underline{\mathbf{H}}}_e \equiv \int_{\mathbb{S}^2} \left(\vec{\Omega}\vec{\Omega}^T\right) \otimes \mathbf{R}_e\mathbf{R}_e^T \mathrm{d}\Omega \quad , \quad \mathbf{H}_{e,u,v} \equiv \int_{\mathbb{S}^2} \Omega_u\Omega_v\mathbf{R}_e\mathbf{R}_e^T \mathrm{d}\Omega. \tag{4.55}$$

The following definitions will also prove to be useful[8]:

$$\boldsymbol{\eta}_e = \mathrm{diag}\Big\{\sigma_{\mathrm{s},\ell} + \nu\sigma_f\delta_{\ell,0}\,,\, m = -\ell, ..., \ell\,;\, \ell = 0, ..., N,\ \ell \text{ even}\Big\}. \tag{4.56}$$

$$\boldsymbol{\sigma}_{s,o} \equiv \mathrm{diag}\Big\{\sigma_{\mathrm{s},\ell}\,,\, m = -\ell, ..., \ell\,;\, \ell = 0, \cdots, N,\ \ell \text{ odd}\Big\}, \tag{4.57}$$

$$\mathbf{S}_o \equiv \int_{\mathbb{S}^2} S\,\mathbf{R}_o\,\mathrm{d}\Omega \quad , \quad \mathbf{S}_e \equiv \int_{\mathbb{S}^2} S\,\mathbf{R}_e\,\mathrm{d}\Omega. \tag{4.58}$$

4.5.2 Existing Parity-Based Methods

In this section, we present a few existing methods based on parity considerations: the even-parity, 'even' SAAF, 'even' SAAF–VT and 'even' LS formulations.

The main advantage of using parity-based methods is the reduced number of unknowns. In the P$_N$ case, for instance, only solving for the even moments reduces the size of the linear system by almost a factor two.[9]

4.5.2.1 Even-Parity Form

The transport equation (4.4) is transformed into an equivalent[10] system:

$$\left\{\begin{array}{l} \vec{\Omega}\cdot\vec{\nabla}\Psi_o + \sigma_\mathrm{t}\Psi_e = H_e\Psi_e + S_e \\ \vec{\Omega}\cdot\vec{\nabla}\Psi_e + \sigma_\mathrm{t}\Psi_o = H_o\Psi_o + S_o \end{array}\right. . \tag{4.59}$$

[8] Note that instead of using the notation $\boldsymbol{\eta}_o$, we simply use $\boldsymbol{\sigma}_{s,o}$ to emphasize that fission neutrons are emitted isotropically and therefore do not appear in the odd-parity transport equation.

[9] See Appendix A.2.2 for the exact number of odd and even moments.

[10] This system is indeed equivalent to Eq. 4.4 although it contains twice as many equations because Ψ_e and Ψ_o only need to be determined on half of the unit sphere, their parity giving their values on the other half.

The first equation gives the following weak formulation: find $\Psi_e \in V_e$ such that for all $\Psi_e^* \in V_e$,

$$-\left(\vec{\Omega}\cdot\vec{\nabla}\Psi_e^*, \Psi_o\right)_{\mathcal{D}} + \left\langle \Psi_e^*, \Psi_o\right\rangle_{\partial\mathcal{D}} + \left(\sigma_\mathrm{t}\Psi_e^*, \Psi_e\right)_{\mathcal{D}} = \left(\Psi_e^*, H_e\Psi_e + S_e\right)_{\mathcal{D}}, \tag{4.60}$$

where the odd-angular flux Ψ_o can be evaluated using:

$$\Psi_o = \left(\sigma_\mathrm{t} - H_o\right)^{-1}\left(S_o - \vec{\Omega}\cdot\vec{\nabla}\Psi_e\right). \tag{4.61}$$

It is interesting to notice that this expression is not void compatible which implies that most parity-based methods will have problems in such regions. Note however that it is valid in purely scattering regions because H_o only contains the odd-parity scattering terms.

4.5.2.2 ’Even’ SAAF

What we coined the ’even’ SAAF method derives from the SAAF weak formulation (given by (4.8)) where we only solve for the even-parity component of Ψ and evaluate its odd-parity component using Eq. 4.61. In practice, it is obtained by only keeping the terms from Eq. 4.8 associated to even test functions. We now show how this is done.

Starting with Eq. 4.8 and expanding Ψ and Ψ^* using Eqs. 4.46 and 4.47, it reads:

$$\begin{aligned}&\left(\vec{\Omega}\cdot\vec{\nabla}\Psi_e^*, \frac{1}{\sigma_\mathrm{t}}\vec{\Omega}\cdot\vec{\nabla}\Psi_e\right)_{\mathcal{D}} + \left(\vec{\Omega}\cdot\vec{\nabla}\Psi_o^*, \frac{1}{\sigma_\mathrm{t}}\vec{\Omega}\cdot\vec{\nabla}\Psi_o\right)_{\mathcal{D}} + \left(\sigma_\mathrm{t}\Psi_e^*, \Psi_e\right)_{\mathcal{D}} + \left(\sigma_\mathrm{t}\Psi_o^*, \Psi_o\right)_{\mathcal{D}}\\ &+ \left\langle \Psi^*, \Psi\right\rangle_{\partial\mathcal{D}}^{+} = \left(\frac{1}{\sigma_\mathrm{t}}\vec{\Omega}\cdot\vec{\nabla}\Psi_o^* + \Psi_e^*, H_e\Psi_e + S_e\right)_{\mathcal{D}} + \left(\frac{1}{\sigma_\mathrm{t}}\vec{\Omega}\cdot\vec{\nabla}\Psi_e^* + \Psi_o^*, H_o\Psi_o + S_o\right)_{\mathcal{D}}.\end{aligned} \tag{4.62}$$

This is because for any function f, g such that fg is an odd function of $\vec{\Omega}$, $(f,g)_{\mathcal{D}} = 0$.

Only the terms corresponding to the even equations (i.e. to the test functions Ψ_e^*) are kept, yielding the 'even' SAAF weak formulation: find $\Psi_e \in V_e$ such that for all $\Psi_e^* \in V_e$,

$$\left(\vec{\Omega} \cdot \vec{\nabla}\Psi_e^*, \frac{1}{\sigma_\mathrm{t}}\vec{\Omega} \cdot \vec{\nabla}\Psi_e\right)_\mathcal{D} + \langle \Psi_e^*, \Psi \rangle_{\partial\mathcal{D}} + (\Psi_e^*, \sigma_\mathrm{t}\Psi_e)_\mathcal{D} = \left(\frac{1}{\sigma_\mathrm{t}}\vec{\Omega} \cdot \vec{\nabla}\Psi_e^*, H_o\Psi_o + S_o\right)_\mathcal{D} + (\Psi_e^*, H_e\Psi_e + S_e)_\mathcal{D} \,. \tag{4.63}$$

Note that we could similarly derive the 'odd' SAAF formulation.

4.5.2.3 'Even' SAAF–VT

The 'even' SAAF–VT weak formulation is similarly derived from Eq. 4.10 and reads: find $\Psi_e \in V_e$ such that for all $\Psi_e^* \in V_e$,

$$\left(\vec{\Omega} \cdot \vec{\nabla}\Psi_e^*, \frac{1}{\sigma_\mathrm{t}}\vec{\Omega} \cdot \vec{\nabla}\Psi_e\right)_\mathcal{D} - \left(\vec{\Omega} \cdot \vec{\nabla}\Psi_e^*, (1-\tau\sigma_\mathrm{t})\Psi_o\right)_\mathcal{D} + \langle \Psi_e^*, \Psi \rangle_{\partial\mathcal{D}} + (\Psi_e^*, \sigma_\mathrm{t}\Psi_e)_\mathcal{D} = \left(\frac{1}{\sigma_\mathrm{t}}\vec{\Omega} \cdot \vec{\nabla}\Psi_e^*, H_o\Psi_o + S_o\right)_\mathcal{D} + (\Psi_e^*, H_e\Psi_e + S_e)_\mathcal{D} \,. \tag{4.64}$$

4.5.2.4 'Even' LS

The 'even' LS weak formulation is likewise derived from Eq. 4.23 and reads: find $\Psi_e \in V_e$ such that for all $\Psi_e^* \in V_e$,

$$\left(\vec{\Omega} \cdot \vec{\nabla}\Psi_e^*, \vec{\Omega} \cdot \vec{\nabla}\Psi_e\right)_\mathcal{D} + \left(\sigma_\mathrm{t}\Psi_e^*, \vec{\Omega} \cdot \vec{\nabla}\Psi_o\right)_\mathcal{D} + \left(\vec{\Omega} \cdot \vec{\nabla}\Psi_e^*, \sigma_\mathrm{t}\Psi_o\right)_\mathcal{D} + (\sigma_\mathrm{t}\Psi_e^*, \sigma_\mathrm{t}\Psi_e)_\mathcal{D} + \langle c\Psi_e^*, \Psi - \Psi^\mathrm{inc} \rangle_{\partial\mathcal{D}}^- = \left(\vec{\Omega} \cdot \vec{\nabla}\Psi_e^*, H_o\Psi_o + S_o\right)_\mathcal{D} + (\sigma_\mathrm{t}\Psi_e^*, H_e\Psi_e + S_e)_\mathcal{D} \,, \tag{4.65}$$

4.5.3 Equivalence of the 'Even' SAAF and Even-Parity Boundary Conditions

We do not assume any angular discretization at this stage and show that the boundary terms in Eqs. 4.60 and 4.62, corresponding respectively to the even-parity and 'even' SAAF formulations, are equivalent.

4.5.3.1 'Even' SAAF

Using parity considerations, it was shown in [69] that it is possible to express the boundary term from Eq. 4.62 as:

$$\langle \Psi^*, \Psi \rangle^+_{\partial\mathcal{D}} = 2\langle \Psi^*_e, \Psi_e \rangle^+_{\partial\mathcal{D}} - 2\langle \Psi^*_e, \Psi^{\text{inc}} \rangle^-_{\partial\mathcal{D}} + 2\langle \Psi^*_o, \Psi_e \rangle^+_{\partial\mathcal{D}}. \tag{4.66}$$

This boundary type – therein referred to as the 'even' type – has the property to decouple the even moments from the odd moments in the case of an incoming Dirichlet boundary condition (in which case Ψ^{inc} only contributes to the linear form) [69].

Keeping only the previous terms corresponding to Ψ^*_e, we then obtain the following boundary term for the 'even' SAAF formulation:

$$\Gamma_{\text{even}} \equiv 2\langle \Psi^*_e, \Psi_e \rangle^+_{\partial\mathcal{D}} - 2\langle \Psi^*_e, \Psi^{\text{inc}} \rangle^-_{\partial\mathcal{D}}. \tag{4.67}$$

4.5.3.2 Even-Parity

Considering Eq. 4.60, the boundary term of the even-parity form is $\langle \Psi^*_e, \Psi_o \rangle_{\partial\mathcal{D}}$. The goal of this paragraph is to show that we can enforce the boundary conditions such that this term is equal to Γ_{even}.

A Marshak boundary condition imposes that, at any point on $\partial\mathcal{D}$:

$$\int_{\vec{\Omega}\cdot\vec{n}_b<0} \Psi\,\Psi_e^*\,|\vec{\Omega}\cdot\vec{n}_b|\,\mathrm{d}\Omega = \int_{\vec{\Omega}\cdot\vec{n}_b<0} \Psi^{\mathrm{inc}}\,\Psi_e^*\,|\vec{\Omega}\cdot\vec{n}_b|\,\mathrm{d}\Omega. \tag{4.68}$$

Yet, because $\Psi_e^*\Psi_o$ and $\Psi_e^*\Psi_e$ are respectively odd and even functions of $\vec{\Omega}$ and using Eq. 4.68, we have:

$$\begin{aligned}
\langle\Psi_e^*,\Psi_o\rangle_{\partial\mathcal{D}} &= -2\langle\Psi_e^*,\Psi_o\rangle_{\partial\mathcal{D}}^-,\\
&= -2\left(\langle\Psi_e^*,\Psi\rangle_{\partial\mathcal{D}}^- - \langle\Psi_e^*,\Psi_e\rangle_{\partial\mathcal{D}}^-\right),\\
&= 2\left(\langle\Psi_e^*,\Psi_e\rangle_{\partial\mathcal{D}}^+ - \langle\Psi_e^*,\Psi^{\mathrm{inc}}\rangle_{\partial\mathcal{D}}^-\right),
\end{aligned} \tag{4.69}$$

which proves that:

$$\langle\Psi_e^*,\Psi_o\rangle_{\partial\mathcal{D}} = \Gamma_{\mathrm{even}}. \tag{4.70}$$

In summary, the boundary terms from the 'even' SAAF and even-parity with a Marshak boundary condition formulations are identical.

4.5.4 'Even' SAAF–P$_N$ and Even-Parity P$_N$ Equivalence

Now that we have shown that the boundary terms from the 'even' SAAF and even-parity forms are identical, we want to show that the weak formulations are actually equivalent in the P$_N$ case.[11]

[11] It may still be true in the S$_N$ case; we have not investigated it though.

4.5.4.1 Even-Parity P_N

Using Eq. 4.70, the even-parity weak formulation given by Eq. 4.60 can be rewritten: find $\Psi_e \in V_e$ such that for all $\Psi_e^* \in V_e$,

$$-\left(\vec{\Omega} \cdot \vec{\nabla}\Psi_e^*, \Psi_o\right)_{\mathcal{D}} + \Gamma_{\text{even}} + \left(\sigma_{\text{t}}\Psi_e^*, \Psi_e\right)_{\mathcal{D}} = \left(\Psi_e^*, H_e\Psi_e + S_e\right)_{\mathcal{D}}, \tag{4.71}$$

and, as a reminder, the odd-angular flux Ψ_o can be evaluated using Eq. 4.61:

$$\Psi_o = (\sigma_{\text{t}} - H_o)^{-1}\left(S_o - \vec{\Omega} \cdot \vec{\nabla}\Psi_e\right), \tag{4.72}$$

which we could directly substitute into Eq. 4.71 to get rid of Ψ_o in the first volumetric term. The problem is that with anisotropic scattering, the operator $(\sigma_{\text{t}} - H_o)^{-1}$ does not reduce to a simple form (with isotropic scattering, it simply reduces to σ_{t}^{-1}) and would in particular depend on the angular discretization. Therefore, we choose to enforce the P_N angular discretization at this stage.

First of all, the P_N counterpart to Eq. 4.61 can be determined by directly integrating the second equation of (4.59) (or odd-parity transport equation) against a test function $\mathbf{R}_o$ and expanding Ψ_e and Ψ_o using P_N expansions, yielding the following linear system:

$$\vec{\mathbf{D}}_e^T \cdot \vec{\nabla}\mathbf{\Phi}_e + \sigma_{\text{t}}\mathbf{\Phi}_o = \boldsymbol{\sigma}_{s,o}\mathbf{\Phi}_o + \mathbf{S}_o. \tag{4.73}$$

The odd-moments can therefore be expressed as:

$$\mathbf{\Phi}_o = (\sigma_{\text{t}}\mathbb{I} - \boldsymbol{\sigma}_{s,o})^{-1}\left(\mathbf{S}_o - \vec{\mathbf{D}}_e^T \cdot \vec{\nabla}\mathbf{\Phi}_e\right), \tag{4.74}$$

where $\mathbb{I}$ is the identity matrix.

Yet, after expanding all variables with their spherical harmonics expansion, it reads:

$$-\left(\vec{\nabla}\mathbf{\Phi}_e^*, \vec{\mathbf{D}}_e\,\mathbf{\Phi}_o\right)_{\mathcal{D}} + \Gamma_{\text{even}} + \left(\mathbf{\Phi}_e^*, \sigma_{\text{t}}\mathbf{\Phi}_e\right)_{\mathcal{D}} = \left(\mathbf{\Phi}_e^*, \boldsymbol{\eta}_e\,\mathbf{\Phi}_e + \mathbf{S}_e\right)_{\mathcal{D}}. \tag{4.75}$$

Using Eq. 4.74, we obtain the even-parity P$_N$ weak formulation: find $\mathbf{\Phi}_e \in V^{P_e}$ such that for all $\mathbf{\Phi}_e \in V^{P_e}$,

$$\begin{aligned}\left(\vec{\nabla}\mathbf{\Phi}_e^*, \vec{\mathbf{D}}_e\,\left(\sigma_{\text{t}}\mathbb{I} - \boldsymbol{\sigma}_{s,o}\right)^{-1}\vec{\mathbf{D}}_e^T \cdot \vec{\nabla}\mathbf{\Phi}_e\right)_{\mathcal{D}} + \Gamma_{\text{even}} + \left(\mathbf{\Phi}_e^*, \sigma_{\text{t}}\mathbf{\Phi}_e\right)_{\mathcal{D}} \\ = \left(\mathbf{\Phi}_e^*, \boldsymbol{\eta}_e\,\mathbf{\Phi}_e + \mathbf{S}_e\right)_{\mathcal{D}} + \left(\vec{\nabla}\mathbf{\Phi}_e^*, \vec{\mathbf{D}}_e\,\left(\sigma_{\text{t}}\mathbb{I} - \boldsymbol{\sigma}_{s,o}\right)^{-1}\mathbf{S}_o\right)_{\mathcal{D}}.\end{aligned} \tag{4.76}$$

We want to show that this weak formulation is identical to that obtained for the 'even' SAAF–P$_N$ form.

4.5.4.2 'Even' SAAF–P$_N$

We can get the 'even' SAAF-P$_N$ weak formulation from Eq. 4.41, knowing from Section 4.5.3 that the boundary terms are equal to Γ_{even}: find $\mathbf{\Phi}_e \in V^{P_e}$ such that for all $\mathbf{\Phi}_e \in V^{P_e}$,

$$\begin{aligned}\left(\vec{\nabla}\mathbf{\Phi}_e^*, \frac{1}{\sigma_{\text{t}}}\underline{\underline{\mathbf{H}}}_e \cdot \vec{\nabla}\mathbf{\Phi}_e\right)_{\mathcal{D}} + \Gamma_{\text{even}} + \left(\mathbf{\Phi}_e^*, \sigma_{\text{t}}\mathbf{\Phi}_e\right)_{\mathcal{D}} \\ = \left(\mathbf{\Phi}_e^*, \boldsymbol{\eta}_e\mathbf{\Phi}_e + \mathbf{S}_e\right)_{\mathcal{D}} + \left(\frac{1}{\sigma_{\text{t}}}\vec{\nabla}\mathbf{\Phi}_e^*, \vec{\mathbf{D}}_e(\boldsymbol{\sigma}_{s,o}\mathbf{\Phi}_o + \mathbf{S}_o)\right)_{\mathcal{D}}.\end{aligned} \tag{4.77}$$

The odd moments $\mathbf{\Phi}_o$ can be substituted from this equation using Eq 4.74:

$$\begin{aligned}&\left(\vec{\nabla}\mathbf{\Phi}_e^*, \frac{1}{\sigma_\mathrm{t}}\underline{\mathbf{H}}_e \cdot \vec{\nabla}\mathbf{\Phi}_e\right)_{\mathcal{D}} + \Gamma_\text{even} + (\mathbf{\Phi}_e^*, \sigma_\mathrm{t}\mathbf{\Phi}_e)_{\mathcal{D}} = (\mathbf{\Phi}_e^*, \boldsymbol{\eta}_e\mathbf{\Phi}_e + \mathbf{S}_e)_{\mathcal{D}} \\ &+ \left(\frac{1}{\sigma_\mathrm{t}}\vec{\nabla}\mathbf{\Phi}_e^*, \vec{\mathbf{D}}_e(\boldsymbol{\sigma}_{s,o}\left(\sigma_\mathrm{t}\mathbb{I} - \boldsymbol{\sigma}_{s,o}\right)^{-1}\left(\mathbf{S}_o - \vec{\mathbf{D}}_e^T \cdot \vec{\nabla}\mathbf{\Phi}_e\right) + \mathbf{S}_o)\right)_{\mathcal{D}}.\end{aligned} \tag{4.78}$$

Although this is far from obvious, we have the following property:

$$\underline{\mathbf{H}}_e = \vec{\mathbf{D}}_e\,\vec{\mathbf{D}}_e^T. \tag{4.79}$$

This is not easy to prove[12], if only because each $\mathbf{D}_{e,u}$, $u \in \{1, ..., d\}$ is not a square matrix but of size $P_e \times P_o$, where P_e and P_o are respectively the number of even and odd moments ($P_e + P_o = P$). This relationship is useful to further simplify the previous formulation:

$$\begin{aligned}&\left(\vec{\nabla}\mathbf{\Phi}_e^*, \frac{1}{\sigma_\mathrm{t}}\vec{\mathbf{D}}_e\left(\mathbb{I} + \boldsymbol{\sigma}_{s,o}\left(\sigma_\mathrm{t}\mathbb{I} - \boldsymbol{\sigma}_{s,o}\right)^{-1}\right)\vec{\mathbf{D}}_e^T \cdot \vec{\nabla}\mathbf{\Phi}_e\right)_{\mathcal{D}} + \Gamma_\text{even} + (\mathbf{\Phi}_e^*, \sigma_\mathrm{t}\mathbf{\Phi}_e)_{\mathcal{D}} \\ &= (\mathbf{\Phi}_e^*, \boldsymbol{\eta}_e\mathbf{\Phi}_e + \mathbf{S}_e)_{\mathcal{D}} + \left(\frac{1}{\sigma_\mathrm{t}}\vec{\nabla}\mathbf{\Phi}_e^*, \vec{\mathbf{D}}_e\left(\boldsymbol{\sigma}_{s,o}\left(\sigma_\mathrm{t}\mathbb{I} - \boldsymbol{\sigma}_{s,o}\right)^{-1} + \mathbb{I}\right)\mathbf{S}_o\right)_{\mathcal{D}}.\end{aligned} \tag{4.80}$$

Yet, one can prove the following:

$$\begin{aligned}\mathbb{I} + \boldsymbol{\sigma}_{s,o}\left(\sigma_\mathrm{t}\mathbb{I} - \boldsymbol{\sigma}_{s,o}\right)^{-1} &= \operatorname{diag}\left\{1 + \frac{\sigma_{s,\ell}}{\sigma_\mathrm{t} - \sigma_{s,\ell}}\,,\, m = -\ell, ..., \ell\,;\, \ell = 0, ..., N,\ \ell \text{ odd}\right\}, \\ &= \operatorname{diag}\left\{\frac{\sigma_\mathrm{t}}{\sigma_\mathrm{t} - \sigma_{s,\ell}}\,,\, m = -\ell, ..., \ell\,;\, \ell = 0, ..., N,\ \ell \text{ odd}\right\}, \\ &= \sigma_\mathrm{t}\left(\sigma_\mathrm{t}\mathbb{I} - \boldsymbol{\sigma}_{s,o}\right)^{-1}.\end{aligned} \tag{4.81}$$

Therefore, the 'even' SAAF-P$_N$ formulation can be rewritten as: find $\mathbf{\Phi}_e \in V^{P_e}$ such

[12] This property was numerically verified for $N \leq 7$ with up to 11 digits of precision. An attempt to prove it can also be found in Appendix C – although it is believed that there is probably a more concise proof.

that for all $\mathbf{\Phi}_e \in V^{P_e}$,

$$\begin{aligned}\Big(\vec{\nabla}\mathbf{\Phi}_e^*, \vec{\mathbf{D}}_e \; (\sigma_{\rm t}\mathbb{I} - \boldsymbol{\sigma}_{s,o})^{-1} \, \vec{\mathbf{D}}_e^T \cdot \vec{\nabla}\mathbf{\Phi}_e\Big)_{\mathcal{D}} + \Gamma_{\rm even} + \big(\mathbf{\Phi}_e^*, \sigma_{\rm t}\mathbf{\Phi}_e\big)_{\mathcal{D}} \\ = \big(\mathbf{\Phi}_e^*, \boldsymbol{\eta}_e \, \mathbf{\Phi}_e + \mathbf{S}_e\big)_{\mathcal{D}} + \Big(\vec{\nabla}\mathbf{\Phi}_e^*, \vec{\mathbf{D}}_e \; (\sigma_{\rm t}\mathbb{I} - \boldsymbol{\sigma}_{s,o})^{-1} \, \mathbf{S}_o\Big)_{\mathcal{D}} ,\end{aligned} \tag{4.82}$$

which is nothing but the even-parity P$_N$ formulation given by Eq. 4.76. We have thus shown that these two are equivalent.

4.5.5 'Even' SAAF–VT–P$_N$ and 'Even' SAAF–P$_N$ Equivalence

In this section, we show that the even-parity P$_N$, 'even' SAAF–VT–P$_N$ and 'even' SAAF–P$_N$ formulations are all equivalent (the equivalence between the first two having been established in the previous section).

The 'even' SAAF–VT–P$_N$ weak formulation is given by: find $\mathbf{\Phi}_e \in V^{P_e}$ such that for all $\mathbf{\Phi}_e \in V^{P_e}$,

$$\begin{aligned}\Big(\vec{\nabla}\mathbf{\Phi}_e^*, \tau \, \underline{\mathbf{H}}_e \cdot \vec{\nabla}\mathbf{\Phi}_e\Big)_{\mathcal{D}} - \Big(\vec{\nabla}\mathbf{\Phi}_e^*, (1-\tau\sigma_{\rm t})\vec{\mathbf{D}}_e \, \mathbf{\Phi}_o\Big)_{\mathcal{D}} + \big(\mathbf{\Phi}_e^*, \sigma_{\rm t}\mathbf{\Phi}_e\big)_{\mathcal{D}} + \Gamma_{\rm even} \\ = \big(\mathbf{\Phi}_e^*, \boldsymbol{\eta}_e \, \mathbf{\Phi}_e + \mathbf{S}_e\big)_{\mathcal{D}} + \Big(\tau \, \vec{\nabla}\mathbf{\Phi}_e^*, \vec{\mathbf{D}}_e \; (\boldsymbol{\sigma}_{s,o} \, \mathbf{\Phi}_o + \mathbf{S}_o)\Big)_{\mathcal{D}} .\end{aligned} \tag{4.83}$$

Using Eq. 4.74 to eliminate $\mathbf{\Phi}_o$ and after some manipulations, it can be expressed as:

$$\begin{aligned}&\Big(\vec{\nabla}\mathbf{\Phi}_e^*, \tau \, \underline{\mathbf{H}}_e \cdot \vec{\nabla}\mathbf{\Phi}_e\Big)_{\mathcal{D}} + \big(\mathbf{\Phi}_e^*, \sigma_{\rm t}\mathbf{\Phi}_e\big)_{\mathcal{D}} + \Gamma_{\rm even} - \big(\mathbf{\Phi}_e^*, \boldsymbol{\eta}_e \, \mathbf{\Phi}_e + \mathbf{S}_e\big)_{\mathcal{D}} \\ &= \Big(\vec{\nabla}\mathbf{\Phi}_e^*, \vec{\mathbf{D}}_e \; \big((1-\tau\sigma_{\rm t}) \, (\sigma_{\rm t}\mathbb{I} - \boldsymbol{\sigma}_{s,o})^{-1} + \tau \, \boldsymbol{\sigma}_{s,o} \; (\sigma_{\rm t}\mathbb{I} - \boldsymbol{\sigma}_{s,o})^{-1} + \tau\big) \, \mathbf{S}_o\Big)_{\mathcal{D}} \\ &- \Big(\vec{\nabla}\mathbf{\Phi}_e^*, \vec{\mathbf{D}}_e \; \big((1-\tau\sigma_{\rm t}) \, (\sigma_{\rm t}\mathbb{I} - \boldsymbol{\sigma}_{s,o})^{-1} + \tau \, \boldsymbol{\sigma}_{s,o} \; (\sigma_{\rm t}\mathbb{I} - \boldsymbol{\sigma}_{s,o})^{-1}\big) \, \vec{\mathbf{D}}_e^T \cdot \vec{\nabla}\mathbf{\Phi}_e\Big)_{\mathcal{D}} .\end{aligned} \tag{4.84}$$

Yet,

$$
\begin{aligned}
&(1-\tau\sigma_{\mathrm{t}})\left(\sigma_{\mathrm{t}}\mathbb{I}-\boldsymbol{\sigma}_{s,o}\right)^{-1}+\tau\,\boldsymbol{\sigma}_{s,o}\,\left(\sigma_{\mathrm{t}}\mathbb{I}-\boldsymbol{\sigma}_{s,o}\right)^{-1}+\tau\\
&\quad=\operatorname{diag}\Big\{\frac{1-\tau\sigma_{\mathrm{t}}}{\sigma_{\mathrm{t}}-\sigma_{s,\ell}}+\frac{\tau\,\sigma_{s,\ell}}{\sigma_{\mathrm{t}}-\sigma_{s,\ell}}+\tau\,,\;m=-\ell,...,\ell\,;\;\ell=0,...,N,\;\ell\text{ odd}\Big\},\\
&\quad=\operatorname{diag}\Big\{\frac{1-\tau\sigma_{\mathrm{t}}+\tau\,\sigma_{s,\ell}+\tau(\sigma_{\mathrm{t}}-\sigma_{s,\ell})}{\sigma_{\mathrm{t}}-\sigma_{s,\ell}}\,,\;m=-\ell,...,\ell\,;\;\ell=0,...,N,\;\ell\text{ odd}\Big\},\\
&\quad=\left(\sigma_{\mathrm{t}}\mathbb{I}-\boldsymbol{\sigma}_{s,o}\right)^{-1}.
\end{aligned}
\tag{4.85}
$$

Therefore,

$$
\begin{aligned}
&\left(\vec{\nabla}\boldsymbol{\Phi}_e^*,\tau\,\underline{\underline{\mathbf{H}}}_e\cdot\vec{\nabla}\boldsymbol{\Phi}_e\right)_{\mathcal{D}}+\left(\boldsymbol{\Phi}_e^*,\sigma_{\mathrm{t}}\boldsymbol{\Phi}_e\right)_{\mathcal{D}}+\Gamma_{\text{even}}-\left(\boldsymbol{\Phi}_e^*,\boldsymbol{\eta}_e\,\boldsymbol{\Phi}_e+\mathbf{S}_e\right)_{\mathcal{D}}\\
&=\left(\vec{\nabla}\boldsymbol{\Phi}_e^*,\vec{\mathbf{D}}_e\,\left(\sigma_{\mathrm{t}}\mathbb{I}-\boldsymbol{\sigma}_{s,o}\right)^{-1}\mathbf{S}_o\right)_{\mathcal{D}}-\left(\vec{\nabla}\boldsymbol{\Phi}_e^*,\vec{\mathbf{D}}_e\,\left(\left(\sigma_{\mathrm{t}}\mathbb{I}-\boldsymbol{\sigma}_{s,o}\right)^{-1}-\tau\right)\vec{\mathbf{D}}_e^T\cdot\vec{\nabla}\boldsymbol{\Phi}_e\right)_{\mathcal{D}}.
\end{aligned}
\tag{4.86}
$$

Using (4.79) on the first term, we then get:

$$
\begin{aligned}
&\left(\vec{\nabla}\boldsymbol{\Phi}_e^*,\vec{\mathbf{D}}_e\,\left(\sigma_{\mathrm{t}}\mathbb{I}-\boldsymbol{\sigma}_{s,o}\right)^{-1}\vec{\mathbf{D}}_e^T\cdot\vec{\nabla}\boldsymbol{\Phi}_e\right)_{\mathcal{D}}+\left(\boldsymbol{\Phi}_e^*,\sigma_{\mathrm{t}}\boldsymbol{\Phi}_e\right)_{\mathcal{D}}+\Gamma_{\text{even}}\\
&\qquad=\left(\boldsymbol{\Phi}_e^*,\boldsymbol{\eta}_e\,\boldsymbol{\Phi}_e+\mathbf{S}_e\right)_{\mathcal{D}}+\left(\vec{\nabla}\boldsymbol{\Phi}_e^*,\vec{\mathbf{D}}_e\,\left(\sigma_{\mathrm{t}}\mathbb{I}-\boldsymbol{\sigma}_{s,o}\right)^{-1}\mathbf{S}_o\right)_{\mathcal{D}}.
\end{aligned}
\tag{4.87}
$$

Eq. 4.74 implies that the 'even' SAAF–VT–P$_N$ with a void treatment formulation can be given by: find $\boldsymbol{\Phi}_e\in V^{P_e}$ such that for all $\boldsymbol{\Phi}_e\in V^{P_e}$,

$$
-\left(\vec{\nabla}\boldsymbol{\Phi}_e^*,\vec{\mathbf{D}}_e\,\boldsymbol{\Phi}_o\right)_{\mathcal{D}}+\left(\boldsymbol{\Phi}_e^*,\sigma_{\mathrm{t}}\boldsymbol{\Phi}_e\right)_{\mathcal{D}}+\Gamma_{\text{even}}=\left(\boldsymbol{\Phi}_e^*,\boldsymbol{\eta}_e\,\boldsymbol{\Phi}_e+\mathbf{S}_e\right)_{\mathcal{D}},
\tag{4.88}
$$

which is exactly the even-parity P$_N$ formulation and thus also the 'even' SAAF–P$_N$ formulation (**without** void treatment).

4.5.6 Interpretation

The previous claim may seem surprising in that we expect the ’even’ SAAF–VT–P$_N$ method to be able to cope with void regions while the even-parity P$_N$ clearly cannot (at least as is). In all reality however, a closer look at the ’even’ SAAF–VT–P$_N$ method shows that it cannot deal with void either.

Indeed, if we use the ’even’ parity option, all the terms containing $\mathbf{\Phi}_o$ must be evaluated using Eq. 4.74 which does not work in void. Therefore, the second term of Eq. 4.83, namely $(-(\vec{\nabla}\mathbf{\Phi}_e^*, (1-\tau\sigma_\mathrm{t})\vec{\mathbf{D}}_e\,\mathbf{\Phi}_o)_\mathcal{D})$ is incompatible with void.

Thus, there is nothing contradictory in the conclusion from the last subsection, quite the opposite.

4.5.7 Conclusion

In summary, it can be concluded that the ’even’ SAAF–VT-P$_N$ method is equivalent to the even-parity P$_N$ method and thus to the ’even’ SAAF-P$_N$ method **without** void treatment. In other words, the void treatment does not do us any good with the ’even’ parity option.

The fundamental reason for that is that the void treatment in the SAAF–VT method relies on splitting Ψ as:

$$\Psi = (1-\tau\sigma_\mathrm{t})\Psi + \tau\sigma_\mathrm{t}\Psi \tag{4.89}$$

and on only using the angular flux equation (AFE) on the second term. As the first term $(1-\tau\sigma_\mathrm{t})\Psi$ eventually becomes $(\vec{\Omega}\cdot\vec{\nabla}\Psi_e^*, (1-\tau\sigma_\mathrm{t})\Psi_o)_{\partial\mathcal{D}}$, Ψ_o – which is not a primal variable with the ’even’ parity option – thus has to be substituted using Eq. 4.74. The effect of that substitution is to make a $1/\sigma_\mathrm{t}$ term reappear.

4.5.8 Void Incompatibility and Loss of Accuracy

Using parity-based methods can be fairly attractive since it reduces the number of unknowns almost by half (see Appendix A.2.2). Nevertheless, two important downsides are to be pointed out.

First, these methods are incompatible with void whenever the odd-parity moments have to be evaluated using Eqs. 4.61 or 4.74. The only second-order form presented above that does not suffer from this flaw is the 'even' LS form because all the Ψ_o terms are weighted by the cross-section (either σ_{t} or σ_s) and thus vanish in void regions.

Second, it may also induce a loss of accuracy because of the gradient operation in Eqs. 4.61 or 4.74: in particular, for 'even' LS, if V is the space of the piecewise polynomials of order 1, the second term in Eq. 4.65 simply vanishes. In practice, we observe that the spatial convergence is then only first order (see Figure 4.1). Similar results are expected for the even-parity form since the boundary term $\langle \Psi_e^*, \Psi_o \rangle_{\partial \mathcal{D}}$ (see Eq. 4.60) requires the evaluation of Ψ_o^* (unless perhaps in the case of reflecting boundary conditions).

4.5.9 Second-Order Filter

Because we just saw that we do not know of any globally conservative parity-based second order methods compatible with void, we present an idea to use the filters studied in Chapter 3 to create such a method. We will however explain why this option does not give satisfying results.

4.5.9.1 Original Idea

As we previously mentioned, (4.59) can be discretized in angle using the P_N approximation, in which case it reads:

$$\left\{\begin{array}{l} \vec{\mathbf{D}}_e \cdot \vec{\nabla}\mathbf{\Phi}_o + (\sigma_{\mathrm{t}}\mathbb{I} - \boldsymbol{\eta}_e)\mathbf{\Phi}_e = \mathbf{S}_e \\ \vec{\mathbf{D}}_e^T \cdot \vec{\nabla}\mathbf{\Phi}_e + (\sigma_{\mathrm{t}}\mathbb{I} - \boldsymbol{\sigma}_{\mathrm{s},o})\mathbf{\Phi}_o = \mathbf{S}_o \end{array}\right. . \tag{4.90}$$

Using the same approach as in Chapter 3, we can add a filtering operator to these equations:

$$\left\{\begin{array}{l} \vec{\mathbf{D}}_e \cdot \vec{\nabla}\mathbf{\Phi}_o + (\sigma_{\mathrm{t}}\mathbb{I} + \sigma_{\mathrm{f}}\mathbf{F}_e - \boldsymbol{\eta}_e)\mathbf{\Phi}_e = \mathbf{S}_e \\ \vec{\mathbf{D}}_e^T \cdot \vec{\nabla}\mathbf{\Phi}_e + (\sigma_{\mathrm{t}}\mathbb{I} + \sigma_{\mathrm{f}}\mathbf{F}_o - \boldsymbol{\sigma}_{\mathrm{s},o})\mathbf{\Phi}_o = \mathbf{S}_o \end{array}\right. , \tag{4.91}$$

where $\mathbf{F}_e$ and $\mathbf{F}_o$ respectively contain the even and odd components of $\mathbf{F}$. The diagonal terms of $\mathbf{F}_o$ are all strictly positive, ensuring that the matrix $(\sigma_{\mathrm{t}}\mathbb{I} + \sigma_{\mathrm{f}}\mathbf{F}_o - \boldsymbol{\sigma}_{\mathrm{s},o})$ is always invertible, even if $\sigma_{\mathrm{t}} = 0$ in some parts of the domain. The even-parity filtered P_N equation therefore reads:

$$\vec{\mathbf{D}}_e \cdot \vec{\nabla}\Big(\big(\sigma_{\mathrm{t}}\mathbb{I} + \sigma_{\mathrm{f}}\mathbf{F}_o - \boldsymbol{\sigma}_{\mathrm{s},o}\big)^{-1}\big(\mathbf{S}_o - \vec{\mathbf{D}}_o \cdot \vec{\nabla}\mathbf{\Phi}_e\big)\Big) + (\sigma_{\mathrm{t}}\mathbb{I} + \sigma_{\mathrm{f}}\mathbf{F}_e - \boldsymbol{\sigma}_{\mathrm{s},e})\mathbf{\Phi}_e = \mathbf{S}_e. \tag{4.92}$$

4.5.9.2 Flaws

Although this method may *a priori* seem to be well suited to deal with void problems, it has a major flaw: unlike the first-order filter, for which it was desirable that the filter contribution vanishes as N goes to infinity (since the unfiltered solution converges to the analytical one), this second-order filter still need be significant, even for large N. Otherwise, the matrix $\big(\sigma_{\mathrm{t}}\mathbb{I} + \sigma_{\mathrm{f}}\mathbf{F}_o - \boldsymbol{\sigma}_{\mathrm{s},o}\big)$ would be close to singular and the solver could not converge any more efficiently, in that limit, than with the standard even-parity form.

An idea would be to look for other second-order forms, better suited for filtering. Unfortunately, a key feature of the filter is that its contribution to the zeroth equation is zero, so as not to break the global conservation of the scheme. This is why such a filtering method seemed only to be compatible with parity-based methods in the first place. Yet, we have just shown that the 'even' SAAF–P$_N$ and 'even' SAAF–VT–P$_N$ methods are equivalent to the even-parity P$_N$ method and thus perform equally as bad in void. There does not exist – to our knowledge – any void compatible, globally conservative second-order form solving only for the even-parity component of the angular flux.

Nonetheless, the proposed second-order filter should still be able to help in the case of strong spatial variations in the solution, to help mitigate negativity or oscillations in the numerical solution.

4.6 Numerical Results

With the driving application of finding a void compatible method, we show numerical results for the LS-P$_N$ method. First, we verify that we obtain the correct solution for an infinite medium with some void regions. Second, we show that we get the expected spatial and angular convergence rates. In particular, we observe the loss of accuracy mentioned in Section 4.5.8. Next, we introduce a heterogeneous multigroup k-eigenvalue problem with a void region and observe that LS does not perform well, highlighting the crucial need for global conservation.

4.6.1 Infinite Medium with Void

First of all, we consider the very simple test problem of an infinite domain composed of two regions: a uniform, pure-absorber material with a volumetric source

($\sigma_{\mathrm{t}} \equiv \sigma_{a,1}$, $S \equiv q$) and a void.

Independently of the geometry (in particular the size and shape of the void region), the analytical solution is given by:

$$\Psi(\vec{r}, \vec{\Omega}) = \int_0^{\infty} q \exp(-\sigma_{a,1}\, s)\, \mathrm{d}s = \frac{q}{\sigma_{a,1}}, \tag{4.93}$$

i.e.

$$\Phi_\ell^m(\vec{r}) = \frac{q\sqrt{w}}{\sigma_{a,1}} \delta_{\ell,0}. \tag{4.94}$$

It was verified that all the methods with results presented below return this infinite solution to machine precision.[13]

4.6.2 Method of Manufactured Solutions

We consider the following pure-absorber slab geometry problem:

$$\mu \frac{\partial \Psi}{\partial x} + \sigma_{\mathrm{t}}(x)\Psi(x,\mu) = S(x,\mu) \quad , \quad 0 \le x \le L \quad , \quad -1 \le \mu \le 1, \tag{4.95}$$

where:

$$\sigma_t = \begin{cases} \sigma_1 & , \quad L/4 < x < 3L/4 \\ \sigma_0 & , \quad \text{otherwise} \end{cases} \tag{4.96}$$

with $\sigma_1 = 5.0$ cm^{-1}, $\sigma_0 = 0.5$ cm^{-1} and $L = 10$ cm. Reflecting boundaries are imposed at $x = 0$ and $x = L$.

The Method of Manufactured Solutions (MMS) consists of choosing an analytical solution $\tilde{\Psi}$ and deriving the source $\tilde{S}$ such that $\tilde{\Psi}$ is indeed solution of the PDE we are interested in.

[13] The only exception is the SAAF–LS method in Chapter 5, which is only mentioned in this work for comparison purposes and to emphasize the need for the conservative fix (see also Appendix D).

We choose our solution in the following form:

$$\tilde{\Psi} = f(x)\, g(\mu) \quad , \quad 0 \leq x \leq L \quad , \quad -1 \leq \mu \leq 1, \tag{4.97}$$

which implies:

$$\tilde{S} = \mu \frac{\mathrm{d}f}{\mathrm{d}x}\, g + \sigma_t f g. \tag{4.98}$$

Expanding g using spherical harmonics:[14]

$$g = \sum_{\ell=0}^{\infty} \alpha_\ell R_\ell, \tag{4.99}$$

it is shown in Appendix A.3 that the ℓ-th moment ($\ell \in \mathbb{N}$) of the source $\tilde{S}$ is given by:

$$\tilde{S}_\ell = \frac{\mathrm{d}f}{\mathrm{d}x} \left(\frac{\ell}{\sqrt{(2\ell+1)(2\ell-1)}} g_{\ell-1} + \frac{\ell+1}{\sqrt{(2\ell+1)(2\ell+3)}} g_{\ell+1} \right) + \sigma_t f\, g_\ell, \tag{4.100}$$

with the convention $g_{-1} = 0$.

In practice, we choose[15]:

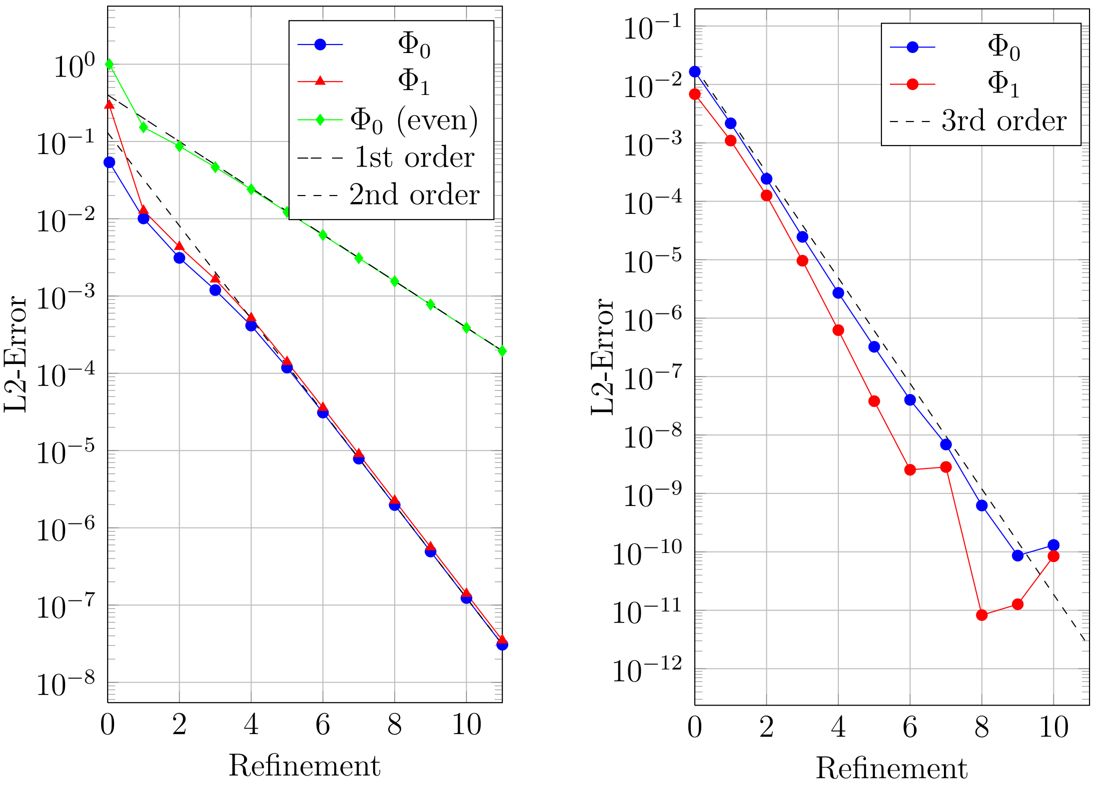
$$f(x) = \phi_0 \left(\cos\left(\frac{\pi x}{L}\right) + a \right), \tag{4.101}$$

$$g(\mu) = R_0 + 0.5 R_2(\mu) + 0.25 R_4(\mu) + 0.125 R_6^0(\mu) + 0.1 R_8(\mu) + 0.05 R_{10}(\mu). \tag{4.102}$$

The convergence results are shown on Figs. 4.1 and 4.2. Several observations can

[14] It is noted that, because the problem only depends on one spatial dimension, the angular dependency can be described exclusively with the spherical harmonics such that $m = 0$ (in other words, the solution does not depend on φ, only on μ). For that reason, we drop the superscript indicating the order m of the spherical harmonics.

[15] The choice of f and g is made such that $\tilde{\Psi}$ satisfies the reflecting boundary conditions. In particular, it would be unphysical to have f' be non-zero on the boundary or to have the odd moments of g be non-zero.



(a) First order Lagrange elements. The green curve shows the error for Φ_0 for the 'even' LS–P$_N$ method.

(b) Second order Lagrange elements.

Figure 4.1: L^2-Error for Φ_0 and Φ_1 for $N = 11$ (4 cells for ref = 0).

be made:

- Fig. 4.1a shows that the spatial convergence rate of the LS–P$_N$ method with piecewise linear Lagrange elements is second order, as expected. In particular, using the 'even' LS–P$_N$ method reduces the number of unknowns but induces the loss of one order of spatial of convergence, as suggested in Section 4.5.8. From Fig. 4.1b, it appears that the spatial convergence rate with piecewise quadratic Lagrange elements is third order.

- The P$_N$ method being a spectral approximation, it is expected to get a spectral convergence, that is exponential convergence with respect to N for infinitely smooth solutions [11]. By construction, g – and therefore $\tilde{\Psi}$ – are infinitely smooth in angle. Fig. 4.2 indeed exhibits such a behavior for $N \leq 10$. The sudden drop in the error at $N = 11$ can be explained by noting that g can – by contruction – be exactly described with moments such that $\ell \leq 10$. The reason why the error drops at $N = 11$ and not $N = 10$ can be found in Eq. 4.100: $\tilde{S}_{11}$ is non-zero because $g_{10} \neq 0$. For $N \geq 11$, the angular error is therefore zero and the error is then exclusively dominated by the spatial error.

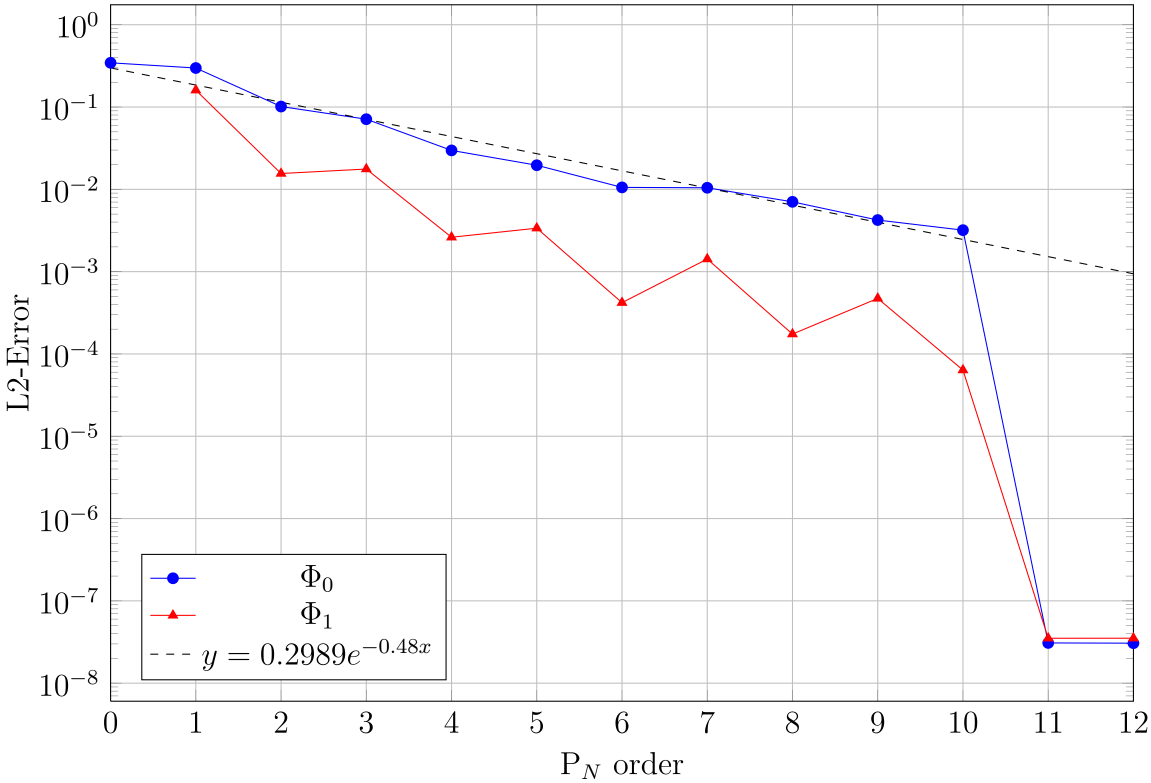


Figure 4.2: L^2-Error for Φ_0 and Φ_1 for a refinement of 11. The dashed line is the interpolation of Φ_0 for N in [0,10].

4.6.3 Heterogenous Multigroup k-Eigenvalue Problem with Void

We consider a test problem described in Fig. 4.3 consisting of 8 pin cells surrounding a void region. The interest is to compare our methods on a multigroup heterogeneous test problem involving a void region. Each of the 9 square subdomains is 1.2598 cm in length, assembled in a 3x3 configuration. The total size of the problem is therefore 3.7794 cm $\times$ 3.7794 cm. The subdomain in the center of the problem is void. The 8 others contain a pin of radius 0.45720 cm. Each pin boundary is approximated by a 20-side polygon. The material properties of the fuel and moderator (shown in blue and yellow on the figure, respectively) are chosen to be identical to the "UO_2 Fuel-Clad mix" and "Moderator" materials from the C5G7 benchmark [1]. This problem is assumed to be infinite along the z direction and therefore only depends on x and y. The same problem[16] was run using MCNP5 [70] with 125 cycles of 10^6 particles (the first 25 cycles being discarded). The reference eigenvalue was estimated to be $\tilde{k}_{\text{eff}} = 1.34745$ with a standard deviation of 5 pcm.[17]

Table 4.1 shows the error with respect to $\tilde{k}_{\text{eff}}$ for the LS–P_N method. Lacking global conservation, it is extremely show to converge as the number of elements and angular moments is increased. In particular, the LS–P_{39} solution with a spatial refinement of 3 has over 5.3×10^8 unknowns[18] but is still over one hundred standard deviations away from the reference.

Table 4.2 compares the same quantity for the LS–S_N and SAAF–VT–S_N methods as a function of the number of polar and azimuthal angles per quadrant, noted N_p and N_a respectively. The quadrature rule used is the Bickley3-Optimized[19] because

[16] For convenience however, the pin boundaries for the MCNP calculation are circles.

[17] We are very thankful to Pablo Vaquer for providing this multigroup MCNP reference solution.

[18] The number of unknowns is the product of the number of nodes n, moments $P = (N+1)(N+2)/2$ and energy groups $G = 7$.

[19] The total number of angles per quadrant is then N_aN_p.

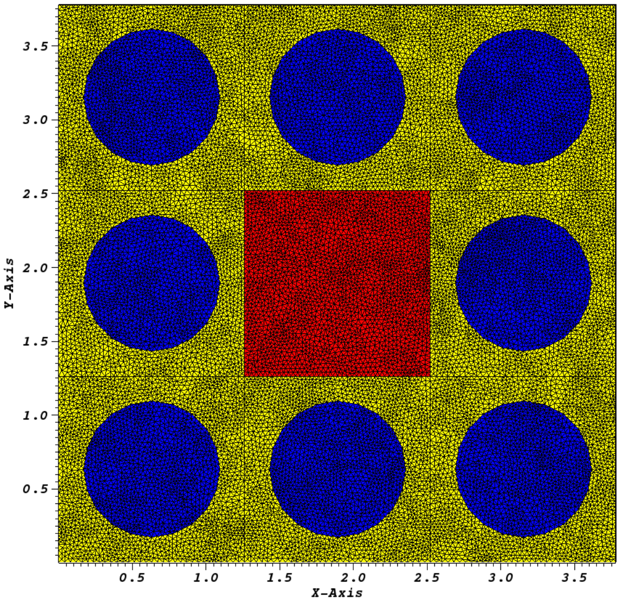


Figure 4.3: Geometry of a 3x3 pin cell test problem. The regions in blue, yellow and red respectively correspond to the fuel, moderator and void. The former two use the cross-sections of the C5G7 benchmark [1]. The latter is in practice chosen such that $\sigma_\mathrm{t} = 10^{-10}$ cm^{-1}. The fuel boundary is approximated with a 20-side polygon. The meshes with a refinement of 0, 1, 2 (shown) and 3 have respectively 1116, 4829, 21090 and 92912 nodes. Besides, they respectively have 2134, 9455, 41776 and 184962 elements and 3249, 14283, 62865 and 277873 sides.

it appeared to converge faster than the Level-Symmetric quadrature rule. The LS method once again exhibits a very slow convergence while the SAAF–VT–S$_N$ method, which is globally conservative, converges significantly faster.

In the next chapter, we will introduce the SAAF–CLS method, which has global conservation and void compatibility, and we will see that the results on that test problem are greatly improved, compared to the LS methods. The updated results can be found in Tables 5.3 and 5.4.

	LS–P$_N$			
N	Ref $=0$	Ref $=1$	Ref $=2$	Ref $=3$
1	56391	52543	51219	50676
3	24370	20159	19083	18771
5	15091	10382	9208	8901
7	12107	7063	5755	5413
9	10705	5407	3974	3586
19	9431	3625	1845	1273
29	9163	3223	1352	722
39	9087	3110	1210	561

Table 4.1: Error $k_{\text{eff}} - \tilde{k}_{\text{eff}}$ (in pcm) for the LS–P$_N$ method. ”Ref” designates the mesh refinement level. The standard deviation on $\tilde{k}_{\text{eff}}$ is 5 pcm.

	SAAF–VT–S$_N$			LS–S$_N$		
(N_p, N_a)	Ref $=0$	Ref $=1$	Ref $=2$	Ref $=0$	Ref $=1$	Ref $=2$
$(2,12)$	67	-2	-8	9042	2955	1045
$(2,24)$	83	9	-2	9033	2988	1075
$(2,48)$	87	16	6	9004	2978	1080
$(2,96)$	87	17	7	8995	2972	1077
$(3,12)$	54	-16	-22	9121	3008	1056
$(3,24)$	71	-4	-16	9112	3042	1087
$(3,48)$	75	3	-8	9083	3032	1091
$(3,96)$	75	3	-7	9074	3025	1089

Table 4.2: Error $k_{\text{eff}} - \tilde{k}_{\text{eff}}$ (in pcm) for the SAAF–VT–S$_N$ and LS–S$_N$ methods. ”Ref” designates the mesh refinement level. The standard deviation on $\tilde{k}_{\text{eff}}$ is 5 pcm.

4.7 Conclusion

In this chapter, we have studied various second-order forms in terms of global conservation and void compatibility, including parity-based methods. We have in particular shown how crucial it is for a scheme to have this former property. At this point, we do not know of any second-order form giving satisfying results for

P_N discretizations and having both of these features. The motivation of Chapter 5 lies in creating one.

5. THE SAAF–CLS METHOD: A GLOBALLY CONSERVATIVE VOID-COMPATIBLE SECOND ORDER FORM*

There does not exist – to our knowledge – any second-order form for P_N that would result in a globally conservative and void compatible scheme. The driving purpose of the present chapter is to develop such a method. However, this work will not be limited to P_N. The basic idea is to decompose our domain into two regions: a non-void region noted $\mathcal{D}_1$ discretized using the standard SAAF formulation, which was shown to be globally conservative in Section 4.3; and a void region noted $\mathcal{D}_0$ using the LS formulation compatible with voids.

In this chapter*, we identify why the LS formulation in void is only globally conservative if the discretization error in $\mathcal{D}_0$ can be neglected, i.e. upon convergence to the analytical solution. A conservative fix is derived, yielding the Conservative LS (CLS) method. It can also be extended to treating regions with uniform cross-section (which is in particular useful for near-void regions). Nevertheless, this additional term – just like the void treatment of the SAAF–VT method [38] – breaks the symmetry of the bilinear form. Next, the variational formulation of the hybrid SAAF–CLS scheme is derived and is shown to be globally conservative.

The remainder of this chapter is structured as follows. After introducing some notation in Section 5.1, we show, in Section 5.2, what term is missing for the LS formulation to be globally conservative in void. We then derive the SAAF–CLS method in Section 5.3, achieving void compatibility and global conservation with an appropriate choice of the scaling between the SAAF and CLS terms. We discuss the

*Part of this chapter has been submitted as "Globally Conservative, Hybrid Self-Adjoint Angular Flux and Least-Squares Method Compatible with Void" by Vincent M. Laboure and Ryan G. McClarren and Yaqi Wang, 2016, to *Nuclear Science and Engineering* [71].

actual implementation of the method and study the results on (i) a slab geometry pure absorber problem, specifically looking at the importance of the conservative fix; and (ii) the multigroup heterogeneous k-eigenvalue problem with a void region, introduced in Section 4.6.3. We then study the possibility of generalizing this method. In Section 5.4.1, we no longer require σ_t to be zero in the void regions but only to be constant therein, particularly addressing the treatment of near-void regions. This is used in Section 5.4.2 to run the dog leg void duct problem, a near-void benchmark introduced by Kobayashi et al [2]. We mention the possibility of extending this work to time-dependent problems in Section 5.4.3.

5.1 Notation

We decompose the spatial domain as $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_0$ (with $\int_{\mathcal{D}_1 \cap \mathcal{D}_0} \mathrm{d}r = 0$) where $\sigma_\mathrm{t} = 0$ in $\mathcal{D}_0$. The interface between $\mathcal{D}_0$ and $\mathcal{D}_1$ is noted $\Gamma = \mathcal{D}_0 \cap \mathcal{D}_1$. We refer to the continuous finite element space corresponding to $\mathcal{D}$, $\mathcal{D}_0$ and $\mathcal{D}_1$ as V, V_0 and V_1, respectively.

As a reminder, we recall the LS weak formulation (see (4.23)) that we apply to $\mathcal{D}_0$: find $\Psi \in V_0$ such that for all $\Psi^* \in V_0$,

$$(L\Psi^*, L\Psi)_{\mathcal{D}_0} + \langle c\Psi^*, (\Psi - \Psi^{\mathrm{inc}})\rangle^-_{\partial\mathcal{D}_0} = (L\Psi^*, H\psi + S)_{\mathcal{D}_0}\,, \tag{5.1}$$

where L and H are respectively representing the streaming plus collision and scattering plus fission operators. Their precise definitions are given by Eqs. 4.2 and 4.3.

In the previous chapter, c designated a parameter with the units of a macroscopic cross-section. In this chapter, we need to further assume that it is a strictly positive constant.

5.2 Conservative Least-Squares Method

In this section, we point out why LS is not globally conservative and propose a fix which constitutes the essence of the CLS method.

5.2.1 Lack of Conservation

We start with the LS formulation in void applied to $\mathcal{D}_0$: find $\Psi \in V_0$ such that for all $\Psi^* \in V_0$,

$$\left(\vec{\Omega} \cdot \vec{\nabla}\Psi^*, \vec{\Omega} \cdot \vec{\nabla}\Psi\right)_{\mathcal{D}_0} + \langle c\Psi^*, (\Psi - \Psi^{\text{inc}})\rangle^-_{\partial\mathcal{D}_0} = 0. \tag{5.2}$$

Using the divergence theorem to transform $\langle c\Psi^\star, \Psi\rangle^-_{\partial\mathcal{D}_0}$, it becomes:

$$\begin{aligned}\left(\vec{\Omega} \cdot \vec{\nabla}\Psi^*, \vec{\Omega} \cdot \vec{\nabla}\Psi\right)_{\mathcal{D}_0} - \left(c\Psi^*, \vec{\Omega} \cdot \vec{\nabla}\Psi\right)_{\mathcal{D}_0} - \left(\vec{\Omega} \cdot \vec{\nabla}\Psi^*, c\Psi\right)_{\mathcal{D}_0} + \langle c\Psi^*, \Psi\rangle^+_{\partial\mathcal{D}_0} \\ - \langle c\Psi^*, \Psi^{\text{inc}}\rangle^-_{\partial\mathcal{D}_0} = 0.\end{aligned} \tag{5.3}$$

In particular, for the constant test function $\Psi^* = 1$, we have:

$$-\left(c, \vec{\Omega} \cdot \vec{\nabla}\Psi\right)_{\mathcal{D}_0} + \langle c, \Psi\rangle^+_{\partial\mathcal{D}_0} - \langle c, \Psi^{\text{inc}}\rangle^-_{\partial\mathcal{D}_0} = 0, \tag{5.4}$$

which is a global conservation statement in $\mathcal{D}_0$ if and only if $(1, \vec{\Omega} \cdot \vec{\nabla}\Psi)_{\mathcal{D}_0} = 0$, i.e. if the discretization error in $\mathcal{D}_0$ is negligible.[1] While the analytical solution does satisfy this relation, nothing can be said about the numerical solution. This LS formulation is thus only globally conservative upon convergence of the numerical solution to the analytical solution, which is expected because of the consistency of the discretization.

[1] Here, we have used the assumption that c is constant. If not, Eq. 5.4 is generally not a conservation statement, even if $(1, \vec{\Omega} \cdot \vec{\nabla}\Psi)_{\mathcal{D}_0} = 0$.

5.2.2 Conservative Fix

From the previous expression, it is clear that the scheme would be globally conservative if we were to add $(c\Psi^*, \vec{\Omega} \cdot \vec{\nabla}\Psi)_{\mathcal{D}_0}$ to the variational formulation. Since this term can be obtained directly by testing the transport equation (see Eq. 4.1) applied to $\mathcal{D}_0$ against $c\Psi^*$, the converged solution would not be affected by this change. We therefore define the CLS formulation applied to $\mathcal{D}_0$ to be: find $\Psi \in V_0$ such that for all $\Psi^* \in V_0$,

$$\left(\vec{\Omega} \cdot \vec{\nabla}\Psi^*, \vec{\Omega} \cdot \vec{\nabla}\Psi\right)_{\mathcal{D}_0} + \left(c\Psi^*, \vec{\Omega} \cdot \vec{\nabla}\Psi\right)_{\mathcal{D}_0} + \langle c\Psi^*, (\Psi - \Psi^{\text{inc}})\rangle^-_{\partial\mathcal{D}_0} = 0. \tag{5.5}$$

Alternatively, splitting the boundary terms depending on whether they belong to $\partial\mathcal{D}$ or to Γ, it can be expressed as:

$$\begin{aligned}\left(\vec{\Omega} \cdot \vec{\nabla}\Psi^*, \vec{\Omega} \cdot \vec{\nabla}\Psi\right)_{\mathcal{D}_0} &- \left(\vec{\Omega} \cdot \vec{\nabla}\Psi^*, c\Psi\right)_{\mathcal{D}_0} \\ &+ \langle c\Psi^*, \Psi\rangle^+_{\partial\mathcal{D}^0} - \langle c\Psi^*, \Psi^{\text{inc}}\rangle^-_{\partial\mathcal{D}^0} + \langle c\Psi^*, \Psi\rangle^{+,0}_{\Gamma} - \langle c\Psi^*, \Psi^{\text{inc}}\rangle^{-,0}_{\Gamma} = 0,\end{aligned} \tag{5.6}$$

where $\partial\mathcal{D}^0 \equiv \partial\mathcal{D} \cap \partial\mathcal{D}_0$. In this expression, we have also used the notation $\langle\cdot,\cdot\rangle^{\pm,0}_{\Gamma}$ to indicate that the angular integration half-range $\pm\vec{\Omega} \cdot \vec{n}(\vec{r}) > 0$ is determined with $\vec{n}$ being the outward unit vector normal to Γ with respect to $\mathcal{D}_0$ (i.e. locally pointing towards $\mathcal{D}_1$) . This formulation is globally conservative but is not symmetric.

5.3 SAAF–CLS Method

In this section, we create a hybrid method combining the CLS terms in $\mathcal{D}_0$ and the SAAF terms in $\mathcal{D}_1$. The variational formulation is derived and numerical results are presented.

5.3.1 Variational Formulation

Choosing $\Psi = \sigma_{\mathrm{t}}^{-1}(-\vec{\Omega}\cdot\vec{\nabla}\Psi + H\Psi + S)$ as the angular flux equation (the so-called first AFE in [38], see also Eq. 4.7), the SAAF formulation applied to $\mathcal{D}_1$ is given by: find $\Psi \in V_1$ such that for all $\Psi^* \in V_1$,

$$\begin{aligned}&\left(\vec{\Omega}\cdot\vec{\nabla}\Psi^*, \frac{1}{\sigma_{\mathrm{t}}}\vec{\Omega}\cdot\vec{\nabla}\Psi\right)_{\mathcal{D}_1} + (\sigma_t\Psi^*, \Psi)_{\mathcal{D}_1} + \langle\Psi^*, \Psi\rangle^+_{\partial\mathcal{D}^1} - \langle\Psi^*, \Psi^{\mathrm{inc}}\rangle^-_{\partial\mathcal{D}^1}\\ &+ \langle\Psi^*, \Psi\rangle^{+,1}_{\Gamma} - \langle\Psi^*, \Psi^{\mathrm{inc}}\rangle^{-,1}_{\Gamma} = \left(\frac{1}{\sigma_{\mathrm{t}}}\vec{\Omega}\cdot\vec{\nabla}\Psi^* + \Psi^*, H\Psi + S\right)_{\mathcal{D}_1},\end{aligned} \tag{5.7}$$

where $\partial\mathcal{D}^1 \equiv \partial\mathcal{D} \cap \partial\mathcal{D}_1$. We scale Eq. 5.6 with a constant $\sigma > 0$ with units of a cross-section, for consistency. We then combine it with Eq. 5.7 and notice that Ψ is continuous across Γ (i.e. $\Psi = \Psi^{\mathrm{inc}}$ on Γ) to end up with:

$$\begin{aligned}&\left(\vec{\Omega}\cdot\vec{\nabla}\Psi^*, \frac{1}{\sigma_{\mathrm{t}}}\vec{\Omega}\cdot\vec{\nabla}\Psi\right)_{\mathcal{D}_1} + (\sigma_{\mathrm{t}}\Psi^*, \Psi)_{\mathcal{D}_1} + \left(\vec{\Omega}\cdot\vec{\nabla}\Psi^*, \frac{1}{\sigma}\vec{\Omega}\cdot\vec{\nabla}\Psi\right)_{\mathcal{D}_0} - \left(\vec{\Omega}\cdot\vec{\nabla}\Psi^*, \frac{c}{\sigma}\Psi\right)_{\mathcal{D}_0}\\ &+ \langle\frac{c}{\sigma}\Psi^*, \Psi\rangle^+_{\partial\mathcal{D}^0} - \langle\frac{c}{\sigma}\Psi^*, \Psi^{\mathrm{inc}}\rangle^-_{\partial\mathcal{D}^0} + \langle\frac{c}{\sigma}\Psi^*, \Psi\rangle^0_{\Gamma}\\ &+ \langle\Psi^*, \Psi\rangle^+_{\partial\mathcal{D}^1} - \langle\Psi^*, \Psi^{\mathrm{inc}}\rangle^-_{\partial\mathcal{D}^1} + \langle\Psi^*, \Psi\rangle^1_{\Gamma} = \left(\frac{1}{\sigma_{\mathrm{t}}}\vec{\Omega}\cdot\vec{\nabla}\Psi^* + \Psi^*, H\Psi + S\right)_{\mathcal{D}_1}.\end{aligned} \tag{5.8}$$

Global conservation imposes the following condition:

$$c = \sigma. \tag{5.9}$$

The SAAF–CLS weak formulation is then given by: find $\Psi \in V$ such that for all $\Psi^* \in V$,

$$\left(\vec{\Omega} \cdot \vec{\nabla}\Psi^*, \frac{1}{\sigma_\mathrm{t}}\vec{\Omega} \cdot \vec{\nabla}\Psi\right)_{\mathcal{D}_1} + \left(\vec{\Omega} \cdot \vec{\nabla}\Psi^*, \frac{1}{c}\vec{\Omega} \cdot \vec{\nabla}\Psi\right)_{\mathcal{D}_0} + (\sigma_t\Psi^*, \Psi)_{\mathcal{D}_1} - \left(\vec{\Omega} \cdot \vec{\nabla}\Psi^*, \Psi\right)_{\mathcal{D}_0}$$
$$+ \langle\Psi^*, \Psi\rangle^+_{\partial\mathcal{D}} - \langle\Psi^*, \Psi^{\mathrm{inc}}\rangle^-_{\partial\mathcal{D}} = \left(\frac{1}{\sigma_\mathrm{t}}\vec{\Omega} \cdot \vec{\nabla}\Psi^* + \Psi^*, H\Psi + S\right)_{\mathcal{D}_1} . \tag{5.10}$$

One can check that this scheme is globally conservative by choosing $\Psi^* = 1$. The non-symmetric term is very similar to the extra term in the SAAF–VT formulation (see Eq. 4.10). The difference between the two formulations however is that in void regions, the second-order streaming term over $\mathcal{D}_0$ in Eq. 5.10 does not vanish, even when the mesh is infinitely refined. This is crucial to avoid having a singular term when using a P_N expansion.

5.3.2 Implementation

The method derived above is implemented in Rattlesnake, the transport solver from the Idaho National Laboratory based on the MOOSE framework [54]. All the results presented below are obtained with the first order LAGRANGE elements from libMesh [56]. We use the Preconditioned Jacobian Free Newton Krylov (PJFNK) method for the nonlinear solves with the PETSc [55] restarted generalized minimal residual (GMRES) solver for the linear solves. Preconditioning is done through the built-in preconditioners in PETSc, either the algebraic multigrid Hypre BoomerAMG [72] or the block jacobi preconditioners.

In practice, the implementation of (5.10) is very simple because it does not require any more kernels and boundary conditions than the SAAF–VT formulation. As a matter of fact, one can notice that (5.10) can also be directly obtained from (4.10) by defining $\tau \equiv 1/\sigma_{\mathrm{t}}$ in $\mathcal{D}_1$ and $\tau \equiv 1/c$ in $\mathcal{D}_0$.

5.3.3 SAAF–CLS Scaling Factor c

Numerically, changing the value of c can have a significant impact on the solver convergence. While its optimal value seems to be problem dependent (as well as dependent on N), it is interesting to note that it has the same units as a cross-section. If σ_{t} is constant on Γ, choosing c in the same order usually works well. If σ_{t} is not constant thereon, its choice is not as obvious but can be typically picked between the extrema. In Section 5.3.5.2, we show that it has little impact on the numerical solution outside the void region, which means in particular that the reaction rates will be minimally affected.

In addition, it is worth noting that, although we have assumed c to be constant for the derivation of the SAAF–CLS method[2], this assumption is actually not required by the final formulation (see Eq. 5.10). In other words, choosing a spatially-dependent c in $\mathcal{D}_0$ does not compromise the global conservation property of the method.[3]

Unless otherwise specified, c is set to 1 cm^{-1}.

5.3.4 Terminology

In the following sections, several methods are being compared. Here, we specify what is precisely meant by each of these. SAAF–CLS refers to Eq. 5.10.[4] To highlight why the conservative fix introduced in Eq. 5.4 is crucial, we also show the results of the same method without the conservative fix and refer to it as SAAF–LS. Because this method is only useful for comparison purposes, its derivation is shown in Ap-

[2] This assumption was then required to claim that Eq. 5.4 was a conservation statement.

[3] The fundamental reason is that c only appears in terms containing $\vec{\Omega}\cdot\vec{\nabla}\Psi^*$: a global conservation statement is obtained with $\Psi^* = 1$, regardless of the value of c.

[4] After we have generalized this method to handle near-void regions, it will refer to Eq. 5.21, Eq. 5.10 being the limiting case as $\sigma_0 \longrightarrow 0$. After extending it to time-dependent problems, it will then refer to Eq. 5.28.

pendix D, the weak formulation being given by (D.8). The SAAF–VT formulation refers to Eq. 4.10.

Lastly, we consider the plain LS method which is obtained using Eq. 5.13 over the whole domain. By default, we then choose $c = 1/\tau$, which is consistent with the SAAF–VT formulation in the case of a constant τ, as we saw in Section 4.2.2.

5.3.5 Numerical Results

5.3.5.1 Slab Geometry Pure Absorber Problem

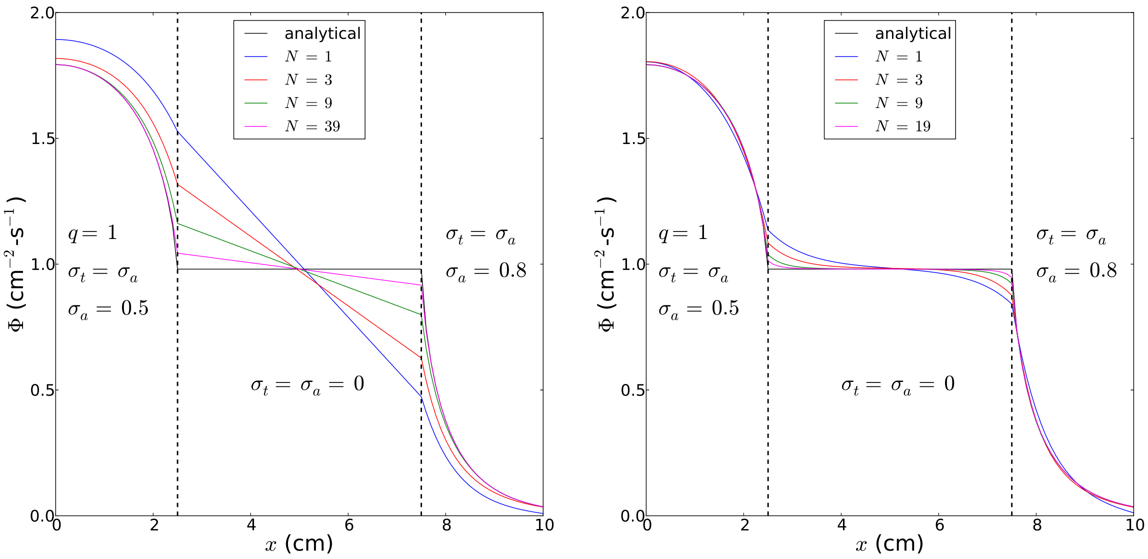
In this section, we consider a slab geometry ($d = 1$) scattering- and fission-free domain ($H = 0$) composed of three distinct uniform regions, defined respectively for $0 \leq x \leq \delta$, $\delta \leq x \leq 3\delta$ and $3\delta \leq x \leq 4\delta$. In the first one, $S \equiv q/w$ and $\sigma_t \equiv \sigma_{a,1}$; the second one is a pure void; in the third one, $S = 0$ and $\sigma_t \equiv \sigma_{a,2}$. The boundaries at $x = 0$ and $x = 4\delta$ respectively are reflecting and vacuum. The analytical scalar flux is given by:

$$\Phi(x) = \frac{q}{w\sigma_{a,1}} \begin{cases} \left(2 - E_2\left(\sigma_{a,1}(\delta - x)\right) - E_2\left(\sigma_{a,1}(\delta + x)\right)\right) & , \quad 0 \leq x < \delta, \\ \left(1 - E_2\left(2\sigma_{a,1}\delta\right)\right) & , \quad \delta \leq x < 3\delta, \\ \left(E_2\left(\sigma_{a,2}(x - 3\delta)\right) - E_2\left(2\sigma_{a,1}\delta + \sigma_{a,2}(x - 3\delta)\right)\right) & , \quad 3\delta \leq x < 4\delta, \end{cases} \tag{5.11}$$

where E_2 represents the following exponential integral:

$$E_2\left(x\right) = \int_1^\infty \frac{\exp(-xz)}{z^2} \mathrm{d}z. \tag{5.12}$$

In practice, we choose q = 1 cm^{-3}–s^{-1}, δ = 2.5 cm, $\sigma_{a,1}$ = 0.5 cm^{-1} and $\sigma_{a,2}$ = 0.8 cm^{-1}. A total of 4096 spatial cells is chosen. Fig. 5.1a highlights why a conservative fix of the LS formulation is necessary (see Eq. 5.5). Without it, the



(a) SAAF–LS–P$_N$

(b) SAAF–CLS–P$_N$

Figure 5.1: Comparison of the scalar flux as a function of x with and without the conservative fix described by Eq. 5.5 for different values of N.

solution in void is clearly inaccurate and the convergence with N is very slow. In particular, we mentioned that the LS formulation is globally conservative if and only if $(1, \vec{\Omega} \cdot \vec{\nabla}\Psi)_{\mathcal{D}_0} = 0$ (see Eq. 5.4), which is clearly not the case in the void region. Fig. 5.1b qualitatively shows the improvement in the results when using the conservative fix. Although we still have $\vec{\Omega} \cdot \vec{\nabla}\Psi \neq 0$ in the void region, the global conservation therein is maintained and the difference with the analytical solution appears to be greatly reduced.

Fig. 5.2 quantifies the L^2-error with the analytical solution. Noteworthy is the fact that using the hybrid SAAF–LS method (without the fix) does not even outperform the plain LS method. However, the SAAF–CLS method clearly does, as a SAAF–CLS–P$_5$ calculation gives an error almost identical to the LS–P$_{59}$ solution. Another interesting feature is that the parity of N matters, odd values of N giving better results for our hybrid method.

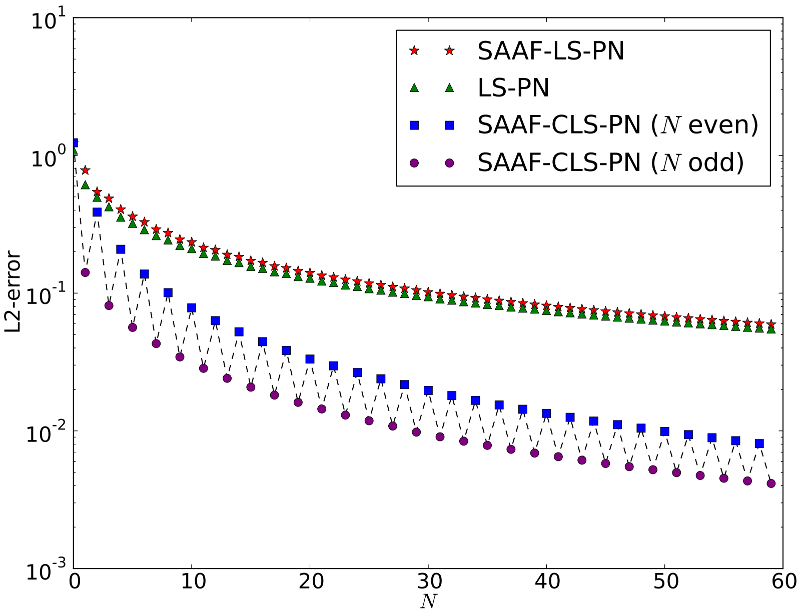


Figure 5.2: Comparison of the L^2-error (in cm$^{-3/2}$–s^{-1}) of the scalar flux Φ for different discretization. In particular, the SAAF–LS–P$_N$ method (i.e. without the conservative fix in the void region) is comparable to the LS–P$_N$ method. The SAAF–CLS–P$_N$ method does much better, especially for odd values of N.

Table 5.1 compares the SAAF–CLS–P$_N$ and SAAF–VT–P$_N$ methods on that same problem. While the L^2-errors of the numerical solution for a given N is close for both methods, the number of GMRES iterations needed to converge grows very quickly for the latter one. This exhibits the conditioning problems it suffers from and makes it impractical for more complicated problems.

5.3.5.2 Modified Reed's Problem

In this section, we wish to study the impact of the scaling factor c on the numerical solution. In particular, we consider the situation where σ_{t} takes very different values

	L^2-error		Iteration count	
N	SAAF–CLS–P$_N$	SAAF–VT–P$_N$	SAAF–CLS–P$_N$	SAAF–VT–P$_N$
0	1.24E-0	1.52E-0	9	7
1	1.41E-1	1.64E-1	35	801
2	3.87E-1	5.96E-1	50	1211
3	8.12E-2	6.58E-2	70	2765
4	2.08E-1	3.68E-1	84	4617
5	5.64E-2	3.77E-2	110	8120

Table 5.1: Comparison of the SAAF–CLS–P$_N$ and SAAF–VT–P$_N$ methods in terms of the L^2-error (in cm$^{-3/2}$–s^{-1}) and the number of GMRES linear iterations. Although the L^2-error is fairly comparable for a given N, the number of iterations rapidly becomes intractable for SAAF–VT–P$_N$.

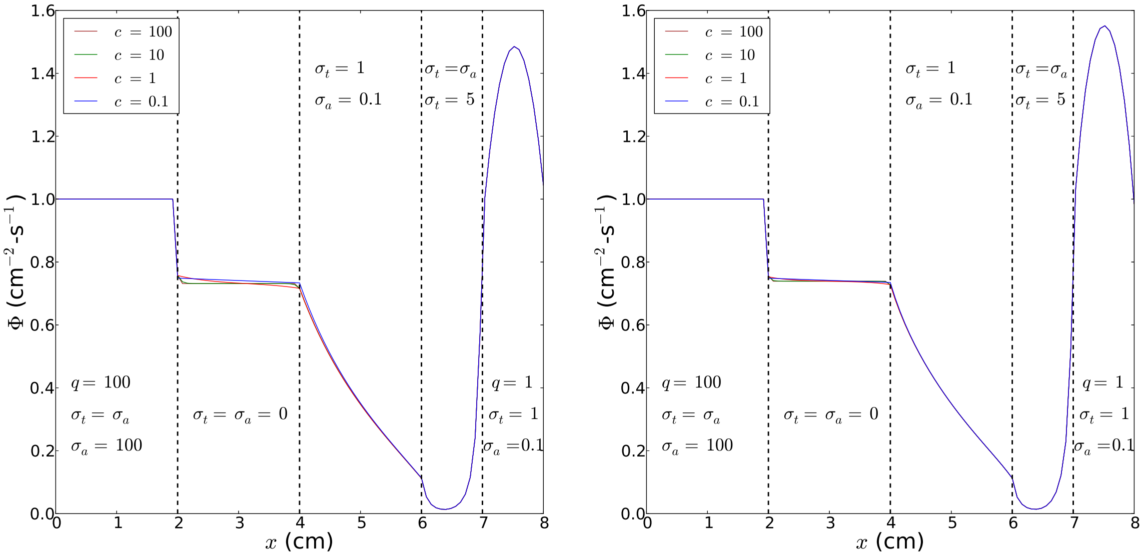
on Γ, in which case the choice of c is not obvious (see Section 5.3.3).

Specifically, we look at a famous test problem, known as the Reed's problem [73] which has significant discontinuities between the different regions of the problem. To accentuate the discontinuity of σ_t on Γ, we slightly modify the problem by reordering the spatial regions. Table 5.2 summarizes the material properties of the problem, defined for $0 \leq x \leq 8$ cm, with reflecting and vacuum boundary conditions respectively imposed at $x = 0$ and at $x = 8$ cm. We also choose 4096 cells, which makes the spatial error negligible for the values of N considered in this section.

	Region 1	Region 2	Region 3	Region 4	Region 5
q	100	0	0	0	1
σ_t	100	0	1	5	1
σ_s	0	0	0.9	0	0.9
Domain	$0 \leq x < 2$	$2 \leq x < 4$	$4 \leq x < 6$	$6 \leq x < 7$	$7 \leq x \leq 8$

Table 5.2: Material properties for the modified Reed's problem: value for the angular–integrated volumetric source q in cm^{-3}–s^{-1} ($S = q/w$) and the total and scattering cross-sections (in cm^{-1}) in each region.

Fig. 5.3 shows the results for the SAAF–CLS–P_N method for different values of c. Fig. 5.3a indicates that it has a small impact on the P_3 numerical solution and that the difference is mostly limited to the void region, the solution elsewhere being virtually identical. Furthermore, the discrepancy is reduced as N is increased, as the P_7 solution given on Fig. 5.3b shows.



(a) P_3

(b) P_7

Figure 5.3: SAAF–CLS–P_N method for different values of c (in cm^{-1}) on the modified Reed's problem.

In conclusion, it indeed appears that the value of c has little influence on the numerical solution[5] outside of the void region, which implies that the reaction rates are minimally affected. Moreover, this impact is further reduced as N is increased.

[5] It can however have an impact on the conditioning of the system.

5.3.5.3 Heterogenous Multigroup k-Eigenvalue Problem with Void

In this section, we go back to the test problem introduced in Section 4.6.3, which exhibited how important having globally conservative schemes can be. We show updated results, by considering those obtained with the SAAF–CLS–P_N and SAAF–CLS–S_N methods.

Table 5.3 shows the error with respect to $\tilde{k}_{\text{eff}}$ for the LS–P_N and the SAAF–CLS–P_N methods. As a reminder, the former, lacking global conservation, is extremely slow to converge. The latter gives much better results as any solution with $N \geq 3$ yields an error smaller than the LS–P_{39} solution, which has over 5.3×10^8 unknowns. In particular, the most refined calculations are within a few standard deviations.[6]

Table 5.4 compares the same quantity for the LS–S_N, SAAF–CLS–S_N and SAAF–VT–S_N methods as a function of the number of polar and azimuthal angles per quadrant, noted N_p and N_a respectively. The same trend can be noticed: the globally conservative methods are much more accurate. In particular, the spatial error dominates the calculations shown if $N_p \geq 2$. It is also interesting to note that SAAF–CLS–S_N and SAAF–VT–S_N give very similar results, which is expected since the two variational formulations are so close.

[6] Recall also that the reference solution uses circles as the pin boundaries while they are approximated with 20-side polygons for the other methods.

	SAAF–CLS–P$_N$				LS–P$_N$			
N	Ref = 0	Ref = 1	Ref = 2	Ref = 3	Ref = 0	Ref = 1	Ref = 2	Ref = 3
1	946	903	894	892	56391	52543	51219	50676
3	312	227	208	238	24370	20159	19083	18771
5	111	14	-8	-14	15091	10382	9208	8901
7	50	-48	-72	-78	12107	7063	5755	5413
9	38	-59	-83	-89	10705	5407	3974	3586
19	66	-23	-42	-47	9431	3625	1845	1273
29	79	-6	-23	-25	9163	3223	1352	722
39	85	1	-16	-18	9087	3110	1210	561

Table 5.3: Error $k_{\text{eff}} - \tilde{k}_{\text{eff}}$ (in pcm) for the SAAF–CLS–P$_N$ and LS–P$_N$ methods. ”Ref” designates the mesh refinement level. The standard deviation on $\tilde{k}_{\text{eff}}$ is 5 pcm.

	SAAF–CLS–S$_N$			SAAF–VT–S$_N$			LS–S$_N$		
(N_p, N_a)	Ref=0	Ref=1	Ref=2	Ref=0	Ref=1	Ref=2	Ref=0	Ref=1	Ref=2
$(1, 12)$	350	276	260	337	276	271	8570	3045	1284
$(1, 24)$	357	291	278	355	289	278	8561	3079	1316
$(1, 48)$	358	294	282	359	296	287	8530	3068	1320
$(1, 96)$	358	294	283	359	296	288	8521	3062	1318
$(2, 12)$	91	2	-17	67	-2	-8	9042	2955	1045
$(2, 24)$	98	15	-2	83	9	-2	9033	2988	1075
$(2, 48)$	98	18	3	87	16	6	9004	2978	1080
$(2, 96)$	98	18	3	87	17	7	8995	2972	1077
$(3, 12)$	81	-11	-31	54	-16	-22	9121	3008	1056
$(3, 24)$	87	3	-16	71	-4	-16	9112	3042	1087
$(3, 48)$	88	6	-11	75	3	-8	9083	3032	1091
$(3, 96)$	88	6	-11	75	3	-7	9074	3025	1089

Table 5.4: Error $k_{\text{eff}} - \tilde{k}_{\text{eff}}$ (in pcm) for the SAAF–CLS–S$_N$, SAAF–VT–S$_N$ and LS–S$_N$ methods. ”Ref” designates the mesh refinement level. The standard deviation on $\tilde{k}_{\text{eff}}$ is 5 pcm.

5.4 Generalized SAAF–CLS Method

In this section, we study the possibility of generalizing the SAAF–CLS method. First, we relax the assumption $\sigma_{\text{t}} = 0$ in $\mathcal{D}_0$ by allowing it to be a non-zero constant.

We then show some results on the dog leg void duct problem. Next, we discuss the eventuality of considering time-dependent problems and of extending the method to parity-based schemes. While the former appears to be a minor complication, the latter most likely induces the loss of void compatibility.

5.4.1 Extension to Near-Void Regions

We show that we can similarly derive a void compatible, globally conservative scheme in the more general setting of a uniform non-void region in $\mathcal{D}_0$. The importance of this result lies in the fact that real-world applications rarely contain pure void regions but more realistically near-void regions. The only different assumption is thus that we no longer require $\sigma_\mathrm{t} = 0$ in $\mathcal{D}_0$ but only to be uniform therein, i.e. $\sigma_\mathrm{t} = \sigma_0$ in $\mathcal{D}_0$. Although the driving application is the treatment of near-void regions, we do not need to assume that σ_0 is small for the reasoning in this section to hold.

5.4.1.1 Generalized CLS Method

The LS formulation compatible with voids applied to $\mathcal{D}_0$ is now given by [44, 67]:

$$(L\Psi^*, L\Psi)_{\mathcal{D}_0} + \langle c\Psi^*, (\Psi - \Psi^{\mathrm{inc}})\rangle^-_{\partial\mathcal{D}_0} = (L\Psi^*, H\psi + S)_{\mathcal{D}_0} . \tag{5.13}$$

As we saw in Section 4.3, LS is globally conservative if and only if σ_t is a strictly positive constant and if we have $c = \sigma_\mathrm{t}$. While the first condition is included in our assumptions, we cannot satisfy the second in all generality because the boundary terms would vanish in void or near-void regions. In summary, this means that, unless $c = \sigma_\mathrm{t}$, Eq. 5.13 is only globally conservative upon convergence to the analytical solution, even though σ_t is constant over $\mathcal{D}_0$.

Just as in Section 5.3, we can define the Conservative Least-Squares formulation on $\mathcal{D}_0$ by adding the term $((c-\sigma_0)\Psi^*, \vec{\Omega}\cdot\vec{\nabla}\Psi+\sigma_0\Psi-H\psi-S)_{\mathcal{D}_0}$ to the LS formulation, which is consistent with the transport equation:

$$\begin{aligned}(L\Psi^*, L\Psi)_{\mathcal{D}_0} &+ \langle c\Psi^*, (\Psi - \Psi^{\mathrm{inc}})\rangle^{-}_{\partial\mathcal{D}_0} \\ &+ \Big((c-\sigma_0)\Psi^*, \vec{\Omega}\cdot\vec{\nabla}\Psi + \sigma_0\Psi - H\psi - S\Big)_{\mathcal{D}_0} = (L\Psi^*, H\Psi + S)_{\mathcal{D}_0}.\end{aligned} \tag{5.14}$$

Equivalently, using the divergence theorem on the first and third term, it can be expressed as:

$$\begin{aligned}&\Big(\vec{\Omega}\cdot\vec{\nabla}\Psi^*, \vec{\Omega}\cdot\vec{\nabla}\Psi\Big)_{\mathcal{D}_0} + (\sigma_0\Psi^*, \sigma_0\Psi)_{\mathcal{D}_0} + \langle\sigma_0\Psi^*, \Psi\rangle_{\partial\mathcal{D}_0} + \langle c\Psi^*, (\Psi - \Psi^{\mathrm{inc}})\rangle^{-}_{\partial\mathcal{D}_0} \\ &+ \langle(c-\sigma_0)\Psi^*, \Psi\rangle_{\partial\mathcal{D}_0} - \Big((c-\sigma_0)\vec{\Omega}\cdot\vec{\nabla}\Psi^*, \Psi\Big)_{\mathcal{D}_0} + ((c-\sigma_0)\Psi^*, \sigma_0\Psi - H\psi - S)_{\mathcal{D}_0} \\ &\qquad = (L\Psi^*, H\Psi + S)_{\mathcal{D}_0},\end{aligned} \tag{5.15}$$

i.e.:

$$\begin{aligned}&\Big(\vec{\Omega}\cdot\vec{\nabla}\Psi^*, \vec{\Omega}\cdot\vec{\nabla}\Psi\Big)_{\mathcal{D}_0} + (\sigma_0\Psi^*, \sigma_0\Psi)_{\mathcal{D}_0} + \langle c\Psi^*, (\Psi - \Psi^{\mathrm{inc}})\rangle^{-}_{\partial\mathcal{D}_0} \\ &+ \langle c\Psi^*, \Psi\rangle_{\partial\mathcal{D}_0} - \Big((c-\sigma_0)\vec{\Omega}\cdot\vec{\nabla}\Psi^*, \Psi\Big)_{\mathcal{D}_0} + ((c-\sigma_0)\Psi^*, \sigma_0\Psi - H\psi - S)_{\mathcal{D}_0} \\ &\qquad = (L\Psi^*, H\Psi + S)_{\mathcal{D}_0}.\end{aligned} \tag{5.16}$$

At the end of the day, the formulation can be expressed as: find $\Psi \in V_0$ such that for all $\Psi^* \in V_0$,

$$\begin{aligned}&\Big(\vec{\Omega}\cdot\vec{\nabla}\Psi^*, \vec{\Omega}\cdot\vec{\nabla}\Psi\Big)_{\mathcal{D}_0} + (c\Psi^*, \sigma_0\Psi)_{\mathcal{D}_0} + \langle c\Psi^*, \Psi\rangle^{+}_{\partial\mathcal{D}_0} - \langle c\Psi^*, \Psi^{\mathrm{inc}}\rangle^{-}_{\partial\mathcal{D}_0} \\ &- \Big((c-\sigma_0)\vec{\Omega}\cdot\vec{\nabla}\Psi^*, \Psi\Big)_{\mathcal{D}_0} = (\vec{\Omega}\cdot\vec{\nabla}\Psi^* + c\Psi^*, H\Psi + S)_{\mathcal{D}_0}.\end{aligned} \tag{5.17}$$

One can check that this scheme is globally conservative by choosing $\Psi^* = 1$ and that it does reduce to Eq. 5.6 in void.

5.4.1.2 Generalized SAAF–CLS Method

Using the fact that $\Psi = \Psi^{\mathrm{inc}}$ on Γ and scaling it with a constant σ for units consistency with the SAAF weak formulation, the previous LS formulation becomes:

$$\begin{aligned}&\left(\frac{1}{\sigma}\vec{\Omega}\cdot\vec{\nabla}\Psi^*, \vec{\Omega}\cdot\vec{\nabla}\Psi\right)_{\mathcal{D}_0} + \left(\frac{c}{\sigma}\Psi^*, \sigma_0\Psi\right)_{\mathcal{D}_0} + \langle\frac{c}{\sigma}\Psi^*, \Psi\rangle^{+}_{\partial\mathcal{D}^0} - \langle\frac{c}{\sigma}\Psi^*, \Psi^{\mathrm{inc}}\rangle^{-}_{\partial\mathcal{D}^0} \\ &+ \langle\frac{c}{\sigma}\Psi^*, \Psi\rangle^{0}_{\Gamma} - \left(\frac{c-\sigma_0}{\sigma}\vec{\Omega}\cdot\vec{\nabla}\Psi^*, \Psi\right)_{\mathcal{D}_0} = (\frac{1}{\sigma}\vec{\Omega}\cdot\vec{\nabla}\Psi^* + \frac{c}{\sigma}\Psi^*, H\Psi + S)_{\mathcal{D}_0},\end{aligned} \tag{5.18}$$

where, as a reminder, $\partial\mathcal{D}^0 \equiv \partial\mathcal{D} \cap \mathcal{D}_0$ and the superscript in the notation $\langle\cdot,\cdot\rangle^0_\Gamma$ is used to indicate that the unit vector normal on Γ is locally pointing towards the outside of $\mathcal{D}_0$.

The generalized SAAF–CLS formulation is then obtained by adding the terms from Eq. 5.7:

$$\begin{aligned}&\left(\vec{\Omega}\cdot\vec{\nabla}\Psi^*, \frac{1}{\sigma_{\mathrm{t}}}\vec{\Omega}\cdot\vec{\nabla}\Psi\right)_{\mathcal{D}_1} + \left(\frac{1}{\sigma}\vec{\Omega}\cdot\vec{\nabla}\Psi^*, \vec{\Omega}\cdot\vec{\nabla}\Psi\right)_{\mathcal{D}_0} + (\sigma_t\Psi^*, \Psi)_{\mathcal{D}_1} + \left(\frac{c}{\sigma}\Psi^*, \sigma_0\Psi\right)_{\mathcal{D}_0} \\ &\quad + \langle\Psi^*, \Psi\rangle^{+}_{\partial\mathcal{D}^1} - \langle\Psi^*, \Psi^{\mathrm{inc}}\rangle^{-}_{\partial\mathcal{D}^1} + \langle\frac{c}{\sigma}\Psi^*, \Psi\rangle^{+}_{\partial\mathcal{D}^0} - \langle\frac{c}{\sigma}\Psi^*, \Psi^{\mathrm{inc}}\rangle^{-}_{\partial\mathcal{D}^0} \\ &\quad + \langle\Psi^*, \Psi\rangle^{1}_{\Gamma} + \langle\frac{c}{\sigma}\Psi^*, \Psi\rangle^{0}_{\Gamma} - \left(\frac{c-\sigma_0}{\sigma}\vec{\Omega}\cdot\vec{\nabla}\Psi^*, \Psi\right)_{\mathcal{D}_0} \\ &\qquad = \left(\frac{1}{\sigma_{\mathrm{t}}}\vec{\Omega}\cdot\vec{\nabla}\Psi^* + \Psi^*, H\Psi + S\right)_{\mathcal{D}_1} + \left(\frac{1}{\sigma}\vec{\Omega}\cdot\vec{\nabla}\Psi^* + \frac{c}{\sigma}\Psi^*, H\Psi + S\right)_{\mathcal{D}_0}.\end{aligned} \tag{5.19}$$

To have global conservation, the boundary terms on Γ need to vanish, i.e.:

$$c = \sigma, \tag{5.20}$$

which the same condition as in the pure void case. The weak formulation is then given by: find $\Psi \in V$ such that for all $\Psi^* \in V$,

$$\begin{aligned}
&\left(\vec{\Omega}\cdot\vec{\nabla}\Psi^*, \frac{1}{\sigma_\mathrm{t}}\vec{\Omega}\cdot\vec{\nabla}\Psi\right)_{\mathcal{D}_1} + \left(\vec{\Omega}\cdot\vec{\nabla}\Psi^*, \frac{1}{c}\vec{\Omega}\cdot\vec{\nabla}\Psi\right)_{\mathcal{D}_0} + (\Psi^*, \sigma_t\Psi)_{\mathcal{D}_1} + (\Psi^*, \sigma_0\Psi)_{\mathcal{D}_0} \\
&- \left(\left(1-\frac{\sigma_0}{c}\right)\vec{\Omega}\cdot\vec{\nabla}\Psi^*, \Psi\right)_{\mathcal{D}_0} + \langle\Psi^*, \Psi\rangle^+_{\partial\mathcal{D}} - \langle\Psi^*, \Psi^{\mathrm{inc}}\rangle^-_{\partial\mathcal{D}} \\
&\quad = \left(\frac{1}{\sigma_\mathrm{t}}\vec{\Omega}\cdot\vec{\nabla}\Psi^* + \Psi^*, H\Psi + S\right)_{\mathcal{D}_1} + \left(\frac{1}{c}\vec{\Omega}\cdot\vec{\nabla}\Psi^* + \Psi^*, H\Psi + S\right)_{\mathcal{D}_0} .
\end{aligned} \tag{5.21}$$

Once again, this formulation is globally conservative, non-symmetric and reduces to Eq. 5.10 if $\mathcal{D}_0$ is pure void (in which case $\sigma_0 = 0$ and $H = S = 0$ in $\mathcal{D}_0$). This formulation is still identical to the SAAF-VT formulation where we have defined $\tau \equiv 1/\sigma_\mathrm{t}$ in $\mathcal{D}_1$ and $\tau \equiv 1/c$ in $\mathcal{D}_0$ (see Eq. 4.10).

5.4.2 Dog Leg Void Duct Problem

In this section, we consider the third benchmark problem introduced by Kobayashi et al in [2], also called the dog leg void duct problem. Here, we only show results for the pure absorber problem. The rectangular spatial domain is defined for $0 \leq x, z \leq 60$ cm and $0 \leq y \leq 100$ cm. The geometry of the problem is shown in Fig. 5.4 and consists of three uniform materials. First, a source region for $\max(x, y, z) \leq 10$ cm with a volumetric source $S = 1$ cm^{-3}–s^{-1} and $\sigma_\mathrm{t} = \sigma_a = 0.1$ cm^{-1}. Second, a near-void region ($\sigma_\mathrm{t} = \sigma_a = 10^{-4}$ cm^{-1}) for $0 \leq x, z \leq 10$ cm and $10 \leq y \leq 60$ cm; $10 \leq x \leq 40$ cm, $50 \leq y \leq 60$ cm and $0 \leq z \leq 10$ cm; $30 \leq x \leq 40$ cm, $50 \leq y \leq 60$ cm and $10 \leq z \leq 40$ cm; $30 \leq x \leq 40$ cm, $60 \leq y \leq 100$ cm and $30 \leq z \leq 40$ cm. Third, a shield region defined everywhere else by $\sigma_\mathrm{t} = \sigma_a = 0.1$ cm^{-1}. Reflecting boundary conditions are imposed at $x = 0$, $y = 0$ and $z = 0$; vacuum boundary conditions are used at $x = 60$ cm, $y = 100$ cm and $z = 60$ cm.

The coarsest mesh used has 6 × 10 × 6 cube elements and is referred to as the "ref = 0" mesh. Increasing the level of mesh refinement by one essentially multiplies the number of elements by eight.

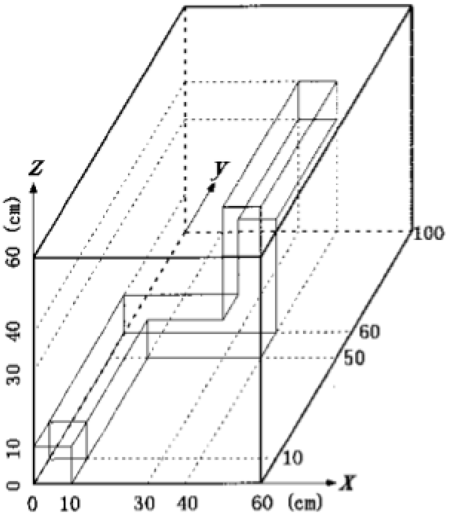


Figure 5.4: Geometry of the dog leg void duct problem (figure taken from "3D radiation transport benchmark problems and results for simple geometries with void region" by Keisuke Kobayashi et. al. [2]).

The interest of this problem lies not only in comparing the different methods on a widely-studied 3-D benchmark problem but also in testing our new method in near-void regions, while the previous two problems only had pure void regions.

In this section, all the S_N simulations use the Level-Symmetric angular quadrature rule and the total number of angles is then given by $N(N+2)$, as the solution depends on all three spatial variables. As a comparison, the total number of moments for a P_N simulation in that case is $(N+1)^2$, which implies that, for a given N, the number of angular unknowns only differs by one between a P_N and a S_N calculation.

In [2], the semi-analytical scalar flux was given at different points of the domain.

In Fig. 5.5a, we compute the SAAF–CLS–P$_N$ error at $(x, y, z) = (5, 5, 5)$, $(5, 35, 5)$, $(5, 55, 5)$ and $(35, 55, 5)$ for different values of N and of the level of mesh refinement. The first point is in the source region whereas the last three are in the near-void region. It appears that the error indeed decreases as the simulation is refined in space and angle, although it becomes less apparent for the spatial point $(35, 55, 5)$, further away from the source. This is not surprising as the magnitude of the scalar flux rapidly decreases in the shield region. As we observed in Fig. 5.2, the error seems to be generally higher for even values of N.

Fig. 5.5b shows the results for the LS–P$_N$ method which are very comparable to SAAF–CLS–P$_N$ at the first and last spatial points but somewhat worse at the second and third point. The difference is not as significant as in Fig. 5.1, most likely because the near-void region is spatially much more limited than it was in Section 5.3.5.1, where the void region accounted for half of the spatial domain.

In Fig. 5.5c, the same quantities are shown for the SAAF–CLS–S$_N$ method, with errors comparable to SAAF–CLS–P$_N$ at the first and last spatial points but not as good at the second and third. It is noted however that the computational time tend to be much lower for the S$_N$ method, in particular because the number of kernels needed to be assembled for the streaming terms is noticeably higher for P$_N$, due to numerous off-diagonal coupling terms.

Lastly, Fig. 5.5d exhibits a behavior for the SAAF–VT–S$_N$ method very close to that of SAAF–CLS–S$_N$, especially for a level of spatial refinement higher than 2. This was expected as both variational formulations look very much alike (see Eqs. 4.10 and 5.10). For instance, the error at the first spatial point for the most refined mesh approaches 10^{-2} cm^{-2}–s^{-1} in both cases. However, it is interesting to point out that the Hypre BoomerAMG preconditioner [72] does not seem to be efficient in the same ranges for both methods, most likely because of the on-diagonal contribution of the $(\vec{\Omega}\cdot\vec{\nabla}\Psi^*, c^{-1}\vec{\Omega}\cdot\vec{\nabla}\Psi)_{\mathcal{D}_0}$ term in the void region which does not vanish as the mesh is refined.

5.4.3 Extension to Time-Dependent Problems

Let us first consider how the transient SAAF formulation can be derived, following the same reasoning as in Section 4.1.1.

The time-dependent transport equation can be written:

$$\frac{1}{v}\frac{\partial\Psi}{\partial t} + \vec{\Omega}\cdot\vec{\nabla}\Psi + \sigma_{\mathrm{t}}\Psi = H\Psi + S, \tag{5.22}$$

which gives the transient AFE:

$$\Psi = \frac{1}{\sigma_{\mathrm{t}}}\left(H\Psi + S - \vec{\Omega}\cdot\vec{\nabla}\Psi - \frac{1}{v}\frac{\partial\Psi}{\partial t}\right). \tag{5.23}$$

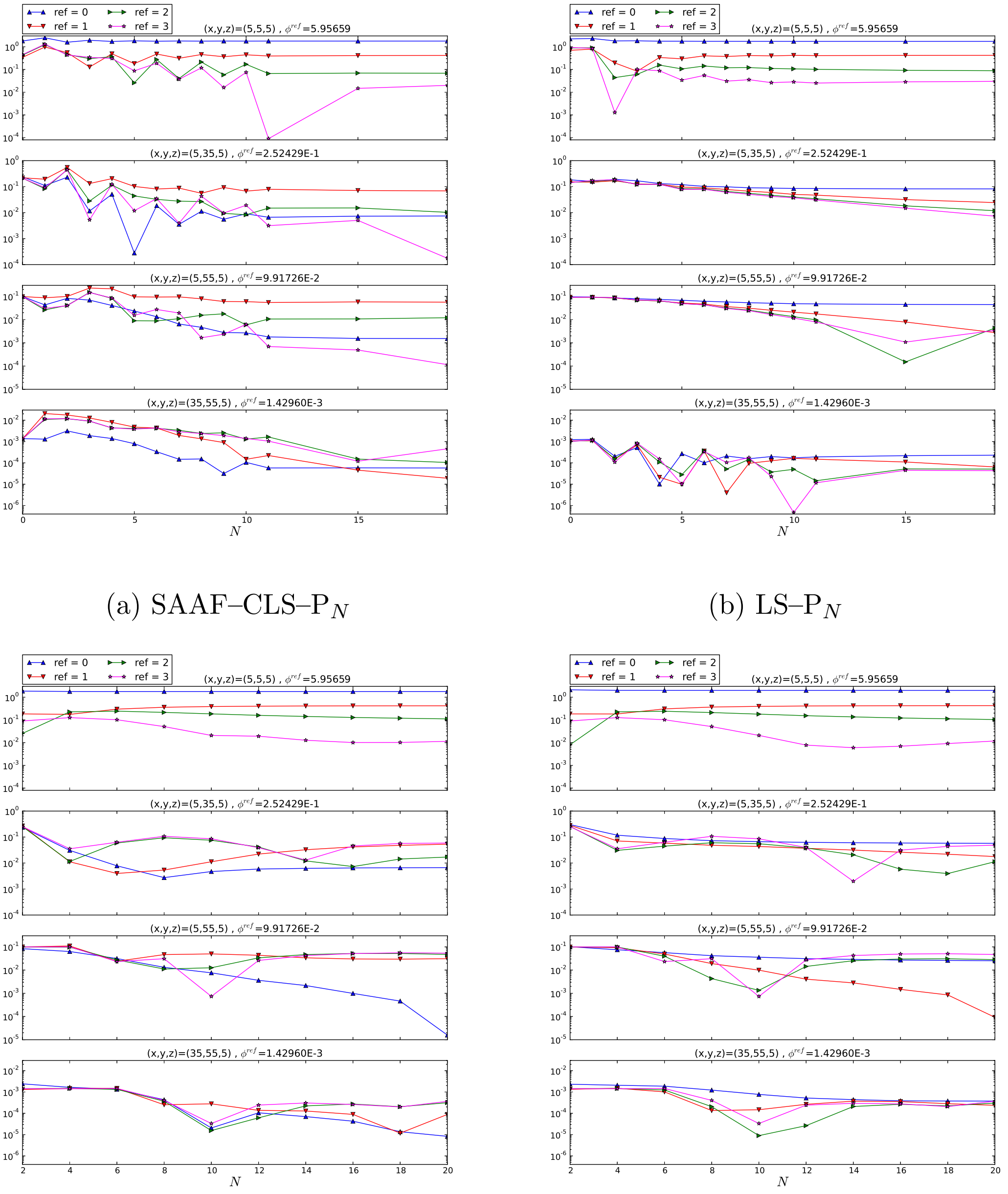
Yet, Eq. 5.22 integrated over $\mathcal{D}_1$ against Ψ^* yields, after an integration by parts on the streaming term:

$$\left(\Psi^*, \frac{1}{v}\frac{\partial\Psi}{\partial t}\right)_{\mathcal{D}_1} - \left(\vec{\Omega}\cdot\vec{\nabla}\Psi^*, \Psi\right)_{\mathcal{D}_1} + \langle\Psi^*, \Psi\rangle_{\partial\mathcal{D}_1} + (\Psi^*, \sigma_{\mathrm{t}}\Psi)_{\mathcal{D}_1} = (\Psi^*, H\Psi + S)_{\mathcal{D}_1}. \tag{5.24}$$

Substituting Eq. 5.23 into the second term gives:

$$\begin{aligned}\left(\frac{1}{\sigma_{\mathrm{t}}}\vec{\Omega}\cdot\vec{\nabla}\Psi^* + \Psi^*, \frac{1}{v}\frac{\partial\Psi}{\partial t}\right)_{\mathcal{D}_1} + \left(\vec{\Omega}\cdot\vec{\nabla}\Psi^*, \frac{1}{\sigma_{\mathrm{t}}}\vec{\Omega}\cdot\vec{\nabla}\Psi\right)_{\mathcal{D}_1} + \langle\Psi^*, \Psi\rangle^+_{\partial\mathcal{D}_1} - \langle\Psi^*, \Psi^{\mathrm{inc}}\rangle^-_{\partial\mathcal{D}_1} \\ + (\Psi^*, \sigma_{\mathrm{t}}\Psi)_{\mathcal{D}_1} = \left(\frac{1}{\sigma_{\mathrm{t}}}\vec{\Omega}\cdot\vec{\nabla}\Psi^* + \Psi^*, H\Psi + S\right)_{\mathcal{D}_1}.\end{aligned} \tag{5.25}$$

If we were to derive the transient LS formulation over $\mathcal{D}_0$ (while still defining $L^\dagger \equiv$



(a) SAAF–CLS–P$_N$

(b) LS–P$_N$

(c) SAAF–CLS–S$_N$

(d) SAAF–VT–S$_N$

Figure 5.5: Error in the scalar flux (in cm^{-2}–s^{-1}) at 4 different spatial points as a function of N and of the level of mesh refinement for various methods. On the top of each graph, the reference value for Φ at the corresponding spatial point is indicated. For a given spatial point, the y-axis range is identical for all methods.

$-\vec{\Omega}\cdot\vec{\nabla}+\sigma_{\mathrm{t}}$) in a similar way as we did in Section 4.1.2, we would obtain:

$$\begin{aligned}\left(\vec{\Omega}\cdot\vec{\nabla}\Psi^{*}+\sigma_{\mathrm{t}}\Psi^{*},\frac{1}{v}\frac{\partial\Psi}{\partial t}\right)_{\mathcal{D}_0}+\left(\vec{\Omega}\cdot\vec{\nabla}\Psi^{*},\vec{\Omega}\cdot\vec{\nabla}\Psi\right)_{\mathcal{D}_0}+\langle c\Psi^{*},\Psi-\Psi^{\mathrm{inc}}\rangle^{-}_{\partial\mathcal{D}}\\ +\left(\sigma_{\mathrm{t}}\Psi^{*},\sigma_{\mathrm{t}}\Psi\right)_{\mathcal{D}_0}=\left(\vec{\Omega}\cdot\vec{\nabla}\Psi^{*}+\sigma_{\mathrm{t}}\Psi^{*},H\Psi+S\right)_{\mathcal{D}_0}.\end{aligned}\tag{5.26}$$

We can then add the term $((c-\sigma_0)\Psi^{*},\frac{1}{v}\frac{\partial\Psi}{\partial t}+\vec{\Omega}\cdot\vec{\nabla}\Psi+\sigma_0\Psi-H\psi-S)_{\mathcal{D}_0}$ to get the transient CLS formulation (and retrieve global conservation over $\mathcal{D}_0$).

Following the same reasoning as presented in Section 5.4, the time-dependent SAAF–CLS formulation eventually reads: find $\Psi\in V$ such that for all $\Psi^{*}\in V$,

$$\begin{aligned}&\left(\frac{1}{c}\vec{\Omega}\cdot\vec{\nabla}\Psi^{*}+\Psi^{*},\frac{1}{v}\frac{\partial\Psi}{\partial t}\right)_{\mathcal{D}_0}+\left(\frac{1}{\sigma_{\mathrm{t}}}\vec{\Omega}\cdot\vec{\nabla}\Psi^{*}+\Psi^{*},\frac{1}{v}\frac{\partial\Psi}{\partial t}\right)_{\mathcal{D}_1}\\ &+\left(\vec{\Omega}\cdot\vec{\nabla}\Psi^{*},\frac{1}{\sigma_{\mathrm{t}}}\vec{\Omega}\cdot\vec{\nabla}\Psi\right)_{\mathcal{D}_1}+\left(\vec{\Omega}\cdot\vec{\nabla}\Psi^{*},\frac{1}{c}\vec{\Omega}\cdot\vec{\nabla}\Psi\right)_{\mathcal{D}_0}+\left(\Psi^{*},\sigma_t\Psi\right)_{\mathcal{D}_1}+\left(\Psi^{*},\sigma_0\Psi\right)_{\mathcal{D}_0}\\ &-\left(\left(1-\frac{\sigma_0}{c}\right)\vec{\Omega}\cdot\vec{\nabla}\Psi^{*},\Psi\right)_{\mathcal{D}_0}+\langle\Psi^{*},\Psi\rangle^{+}_{\partial\mathcal{D}}-\langle\Psi^{*},\Psi^{\mathrm{inc}}\rangle^{-}_{\partial\mathcal{D}}\\ &\quad=\left(\frac{1}{\sigma_{\mathrm{t}}}\vec{\Omega}\cdot\vec{\nabla}\Psi^{*}+\Psi^{*},H\Psi+S\right)_{\mathcal{D}_1}+\left(\frac{1}{c}\vec{\Omega}\cdot\vec{\nabla}\Psi^{*}+\Psi^{*},H\Psi+S\right)_{\mathcal{D}_0},\end{aligned}\tag{5.27}$$

which could be rewritten: find $\Psi\in V$ such that for all $\Psi^{*}\in V$,

$$\begin{aligned}&\left(\tau\,\vec{\Omega}\cdot\vec{\nabla}\Psi^{*}+\Psi^{*},\frac{1}{v}\frac{\partial\Psi}{\partial t}\right)_{\mathcal{D}}+\left(\vec{\Omega}\cdot\vec{\nabla}\Psi^{*},\tau\,\vec{\Omega}\cdot\vec{\nabla}\Psi\right)_{\mathcal{D}}+\left(\Psi^{*},\sigma_t\Psi\right)_{\mathcal{D}}\\ &-\left((1-\tau\sigma_{\mathrm{t}})\,\vec{\Omega}\cdot\vec{\nabla}\Psi^{*},\Psi\right)_{\mathcal{D}}+\langle\Psi^{*},\Psi\rangle^{+}_{\partial\mathcal{D}}-\langle\Psi^{*},\Psi^{\mathrm{inc}}\rangle^{-}_{\partial\mathcal{D}}\\ &\qquad=\left(\tau\,\vec{\Omega}\cdot\vec{\nabla}\Psi^{*}+\Psi^{*},H\Psi+S\right)_{\mathcal{D}},\end{aligned}\tag{5.28}$$

where τ is defined by:

$$\tau \equiv \begin{cases} \dfrac{1}{c} & , \quad \vec{r} \in \mathcal{D}_0 \\ \dfrac{1}{\sigma_{\mathrm{t}}} & , \quad \vec{r} \in \mathcal{D}_1 \end{cases} . \tag{5.29}$$

Any type of time-discretization schemes could then be applied to this formulation (Backward Euler, BDF–2,...). Looking at this variational formulation, it appears that – despite having assumed c to be constant and σ_{t} to be uniform in the void region $\mathcal{D}_0$ – relaxing these two assumptions does not affect the global conservation of the method and is therefore perfectly acceptable in practice.

5.4.4 Parity Option

As we saw in Section 4.5, there does not seem to exist any void compatible and globally conservative parity-based method. A legitimate question is whether the newly-derived SAAF–CLS method could provide such a method.

The 'even' SAAF–CLS formulation would read: find $\Psi_e \in V_e$ such that for all $\Psi_e^* \in V_e$,

$$\begin{aligned} &\left(\tau\,\vec{\Omega}\cdot\vec{\nabla}\Psi_e^*, \frac{1}{v}\frac{\partial \Psi_o}{\partial t}\right)_{\mathcal{D}} + \left(\Psi_e^*, \frac{1}{v}\frac{\partial \Psi_e}{\partial t}\right)_{\mathcal{D}} + \left(\vec{\Omega}\cdot\vec{\nabla}\Psi_e^*, \tau\,\vec{\Omega}\cdot\vec{\nabla}\Psi_e\right)_{\mathcal{D}} \\ &+ \left(\Psi_e^*, \sigma_t \Psi_e\right)_{\mathcal{D}} - \left((1-\tau\sigma_{\mathrm{t}})\,\vec{\Omega}\cdot\vec{\nabla}\Psi_e^*, \Psi_o\right)_{\mathcal{D}} + \Gamma_{\mathrm{even}} \\ &\qquad = \left(\tau\,\vec{\Omega}\cdot\vec{\nabla}\Psi_e^*, H\Psi_o + S_o\right)_{\mathcal{D}} + \left(\Psi_e^*, H\Psi_e + S_e\right)_{\mathcal{D}} , \end{aligned} \tag{5.30}$$

where Γ_{even} is defined by Eq. 4.67. Unfortunately, even in the steady-state, isotropic scattering cases, Ψ_o has to be evaluated in the void/near-void region. This is necessarily done using Eq. 4.61, which is not valid in void regions. Therefore, the 'even' SAAF–CLS formulation is not compatible with void.

6. CONCLUSION

Now that this thesis is coming to a close, let us summarize what has been accomplished therein.

6.1 Implicit Filtered P_N for First-Order Forms

In Chapter 3, we have presented and implemented a fully-implicit, discontinuous Galerkin finite element method for simulating filtered spherical harmonic (P_N) equations in the context of thermal radiation transport and provided guidelines to determine filtering strategies for general problems.

We have studied the eigenspectrum of the filtered P_N equations. Interestingly, the conditioning of underlying linear systems improves for moderate values of the filter strength σ_{f}. Indeed, it was confirmed that such values led to a significant reduction in the number of GMRES iterations needed to solve the Crooked Pipe benchmark problem. We have also tested numerically the convergence properties of the filter and have found that the features of the linear, pure transport problem carry over to the non-linear, thermal radiation transport problem. Roughly speaking, the filter order determines the convergence rate for smooth solutions, while for non-smooth problems, the filter has little impact.

Finally, we have performed detailed simulations of the Crooked Pipe problem and used it as a test case to compare different filtering strategies. We observe that filtering improves numerical solutions significantly, especially for small values of N. For the most part, it is a local filtering strategy that works best.

6.2 Conditional Equivalence of Second-Order Forms

In Chapter 4, we have shown how the choice of the weakly-imposed boundary conditions for LS can have a significant impact on the properties of the scheme. In particular, we have shown how LS can – under certain limiting conditions – be made equivalent to SAAF or consistent with SAAF–VT.

We have also proven the equivalences between several parity-based P_N methods: the even-parity P_N, 'even' SAAF-P_N and SAAF–VT–P_N. The practical corollary is that these methods are not well-suited for second-order filters with the driving application of allowing for void regions.

Numerical results on a heterogeneous multigroup k-eigenvalue problem with void highlighted how global conservation can be crucial to achieve rapid convergence.

An important lesson from this chapter is that LS does not have this conservation property unless the boundary scaling term and the total cross-section are strictly positive constants, equal to each other. In particular, this cannot be achieved if there is void anywhere in the domain of interest. This observation was key to the conservative fix derived in the following chapter.

6.3 SAAF–CLS Method

In Chapter 5, we have derived a second-order method compatible with void and globally conservative working with both P_N and S_N. This is achieved using LS terms in the void region with a non-symmetric correction to retrieve global conservation. It is then combined with SAAF terms in the non-void regions with a scaling chosen such that the interface terms vanish, thereby maintaining global conservation. We have observed that this conservative fix is crucial to gain any benefit, compared to the plain LS method. Overall, this SAAF–CLS method has shown much improvement for problems for which global conservation is key, such as configurations with large void regions or k-eigenvalue calculations. Particularly, we have obtained very satisfying results for both the P_N and S_N versions of our method on a multigroup k-eigenvalue problem with significant heterogeneity and void regions. These results have been in good agreement with a MCNP reference calculation but also have been very comparable to those obtained with the SAAF–VT–S_N method.

While the SAAF–CLS and SAAF–VT variational formulations formally look very similar, both sacrificing the symmetry of the bilinear form, our method presents the advantage of being compatible with both P_N and S_N angular discretizations, unlike the SAAF–VT method which has only shown success with S_N. The reason is that the latter method tends to reduce to a first-order form in void regions, which results in a singular system following a P_N discretization for a steady-state calculation with CGFEM.

Further, we have generalized the SAAF–CLS method to near-void regions and time-dependent problems. We showed that global conservation could be preserved, providing a slightly different correction to the LS formulation. We have then tested this method on the dog leg void duct benchmark problem by Kobayashi et al [2].

6.4 Future Studies

Future work will include some deeper studies on time-dependent problems where void compatibility is required.

Furthermore, based on the observation that the Simplified P_N (SP_N) method is equivalent to P_N for a constant total cross-section and an infinite medium [74], another idea would be to try to develop a hybrid scheme combining the CLS–P_N and SP_N methods.

REFERENCES

[1] E. E. Lewis, M. A. Smith, G. Palmiotti, T. A. Taiwo, and N. Tsoul-Fanidis. Benchmark specification for deterministic 2-D/3-D MOX fuel assembly transport calculations without spatial homogenisation (C5G7 MOX). Technical Report NEA/NSC/DOC(2001)4, OECD/NEA Expert Group on 3-D Radiation Transport Benchmarks, 2001.

[2] Keisuke Kobayashi, Naoki Sugimura, and Yasunobu Nagaya. 3D radiation transport benchmark problems and results for simple geometries with void region. *Progress in Nuclear Energy*, 39(2):119–114, 2001.

[3] J.A. Fleck Jr., J.D. Cummings. An implicit Monte Carlo scheme for calculating time and frequency dependent nonlinear radiation transport. *Journal of Computational Physics*, 8:313–342, 1971.

[4] Ryan G. McClarren, Todd J. Urbatsch. A modified implicit Monte Carlo method for time-dependent radiative transfer with adaptive material coupling. *Journal of Computational Physics*, 228:5669–5686, 2009.

[5] Edward W. Larsen and Jim E. Morel. Advances in discrete-ordinates methodology. pages 1–84. Springer Netherlands, 2010.

[6] Thomas A. Brunner, James P. Holloway. Two-dimensional time dependent Riemann solvers for neutron transport. *Journal of Computational Physics*, 210:386–399, 2005.

[7] Guido Kanschat. Solution of radiative transfer problems with finite elements. In Guido Kanschat, Erik Meinkhn, Rolf Rannacher, and Rainer Wehrse, edi-

tors, *Numerical Methods in Multidimensional Radiative Transfer*, pages 49–98. Springer Berlin Heidelberg, 2009.

[8] T. A. Brunner and J. P. Holloway. One-dimensional Riemann solvers and the maximum entropy closure. *Journal of Quantitative Spectroscopy & Radiative Transfer*, 69:543–566, 2001.

[9] C. D. Hauck. High-order entropy-based closures for linear transport in slab geometries. *Commun. Math. Sci.*, 9:187–205, 2011.

[10] B. Dubroca and J.-L. Fuegas. Étude théorique et numérique d'une hiérarchie de modèles aux moments pour le transfert radiatif. *C.R. Acad. Sci. Paris*, I. 329:915–920, 1999.

[11] Martin Frank, Cory Hauck, and Kerstin Kuepper. Convergence of filtered spherical harmonic equations for radiation transport. *Communications in Mathematical Sciences*, 14:1443–1465, 2016.

[12] Ryan G. McClarren, Cory D. Hauck. Robust and accurate filtered spherical harmonics expansions for radiative transfer. *Journal of Computational Physics*, 229:5597–5614, 2010.

[13] C. Kristopher Garrett and Cory D. Hauck. A comparison of moment closures for linear kinetic transport equations: The line source benchmark. *Transport Theory and Stastical Physics*, 42:203–235, 2015.

[14] Cory Hauck, Ryan G. McClarren. Positive P_N Closures. *SIAM Journal on Scientific Computing*, 32(5):2603, 2010.

[15] Weixiong Zheng and Ryan G. McClarren. Moment closures based on minimizing the residual of the PN angular expansion in radiation transport. *Journal of Computational Physics*, 314:682–699, 2016.

[16] Ryan G. McClarren, Cory Hauck. Simulating radiative transfer with filtered spherical harmonics. *Physics Letters A*, 374:2290–2296, 2010.

[17] Ryan G McClarren, Cory D Hauck, and Robert B Lowrie. Filtered spherical harmonics methods for transport problems. In *Proceedings of the 2009 international conference on mathematics and computational methods and reactor physics*, 2008.

[18] Simon Merton Cory Ahrens. An improved filtered spherical harmonic method for transport calculations. In *Proceedings of the 2013 International Conference on Mathematics and Computational Methods Applied to Nuclear Science and Engineering-M&C 2013*, 2013.

[19] David Radice, Ernazar Abdikamalov, Luciano Rezzolla, Christian D. Ott. A new spherical harmonics scheme for multi-dimensional radiation transport I. static matter configurations. *Journal of Computational Physics*, 242:648–669, 2013.

[20] Eleuterio F. Toro. *Riemann Solvers and Numerical Methods for Fluid Dynamics: A Practical Introduction, 2nd edn.* Springer, Berlin, 2009.

[21] CC Pain, MD Eaton, RP Smedley-Stevenson, AJH Goddard, MD Piggott, and CRE de Oliveira. Streamline upwind Petrov–Galerkin methods for the steady-state boltzmann transport equation. *Computer methods in applied mechanics and engineering*, 195(33):4448–4472, 2006.

[22] CC Pain, MD Eaton, RP Smedley-Stevenson, AJH Goddard, MD Piggott, and CRE de Oliveira. Space–time streamline upwind Petrov–Galerkin methods for the boltzmann transport equation. *Computer methods in applied mechanics and engineering*, 195(33):4334–4357, 2006.

[23] Herbert Egger and Matthias Schlottbom. A mixed variational framework for the radiative transfer equation. *Mathematical Models and Methods in Applied Sciences*, 22(3), 2012.

[24] Stephen Wright, Simon Arridge, and Martin Schweiger. A finite element method for the even-parity radiative transfer equation using the PN approximation. In Guido Kanschat, Erik Meinkhn, Rolf Rannacher, and Rainer Wehrse, editors, *Numerical Methods in Multidimensional Radiative Transfer*, pages 39–48. Springer Berlin Heidelberg, 2009.

[25] J. L. Guermond, R. Pasquetti, B. Popov. Entropy viscosity method for nonlinear conservation laws. *Journal of Computational Physics*, 230:4248–4267, 2011.

[26] Wm H Reed and TR Hill. Triangular mesh methods for the neutron transport equation. *Los Alamos Report LA-UR-73-479*, 1973.

[27] J.L. Guermond, G. Kanschat. Asymptotic analysis of upwind discontinuous Galerkin approximation of the radiative transport equation in the diffusion limit. *SIAM Journal on Numerical Analysis*, 48(1):53–78, 2010.

[28] E.W. Larsen, J.E. Morel, and W.F. Miller, Jr. Asymptotic solutions of numerical transport problems in optically thick, diffusive regimes. *Journal of Computational Physics*, 69:283–324, 1987.

[29] Adams, M. L. Discontinuous finite element transport solutions in thick diffusive problems. *Nuclear Science and Engineering*, 137(3):298–333, 2001.

[30] G. J. Habetler and B. J. Matkowsky. Uniform asymptotic expansions in transport theory with small mean free paths, and the diffusion approximation. *Journal of Mathematical Physics*, 16:846–854, April 1975.

[31] E. W. Larsen and J. B. Keller. Asymptotic solution of neutron transport problems for small mean free paths. *Journal of Mathematical Physics*, 15:75–81, January 1974.

[32] McClarren, R. G., Evans, T. M., Lowrie, R. B., and Densmore, J. D. Semi-implicit time integration for thermal radiative transfer. *Journal of Computational Physics*, 227(16):7561–7586, 2008.

[33] J. E. Morel and J. M. McGhee. A self-adjoint angular flux equation. *Nuclear Science and Engineering*, 132:312–325, 1999.

[34] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, 1996.

[35] Jim E. Morel, B. Todd Adams, Taewan Noh, John M. McGhee, Thomas M. Evans, and Todd J. Urbatsch. Spatial discretizations for self-adjoint forms of the radiative transfer equations. *Journal of Computational Physics*, 214:12–40, 2006.

[36] Jennifer Liscum-Powell. Finite element numerical solution of a self-adjoint transport equation for coupled electron-photon problems. Technical Report SAND2000-2017, Sandia National Laboratory, 2000.

[37] Jennifer L. Liscum-Powell, Anil K. Prinja, Jim E. Morel, and Leonard J. Lorence, Jr. Finite element solution of the self-adjoint angular flux equation for coupled electron-photon transport. *Nuclear Science and Engineering*, 142:270–291, 2002.

[38] Yaqi Wang, Hongbin Zhang, and Richard Martineau. Diffusion acceleration schemes for the self-adjoint angular flux formulation with a void treatment. *Nuclear Science and Engineering*, 176:201–225, 2014.

[39] E. E. Lewis and Jr. W. F. Miller. *Computational Methods of Neutron Transport*. Wiley, 1984.

[40] Clif Drumm, Wesley Fan, Andrew Bielen, and Jeffrey Chenhall. Least squares finite elements algorithms in the SCEPTRE radiation transport code. In *International Conference on Mathematics and Computational Methods Applied to Nuclear Science and Engineering*, Rio de Janeiro, RJ, Brazil, May 8–12 2011. Latin American Section / American Nuclear Society.

[41] Christopher J. Gesh. *Finite Element Methods for Second Order Forms of the Transport Equation*. PhD thesis, Texas A&M University, 1999.

[42] Thomas A Manteuffel and Klaus J Ressel. Least-squares finite-element solution of the neutron transport equation in diffusive regimes. *SIAM Journal on Numerical Analysis*, 35(2):806–835, 1998.

[43] Thomas A Manteuffel, Klaus J Ressel, and Gerhard Starke. A boundary functional for the least-squares finite-element solution of neutron transport problems. *SIAM Journal on Numerical Analysis*, 37(2):556–586, 1999.

[44] Jon Hansen, Jacob Peterson, Jim Morel, Jean Ragusa, and Yaqi Wang. A least-squares transport equation compatible with voids. *Journal of Computational and Theoretical Transport Special Issue: Papers from the 23rd International Conference on Transport Theory*, 43(1–7), 2014.

[45] F. Graziani and J. LeBlanc. Technical Report UCRL-MI-143393, 2000.

[46] G.C. Pomraning. *The equations of radiation hydrodynamics*. International series of monographs in natural philosophy. Pergamon Press, 1973.

[47] Ryan G. McClarren, James Paul Holloway, Thomas A. Brunner. On solutions to the Pn equations for thermal radiative transfer. *Journal of Computational*

Physics, 227:2864–2885, 2008.

[48] Thomas A. Brunner. *Riemann Solvers for Time-Dependent Transport Based on the Maximum Entropy and Spherical Harmonics Closures*. PhD thesis, University of Michigan, 2000.

[49] Vincent M. Laboure, Ryan G. McClarren, and Cory D. Hauck. Implicit filtered P_N for high-energy density thermal radiation transport using discontinuous galerkin finite elements. *Journal of Computational Physics*, 2016.

[50] Bernardo Cockburn and Chi-Wang Shu. Rungekutta discontinuous Galerkin methods for convection-dominated problems. *Journal of Scientific Computing,*, 16:173–261, 2001.

[51] Zhangxin Chen. *Finite Element Methods and Their Applications*. Springer, 2005.

[52] Ryan G. McClarren. *Spherical Harmonics Methods for Thermal Radiation Transport*. PhD thesis, The University of Michigan, Nuclear Engineering and Radiological Sciences, 2006.

[53] MATLAB. The MathWorks Inc., Natick, Massachusetts.

[54] Derek Gaston, Chris Newman, Glen Hansen, and Damien Lebrun-Grandié. MOOSE: A parallel computational framework for coupled systems of nonlinear equations. *Nuclear Engineering and Design*, 239(10):1768–1778, 2009.

[55] Satish Balay, Shrirang Abhyankar, Mark F. Adams, Jed Brown, Peter Brune, Kris Buschelman, Victor Eijkhout, William D. Gropp, Dinesh Kaushik, Matthew G. Knepley, Lois Curfman McInnes, Karl Rupp, Barry F. Smith, and Hong Zhang. PETSc users manual. Technical Report ANL-95/11 - Revision

3.5, Argonne National Laboratory, Computer, Computational, and Statistical Sciences Division, 2014.

[56] B. S. Kirk, J. W. Peterson, R. H. Stogner, and G. F. Carey. `libMesh`: A C++ Library for Parallel Adaptive Mesh Refinement/Coarsening Simulations. *Engineering with Computers*, 22(3–4):237–254, 2006.

[57] C. Geuzaine and J.-F. Remacle. Gmsh: a three-dimensional finite element mesh generator with built-in pre- and post-processing facilities. *International Journal for Numerical Methods in Engineering*, 79(11):1309–1331, 2009.

[58] Hank Childs, Eric Brugger, Brad Whitlock, Jeremy Meredith, Sean Ahern, David Pugmire, Kathleen Biagas, Mark Miller, Cyrus Harrison, Gunther H. Weber, Hari Krishnan, Thomas Fogal, Allen Sanderson, Christoph Garth, E. Wes Bethel, David Camp, Oliver Rübel, Marc Durant, Jean M. Favre, and Paul Navrátil. VisIt: An End-User Tool For Visualizing and Analyzing Very Large Data. In *High Performance Visualization–Enabling Extreme-Scale Scientific Insight*, pages 357–372. October 2012.

[59] C. R. E. de Oliveira, C. C. Pain, and M. D. Eaton. Hierarchical angular preconditioning for the finite element-spherical harmonics radiation transport method. *Proceedings of PHYSOR 2000 ANS International Topical Meeting on Advances in Reactor Physics and Mathematics and Computation into the Next Millenium, Pittsburgh, USA*, 2000.

[60] S. L. Campbell, I. C. F. Ipsen, C. T. Kelley, and C. D. Meyer. GMRES and the minimal polynomial. *BIT*, 36(664), 1996.

[61] Robert B. Lowrie. A comparison of implicit time integration methods for nonlinear relaxation and diffusion. *Journal of Computational Physics*, 196:566–590, 2004.

[62] Ryan G. McClarren, James Paul Holloway, Thomas A. Brunner. Analytic P1 solutions for time-dependent, thermal radiative transfer in several geometries. *Journal of Quantitative Spectroscopy & Radiative Transfer*, 109:389–403, 2008.

[63] Marvin L. Adams and Paul F. Nowak. Asymptotic analysis of a computational method for time- and frequency- dependent radiative transfer. *Journal of Computational Physics*, 146:366–403, 1998.

[64] Vincent M. Laboure, Yaqi Wang, and Mark D. DeHart. Least-squares P$_N$ formulation of the transport equation using self-adjoint-angular-flux consistent boundary conditions. In *PHYSOR 2016 - Unifying Theory and Experiments in the 21st Century*, Sun Valley, ID, May 2016. ANS.

[65] Yaqi Wang. *Adaptative Mesh Refinement Solution Techniques for the Multigroup S_N Transport Equation Using a Higher-Order Discontinuous Finite Element Method*. PhD thesis, Texas A&M University, 2009.

[66] Thomas A. Manteuffel and Klaus J. Ressel. Least-squares finite-element solution of the neutron transport equation in diffusive regimes. *SIAM Journal on Numerical Analysis*, 35(2):806–835, 1998.

[67] J.R. Peterson, H.R. Hammer, J.E. Morel, J.C. Ragusa, and Y. Wang. Conservative nonlinear diffusion acceleration applied to the unweighted least-squares transport equation in moose. In *MC2015 - Joint International Conference on Mathematics and Computation (M&C), Supercomputing in Nuclear Applications (SNA) and the Monte Carlo (MC) method*, Nashville, TN, April 2015. ANS.

[68] Thomas J. R. Hugues, Gerald Engel, Luca Mazzei, and Mats G. Larson. The continuous galerkin method is locally conservative. *Journal of Computational Physics*, 163(2):467–488, 2000.

[69] Yaqi Wang and Frederick N. Gleicher. Revisit boundary conditions for the self-adjoint angular flux formulation. In *PHYSOR 2014 - The Role of Reactor Physics toward a Sustainable Future*, The Westin Miyako, Kyoto, Japan, September 28 - October 3, 2014. ANS.

[70] X-5 Monte Carlo Team. MCNP – A General Monte Carlo N-Particle Transport Code, Version 5. Technical report, Los Alamos National Laboratory, 2003.

[71] Vincent M. Laboure, Ryan G. McClarren, and Yaqi Wang. Globally conservative, hybrid self-adjoint angular flux and least-squares method compatible with void. In *(submitted to Nuclear Science and Engineering)*, 2016.

[72] *hypre*: High performance preconditioners.

[73] W.H. Reed. New difference schemes for the neutron transport equation. *Nuclear Science and Engineering*, 46:309–314, 1971.

[74] Ryan G. McClarren. Theoretical aspects of the Simplified P_N equations. *Transport Theory and Statistical Physics*, 39(2–4):73–109, 2010.

[75] Edward Condon and George Shortley. *The theory of atomic spectra*. Cambridge University Press, 1935.

[76] G. Goertzel and N. Tralli. *Some Mathematical Methods of Physics*. McGraw-Hill, 1960.

APPENDIX A

SPHERICAL HARMONICS

A.1 Real vs. Complex Form

In this section, we compare the complex and real-form spherical harmonics and explain why we believe that the latter form is preferable for transport applications.

A.1.1 Complex Spherical Harmonics

The complex spherical harmonics are defined for all (ℓ, m) such that $0 \leq |m| \leq \ell$ as follows:

$$Y_\ell^m(\mu,\varphi) = (-1)^m \sqrt{\frac{(2\ell+1)}{4\pi}\frac{(\ell-m)!}{(\ell+m)!}}\,\frac{(1-\mu^2)^{m/2}}{2^\ell \ell!} e^{im\varphi} \frac{\mathrm{d}^{\ell+m}}{\mathrm{d}\mu^{\ell+m}}\big((\mu^2-1)^\ell\big) \quad \text{for } m \geq 0, \tag{A.1}$$

$$\overline{Y_\ell^m}(\mu,\varphi) = (-1)^m Y_\ell^{-m}(\mu,\varphi) \qquad \text{for } m \leq -1, \tag{A.2}$$

where $i = \sqrt{-1}$ and $\overline{X}$ represents the complex conjugate of X.

A.1.2 Real-Form Spherical Harmonics

The real-form spherical harmonics defined by Eq. 2.12 satisfy:

$$\begin{aligned} R_\ell^m &= \begin{cases} \frac{i}{\sqrt{2}}\left(Y_\ell^m - (-1)^m Y_\ell^{-m}\right) & , \quad m < 0 \\ Y_\ell^0 & , \quad m = 0 \\ \frac{1}{\sqrt{2}}\left(Y_\ell^{-m} + (-1)^m Y_\ell^m\right) & , \quad m > 0 \end{cases} , \\ &= \begin{cases} \sqrt{2}\,(-1)^m\,\Im(Y_\ell^{|m|}) & , \quad \text{if } m < 0 \\ Y_\ell^0 & , \quad \text{if } m = 0 \\ \sqrt{2}\,(-1)^m\,\Re\left(Y_\ell^m\right) & , \quad \text{if } m > 0 \end{cases} , \end{aligned} \tag{A.3}$$

where $\Re(X)$ and $\Im(X)$ designates the real and imaginary parts of a variable X, respectively. It can be shown that the complex spherical harmonics functions also form an orthonormal set of functions.

A.1.3 Recurrence Relationships

One way yo evaluate the $\vec{\mathbf{D}}$ matrices (see Eq. 3.4) is by using recurrence relationships. Defining ϑ to be the polar angle ($\mu = \cos\vartheta$), the following relationships are true for any ℓ, m such that $0 \leq |m| \leq \ell \leq N$ [48, 75, 76]:

$$\cos\vartheta\, Y_\ell^m = A_\ell^m Y_{\ell+1}^m + B_\ell^m Y_{\ell-1}^m, \tag{A.4}$$

$$\sin\vartheta\cos\varphi\, Y_\ell^m = \frac{1}{2}(-C_\ell^m Y_{\ell+1}^{m+1} + D_\ell^m Y_{\ell-1}^{m+1} + E_\ell^m Y_{\ell+1}^{m-1} - F_\ell^m Y_{\ell-1}^{m-1}), \tag{A.5}$$

$$\sin\vartheta\sin\varphi\, Y_\ell^m = \frac{i}{2}(C_\ell^m Y_{\ell+1}^{m+1} - D_\ell^m Y_{\ell-1}^{m+1} + E_\ell^m Y_{\ell+1}^{m-1} - F_\ell^m Y_{\ell-1}^{m-1}), \tag{A.6}$$

where:

$$A_\ell^m = \sqrt{\frac{(\ell-m+1)(\ell+m+1)}{(2\ell+3)(2\ell+1)}} \quad , \quad B_\ell^m = \sqrt{\frac{(\ell-m)(\ell+m)}{(2\ell+1)(2\ell-1)}}, \tag{A.7}$$

$$C_\ell^m = \sqrt{\frac{(\ell+m+1)(\ell+m+2)}{(2\ell+3)(2\ell+1)}}, \quad , \quad D_\ell^m = \sqrt{\frac{(\ell-m)(\ell-m-1)}{(2\ell+1)(2\ell-1)}}, \tag{A.8}$$

$$E_\ell^m = \sqrt{\frac{(\ell-m+1)(\ell-m+2)}{(2\ell+3)(2\ell+1)}} \quad , \quad F_\ell^m = \sqrt{\frac{(\ell+m)(\ell+m-1)}{(2\ell+1)(2\ell-1)}}. \tag{A.9}$$

It is also possible to derive similar recurrence relationships for the real-form spherical harmonics R_ℓ^m but it gets a lot messier. We show how to do it for each component for the reader's entertainment but we will see later that using quadrature rules is much more simple.

A.1.3.1 z-Component

The z-component is the easiest to derive as it does not couple the moments with different m. Let us first consider the $m = 0$ case:

$$\cos\vartheta\, R_\ell^0 = \cos\vartheta\, Y_\ell^0 = A_\ell^0 Y_{\ell+1}^0 + B_\ell^0 Y_{\ell-1}^0 = A_\ell^0 R_{\ell+1}^0 + B_\ell^0 R_{\ell-1}^0. \tag{A.10}$$

If $m < 0$:

$$\begin{aligned}\cos\vartheta\, R_\ell^m &= \frac{i\cos\vartheta}{\sqrt{2}}(Y_\ell^m - (-1)^m Y_\ell^{-m}) \\ &= \frac{i}{\sqrt{2}}(A_\ell^m Y_{\ell+1}^m + B_\ell^m Y_{\ell-1}^m - (-1)^m(A_\ell^{-m} Y_{\ell+1}^{-m} + B_\ell^{-m} Y_{\ell-1}^{-m})) \\ &= A_\ell^m R_{\ell+1}^m + B_\ell^m R_{\ell-1}^m\end{aligned} \tag{A.11}$$

because $A_\ell^{-m} = A_\ell^m$ and $B_\ell^{-m} = B_\ell^m$. We can show a similar result for $m > 0$. In summary:

$$\cos\vartheta\, R_\ell^m = A_\ell^m R_{\ell+1}^m + B_\ell^m R_{\ell-1}^m \tag{A.12}$$

A.1.3.2 x-Component

We want to express $\sin\vartheta\cos\varphi\, R_\ell^m$ as a function of the real-form spherical harmonics.

If $m = 0$,

$$\begin{aligned}\sin\vartheta\cos\varphi\, R_\ell^0 &= \sin\vartheta\cos\varphi\, Y_\ell^0 \\ &= \frac{1}{2}\left(-C_\ell^0 Y_{\ell+1}^1 + D_\ell^0 Y_{\ell-1}^1 + E_\ell^0 Y_{\ell+1}^{-1} - F_\ell^0 Y_{\ell-1}^{-1}\right).\end{aligned} \tag{A.13}$$

Because $C_\ell^0 = E_\ell^0$ and $D_\ell^0 = F_\ell^0$, we have:

$$Y_{\ell+1}^{-1} - Y_{\ell+1}^1 = \sqrt{2} R_{\ell+1}^1 \quad , \quad Y_{\ell-1}^{-1} - Y_{\ell-1}^1 = \sqrt{2} R_{\ell-1}^1, \tag{A.14}$$

that is[1]:

$$\sin\vartheta\cos\varphi\, R_\ell^0 = \frac{1}{\sqrt{2}}\left(C_\ell^0 R_{\ell+1}^1 - D_\ell^0 R_{\ell-1}^1\right). \tag{A.15}$$

If $m > 0$,

$$\begin{aligned}\sin\vartheta\cos\varphi\, R_\ell^0 &= \frac{\sin\vartheta\cos\varphi}{\sqrt{2}}\ \left(Y_\ell^{-m} + (-1)^m Y_\ell^m\right)\\ &= \frac{1}{2\sqrt{2}}\Big(\left(-C_\ell^{-m}Y_{\ell+1}^{-m+1} + D_\ell^{-m}Y_{\ell-1}^{-m+1} + E_\ell^{-m}Y_{\ell+1}^{-m-1} - F_\ell^{-m}Y_{\ell-1}^{-m-1}\right)\\ &+ (-1)^m\left(-C_\ell^m Y_{\ell+1}^{m+1} + D_\ell^m Y_{\ell-1}^{m+1} + E_\ell^m Y_{\ell+1}^{m-1} - F_\ell^m Y_{\ell-1}^{m-1}\right)\Big).\end{aligned}$$

Yet,

$$C_\ell^{-m} = E_\ell^m \quad , \quad D_\ell^m = F_\ell^{-m}, \tag{A.16}$$

thus, using the fact that $(-1)^{m-1} = (-1)^{m+1} = -(-1)^m$,

$$\begin{aligned}\sin\vartheta\cos\varphi\, R_\ell^m = \frac{1}{2\sqrt{2}}\Big(& - E_\ell^m(Y_{\ell+1}^{-m+1} + (-1)^{m-1}Y_{\ell+1}^{m-1}) + C_\ell^m(Y_{\ell+1}^{-m-1} + (-1)^{m-1}Y_{\ell+1}^{m+1})\\ & + F_\ell^m(Y_{\ell-1}^{-m+1} + (-1)^{m-1}Y_{\ell-1}^{m-1}) - D_\ell^m(Y_{\ell-1}^{-m-1} + (-1)^{m-1}Y_{\ell-1}^{m+1})\Big)\end{aligned}$$

If $m > 1$, we simply have:

$$\sin\vartheta\cos\varphi\, R_\ell^m = \frac{1}{2}\Big(- E_\ell^m R_{\ell+1}^{m-1} + C_\ell^m Y_{l+1,+m+1} + F_\ell^m R_{\ell-1}^{m-1} - D_\ell^m R_{\ell-1}^{m+1}\Big) \tag{A.17}$$

If $m = 1$:

$$Y_{\ell+1}^{-m+1} + (-1)^{m-1}Y_{\ell+1}^{m-1} = 2Y_{\ell+1}^0 \quad , \quad Y_{\ell-1}^{-m+1} + (-1)^{m-1}Y_{\ell-1}^{m-1} = 2Y_{\ell-1}^0.$$

[1] Note that this expression does not have any problem for $\ell = 0$ because $D_0^0 = 0$.

Therefore:

$$\sin\vartheta\cos\varphi\, R_\ell^m = \frac{1}{2}\Big(C_\ell^m Y_{l+1,+m+1} - D_\ell^m R_{\ell-1}^{m+1}\Big) + \frac{1}{\sqrt{2}}\Big(-E_\ell^m R_{\ell+1}^0 + F_\ell^m R_{\ell-1}^0\Big). \tag{A.18}$$

If $m < 0$, the treatment is similar:

$$\begin{aligned}\sin\vartheta\cos\varphi\, R_\ell^0 &= \frac{i\ \sin\vartheta\cos\varphi}{\sqrt{2}}\ \left(Y_\ell^m - (-1)^m Y_\ell^{-m}\right), \\ &= \frac{i}{2\sqrt{2}}\Big(\left(-C_\ell^m Y_{\ell+1}^{m+1} + D_\ell^m Y_{\ell-1}^{m+1} + E_\ell^m Y_{\ell+1}^{m-1} - F_\ell^m Y_{\ell-1}^{m-1}\right), \\ &\quad -(-1)^m\left(-C_\ell^{-m} Y_{\ell+1}^{-m+1} + D_\ell^{-m} Y_{\ell-1}^{m+1} + E_\ell^{-m} Y_{\ell+1}^{m-1} - F_\ell^{-m} Y_{\ell-1}^{m-1}\right)\Big), \\ &= \frac{i}{2\sqrt{2}}\Big(-C_\ell^m(Y_{\ell+1}^{m+1} - (-1)^{m-1}Y_{\ell+1}^{-m-1}) + D_\ell^m(Y_{\ell-1}^{m+1} - (-1)^{m-1}Y_{\ell-1}^{-m-1}) \\ &\qquad + E_\ell^m(Y_{\ell+1}^{m-1} - (-1)^{m-1}Y_{\ell+1}^{-m+1}) - F_\ell^m(Y_{\ell-1}^{m-1} - (-1)^{m-1}Y_{\ell-1}^{-m+1})\Big).\end{aligned}$$

If $m < -1$,

$$\sin\vartheta\cos\varphi\, R_\ell^m = \frac{1}{2}\Big(-C_\ell^m R_{\ell+1}^{m+1} + D_\ell^m R_{\ell-1}^{m+1} + E_\ell^m R_{\ell+1}^{m-1} - F_\ell^m R_{\ell-1}^{m-1}\Big). \tag{A.19}$$

If $m = -1$,

$$Y_{\ell+1}^{m+1} - (-1)^{m-1}Y_{\ell+1}^{-m-1} = 0 \quad , \quad Y_{\ell-1}^{m+1} - (-1)^{m-1}Y_{\ell-1}^{-m-1} = 0. \tag{A.20}$$

Thus,

$$\sin\vartheta\cos\varphi\, R_\ell^m = \frac{1}{2}\Big(E_\ell^m R_{\ell+1}^{m-1} - F_\ell^m R_{\ell-1}^{m-1}\Big). \tag{A.21}$$

A.1.3.3 y-Component

The derivation for the y-component is very much similar to that of the x-component and is not detailed here. The results are summarized in the next section.

A.1.3.4 Summary

To sum up, recursion relationships can be derived for the real-form spherical harmonics but are a lot more complicated:

$$\cos\vartheta\,R_\ell^m = A_\ell^m R_{\ell+1}^m + B_\ell^m R_{\ell-1}^m \tag{A.22}$$

$$\boxed{\begin{array}{l} \sin\vartheta\cos\varphi\,R_\ell^m = \\[2em] = \begin{cases} \dfrac{1}{\sqrt{2}}(C_\ell^0 R_{\ell+1}^1 - D_\ell^0 R_{\ell-1}^1) \quad , \quad \text{if } m=0 \\[2em] \dfrac{1}{2}(C_\ell^m R_{\ell+1}^{m+1} - D_\ell^m R_{\ell-1}^{m+1}) + \begin{cases} \dfrac{1}{2}(-E_\ell^m R_{\ell+1}^{m-1} - F_\ell^m R_{\ell-1}^{m-1}) & , \quad \text{if } m>1 \\ \dfrac{1}{\sqrt{2}}(-E_\ell^m R_{\ell+1}^{m-1} - F_\ell^m R_{\ell-1}^{m-1}) & , \quad \text{if } m=1 \end{cases} \\[2em] \dfrac{1}{2}(E_\ell^m R_{\ell+1}^{m-1} - F_\ell^m R_{\ell-1}^{m-1}) + \begin{cases} \dfrac{1}{2}(-C_\ell^m R_{\ell+1}^{m+1} + D_\ell^m R_{\ell-1}^{m+1}) & , \quad \text{if } m<-1 \\ 0 & , \quad \text{if } m=-1 \end{cases} \end{cases} \end{array}}$$

$$\boxed{\begin{aligned}&\sin\vartheta\sin\varphi\, R_\ell^m = \\ &\begin{cases} \frac{1}{\sqrt{2}}(C_\ell^0 R_{\ell+1}^{-1} - D_\ell^0 R_{\ell-1}^{-1}) \quad , \quad \text{if } m = 0 \\ \frac{1}{2}(C_\ell^m R_{\ell+1}^{-m-1} - D_\ell^m R_{\ell-1}^{-m-1}) + \begin{cases} \frac{1}{2}(E_\ell^m R_{\ell+1}^{-m+1} - F_\ell^m R_{\ell-1}^{-m+1}) & , \quad \text{if } m > 1 \\ 0 & , \quad \text{if } m = 1 \end{cases} \\ \frac{1}{2}(-E_\ell^m R_{\ell+1}^{-m+1} + F_\ell^m R_{\ell-1}^{-m+1}) + \begin{cases} \frac{1}{2}(-C_\ell^m R_{\ell+1}^{-m-1} + D_\ell^m R_{\ell-1}^{-m-1}) & , \text{ if } m < -1 \\ \frac{1}{\sqrt{2}}(-C_\ell^m R_{\ell+1}^{-m-1} + D_\ell^m R_{\ell-1}^{-m-1}) & , \text{ if } m = -1 \end{cases} \end{cases}\end{aligned}}$$

A.1.4 Computing $\vec{\mathbf{D}}$

Although the previous section seems to indicate that the recurrence relationship are more complicated for the real-form spherical harmonics, it should be pointed out that the $\vec{\mathbf{D}}$ matrices can be easily computed using Gauss-Legendre or Gauss-Jacobi quadrature rules. As a matter of fact, $\mathbf{D}_z$ can be exactly calculated using the former type while $\mathbf{D}_x$ and $\mathbf{D}_y$ can be deduced from the expression of $\mathbf{D}_z$ and applying rotation matrices.

More specifically, the element of the matrix $\mathbf{D}_z$ corresponding to the (ℓ, m)-th row and the (ℓ', m')-th column is given by:

$$(\mathbf{D}_z)_{(\ell,m),(\ell',m')} = \int_{\mathbb{S}^2} \mu R_\ell^m R_{\ell'}^{m'} \,\mathrm{d}\Omega. \tag{A.23}$$

For simplicity, let us assume $m, m' > 0$ and $d = 3$, in which case, we have:

$$\begin{aligned}(\mathbf{D}_z)_{(\ell,m),(\ell',m')} &= \int_{\mathbb{S}^2} 2C_\ell^m C_{\ell'}^{m'} \, \mu P_\ell^m P_{\ell'}^{m'} \, \cos(m\varphi)\cos(m'\varphi)\,\mathrm{d}\Omega, \\ &= 2C_\ell^m C_{\ell'}^{m'} \int_{-1}^{1} \mu P_\ell^m P_{\ell'}^{m'}\,\mathrm{d}\mu \int_0^{2\pi} \cos(m\varphi)\cos(m'\varphi)\,\mathrm{d}\varphi, \\ &= 2C_\ell^m C_{\ell'}^{m'} \int_{-1}^{1} \mu P_\ell^m P_{\ell'}^{m'}\,\mathrm{d}\mu \int_0^{2\pi} \frac{\cos((m+m')\varphi) + \cos((m-m')\varphi)}{2}\mathrm{d}\varphi.\end{aligned} \tag{A.24}$$

Because the integral over φ is non-zero if and only if $m = m'$, we obtain:

$$(\mathbf{D}_z)_{(\ell,m),(\ell',m')} = 2\pi\, C_\ell^m C_{\ell'}^m \delta_{m,m'} \int_{-1}^{1} \mu(1-\mu^2)^m \frac{\mathrm{d}^m}{\mathrm{d}\mu^m} P_\ell^0 \frac{\mathrm{d}^m}{\mathrm{d}\mu^m} P_{\ell'}^0 \mathrm{d}\mu. \tag{A.25}$$

Since P_ℓ^0 and $P_{\ell'}^0$ are polynomials of degree ℓ and ℓ', respectively, this implies that the function integrated over μ is a polynomial of degree $\ell + \ell' + 1$, i.e. at most $2N+1$. A $(N+1)$-point Gauss-Legendre quadrature rule can therefore be used to compute the coefficients of $\mathbf{D}_z$ exactly.

With a similar reasoning, we can show that the matrices $\mathbf{L}^\oplus$ and $\mathbf{Q}^\oplus$ can be computed exactly using a $(N+1)$-point Gauss-Jacobi quadrature rule and, if need be, rotation matrices.

A.1.5 Advantages of the Real-Form Spherical Harmonics

Using complex or real-form spherical harmonics does not change the number of unknowns. However there are reasons why the real-form spherical harmonics are more appealing than their complex counterparts, which all come down to a simple observation: the solution of the transport equation being intrinsically a real quantity, trying to represent it with complex functions seems to be asking for trouble.

More specifically, expanding the angular flux as:

$$\Psi = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} \Psi_\ell^m Y_\ell^m, \tag{A.26}$$

a priori yields complex coefficients Ψ_ℓ^m, unless the solution spatially only depends on z: all the moments with $m > 0$ have a real part and a non-zero imaginary part. This admittedly does not increase the number of non-redundant moments because it is possible to reorder the expansion:

$$\begin{aligned}
\Psi &= \sum_{\ell=0}^{\infty} \Big(\Psi_\ell^0 Y_\ell^0 + \sum_{m=1}^{\ell} \left(\Psi_\ell^m Y_\ell^m + \Psi_\ell^{-m} Y_\ell^{-m} \right) \Big) \\
&= \sum_{\ell=0}^{\infty} \Big(\Psi_\ell^0 Y_\ell^0 + \sum_{m=1}^{\ell} \left(\Psi_\ell^m Y_\ell^m + \overline{\Psi_\ell^m Y_\ell^m} \right) \Big) \\
&= \sum_{\ell=0}^{\infty} \Big(\Psi_\ell^0 Y_\ell^0 + 2 \sum_{m=1}^{\ell} \Re(\Psi_\ell^m Y_\ell^m) \Big) \\
&= \sum_{\ell=0}^{\infty} \Big(\Psi_\ell^0 Y_\ell^0 + 2 \sum_{m=1}^{\ell} \left(\Re(\Psi_\ell^m) \Re(Y_\ell^m) - \Im(\Psi_\ell^m) \Im(Y_\ell^m) \right) \Big),
\end{aligned} \tag{A.27}$$

using the fact that Ψ is a real function. There are still $(N+1)^2$ unknowns in a general case: $\Re(\Psi_\ell^m)$ for $0 \leq m \leq \ell$ and $\Im(\Psi_\ell^m)$ for $1 \leq m \leq \ell$. Nonetheless, there are several reasons that make the use of this formula less attractive.

A.1.5.1 Physical Meaning

Let alone the fact that it is less intuitive to work with imaginary quantities, some expansion coefficients with $m > 1$ lose their physical meaning (they are unchanged for $m = 0$). In particular, Ψ_0^0 and Ψ_0^1 still represent the scalar flux and the partial current along the z-direction, respectively. However, Ψ_1^1 and Ψ_1^{-1} with real-form spherical harmonics can be identified as the partial current along the x- and the

y-axis, respectively, whereas $\Re(\Psi_1^1)$ and $\Im(\Psi_1^1)$ have no such meaning with complex spherical harmonics.

A.1.5.2 Symmetric Matrices

Using the complex spherical harmonics makes the $\mathbf{D}_x$ and $\mathbf{D}_y$ matrices non-symmetric. This is because, for $d > 1$, the $m < 0$ expansion coefficients are non-zero, although they can be expressed in terms of the positive m expansion terms. The equations for $m = 0$ (see Eqs. A.4–A.6) therefore still couples the negative and positive m terms which doubles the corresponding coefficients and breaks the symmetry of the matrix.[2]

For the real form spherical harmonics, although it is not obvious to see it directly from the more complicated recurrence relationships (see Section A.1.3.4), $\mathbf{D}_x$ and $\mathbf{D}_y$ are symmetric. It can actually be directly seen in their definitions.

A.1.5.3 Reflecting Boundary Conditions

Lastly, if the boundary conditions are determined using ghost cells (in the same way as in [52]), it may be more complicated using complex spherical harmonics.

A.2 Parity Considerations

A.2.1 Spherical Harmonics Parity

As a reminder, the real-form spherical harmonics are defined as:

$$R_\ell^m(\vec{\Omega}) = \begin{cases} \sqrt{2}\, C_\ell^m\, P_\ell^m(\mu)\cos(m\varphi), & 0 < m \leq \ell \leq N \\ C_\ell^0\, P_\ell^0(\mu), & 0 \leq \ell \leq N \\ \sqrt{2}\, C_\ell^{|m|}\, P_\ell^{|m|}(\mu)\sin(|m|\varphi), & 0 < -m \leq \ell \leq N \end{cases}, \tag{A.28}$$

[2] This is very much similar to what happens for instance to a mass or stiffness matrix when a reflecting boundary condition is applied: the off-diagonal elements of the row corresponding to the reflecting nodes are multiplied by 2.

P_ℓ^m designating the associated Legendre polynomial of degree ℓ and order m and $C_\ell^m = \sqrt{\frac{(2\ell+1)}{w}\frac{(\ell-m)!}{(\ell+m)!}}$ being a normalization constant. The associated Legendre polynomial is further defined to be:

$$P_\ell^m(\mu) = (-1)^m \frac{(1-\mu^2)^{m/2}}{2^\ell \ell!} \frac{\mathrm{d}^{\ell+m}}{\mathrm{d}\mu^{\ell+m}}\big((\mu^2-1)^\ell\big), \tag{A.29}$$

the $(-1)^m$ term – also known as the Condon-Shortley phase – may or may not be included.[3] Yet,

$$\frac{\mathrm{d}^{\ell+m}}{\mathrm{d}\mu^{\ell+m}}\big((\mu^2-1)^\ell\big) = \frac{\mathrm{d}^{\ell+m}}{\mathrm{d}\mu^{\ell+m}}\left(\sum_{k=0}^{\ell}\binom{\ell}{k}(-1)^{\ell-k}\mu^{2k}\right), \tag{A.30}$$

where, for all $a, b \in \mathbb{N}$, $\binom{a}{b} \equiv \frac{a!}{b!(a-b)!}$. Besides, since μ^{2k} is an even function of μ, $\frac{\mathrm{d}^{\ell+m}}{\mathrm{d}\mu^{\ell+m}}\mu^{2k}$ is an odd function of μ if and only if $(\ell+m)$ is odd (if we disregard the trivial cases $(\ell+m) > 2k$). This implies that the parity of P_ℓ^m is that of $(\ell+m)$.

In addition, for a given $\vec{\Omega} = \sqrt{1-\mu^2}\cos\varphi\,\vec{e}_x + \sqrt{1-\mu^2}\sin\varphi\;\vec{e}_y + \mu\,\vec{e}_z$, we have:

$$-\vec{\Omega} = \begin{pmatrix} \sqrt{1-\mu^2}\cos(\varphi+\pi) \\ \sqrt{1-\mu^2}\sin(\varphi+\pi) \\ -\mu \end{pmatrix}. \tag{A.31}$$

In other words, $\vec{\Omega}$ is changed in $-\vec{\Omega}$ if μ and φ are respectively changed in $-\mu$ and

[3] This typically has very little impact in a code. In particular, most matrices (rotation matrices, $\vec{\mathbf{D}}$, $\mathbf{H}_{u,v}$ etc.) are unchanged whether it is included or not because it will only change the sign of the corresponding moment for odd m.

$\varphi + \pi$. Therefore, for $m > 0$,

$$\begin{aligned} R_\ell^m(-\vec{\Omega}) &= \sqrt{2}\, C_\ell^m\, P_\ell^m(-\mu)\cos(m(\varphi+\pi)), \\ &= \sqrt{2}\, C_\ell^m\, (-1)^{\ell+m} P_\ell^m(\mu)\, (-1)^m \cos(m\varphi), \\ &= (-1)^\ell R_\ell^m(\vec{\Omega}). \end{aligned} \tag{A.32}$$

The same property can be established for $m \leq 0$, which implies that the parity of R_ℓ^m is entirely determined by the parity of ℓ. More specifically, R_ℓ^m is an even function of $\vec{\Omega}$ if and only if ℓ is even.

A.2.2 Number of Even and Odd Moments

In this section, we derive the number of 'even' and 'odd' moments for N odd, respectively noted P_e and P_o. The total number of moments is noted $P = P_e + P_o$.

A.2.2.1 One Dimension

Let us derive the number of moments for a problem depending only on $d = 1$ spatial dimension, in which case we only need the moments Φ_ℓ^m such that $0 \leq \ell \leq N$ and $m = 0$. It follows:

$$P_o = \sum_{\substack{\ell=1 \\ \ell \text{ odd}}}^{N} 1 = \frac{N+1}{2}, \tag{A.33}$$

$$P_e = \sum_{\substack{\ell=0 \\ \ell \text{ even}}}^{N} 1 = \frac{N+1}{2}, \tag{A.34}$$

We indeed have $P = P_e + P_o = N + 1$, which is expected. Furthermore, for N odd, we thus have:

$$\frac{P_e}{P} = \frac{P_o}{P} = \frac{1}{2}. \tag{A.35}$$

A.2.2.2 Two Dimensions

Let us derive the number of moments for a problem depending only on $d = 2$ spatial dimensions, in which case we only need the moments Φ_ℓ^m such that $0 \leq m \leq \ell \leq N$. The number of odd moments is given by[4]:

$$\begin{aligned} P_o &= \sum_{\substack{\ell=1 \\ \ell \text{ odd}}}^{N} \sum_{m=0}^{\ell} 1 = \sum_{\substack{\ell=1 \\ \ell \text{ odd}}}^{N} (\ell+1), \\ &= \sum_{\ell'=0}^{(N-1)/2} (2\ell'+2) = (N+1) + 2 \sum_{\ell'=0}^{(N-1)/2} \ell', \\ &= (N+1) + \frac{(N-1)}{2}\frac{(N+1)}{2}, \\ &= \frac{(N+1)(N+3)}{4}. \end{aligned} \tag{A.36}$$

Similarly, the number of even moments is given by[5]:

$$\begin{aligned} P_e &= \sum_{\substack{\ell=0 \\ \ell \text{ even}}}^{N} \sum_{m=0}^{\ell} 1 = \sum_{\substack{\ell=0 \\ \ell \text{ even}}}^{N} (\ell+1), \\ &= \sum_{\ell''=0}^{(N-1)/2} (2\ell''+1) = \frac{(N+1)}{2} + 2 \sum_{\ell''=0}^{(N-1)/2} \ell'', \\ &= \frac{(N+1)}{2} + \frac{(N-1)}{2}\frac{(N+1)}{2}, \\ &= \frac{(N+1)^2}{4}. \end{aligned} \tag{A.37}$$

As a sanity check, we can see that the total number of moments is:

$$P = P_e + P_o = \frac{(N+1)^2}{4} + \frac{(N+1)(N+3)}{4} = \frac{(N+1)(N+2)}{2}, \tag{A.38}$$

which is the well-know value.

[4] Using the substitution $\ell \equiv 2\ell' + 1$

[5] Using the substitution $\ell \equiv 2\ell''$

This implies that:

$$\frac{P_e}{P} = \frac{N+1}{2(N+2)} \quad , \quad \frac{P_o}{P} = \frac{N+3}{2(N+2)}. \tag{A.39}$$

A.2.2.3 Three Dimensions

Let us derive the number of moments for a problem depending only on $d = 3$ spatial dimensions, in which case we need all the moments, i.e. Φ_ℓ^m such that $0 \leq |m| \leq \ell \leq N$. The number of odd moments is given by:

$$\begin{aligned} P_o &= \sum_{\substack{\ell=1 \\ \ell \text{ odd}}}^{N} \sum_{m=-\ell}^{\ell} 1 = \sum_{\substack{\ell=1 \\ \ell \text{ odd}}}^{N} (2\ell+1), \\ &= \sum_{\ell'=0}^{(N-1)/2} (4\ell'+3) = \frac{3(N+1)}{2} + 4 \sum_{\ell'=0}^{(N-1)/2} \ell', \\ &= \frac{3(N+1)}{2} + \frac{(N-1)(N+1)}{2}, \\ &= \frac{(N+1)(N+2)}{2}. \end{aligned} \tag{A.40}$$

Similarly, the number of even moments is given by:

$$\begin{aligned} P_e &= \sum_{\substack{\ell=0 \\ \ell \text{ even}}}^{N} \sum_{m=-\ell}^{\ell} 1 = \sum_{\substack{\ell=0 \\ \ell \text{ even}}}^{N} (2\ell+1), \\ &= \sum_{\ell''=0}^{(N-1)/2} (4\ell''+1) = \frac{(N+1)}{2} + 4 \sum_{\ell''=0}^{(N-1)/2} \ell'', \\ &= \frac{(N+1)}{2} + (N-1)\frac{(N+1)}{2}, \\ &= \frac{N(N+1)}{2}, \end{aligned} \tag{A.41}$$

As a sanity check, we can see that the total number of moments is:

$$P = P_e + P_o = \frac{N(N+1)}{2} + \frac{(N+1)(N+2)}{2} = (N+1)^2, \qquad \text{(A.42)}$$

as expected.

A.2.2.4 Summary

Table A.1 summarizes the number of moments for $d = 1, 2, 3$ for N odd. In all three cases, we have:

$$\frac{P_e}{P} \xrightarrow[N \longrightarrow \infty]{} \frac{1}{2} \quad , \quad \frac{P_o}{P} \xrightarrow[N \longrightarrow \infty]{} \frac{1}{2}, \qquad \text{(A.43)}$$

with these limits being an equality for $d = 1$.

It seems thus acceptable to say that the number of moments is roughly speaking divided by two when solving only for the even-parity moments.

Dimension d	P_e	P_o	P	$\frac{P_e}{P}$	$\frac{P_o}{P}$
1	$\frac{N+1}{2}$	$\frac{N+1}{2}$	$N+1$	$\frac{1}{2}$	$\frac{1}{2}$
2	$\frac{(N+1)^2}{4}$	$\frac{(N+1)(N+3)}{4}$	$\frac{(N+1)(N+2)}{2}$	$\frac{N+1}{2(N+2)}$	$\frac{N+3}{2(N+2)}$
3	$\frac{N(N+1)}{2}$	$\frac{(N+1)(N+2)}{2}$	$(N+1)^2$	$\frac{N}{2(N+1)}$	$\frac{N+2}{2(N+1)}$

Table A.1: Number of even, odd and all moments for $d = 1, 2, 3$, for N odd. For N even, the value for P is unchanged.

A.3 Moments of the Volumetric Source

This section is dedicated to show how to obtain the moments of the volumetric source given by Eq. 4.100. The details of the problem and notation can be found in Section 4.6.2.

As a reminder, because of the 1-D nature of the problem, the superscript indicating the order of the spherical harmonics m are omitted, implying $m = 0$. Besides, the volumetric source is given by:

$$\tilde{S} = \mu \frac{\mathrm{d}f}{\mathrm{d}x} g + \sigma_t f g, \tag{A.44}$$

and g is expanded as:

$$g = \sum_{\ell=0}^{\infty} g_\ell R_\ell \tag{A.45}$$

where the spherical harmonic R_ℓ and the Legendre polynomial P_ℓ of degree ℓ satisfy[6]:

$$R_\ell = \sqrt{\frac{2\ell+1}{2}} P_\ell. \tag{A.46}$$

A.3.1 Property

We have the well-known recurrence relation for any $\ell \in \mathbb{N}$:

$$\mu P_\ell = \frac{\ell+1}{2\ell+1} P_{\ell+1} + \frac{\ell}{2\ell+1} P_{\ell-1} \tag{A.47}$$

[6] See Eq. 2.12 with $w = 2$

with the convention $P_{-1} = 0$. Similarly,

$$\begin{aligned}
\mu R_\ell &= \frac{\ell+1}{\sqrt{2(2\ell+1)}} P_{\ell+1} + \frac{\ell}{\sqrt{2(2\ell+1)}} P_{\ell-1} \\
&= \frac{\ell+1}{\sqrt{(2\ell+1)(2\ell+3)}} R_{\ell+1} + \frac{\ell}{\sqrt{(2\ell+1)(2\ell-1)}} R_{\ell-1}
\end{aligned} \tag{A.48}$$

with the convention $R_{-1} = 0$.

A.3.2 Moments of $\tilde{S}$

The ℓ-th ($\ell \in \mathbb{N}$) moment of $\tilde{S}$ is then given by:

$$\begin{aligned}
\tilde{S}_\ell &= \int_{-1}^{1} \tilde{S}\, R_\ell \,\mathrm{d}\mu, \\
&= \int_{-1}^{1} \left(\mu \frac{\mathrm{d}f}{\mathrm{d}x}\, g + \sigma_t f g \right) R_\ell \,\mathrm{d}\mu, \\
&= \frac{\mathrm{d}f}{\mathrm{d}x} \int_{-1}^{1} \mu \sum_{\ell'=0}^{\infty} g_{\ell'} R_{\ell'}\, R_\ell \,\mathrm{d}\mu + \sigma_t f \int_{-1}^{1} \sum_{\ell'=0}^{\infty} g_{\ell'} R_{\ell'}\, R_\ell \,\mathrm{d}\mu, \\
&= \frac{\mathrm{d}f}{\mathrm{d}x} \int_{-1}^{1} \sum_{\ell'=0}^{\infty} g_{\ell'} \left(\frac{\ell'+1}{\sqrt{(2\ell'+1)(2\ell'+3)}} R_{\ell'+1} + \frac{\ell'}{\sqrt{(2\ell'+1)(2\ell'-1)}} R_{\ell'-1} \right) R_\ell \,\mathrm{d}\mu \\
&\qquad + \sigma_t f\, g_\ell .
\end{aligned} \tag{A.49}$$

Therefore, with the convention $g_{-1} = 0$,

$$\tilde{S}_\ell = \frac{\mathrm{d}f}{\mathrm{d}x} \left(\frac{\ell}{\sqrt{(2\ell+1)(2\ell-1)}} g_{\ell-1} + \frac{\ell+1}{\sqrt{(2\ell+1)(2\ell+3)}} g_{\ell+1} \right) + \sigma_t f\, g_\ell . \tag{A.50}$$

APPENDIX B

ROSSELAND-AVERAGED CROSS-SECTIONS FOR THE MULTIGROUP CROOKED PIPE TEST PROBLEM

The purpose of this appendix is to show how we determine the temperature-dependent 2-group cross-sections for the multigroup Crooked Pipe test problem in Section 3.4.4.

B.1 Model Opacity

We follow the model opacity given in [63]:

$$\sigma_a(\nu, T) = \sigma_0 \frac{1 - \exp\left(-h\nu/kT\right)}{(h\nu)^3}, \tag{B.1}$$

where σ_0 is an arbitrary constant.[1] In addition, ν is the photon frequency.

Considering the g-th group with an energy interval $[E_{g-1}, E_g] = [h\nu_{g-1}, h\nu_g]$, we compute the multigroup Rosseland cross-sections:

$$\sigma_{a,g} = \frac{\displaystyle\int_{\nu_{g-1}}^{\nu_g} \frac{\partial B_\nu}{\partial T} \mathrm{d}\nu}{\displaystyle\int_{\nu_{g-1}}^{\nu_g} \frac{1}{\sigma} \frac{\partial B_\nu}{\partial T} \mathrm{d}\nu}, \tag{B.2}$$

where the frequency-dependent Planckian B_ν is given by:

$$B_\nu = \frac{2h\nu^3}{c^2} \frac{1}{\exp\left(\frac{h\nu}{kT}\right) - 1}, \tag{B.3}$$

[1] Note that σ_0 does not have the same units as σ_a but we keep the same notations as in [63].

which implies:

$$\frac{\partial B_\nu}{\partial T} = \frac{2h^2\nu^4}{c^2kT^2}\frac{\exp\left(\frac{h\nu}{kT}\right)}{\left(\exp\left(\frac{h\nu}{kT}\right)-1\right)^2}. \tag{B.4}$$

For numerical integration purposes, it can be useful to note that:

$$\begin{aligned}\left(\exp\left(\alpha\right)-1\right)^2 &= \exp\left(\alpha\right)\left(\exp\left(\alpha/2\right)-\exp\left(-\alpha/2\right)\right)^2, \\ &= 4\exp\left(\alpha\right)\left(\sinh\left(\alpha/2\right)\right)^2, \\ &= \frac{4\exp\left(\alpha\right)}{\operatorname{csch}^2\left(\alpha/2\right)},\end{aligned} \tag{B.5}$$

and thus that, defining $\alpha = h\nu/kT$:

$$\frac{\partial B_\nu}{\partial T} = \frac{h^2\nu^4}{2c^2kT^2}\operatorname{csch}^2\left(\frac{h\nu}{2kT}\right), \tag{B.6}$$

this final expression avoiding to have potentially huge values on both the numerator and denominator. Furthermore, for code verification, it is useful to recall that:

$$\int_0^\infty B_\nu \mathrm{d}\nu = \frac{acT^4}{4\pi} \quad , \quad \int_0^\infty \frac{\partial B_\nu}{\partial T}\mathrm{d}\nu = \frac{acT^3}{\pi}. \tag{B.7}$$

B.2 Application to the Multigroup Crooked Pipe Test Problem

In this section, we explain how we chose the energy structure that led to the determination of the multigroup cross-sections for the crooked pipe. Because of the large number of unknowns in the one-group calculation, we choose to limit ourselves to two energy groups, without loss of generality.

We will refer to the groups $g = 1$ and $g = 2$ as the thermal and fast groups, respectively. Besides, the red and blue regions from Fig. 3.6 are called *thick* and *thin* regions, respectively.

B.2.1 Energy Structure

To determine the energy bounds of the two groups, let us first look at the Planckian distributions for particular temperatures. Physically speaking, the maximum temperature in the Crooked Pipe test problem is $T = 0.3$ keV while the minimum[2] temperature is $T = 0.05$ keV. Fig. B.1 shows the distributions for these two temperatures. It therefore appears that the vast majority of the Planckian re-emission will occur between $h\nu_{\mathrm{min}} \equiv 1$ eV and $h\nu_{\mathrm{max}} \equiv 10$ keV.

Now, the question is to determine what the frequency limit ν_1 between the two energy groups should be.

First, we want to have the energy mostly injected in the fast group so that the filtering is mainly needed therein. The curve corresponding to $T = 0.3$ keV on Fig. B.1 actually also represents the emission profile imposed at the left entrance of the pipe.

Second, we of course need the Planckian re-emission inside the pipe to have some contribution to the thermal group (otherwise, we might as well run a one-group calculation). The curve corresponding to $T = 0.05$ keV and $T = 0.1$ keV on Figure B.1 gives us a good idea of what the re-emission profile will look like at early times.

Therefore, if we want to have most of the re-emission corresponding to $T < 0.1$ keV and to $T = 0.3$ keV to occur mostly in the thermal and fast groups, respectively, a good trade-off could be:

$$h\nu_1 \equiv 0.3\,\mathrm{keV}. \tag{B.8}$$

[2] The temperature can actually go slightly below 50 eV near the boundaries of the domain and in particular near the right entrance of the pipe. This, however, happens far from the main regions of interest, which are located after the first elbow in the pipe.

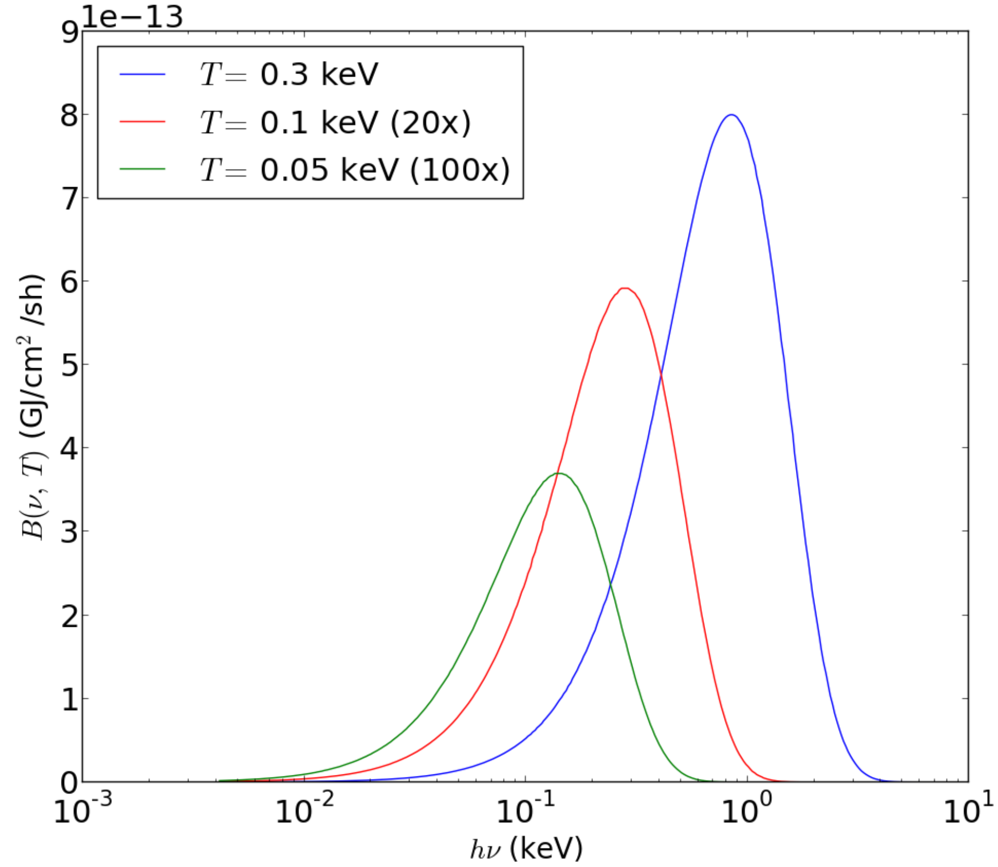


Figure B.1: Value of $B(\nu, T)$ as a function of $h\nu$ for several values of T. The curves for $T = 0.05$ keV and $T = 0.1$ keV have been respectively been multiplied by 100 and 20.

In summary, our energy structure is chosen to be:

$$(h\nu_{\min}, h\nu_1, h\nu_{\max}) = (0.001\,\mathrm{keV},\ 0.3\,\mathrm{keV},\ 10\,\mathrm{keV}). \tag{B.9}$$

B.2.2 Initial and Boundary Conditions

For the one-group – or gray – calculation, we imposed the following incoming Dirichlet boundary condition at the left entrance of the pipe:

$$\Psi^{\mathrm{d}} = \frac{acT_L^4}{4\pi}, \quad T_L = 0.3\,\mathrm{keV}. \tag{B.10}$$

In other words, we enforced that the radiation field and the material internal energy in the region for $x < 0$ are at equilibrium.

Similarly, for a multigroup calculation, we then have (from the transport equation in steady-state and infinite medium):

$$\Psi_g^{\mathrm{d}} = B_g(T_L) = \int_{\nu_{g-1}}^{\nu_g} B(\nu, T_L)\mathrm{d}\nu. \tag{B.11}$$

Likewise, the initial conditions should be given by:

$$\Psi_g(t=0) = B_g(T_0) = \int_{\nu_{g-1}}^{\nu_g} B(\nu, T_0)\mathrm{d}\nu, \tag{B.12}$$

where $T_0 = 0.05$ keV is the initial temperature.

B.2.3 Temperature-Dependent Cross-Sections

We choose σ_0 in Eq. B.1 such that the initial (i.e. for $T = 0.05$ keV) cross-section for the thermal group is equal to the one-group cross-sections of the Crooked Pipe. Table B.1 then gives the values for the fast and thermal groups in the thick and thin regions and different values of the temperature.

Using a power interpolation, we obtain:

$$\sigma_{a,\text{thick, thermal}}(T) = 41.00\,T^{-0.5163}, \tag{B.13}$$

	Thin region		Thick region		
T (keV)	Thermal group	Fast group	Thermal group	Fast group	Ratio
0.05	0.2000	2.576E-2	200.0	25.76	7.76
0.06	0.1734	1.980E-2	173.4	19.80	8.76
0.07	0.1577	1.515E-2	157.7	15.15	10.4
0.08	0.1472	1.162E-2	147.2	11.62	12.7
0.09	0.1392	8.989E-3	139.2	8.99	15.5
0.1	0.1328	7.026E-3	132.8	7.03	18.9
0.11	0.1274	5.557E-3	127.4	5.56	22.9
0.12	0.1226	4.447E-3	122.6	4.45	27.6
0.13	0.1183	3.601E-3	118.3	3.60	32.8
0.14	0.1143	2.949E-3	114.3	2.95	38.8
0.15	0.1106	2.440E-3	110.6	2.44	45.3
0.16	0.1072	2.039E-3	107.2	2.04	52.6
0.17	0.1040	1.719E-3	104.0	1.72	60.5
0.18	0.1010	1.461E-3	101.0	1.46	69.1
0.19	0.0981	1.251E-3	98.1	1.25	78.4
0.2	0.0954	1.079E-3	95.4	1.08	88.4
0.21	0.0928	9.371E-4	92.8	0.937	99.1
0.22	0.0904	8.185E-4	90.4	0.818	110
0.23	0.0881	7.188E-4	88.1	0.719	123
0.24	0.0859	6.346E-4	85.9	0.635	135
0.25	0.0838	5.629E-4	83.8	0.563	149
0.26	0.0818	5.015E-4	81.8	0.502	163
0.27	0.0799	4.487E-4	79.9	0.449	178
0.28	0.0780	4.030E-4	78.0	0.403	194
0.29	0.0763	3.632E-4	76.3	0.363	210
0.3	0.0746	3.285E-4	74.6	0.329	227

Table B.1: Rosseland-averaged temperature-dependent cross-sections for the 2-group Crooked Pipe test problem in cm^{-1}. The last column shows the ratio of the thermal to fast cross-sections (for either the thin or the thick region).

$$\sigma_{a,\text{thick, fast}}(T) = 0.01702\, T^{-2.564}, \tag{B.14}$$

where T is expressed in keV and σ_a in cm^{-1} and with the coefficients of determination being $R^2_{\text{thermal}} = 0.996$ and $R^2_{\text{fast}} = 0.993$, respectively, These are used as definitions for σ_a in the thick region in Section 3.4.4. The values in the thin regions are 1000 times less.

B.2.4 Implementation

From an implementation point of view, it was noticed that lagging the temperature dependency of the cross-sections (i.e. computing them with the value of the temperature from the previous time step) accelerated the solver convergence and minimally changed the results.

APPENDIX C

USEFUL MATRIX PROPERTY FOR THE EQUIVALENCE OF THE EVEN-PARITY P$_N$ AND THE 'EVEN' SAAF–P$_N$ FORMULATIONS

In this appendix, we consider Eq. 4.4 where we disregard – for simplicity – the fission terms. As a reminder, it can then be written:

$$\vec{\Omega} \cdot \vec{\nabla}\Psi + \sigma_{\mathrm{t}}(\vec{r})\Psi(\vec{r},\vec{\Omega}) = \sum_{\ell=0}^{N_s} \sigma_{s,\ell}(\vec{r}) \sum_{m=-\ell}^{\ell} \Phi_\ell^m(\vec{r}) R_\ell^m(\vec{\Omega}) + S(\vec{r},\vec{\Omega}) = H\Psi + S. \quad \text{(C.1)}$$

We attempt to prove Eq. 4.79, although it would not be surprising if a more concise proof could be derived. First, we define some operators that can be used to derive Eq. 4.74 in a different way. Second, we use those operators to prove Eq. 4.79.

C.1 Alternative Derivation

We define the following operator[1], for any function $f = f(\vec{\Omega})$:

$$\bar{G}^{-1}\,(\vec{\Omega}) \equiv \sigma_{\mathrm{t}} f - H f. \quad \text{(C.2)}$$

In particular,

$$\bar{G}\,f = (\sigma_{\mathrm{t}}\mathbb{I} - H)^{-1} f = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} \frac{1}{\sigma_{\mathrm{t}} - \sigma_{s,\ell}} R_\ell^m(\vec{\Omega}) \int_{\mathbb{S}^2} R_\ell^m(\vec{\Omega}') f(\vec{\Omega}')\,\mathrm{d}\Omega'. \quad \text{(C.3)}$$

Therefore, further defining:

$$G\,f(\vec{\Omega}) \equiv \bar{G}\,f(\vec{\Omega}) - \frac{1}{\sigma_{\mathrm{t}}} f(\vec{\Omega}), \quad \text{(C.4)}$$

[1] Part of this derivation comes from a personal communication with Dr. Yaqi Wang.

and using the fact that:

$$f(\vec{\Omega}) = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} R_\ell^m(\vec{\Omega}) \int_{\mathbb{S}^2} R_\ell^m(\vec{\Omega}') f(\vec{\Omega}') \, \mathrm{d}\Omega', \tag{C.5}$$

we have:

$$G\, f(\vec{\Omega}) = \sum_{\ell=0}^{N_s} \sum_{m=-\ell}^{\ell} \frac{\sigma_{s,\ell}}{(\sigma_\mathrm{t} - \sigma_{s,\ell})\sigma_\mathrm{t}} R_\ell^m(\vec{\Omega}) \int_{\mathbb{S}^2} R_\ell^m(\vec{\Omega}') f(\vec{\Omega}') \, \mathrm{d}\Omega', \tag{C.6}$$

since $\sigma_{s,\ell} = 0$ for $\ell > N_s$. Yet, Eq. 4.61 can then be expressed as:

$$\begin{aligned} \Psi_o &= (\sigma_\mathrm{t} - H)^{-1} \left(S_o - \vec{\Omega} \cdot \vec{\nabla} \Psi_e \right) = \bar{G} S_o - \bar{G}\, \vec{\Omega} \cdot \vec{\nabla} \Psi_e, \\ &= \bar{G} S_o - G\, \vec{\Omega} \cdot \vec{\nabla} \Psi_e - \frac{1}{\sigma_\mathrm{t}} \vec{\Omega} \cdot \vec{\nabla} \Psi_e. \end{aligned} \tag{C.7}$$

Besides, since $\vec{\Omega} \cdot \vec{\nabla} \Psi_e$ is odd,

$$\begin{aligned} G\, \vec{\Omega} \cdot \vec{\nabla} \Psi_e &= \sum_{\substack{\ell=0 \\ \ell \text{ odd}}}^{N_s} \sum_{m=-\ell}^{\ell} \frac{\sigma_{s,\ell}}{(\sigma_\mathrm{t} - \sigma_{s,\ell})\sigma_\mathrm{t}} R_\ell^m(\vec{\Omega}) \int_{\mathbb{S}^2} R_\ell^m(\vec{\Omega}')\, \vec{\Omega}' \cdot \vec{\nabla} \Psi_e \, \mathrm{d}\Omega', \\ &= \sum_{\substack{\ell=0 \\ \ell \text{ odd}}}^{N} \sum_{m=-\ell}^{\ell} \frac{\sigma_{s,\ell}}{(\sigma_\mathrm{t} - \sigma_{s,\ell})\sigma_\mathrm{t}} R_\ell^m(\vec{\Omega}) \int_{\mathbb{S}^2} R_\ell^m(\vec{\Omega}')\, \vec{\Omega}' \cdot \vec{\nabla} \left(\mathbf{R}_e^T \, \boldsymbol{\Phi}_e \right) \, \mathrm{d}\Omega', \\ &= \mathbf{R}_o^T \operatorname*{diag}_{\ell \text{ odd}} \{ \frac{\sigma_{s,\ell}}{(\sigma_\mathrm{t} - \sigma_{s,\ell})\sigma_\mathrm{t}}, \ 0 \le |m| \le \ell \le N; \} \, \vec{\mathbf{D}}_e^T \cdot \vec{\nabla} \boldsymbol{\Phi}_e. \end{aligned} \tag{C.8}$$

It follows:

$$\int_{\mathbb{S}^2} \mathbf{R}_o \, G\, \vec{\Omega} \cdot \vec{\nabla} \Psi_e \, \mathrm{d}\Omega = \operatorname*{diag}_{\ell \,\text{odd}} \{ \frac{\sigma_{s,\ell}}{(\sigma_\mathrm{t} - \sigma_{s,\ell})\sigma_\mathrm{t}}, \ 0 \le |m| \le \ell \le N; \} \, \vec{\mathbf{D}}_e^T \cdot \vec{\nabla} \boldsymbol{\Phi}_e. \tag{C.9}$$

Similarly, it yields:

$$\int_{\mathbb{S}^2} \mathbf{R}_o \left(G\vec{\Omega}\cdot\vec{\nabla}\Psi_e + \frac{1}{\sigma_\mathrm{t}}\vec{\Omega}\cdot\vec{\nabla}\Psi_e \right) \mathrm{d}\Omega = \operatorname*{diag}_{\ell\,\mathrm{odd}}\{\frac{1}{(\sigma_\mathrm{t}-\sigma_{s,\ell})},\, 0 \le |m| \le \ell \le N\}\, \vec{\mathbf{D}}_e^T\cdot\vec{\nabla}\boldsymbol{\Phi}_e. \tag{C.10}$$

In other words:

$$\int_{\mathbb{S}^2} \mathbf{R}_o\, \bar{G}\, \vec{\Omega}\cdot\vec{\nabla}\Psi_e \,\mathrm{d}\Omega = (\sigma_\mathrm{t}\mathbb{I} - \boldsymbol{\sigma}_{s,o})^{-1}\, \vec{\mathbf{D}}_e^T \cdot \vec{\nabla}\boldsymbol{\Phi}_e. \tag{C.11}$$

Likewise:

$$\int_{\mathbb{S}^2} \mathbf{R}_o\, \bar{G}\, S_o \,\mathrm{d}\Omega = (\sigma_\mathrm{t}\mathbb{I} - \boldsymbol{\sigma}_{s,o})^{-1}\, \mathbf{S}_o, \tag{C.12}$$

which means that (using Eq. C.7 multiplied with $\mathbf{R}_o$ and integrated over $\mathbb{S}^2$):

$$\boldsymbol{\Phi}_o = \int_{\mathbb{S}^2} \mathbf{R}_o\, \Psi_o \,\mathrm{d}\Omega = (\sigma_\mathrm{t}\mathbb{I} - \boldsymbol{\sigma}_{s,o})^{-1} \Big(\mathbf{S}_o - \vec{\mathbf{D}}_e^T \cdot \vec{\nabla}\boldsymbol{\Phi}_e \Big). \tag{C.13}$$

This is another way of deriving Eq. (4.74).

C.2 Property

In addition, we have:

$$\begin{aligned}
-\Big(\vec{\Omega}\cdot\vec{\nabla}\Psi_e^*, \Psi_o\Big)_\mathcal{D} &= -\left(\vec{\Omega}\cdot\vec{\nabla}\Psi_e^*, \bar{G}S_o - G\,\vec{\Omega}\cdot\vec{\nabla}\Psi_e - \frac{1}{\sigma_\mathrm{t}}\vec{\Omega}\cdot\vec{\nabla}\Psi_e\right)_\mathcal{D}, \\
&= \left(\vec{\Omega}\cdot\vec{\nabla}\Psi_e^*, \frac{1}{\sigma_\mathrm{t}}\vec{\Omega}\cdot\vec{\nabla}\Psi_e\right)_\mathcal{D} - \Big(\vec{\Omega}\cdot\vec{\nabla}\Psi_e^*, \bar{G}S_o\Big)_\mathcal{D} + \Big(\vec{\Omega}\cdot\vec{\nabla}\Psi_e^*, G\,\vec{\Omega}\cdot\vec{\nabla}\Psi_e\Big)_\mathcal{D}, \\
&= \left(\vec{\Omega}\cdot\vec{\nabla}\Psi_e^*, \frac{1}{\sigma_\mathrm{t}}\vec{\Omega}\cdot\vec{\nabla}\Psi_e\right)_\mathcal{D} - \Big(\vec{\Omega}\cdot\vec{\nabla}\Psi_e^*, \bar{G}S_o\Big)_\mathcal{D} \\
&\qquad + \left(\vec{\Omega}\cdot\vec{\nabla}\Psi_e^*, \mathbf{R}_o^T \operatorname*{diag}_{\ell\,\mathrm{odd}}\{\frac{\sigma_{s,\ell}}{(\sigma_\mathrm{t}-\sigma_{s,\ell})\sigma_\mathrm{t}}\}\vec{\mathbf{D}}_e^T\cdot\vec{\nabla}\boldsymbol{\Phi}_e\right)_\mathcal{D}.
\end{aligned} \tag{C.14}$$

Expanding Ψ_e and Ψ_e^* respectively as $\mathbf{R}_e^T\mathbf{\Phi}_e$ and $\mathbf{R}_e^T\mathbf{\Phi}_e$, we have:

$$\left(\vec{\Omega}\cdot\vec{\nabla}\Psi_e^*, \bar{G}S_o\right)_{\mathcal{D}} = \int_{\mathcal{D}} \left(\vec{\nabla}\mathbf{\Phi}_e^*\right)^T \left(\int_{\mathbb{S}^2} \vec{\Omega}\,\mathbf{R}_e\mathbf{R}_o^T \mathrm{d}\Omega\right) \left(\sigma_{\mathrm{t}}\mathbb{I} - \boldsymbol{\sigma}_{s,o}\right)^{-1}\mathbf{S}_o\,\mathrm{d}r. \tag{C.15}$$

It follows:

$$\begin{aligned} -\left(\vec{\Omega}\cdot\vec{\nabla}\Psi_e^*, \Psi_o\right)_{\mathcal{D}} = \left(\vec{\nabla}\mathbf{\Phi}_e^*, \frac{1}{\sigma_{\mathrm{t}}}\underline{\mathbf{H}}_e\vec{\nabla}\mathbf{\Phi}_e\right)_{\mathcal{D}} &- \left(\vec{\nabla}\mathbf{\Phi}_e^*, \vec{\mathbf{D}}_e\left(\sigma_{\mathrm{t}}\mathbb{I} - \boldsymbol{\sigma}_{s,o}\right)^{-1}\mathbf{S}_o\right)_{\mathcal{D}} \\ &+ \left(\vec{\nabla}\mathbf{\Phi}_e^*, \vec{\mathbf{D}}_e \operatorname*{diag}_{\ell\,\mathrm{odd}}\{\frac{\sigma_{s,\ell}}{(\sigma_{\mathrm{t}} - \sigma_{s,\ell})\sigma_{\mathrm{t}}}\}\vec{\mathbf{D}}_e^T\cdot\vec{\nabla}\mathbf{\Phi}_e\right)_{\mathcal{D}}. \end{aligned} \tag{C.16}$$

Yet, expanding Ψ_o as $\mathbf{R}_o^T\mathbf{\Phi}_o$ and using Eq. 4.74 on the left hand-side, it reads:

$$\begin{aligned} &\left(\vec{\nabla}\mathbf{\Phi}_e^*, \vec{\mathbf{D}}_e\ \left(\sigma_{\mathrm{t}}\mathbb{I} - \boldsymbol{\sigma}_{s,o}\right)^{-1}\vec{\mathbf{D}}_e^T\cdot\vec{\nabla}\mathbf{\Phi}_e\right)_{\mathcal{D}} - \left(\vec{\nabla}\mathbf{\Phi}_e^*, \vec{\mathbf{D}}_e\ \left(\sigma_{\mathrm{t}}\mathbb{I} - \boldsymbol{\sigma}_{s,o}\right)^{-1}\mathbf{S}_o\right)_{\mathcal{D}} \\ &\qquad = \left(\vec{\nabla}\mathbf{\Phi}_e^*, \frac{1}{\sigma_{\mathrm{t}}}\underline{\mathbf{H}}_e\vec{\nabla}\mathbf{\Phi}_e\right)_{\mathcal{D}} - \left(\vec{\nabla}\mathbf{\Phi}_e^*, \vec{\mathbf{D}}_e\left(\sigma_{\mathrm{t}}\mathbb{I} - \boldsymbol{\sigma}_{s,o}\right)^{-1}\mathbf{S}_o\right)_{\mathcal{D}} \\ &\qquad\quad + \left(\vec{\nabla}\mathbf{\Phi}_e^*, \vec{\mathbf{D}}_e \operatorname*{diag}_{\ell\,\mathrm{odd}}\{\frac{\sigma_{s,\ell}}{(\sigma_{\mathrm{t}} - \sigma_{s,\ell})\sigma_{\mathrm{t}}}\}\vec{\mathbf{D}}_e^T\cdot\vec{\nabla}\mathbf{\Phi}_e\right)_{\mathcal{D}}, \end{aligned} \tag{C.17}$$

i.e.:

$$\begin{aligned} &\left(\vec{\nabla}\mathbf{\Phi}_e^*, \vec{\mathbf{D}}_e\ \left(\sigma_{\mathrm{t}}\mathbb{I} - \boldsymbol{\sigma}_{s,o}\right)^{-1}\vec{\mathbf{D}}_e^T\cdot\vec{\nabla}\mathbf{\Phi}_e\right)_{\mathcal{D}} \\ &\qquad = \left(\vec{\nabla}\mathbf{\Phi}_e^*, \frac{1}{\sigma_{\mathrm{t}}}\underline{\mathbf{H}}_e\vec{\nabla}\mathbf{\Phi}_e\right)_{\mathcal{D}} + \left(\vec{\nabla}\mathbf{\Phi}_e^*, \vec{\mathbf{D}}_e \operatorname*{diag}_{\ell\,\mathrm{odd}}\{\frac{\sigma_{s,\ell}}{(\sigma_{\mathrm{t}} - \sigma_{s,\ell})\sigma_{\mathrm{t}}}\}\vec{\mathbf{D}}_e^T\cdot\vec{\nabla}\mathbf{\Phi}_e\right)_{\mathcal{D}}. \end{aligned} \tag{C.18}$$

Thus, because:

$$\begin{aligned}
&\left(\vec{\nabla}\mathbf{\Phi}_e^*, \frac{1}{\sigma_\mathrm{t}}\underline{\mathbf{H}}_e\vec{\nabla}\mathbf{\Phi}_e\right)_{\mathcal{D}}, \\
&\quad= \left(\vec{\nabla}\mathbf{\Phi}_e^*, \vec{\mathbf{D}}_e\ (\sigma_\mathrm{t}\mathbb{I} - \boldsymbol{\sigma}_{s,o})^{-1}\,\vec{\mathbf{D}}_e^T \cdot \vec{\nabla}\mathbf{\Phi}_e\right)_{\mathcal{D}} - \left(\vec{\nabla}\mathbf{\Phi}_e^*, \vec{\mathbf{D}}_e \operatorname*{diag}_{\ell\,\mathrm{odd}}\{\frac{\sigma_{s,\ell}}{(\sigma_\mathrm{t} - \sigma_{s,\ell})\sigma_\mathrm{t}}\}\vec{\mathbf{D}}_e^T \cdot \vec{\nabla}\mathbf{\Phi}_e\right)_{\mathcal{D}}, \\
&\quad= \left(\vec{\nabla}\mathbf{\Phi}_e^*, \vec{\mathbf{D}}_e\ \operatorname*{diag}_{\ell\,\mathrm{odd}}\{\frac{1}{(\sigma_\mathrm{t} - \sigma_{s,\ell})} - \frac{\sigma_{s,\ell}}{(\sigma_\mathrm{t} - \sigma_{s,\ell})\sigma_\mathrm{t}}\}\vec{\mathbf{D}}_e^T \cdot \vec{\nabla}\mathbf{\Phi}_e\right)_{\mathcal{D}}, \\
&\quad= \left(\vec{\nabla}\mathbf{\Phi}_e^*, \frac{1}{\sigma_\mathrm{t}}\vec{\mathbf{D}}_e\,\vec{\mathbf{D}}_e^T \cdot \vec{\nabla}\mathbf{\Phi}_e\right)_{\mathcal{D}},
\end{aligned} \tag{C.19}$$

we indeed have:

$$\underline{\mathbf{H}}_e = \vec{\mathbf{D}}_e\,\vec{\mathbf{D}}_e^T, \tag{C.20}$$

which is Eq. 4.79. It can also be rewritten:

$$\mathbf{H}_{e,u,v} = \mathbf{D}_{e,u}\,\mathbf{D}_{e,v}^T, \tag{C.21}$$

for any $u, v \in \{1, ..., d\}$.

APPENDIX D

SAAF–LS METHOD

In this section, we present the variational formulation for the SAAF–LS method whose results are shown on Figures 5.1 and 5.2. It is noted that this method is only mentioned for comparison purposes and to highlight why the conservative correction yielding is so necessary.

D.1 Variational Formulation

D.1.1 LS on $\mathcal{D}_0$

As a reminder, the LS formulation applied to $\mathcal{D}_0$ is given by (see Eq. 4.23):

$$\left(L\Psi^\star, L\Psi\right)_{\mathcal{D}_0} + \langle c\Psi^\star, (\Psi - \Psi^{\mathrm{inc}})\rangle^-_{\partial\mathcal{D}_0} = \left(L\Psi^\star, H\psi + S\right)_{\mathcal{D}_0}, \tag{D.1}$$

which gives, in void:

$$\left(\vec{\Omega}\cdot\vec{\nabla}\Psi^\star, \vec{\Omega}\cdot\vec{\nabla}\Psi\right)_{\mathcal{D}_0} + \langle c\Psi^\star, (\Psi - \Psi^{\mathrm{inc}})\rangle^-_{\partial\mathcal{D}_0} = 0. \tag{D.2}$$

Assuming continuity of Ψ across Γ (i.e. $\Psi = \Psi^{\mathrm{inc}}$ for any incoming direction), it implies:

$$\langle c\Psi^\star, (\Psi - \Psi^{\mathrm{inc}})\rangle^-_{\Gamma} = 0, \tag{D.3}$$

So the LS formulation on $\mathcal{D}_0$ simply reduces to:

$$\left(\vec{\Omega}\cdot\vec{\nabla}\Psi^\star, \vec{\Omega}\cdot\vec{\nabla}\Psi\right)_{\mathcal{D}_0} + \langle c\Psi^\star, (\Psi - \Psi^{\mathrm{inc}})\rangle^-_{\partial\mathcal{D}^0} = 0, \tag{D.4}$$

where, as a reminder, $\partial\mathcal{D}^0 = \partial\mathcal{D}^0 \cap \partial\mathcal{D}$.

D.1.2 SAAF on $\mathcal{D}_1$

Likewise, the SAAF formulation applied to $\mathcal{D}_1$ is given by (see Eq. 4.8):

$$\begin{aligned}\left(\vec{\Omega}\cdot\vec{\nabla}\Psi^\star, \frac{1}{\sigma_\mathrm{t}}\vec{\Omega}\cdot\vec{\nabla}\Psi\right)_{\mathcal{D}_1} + (\sigma_t\Psi^\star, \Psi)_{\mathcal{D}_1} + \langle\Psi^\star, \Psi\rangle^+_{\partial\mathcal{D}_1} - \langle\Psi^\star, \Psi^\mathrm{inc}\rangle^-_{\partial\mathcal{D}_1} \\ = \left(\frac{1}{\sigma_\mathrm{t}}\vec{\Omega}\cdot\vec{\nabla}\Psi^\star + \Psi^\star, H\Psi + S\right)_{\mathcal{D}_1}.\end{aligned} \tag{D.5}$$

The boundary terms can be split between the terms on $\partial\mathcal{D}$ and the terms on Γ:

$$\begin{aligned}\left(\vec{\Omega}\cdot\vec{\nabla}\Psi^\star, \frac{1}{\sigma_\mathrm{t}}\vec{\Omega}\cdot\vec{\nabla}\Psi\right)_{\mathcal{D}_1} + (\sigma_t\Psi^\star, \Psi)_{\mathcal{D}_1} + \langle\Psi^\star, \Psi\rangle^+_{\partial\mathcal{D}^1} - \langle\Psi^\star, \Psi^\mathrm{inc}\rangle^-_{\partial\mathcal{D}^1} \\ + \langle\Psi^\star, \Psi\rangle^{+,1}_\Gamma - \langle\Psi^\star, \Psi^\mathrm{inc}\rangle^{-,1}_\Gamma = \left(\frac{1}{\sigma_\mathrm{t}}\vec{\Omega}\cdot\vec{\nabla}\Psi^\star + \Psi^\star, H\Psi + S\right)_{\mathcal{D}_1}.\end{aligned} \tag{D.6}$$

In this expression, we have used the notation $\langle\cdot,\cdot\rangle^{\pm,1}_\Gamma$ to indicate that the angular integration half-range $\pm\vec{\Omega}\cdot\vec{n}(\vec{r}) > 0$ is determined with $\vec{n}$ being the outward unit vector normal to Γ with respect to $\mathcal{D}_1$ (i.e. locally pointing towards $\mathcal{D}_0$).

Again, assuming that Ψ is continuous across Γ, the terms on Γ are such that:

$$\langle\Psi^\star, \Psi\rangle^{+,1}_\Gamma - \langle\Psi^\star, \Psi^\mathrm{inc}\rangle^{-,1}_\Gamma = \langle\Psi^\star, \Psi\rangle^1_\Gamma \tag{D.7}$$

where $\langle\cdot,\cdot\rangle^1_\Gamma$ means that the unit normal vector is pointing towards $\mathcal{D}_0$.

D.1.3 Hybrid SAAF-LS Weak Formulation

Scaling the LS terms with $1/\sigma$, we end up with the following hybrid SAAF-LS weak formulation: find $\Psi \in V$ such that for all $\Psi^* \in V$,

$$\begin{aligned} &\left(\vec{\Omega} \cdot \vec{\nabla}\Psi^\star, \frac{1}{\sigma_\mathrm{t}}\vec{\Omega} \cdot \vec{\nabla}\Psi\right)_{\mathcal{D}_1} + \left(\vec{\Omega} \cdot \vec{\nabla}\Psi^\star, \frac{1}{\sigma}\vec{\Omega} \cdot \vec{\nabla}\Psi\right)_{\mathcal{D}_0} + (\sigma_t\Psi^\star, \Psi)_{\mathcal{D}_1} \\ &+ \langle \frac{c}{\sigma}\Psi^\star, (\Psi - \Psi^{\mathrm{inc}})\rangle^-_{\partial\mathcal{D}^0} + \langle \Psi^\star, \Psi\rangle^+_{\partial\mathcal{D}^1} - \langle \Psi^\star, \Psi^{\mathrm{inc}}\rangle^-_{\partial\mathcal{D}^1} + \langle \Psi^\star, \Psi\rangle^1_\Gamma \\ &\qquad = \left(\frac{1}{\sigma_\mathrm{t}}\vec{\Omega} \cdot \vec{\nabla}\Psi^\star + \Psi^\star, H\Psi + S\right)_{\mathcal{D}_1}. \end{aligned} \tag{D.8}$$

D.2 Properties

Although the resulting bilinear form is symmetric, it lacks global conservation – even if we choose $c = \sigma$, which is the main reason why this method is pretty poor. In particular, it does not even preserve the infinite solution presented in Section 4.6.1.