DISSERTATION

submitted

to the

Combined Faculty for the Natural Sciences and Mathematics

of

Heidelberg University, Germany

for the degree of

Doctor of Natural Sciences

Put forward by

Engineer José Pablo Lucero Lorca

Born in Tunuyán, Provincia de Mendoza, República Argentina

Oral examination:

Multilevel Schwarz methods for multigroup radiation transport problems

Advisor: Prof. Dr. Guido Kanschat

A mis padres,

Mirta y José.

Acknowledgments

I will be forever beholden to Prof. Dr. Guido Kanschat, from the Faculty of Mathematics and Computer Science at Universität Heidelberg, for the chance to work with him throughout this doctorate. His knowledge, experience, generosity, openness and patience through our candid discussions made his guidance invaluable and provided a lasting mathematical intuition that I am fortunate to be able to carry for the rest of my career.

My gratitude goes to Prof. Dr. Jean Ragusa as well, from the Institute of Scientific Computation at Texas A&M University, for providing fruitful discussions and ideas from the beginning of my thesis. It is always a pleasure to discuss and learn from such a knowledgeable, humble and insightful scientist.

I owe a big part of this thesis to Prof. Dr. Martin Gander, from the Section de Mathématiques at Université de Genève, who introduced me to some of the methods we used. Prof. Dr. Gander's unassuming, contagious love for precise and deep analysis of numerical methods has left a mark on me as a researcher.

To all the members of Prof. Dr. Kanschat's group, who helped me through many steps of the way, and finally to my friends, for their support in times of need.

Computations in this thesis were done using release 8.5 of the deal.II finite element library [**1**, **2**].

The author was supported by the Heidelberg Graduate School of Mathematical and Computational Methods for the Sciences (HGS MathComp), DFG grant GSC 220 in the German Universities Excellence Initiative.

Contents

Abstract

The development of advanced discretization methods for the radiation transport equation is of fundamental importance, since the numerical effort of modeling increasingly complex multidimensional problems with increasing accuracy is extremely challenging. Different expressions of this equation arise in several science fields, from nuclear fission and fusion to astrophysics, climatology and combustion.

Mathematically, the radiation intensity is usually a rapidly changing function, causing a considerable loss in accuracy for many discretization methods. Depending on the coefficient ranges, the equation behaves like totally different equation types, making it very difficult to find a discretization method that is efficient in all regimes. Computationally, the huge amount of unknowns involved demands not only extremely powerful computers, but also efficient numerical methods and optimized implementations. Today, solvers covering all the coefficient ranges and still being robust in the diffusion dominated case are very scarce.

In the last 20 years, Discontinous Galerkin (DG) methods have been studied for the monoenergetic problem, unsuccessfully, due to lack of stability for diffusion-dominated cases. Recently, new mathematical developments have fully explained the instability and provided a remedy by using a numerical flux depending on the scattering cross section and the mesh size. The new formulation has proven to be stable and allows the application of multigrid, matrix-free methods, reducing the memory needed for such an amount of unknowns.

We use these numerical methods to address the solution of a energy dependent problem with a multigroup approach. We study the diffusion approximation to the transport problem, obtaining convergence proofs for the symmetric scattering case and advances in the nonsymmetric case, using field of values analysis.

For the full transport case, we discretize by means of an asymptotic preserving, weakly penalized discontinuous Galerkin method that we solve with a multigrid preconditioned GMRES solver, using nonoverlapping Schwarz smoothers for the energy and direction dependent radiative transfer problem.

To address the local thermodynamic equilibrium (LTE) constraint, we use a nonlinear additive Schwarz method to precondition the Newton solver. By solving full local radiative transfer problems for each grid cell, performed in parallel on a matrix-free implementation, we achieve a method capable to address large scale calculations arising from applications such as astrophysics, atmospheric radiation calculations and nuclear applications.

To the best of our knowledge, this is the first time this preconditioner combination has been used in LTE radiation transport and in several tests we show the robustness of the approach for different mesh sizes, cross sections, energy distributions and anisotropic regimes, both in the linear and nonlinear cases.

Zusammenfassung

Die Entwicklung von fortgeschrittenen Methoden für die Lösung der Strahlungstransportgleichung ist von elementarer Bedeutung, da die numerische Berechnung von immer komplexeren, multidimensionalen Problemen mit ausreichender Genauigkeit sehr anspruchsvoll ist. Die Strahlungstransportgleichung existiert in verschiedenen Formen - abhängig von welchem wissenschaftlichen Gebiet sie betrachten wird, von der Kernspaltung und Kernfusion bis zu der Astrophysik, Klimatologie und Verbrennungstechnik.

Mathematisch gesehen, ist die Strahlungsintensität eine sich meist schnell ändernde Funktion, die eine beträchtliche Genauigkeitsabnahme für viele Diskretisierungsmethoden zur Folge hat. Abhängig von der Größenordnung der Koeffizient, kann sich diese Gleichung wie verschiedene Gleichungstypen verhalten, was es erschwert, eine einzige Diskretisierungsmethode zu finden, welche für alle Wertebereiche hinweg effizient ist. Rechnerisch gesehen, wird nicht nur eine enorme Rechenleistung zum Lösen der zahlreichen Unbekannten benötigt, sondern auch effiziente, numerische Methoden und einhergehende, optimierte Implementierungen. Bis heute gibt es sehr wenige, universelle Algorithmen, die obiges Problem für alle Wertebereiche robust lösen. Insbesondere lässt sich das diffusionsdominierte Regime mit unserem Lösungsverfahren im Unterschied zu bestehenden Algorithmen effizient berechnen.

In den letzten 20 Jahren wurden diskontinuierliche Galerkin-Methoden (DG) für das monoenergetische Problem untersucht, jedoch aufgrund von numerischen Instabilitäten für den diffusionsdominanten Fall ohne Erfolg. Die Instabilitäten konnten kürzlich anhand eines stabilisierten numerischen Flusses, welcher vom Streuquerschnitt und der Gitterweite abhängt, korrigiert werden. Die neue Formulierung hat sich als robust erwiesen und erlaubt die Verwendung von matrix-freien Mehrgitter Methoden, die den benötigten Arbeitsspeicher erheblich reduzieren.

Wir benutzen diese numerische Methoden um ein energieabhängiges Problem mit einem Multigruppen-Verfahren zu lösen. Wir approximieren das Transportproblem asymptotisch anhand eines Diffusionsproblems und führen eine Konvergenztheorie für den Fall einer symmetrischen Streuung durch. Für den Fall der nichtsymmetrischen Streuung beschreiben wir unseren analytischen Fortschritt basierend auf der numerischen Wertebereich Methode.

Die vollständige Transportgleichung diskretisieren wir mittels asymptotisch erhaltenden, schwach bestrafenden, diskontinuierlichen Galerkin-Methode. Das resultierende Gleichungssystem lösen wir mithilfe eines Mehrgitter-vorkonditionierten GMRES Verfahrens, in welchem wir nicht-überlappende Schwarz-Glätter für das energie- und richtungsabhängige Strahlungstransferproblem verwenden.

Um die lokale thermodynamische Gleichgewichtsbedingung (LTG) zu erfüllen, benutzen wir eine nicht-lineare, additive Schwarz-Methode um das Newton-Verfahren zu präkonditionieren. Indem wir das vollständige lokale Strahlungstransportproblem für jede Gitterzelle parallel basierend auf einer matrix-freien Implementierung lösen, erhalten wir eine Methode, welche umfangreiche Berechnungen, die notwendig in den Bereichen der Astrophysik, atmosphärischen Strahlungsberechnung und der Kerntechnik Anwendung sind, lösen kann.

Nach bestem Wissen, ist dies das erste Mal, dass diese Kombination aus oben geschilderten Präkonditionierern für ein LTG Strahlungstransportproblem benutzt wurde. Wir konnten die Robustheit in mehreren Tests, das bedeutet für variierende Gittergrößen, Wirkungsquerschnitte, Energieverteilungen und anisotropische Wertebereiche, sowohl im linearen als auch im nicht-linearen Fall demonstrieren.

Introduction

The development of advanced discretization methods for the radiation transport equation is of fundamental importance, since the numerical effort of modeling increasingly complex multidimensional problems with increasing accuracy is still extremely challenging.

The steady-state classical three-dimensional radiative transport equation in local thermodynamical equilibrium (LTE) including scattering for the invariant radiation density $\mathcal{I} = \mathcal{I}(\mathbf{x}, \mathbf{\Omega}, \nu)$, depending on the spatial variable $\mathbf{x}$, photon propagation direction $\mathbf{\Omega}$ and frequency ν reads:

$$\mathbf{\Omega} \cdot \nabla I_\nu(\mathbf{x}, \nu, \mathbf{\Omega}) + \rho(\kappa_{\nu,s} + \kappa_{\nu,a}) I_\nu(\mathbf{x}, \nu, \mathbf{\Omega}) - \int_0^\infty \int_{4\pi} \rho \kappa_{\nu,s}(\nu' \to \nu, \mathbf{\Omega}' \to \mathbf{\Omega}) I_\nu(\mathbf{x}, \mathbf{\Omega}', \nu') d\mathbf{\Omega}' d\nu' - \rho \kappa_{\nu,a} B_\nu = S_\nu(\mathbf{x}, \nu, \mathbf{\Omega}).$$

All the absorbed photons are re-emitted with a Planck's spectrum as follows

$$\int_0^\infty \int_{4\pi} \rho \kappa_{\nu,a} I_\nu d\mathbf{\Omega} d\nu = \int_0^\infty \int_{4\pi} \rho \kappa_{\nu,a} B_\nu d\mathbf{\Omega} d\nu,$$

where $\kappa_{\nu,a}$ is the absorption opacity and $\kappa_{\nu,s}$ is the scattering opacity, with

$$\kappa_{\nu,s}(\mathbf{x}, \nu, \mathbf{\Omega}) = \int_0^\infty \int_{4\pi} \kappa_{\nu,s}(\nu \to \nu', \mathbf{\Omega} \to \mathbf{\Omega}') d\mathbf{\Omega}' d\nu',$$

and appropriate boundary conditions.

Different expressions of the transport equation arise in several science fields, from nuclear fission and fusion to astrophysics, climatology and combustion. In this regard, using different nomenclatures, different laws govern the behavior of each parameter involved.

The radiation intensity is in general (neglecting polarization) a function of seven variables if we include time depencence. Even if only a moderate discretization of 10^2 grid points for each independent variable is taken into consideration, it results in a huge discrete problem with 10^{14} unknowns. This amount of data demands not only extremely powerful computers, but also efficient numerical methods to reduce the memory and CPU requirements. Especially for the latter point, an appropriate discretization method is of fundamental importance.

The intensity is usually a rapidly changing function of the spatial, angular and frequency variables yielding jumps of the intensity or its derivatives within small parts of the corresponding computational domain. These jumps usually cause a considerable loss in accuracy for many discretization methods.

Depending on the coefficient ranges, the linear Boltzmann equation behaves like totally different equation types: in material free areas it behaves like a hyperbolic equation; in scattering dominant, optically thick media it behaves like an elliptic

equation (steady-state) and in regions with highly forward-peaked phase function, it can behave like a parabolic equation. It is extremely difficult to find a discretization method efficiently dealing with these different regimes.

Moreover, available solvers covering all the coefficient ranges and still robust in the diffusion dominated case, where scattering is the main interaction, are very scarce.

Some years ago, E. Meinköhn, G. Kanschat, R. Rannacher and R. Wehrse in collaboration between University of Heidelberg's Institute of Applied Mathematics, Institute for Theoretical Astrophysics and Interdisciplinary Center for Scientific Computing developed a robust methodology based in continuous finite elements to solve this equation in the frame of astrophysics, to model 3D radiation fields in gas clouds from the early universe, in particular as to the influence of varying distributions of density and velocity [**3**].

In 2001, M. L. Adams analyzed the application of a Discontinuous Galerkin (DG) approximation to the linear Boltzmann equation, finding that this technique suffers from several defects: their leading-order solutions are in general discontinuous, they satisfy diffusion discretizations that can be ill-behaved, and they may not be accurate given boundary layers that are not resolved by the spatial mesh [**4**].

More recently, J-L Guermond and G. Kanschat found a necessary and sufficient condition for the standard upwind DG approximation to converge to the correct limit solution in the diffusive regime, is that the approximation space should contain a linear space of continuous functions, and that the restrictions of the functions of this space to each mesh cell should contain the linear polynomials [**5**]. This work led to the development of a robust, multigrid, unconditionally stable DG discretization for radiation transport problems in optically thick and diffusive media [**6**] [**7**].

With these advancements, we included an energy description of the transport phenomena by using a multigroup discretization. Energy is divided into *bins* and the exchange of particles between bins when undergoing a scattering event is represented in the scattering kernel, together with the exchange in flight direction. A key ingredient of the solver is that in the smoothers, a full multigroup transport problem is solved, effectively filtering out the residual components that are, in some sense, local. This locality is characteristic of diffusive regimes, as particles have very short mean free paths. On the other hand, in cases where the transport component is dominant, the information from each point needs to be communicated to the parts of the domain that are downstream along the characteristics, which makes multiplicative smoothers particularly useful since they can be ordered in a downstream fashion.

We use a multigrid preconditioner with nonoverlapping Schwarz smoothers and cell-wise subdomains, in order to increase parallelization, and test the behavior of the additive and multiplicative versions of the smoothers depending on the regime of the problem being considered.

We consider the diffusion approximation to the transport problem and study the behavior of the preconditioner for an interior penalty DG discretization of the multigroup diffusion equation. Proofs are provided for the case of symmetric scattering, and a thorough study, based on a field of values analysis, is performed for the nonsymmetric case. We show results for a wide range of parameters and smoother types.

Following, we tackle the full multigroup LTE transport problem for which we used a Newton solver, that we will precondition making use of a nonlinear preconditioner based in the solving of local nonlinear problems to a relatively high degree of accuracy, limiting the number of nonlinear iterations to be performed by the Newton solver. This preconditioning extends the idea of solving *complete* multigroup transport problems in each subdomain to solve complete *nonlinear* multigroup transport problems in each subdomain. As we described before, the efficiency of such an algorithm depends, in some sense, on the locality of the problem which is expected from local thermodynamic equilibrium constraints.

Non-linear preconditioning was suggested in [**8**] with the Additive Schwarz Preconditioned Inexact Newton method (ASPIN) (see also [**9**]). In our case, using cell-wise subdomains, the method is identical to solving with an inexact Newton method the nonlinear block Jacobi iteration equations at the fixed point. To the best of our knowledge, this is the first time these preconditioners are used for LTE radiation transport and in several tests we show its efficiency to limit iteration counts.

The thesis is structured as follows:

- Chapter 1 we describe the bulk of transport theory we will use in the thesis in detail. We also put transport in the context of kinetic theory, as described in bibliography, and we show the multigroup discretization, both for the scattering term and Planck's law.
- Chapter 2 addresses the diffusion approximation, where a significant part of the contribution from this thesis is contained. We introduce the discontinuous Galerkin discretization of interior penalty type for a symmetric collision kernel and address the numerical analysis of several multilevel algorithms, including numerical experiments.
- Chapter 3 includes work on the extension of the analysis from chapter 3 to a nonsymmetric kernel, which is also a contribution from this thesis. We use Fourier and field of values analysis to study the effect of the nonsymmetry on the efficiency of smoothers.
- Chapter 4 introduces previous results on the discretization of linear transport, for which we contribute with numerical experiments to study of the dependence of the stabilization parameters, anisotropic scattering and different combinations of smoothers between multigrid levels.
- Chapter 5 contains the last contribution from this thesis, using the theories and methods from previous chapters to address the radiation transport equations in local thermodynamic equilibrium with scattering, introducing the nonlinear preconditioner algorithm as well. We show numerical results for a wide range of regimes and some examples of the performance of the code in cases with a density distribution throughout the domain, both in 2D and 3D.

CHAPTER 1

Transport Theory

Abstract. In this chapter we describe the transport theory we will use in the rest of the thesis. We begin by explaining the physical process and its applications, together with a brief description of the location of linear transport theory in the scheme of the more general topic of kinetic theory, ranging from classical (or quantum) mechanics, to hydrodynamics. We finish by describing the multigroup discretization.

1. Introduction

The following chapter is largely inspired by the Transport Theory book from J.J. Duderstadt and W.R. Martin [**10**] which provides a larger description of nonequilibrium statistical mechanics and a wide range of applications.

In literature, the term "transport theory" is used to refer to slightly different subjects when used to describe a physical process, but we can say in general it concerns the *streaming* of particles through a host medium.

Many applications require such calculations, e.g.

- light streams from the sun and through the atmosphere before reaching the surface of the earth. In the process, it streams through air particles and scatters exchanging energy with them and increasing its temperature. This process repeats everywhere in the known universe where stars expel photons that collide onto planets and stream through dust;
- people, inside vehicles moving along highways or walking through a train station;
- neutrons stream through fuel in the core of nuclear reactors, interacting with fissile isotopes which, in turn, release energy and more neutrons, thus maintaining a controlled chain reaction;
- ions stream through plasmas, in many applications ranging from everyday electronics and fundamental research to astrophysics, where it constitutes 99% of the visible matter in the observable universe.

From this vantage point, transport is by nature a highly nonlinear process and effectively a very difficult problem to solve even with present day computer power. The sheer size of the problem to be solved, given the amount of variables involved, is very big.

On the other hand, the interactions both between particles and between particles and the host medium usually occur at different speeds, in those *regimes*, it is possible to approximate their behavior by using a linear problem.

There exist different approaches for transport processes, *stochastic* and *deterministic*. The former is the modeling of particles explicitly and follow the interactions with the media or between particles from a probabilistic approach, the latter

models the random nature of particle interactions as a field of probability densities or distribution functions that we describe in the following section.

2. Particle distribution functions

We concern ourselves with the deterministic approach where instead of addressing the exact amount of particles in a certain region we only model the *expected* particle *phase space density* defined by:

$$n(\mathbf{x},\mathbf{v},t)dxd\mathbf{v} = \begin{array}{l}\text{expected number of particles streaming with velocity in}\\ d\mathbf{v} \text{ around } \mathbf{v}, \text{ with an energy in } dE \text{ around } E,\\ \text{and position } x \text{ around } dx \text{ at time } t.\end{array}$$

It is found often useful to define the phase space density as a function of the solid angle $d\mathbf{\Omega}$ and the kinetic energy E as follows

$$n(\mathbf{x},\mathbf{\Omega},E,t)dxd\mathbf{\Omega}dE = \begin{array}{l}\text{expected number of particles streaming with direction in}\\ d\mathbf{\Omega} \text{ around } \mathbf{\Omega}, \text{ with an energy in } dE \text{ around } E,\\ \text{and position } x \text{ around } dx \text{ at time } t.\end{array}$$

It is possible to change between sets of variables as follows

$$\begin{aligned} n(\mathbf{x},\mathbf{\Omega},E,t) &= \left(\frac{v}{m}\right) n(\mathbf{x},\mathbf{v},t),\\ n(\mathbf{x},v,\mathbf{\Omega},t) &= v^2 n(\mathbf{x},\mathbf{v},t), \text{ and}\\ n(\mathbf{x},\mathbf{\Omega},E,t) &= \left(\frac{1}{mv}\right) n(\mathbf{x},v,\mathbf{\Omega},t), \end{aligned}$$

where we have considered $\mathbf{\Omega} = \frac{\mathbf{v}}{|\mathbf{v}|}$ and a non-relativistic kinetic energy as $E = \frac{1}{2}mv^2$.

The phase space particle density contains all the information we need to describe the particle distribution, we can obtain the usual particle density (independent of energy and direction) by integration

$$N(\mathbf{x},t) = \int n(\mathbf{x},\mathbf{v},t)d\mathbf{v} = \int_0^\infty \int_{4\pi} n(\mathbf{x},\mathbf{\Omega},E,t)d\mathbf{\Omega}dE.$$

Related to the concept of particle density is the particle *phase space current density* or *angular current density* defined as

$$\boldsymbol{j}(\mathbf{x},\mathbf{v},t)\cdot d\boldsymbol{S}d\mathbf{v} = \mathbf{v}n(\mathbf{x},\mathbf{v},t)\cdot d\boldsymbol{S}d\mathbf{v} = \begin{array}{l}\text{expected number of particles crossing}\\ \text{an area } dS \text{ per unit time with velocity}\\ \text{in } \mathbf{v} \text{ around } d\mathbf{v},\end{array}$$

where similarly to the phase space density we can define

$$\boldsymbol{J}(\mathbf{x},t) = \int \boldsymbol{j}(\mathbf{x},\mathbf{v},t)d\mathbf{v},$$

and $\boldsymbol{J}(\mathbf{x},t)\cdot d\boldsymbol{S}$ can be interpreted as the rate at which particles pass through a differential surface area dS.

Finally, the *partial current density* $J_\pm(\mathbf{x})$ characterizes the rate at which particles flow through an area in a given direction, defined as follows

$$J_\pm(\mathbf{x},t) = \pm\int_\pm \hat{\boldsymbol{e}}_S\cdot\boldsymbol{j}(\mathbf{x},\mathbf{v},t)d\mathbf{v},$$

where $\hat{\boldsymbol{e}}_S$ is the unit normal to the surface being considered and the velocity integration is taken over only those particle directions in the positive or negative direction to the surface. Defined in this way we observe that

$$\hat{\boldsymbol{e}}_S \cdot \boldsymbol{J}(\mathbf{x}, t) = J_+(\mathbf{x}, t) - J_-(\mathbf{x}, t),$$

showing the character of "net" current density of $\boldsymbol{J}(\mathbf{x}, t)$.

3. Collision rate with the background medium

In this section we model the interaction of particles with the host medium composed of *scattering centers*, we will assume that such interactions occur instantaneously at a point in space (i.e. particles stream until they suffer a collision, after which they are either absorbed or scattered to a new velocity). It is clear at this point that such an assumption would not be valid for long distance forces interaction or in situations where a particle may be absorbed at some point and re-emitted some time later.

We introduce the concept of *mean free path* (mfp) to characterize collisions as a local interaction as follows

$$(mfp)^{-1} \coloneqq \Sigma(\mathbf{x}, \mathbf{v}) \coloneqq \text{probability of a particle interaction per unit distance traveled by particle of velocity } \mathbf{v} \text{ at position } x.$$

The quantity Σ is referred to in the radiation transport terminology as *macroscopic cross section*, where the concept is related to another quantity, the *microscopic cross section* by

$$\Sigma(\mathbf{x}, \mathbf{v}) = N_B(\mathbf{x})\sigma(\mathbf{v}),$$

and $N_B(\mathbf{x})$ is the number density of the background medium.

When a collision occurs, the particle can be either absorbed or scattered away from the scattering center. After the absorption the excess energy provided to the scattering center can be re-emitted as another particle with a different velocity (e.g. nuclear fission). Following a scattering event, some of the energy carried by the incident particle can also be provided to the scattering center, and other particles generated. To account for these interactions we will need a finer description of the macroscopic cross section.

We define the concept of *scattering probability function* $f(\mathbf{v}' \to \mathbf{v})$ as

$$f(\mathbf{v}' \to \mathbf{v})d\mathbf{v} \coloneqq \text{probability that in a collision, the incident particle with velocity } \mathbf{v}' \text{ will induce the emission of any secondary particles with velocity } \mathbf{v} \text{ in } d\mathbf{v}.$$

In order to characterize the generation of secondary particles we will define the mean number of secondary particles emitted per collision event, $c(\mathbf{x}, \mathbf{v})$, by

$$c(\mathbf{x}, \mathbf{v}) \coloneqq \text{mean number of secondary particles emitted in a collision event experienced by an incident particle with velocity } \mathbf{v} \text{ at position } x.$$

With this information we can define the *collision kernel* as

$$\Sigma(\mathbf{x}, \mathbf{v}' \to \mathbf{v}) \coloneqq \Sigma(\mathbf{x}, \mathbf{v}')c(\mathbf{x}, \mathbf{v}')f(\mathbf{x}, \mathbf{v}' \to \mathbf{v}).$$

We note that by definition we have

$$\Sigma(\mathbf{x},\mathbf{v}) = \int \Sigma(\mathbf{x},\mathbf{v}\to\mathbf{v}')d\mathbf{v}'$$

It is important to remark here that we do *assume that all collision events are localized and uncorrelated* and in the case of photons or quantum mechanical particles the interaction events have to be sufficiently well separated to ensure the loss of phase information from one event to another (mean free paths must be longer than particle wavelengths).

We can at this point, begin the definition of the collision term noted $\left(\frac{\partial n}{\partial t}\right)_{\text{coll}}$ by realizing that

$$v\Sigma(\mathbf{x},\mathbf{v}) \coloneqq \text{collision frequency of a particle of velocity } \mathbf{v}.$$

Hence the rate at which such reactions will occur in a unit volume can be written as

$$v\Sigma(\mathbf{x},\mathbf{v})n(\mathbf{x},\mathbf{v},t) \coloneqq \text{reaction rate density}.$$

Noting that particles disappearing from velocity $\mathbf{v}$ at point x by any kind of interaction is accounted by $v\Sigma(\mathbf{x},\mathbf{v})n(\mathbf{x},\mathbf{v},t)$ and secondary particles appearing at velocity $\mathbf{v}$ from velocity $\mathbf{v}'$ at point x are accounted by $\int \Sigma(\mathbf{x},\mathbf{v}'\to\mathbf{v})n(\mathbf{x},\mathbf{v}',t)d\mathbf{v}$, we can write

$$\left(\frac{\partial n}{\partial t}\right)_{\text{coll}} = \int \Sigma(\mathbf{x},\mathbf{v}'\to\mathbf{v})n(\mathbf{x},\mathbf{v}',t)d\mathbf{v}' - v\Sigma(\mathbf{x},\mathbf{v})n(\mathbf{x},\mathbf{v},t).$$

4. The transport equation

With the quantities defined in previous sections we are now able to write down the transport equation as the balance between sources and sinks in every point in space, velocity and time.

Consider an arbitrary volume V, surrounded by a surface area S, the balance of neutrons of velocity $\mathbf{v}$ in such a volume is

$$\begin{pmatrix}\text{time rate}\\ \text{of change}\\ \text{of } n\end{pmatrix} = \begin{pmatrix}\text{change due}\\ \text{to leakage}\\ \text{through } S\end{pmatrix} + \begin{pmatrix}\text{change due}\\ \text{to}\\ \text{collisions}\end{pmatrix} + \begin{pmatrix}\text{sources}\end{pmatrix}.$$

Using the definitions in previous sections we obtain

$$\frac{\partial}{\partial t}\int_V n(\mathbf{x},\mathbf{v},t)dx = -\int_S \boldsymbol{j}(\mathbf{x},\mathbf{v},t)\cdot d\boldsymbol{S} + \int_V \left(\frac{\partial n}{\partial t}\right)_{\text{coll}} dx + \int_V s(\mathbf{x},\mathbf{v},t)dx.$$

Using Gauss' theorem we see that

$$\begin{aligned}\int_S \boldsymbol{j}(\mathbf{x},\mathbf{v},t)\cdot d\boldsymbol{S} &= \int_V \nabla\cdot\boldsymbol{j}(\mathbf{x},\mathbf{v},t)dx = \int_V \nabla\cdot\left(\mathbf{v}n(\mathbf{x},\mathbf{v},t)\right)dx\\ &= \int_V \mathbf{v}\cdot\nabla n(\mathbf{x},\mathbf{v},t)dx\end{aligned}$$

where the in the last step we can extract $\mathbf{v}$ from the gradient because it is an independent variable.

Replacing in the previous expression

$$\int_V \left(\frac{\partial n}{\partial t} + \mathbf{v}\cdot\nabla n(\mathbf{x},\mathbf{v},t) - \left(\frac{\partial n}{\partial t}\right)_{\text{coll}} - s(\mathbf{x},\mathbf{v},t)\right)dx = 0$$

where because V is arbitrary, the integrand must be identically zero

$$\frac{\partial n}{\partial t} + \mathbf{v} \cdot \nabla n(\mathbf{x}, \mathbf{v}, t) - \left(\frac{\partial n}{\partial t}\right)_{\text{coll}} - s(\mathbf{x}, \mathbf{v}, t) = 0,$$

replacing the expression for $\left(\frac{\partial n}{\partial t}\right)_{\text{coll}}$ we obtain

$$\frac{\partial n}{\partial t} + \mathbf{v} \cdot \nabla n(\mathbf{x}, \mathbf{v}, t) + v\Sigma(\mathbf{x}, \mathbf{v}) n(\mathbf{x}, \mathbf{v}, t) - \int \Sigma(\mathbf{x}, \mathbf{v}' \to \mathbf{v}) n(\mathbf{x}, \mathbf{v}', t) d\mathbf{v}' - s(\mathbf{x}, \mathbf{v}, t) = 0.$$

REMARK 1.1. Using the definition of *material or substantial derivative* describing the time rate of change of the local particle density along the particle trajectory to the change in the local density due to collisions and sources we obtain

$$\begin{aligned} \frac{Dn}{Dt} &= \left(\frac{\partial n}{\partial t}\right)_{\text{coll}} + s \\ \frac{\partial n}{\partial t} + \frac{\partial x}{\partial t}\frac{\partial n}{\partial x} + \frac{\partial \mathbf{v}}{\partial t}\frac{\partial n}{\partial \mathbf{v}} &= \left(\frac{\partial n}{\partial t}\right)_{\text{coll}} + s \\ \frac{\partial n}{\partial t} + \mathbf{v} \cdot \frac{\partial n}{\partial x} + \frac{\boldsymbol{F}}{m} \cdot \frac{\partial n}{\partial \mathbf{v}} &= \left(\frac{\partial n}{\partial t}\right)_{\text{coll}} + s. \end{aligned}$$

We observe in the last expression that under the absence of external forces, the previous transport equation expression is obtained.

The product $v\ n(\mathbf{x}, \mathbf{v}, t)$ arises so frequently in the calculation of the reaction rate densities in transport theory that it has become customary to introduce a special notation:

$$\varphi(\mathbf{x}, \mathbf{v}, t) = vn(\mathbf{x}, \mathbf{v}, t) = \text{angular flux or phase space flux.}$$

In a similar sense we can introduce the velocity-integrated flux

$$\phi(\mathbf{x}, t) = \int \varphi(\mathbf{x}, \mathbf{v}, t) d\mathbf{v} = \int vn(\mathbf{x}, \mathbf{v}, t) d\mathbf{v}.$$

With these definitions, the angular current density can be easily expressed as

$$\boldsymbol{j}(\mathbf{x}, \mathbf{v}, t) = \boldsymbol{\Omega}\varphi(\mathbf{x}, \mathbf{v}, t).$$

Finally we provide the most used expression of the transport equation, in terms of the angular flux $\varphi(\mathbf{x}, E, \boldsymbol{\Omega}, t)$ as

$$\frac{1}{v}\frac{\partial n}{\partial t} + \boldsymbol{\Omega} \cdot \nabla\varphi + \Sigma\varphi - \int_0^\infty \int_{4\pi} \Sigma(E' \to E, \boldsymbol{\Omega}' \to \boldsymbol{\Omega})\varphi(\mathbf{x}, E', \boldsymbol{\Omega}', t) + s = 0,$$

where to obtain a complete description we must include initial and boundary conditions, since only a single time derivative appears in the equation, we can simply choose the initial condition to be the specification of the initial value of the phase space flux for all positions and velocities:

$$\text{initial condition: } \varphi(\mathbf{x}, \mathbf{v}, 0) = \varphi_0(\mathbf{x}, \mathbf{v}) \qquad \text{for all } x \text{ and } \mathbf{v}$$

In the context of this thesis, we reduce it to the steady state form as follows

$$\boldsymbol{\Omega} \cdot \nabla\varphi + \Sigma\varphi - \int_0^\infty \int_{4\pi} \Sigma(E' \to E, \boldsymbol{\Omega}' \to \boldsymbol{\Omega})\varphi(E', \boldsymbol{\Omega}') d\boldsymbol{\Omega}' dE + s = 0,$$

and boundary conditions will be detailed hereafter.

5. Boundary conditions

To complete our mathematical description of particle transport we must specify boundary conditions that accompany the transport equation. Several of the more common boundary conditions include the following:

5.1. Free surface. A free surface is defined such that particles can only escape a body through the surface; they cannot reenter it. Hence we would demand that the density vanish on the surface for all inward directions:

$$\varphi(\mathbf{x}_S, \mathbf{v}, t) = 0 \qquad \text{for all } \mathbf{v} \text{ such that } \mathbf{v} \cdot \hat{\boldsymbol{e}}_S < 0.$$

We must be careful here to avoid *reentrant geometries* in which the escaping particle can reenter the body at a different point. These can usually be "patched up" by incorporating parts of the surrounding into the specification of the system of interest. One can also impose inhomogeneous boundary conditions at the surface by specifying the incoming density

$$\varphi(\mathbf{x}_S, \mathbf{v}, t) = h(\mathbf{x}_S, \mathbf{v}, t) \qquad \text{for all } \mathbf{v} \text{ such that } \mathbf{v} \cdot \hat{\boldsymbol{e}}_S < 0.$$

We can always replace such inhomogeneous boundary conditions by equivalent fictitious surface sources on the boundary and return to homogeneous boundary conditions however.

5.2. Reflecting boundary. Should we wish to impose a reflecting boundary condition that essentially assumes that particles are reflected at the boundary in a billiard ball fashion (i.e. angles of incidence and reflection are equal). Then one would require

$$\varphi(\mathbf{x}_S, \mathbf{v}, t) = \varphi(\mathbf{x}_S, \mathbf{v}_r, t) \qquad \text{for } \mathbf{v} \cdot \hat{\boldsymbol{e}}_S < 0,$$

where $\mathbf{v}_r$ is defined such that $\mathbf{v} \cdot \hat{\boldsymbol{e}}_S = \mathbf{v}_r \cdot \hat{\boldsymbol{e}}_S$ and $\mathbf{v} \times \mathbf{v}_r \cdot \hat{\boldsymbol{e}}_S = 0$. Such reflecting boundary conditions are occasionally used to express a symmetry property of the solution.

A variation of this theme is the albedo boundary condition in which the incoming density is reduced by a specified factor α (the "albedo").

$$\varphi(\mathbf{x}_S, \mathbf{v}, t) = \alpha\varphi(\mathbf{x}_S, \mathbf{v}_r, t) \qquad \text{for } \mathbf{v} \cdot \hat{\boldsymbol{e}}_S < 0.$$

5.3. Periodic boundary conditions. In systems with periodic symmetry it occasionally becomes desirable to impose periodic boundary conditions in which the outgoing density on certain boundaries is equated with the incoming density on other boundaries that are related by symmetric conditions.

5.4. Infinity. We usually demand that the density be well behaved at infinity, for example

$$\lim_{\|x\| \to \infty} \varphi(\mathbf{x}, \mathbf{v}, t) < \infty.$$

6. Local thermodynamic equilibrium

To describe the transport of low energy photons (light) through matter (e.g., a stellar or planetary atmosphere), we define the photon energy intensity as the product of the photon energy $h\nu$ and photon flux $cn(\mathbf{x}, E, \mathbf{\Omega})$

$$I_\nu(\mathbf{x}, \mathbf{\Omega}) = (h\nu)cn(\mathbf{x}, E, \mathbf{\Omega}).$$

The corresponding form of the transport equation (now referred to as radiative transfer equation) is

$$\frac{1}{c}\frac{\partial I_\nu}{\partial t} + \mathbf{\Omega} \cdot \nabla I_\nu = \rho(\mathbf{x}, t)\left[-\kappa'_\nu(\mathbf{x}, \mathbf{\Omega}, t)I_\nu(\mathbf{x}, \mathbf{\Omega}, t) + \varepsilon_\nu(\mathbf{x}, \mathbf{\Omega}, t)\right],$$

where $\rho(\mathbf{x}, t)$ is the local matter density, $\kappa'_\nu(\mathbf{x}, \mathbf{\Omega}, t)$ is the absorption coefficient and $\varepsilon_\nu(\mathbf{x}, \mathbf{\Omega}, t)$ is the emission coefficient.

In local thermodynamic equilibrium, one can simplify this to write

$$\frac{1}{c}\frac{\partial I_\nu}{\partial t} + \mathbf{\Omega} \cdot \nabla I_\nu = \rho\kappa'_\nu\left[-I_\nu + S_\nu\right],$$

where the emission term is given by

$$S_\nu = \frac{2h\nu^3}{c^2}\left[e^{\frac{h\nu}{kT}} - 1\right]^{-1} =: B_\nu.$$

If we wish to include the process of photon scattering and cross section frequency dependence, we can generalize this equation to

$$\frac{1}{c}\frac{\partial I_\nu(\mathbf{x}, \nu, \mathbf{\Omega})}{\partial t} + \mathbf{\Omega} \cdot \nabla I_\nu(\mathbf{x}, \nu, \mathbf{\Omega}) + \rho(\kappa_{\nu,s} + \kappa_{\nu,a})I_\nu(\mathbf{x}, \nu, \mathbf{\Omega})$$
$$- \int_0^\infty \int_{4\pi} \rho\kappa_{\nu,s}(\nu' \to \nu, \mathbf{\Omega}' \to \mathbf{\Omega})I_\nu(\mathbf{x}, \mathbf{\Omega}', \nu')d\mathbf{\Omega}'d\nu' - \rho\kappa_{\nu,a}B_\nu = S_\nu(\mathbf{x}, \nu, \mathbf{\Omega}).$$

All the absorbed photons are re-emitted with a Planck's spectrum as follows

$$\int_0^\infty \int_{4\pi} \rho\kappa_{\nu,a}I_\nu d\mathbf{\Omega}d\nu = \int_0^\infty \int_{4\pi} \rho\kappa_{\nu,a}B_\nu d\mathbf{\Omega}d\nu$$

where $\kappa_{\nu,a}$ is the absorption opacity and $\kappa_{\nu,s}$ is the scattering opacity, with

$$\kappa_{\nu,s}(\mathbf{x}, \nu, \mathbf{\Omega}) = \int_0^\infty \int_{4\pi} \kappa_{\nu,s}(\nu \to \nu', \mathbf{\Omega} \to \mathbf{\Omega}')d\mathbf{\Omega}'d\nu'.$$

In the context of this thesis we will address the solution of the time-independent case of the aforementioned equation.

7. Some generalizations of the transport equation

It is possible to demonstrate (using techniques from statistical mechanics) that the most general form of the transport equation can be written as

$$\frac{\partial n}{\partial t} + \mathbf{v} \cdot \frac{\partial n}{\partial x} + \frac{\boldsymbol{F}}{m} \cdot \frac{\partial n}{\partial \mathbf{v}} = \int_0^t \int\int \Sigma(\mathbf{v}' \to \mathbf{v}, x' \to x, t - \tau)n(\mathbf{x}', \mathbf{v}', \tau) + \mathcal{D}(\mathbf{x}, \mathbf{v}, t)$$

where the collision term has been generalized to account for nonlocal processes. The term $\mathcal{D}(\mathbf{x}, \mathbf{v}, t)$ which appears in this equation depends on the initial value of the density and frequently vanishes in time rapidly enough that it can be ignored. This equation is of only formal interest until one can determine the generalized collision kernels $\Sigma(\mathbf{v}' \to \mathbf{v}, x' \to x, t - \tau)$. Although explicit expressions for these kernels

can be written down in terms of the microscopic behavior of the host material, the specific determination of theses quantities remains a formidable task, and they are usually approximated using modeled calculations or experimental measurements. In many cases the collision kernels will depend in a nonlinear fashion on the particle distribution function $n(\mathbf{x}, \mathbf{v}, t)$.

7.1. The place of transport in kinetic theory. Transport theory is actually a very restricted subset of the more general topic of kinetic theory. One can distinguish between the two subjects by noting that *kinetic theory* is usually regarded as the aspect of statistical mechanics that is concerned with the *derivation* and study of equations for the particle phase space density $n(\mathbf{x}, \mathbf{v}, t)$. Such "kinetic equations" or "transport equations" are typified by the Boltzmann equation for a dilute gas, the neutron transport equation, or the Vlasov equation for a plasma. We regard *transport theory* as the more restricted mathematical discipline concerned with the *solution* of such kinetic equations and the application of such solutions to the study of particle transport processes.

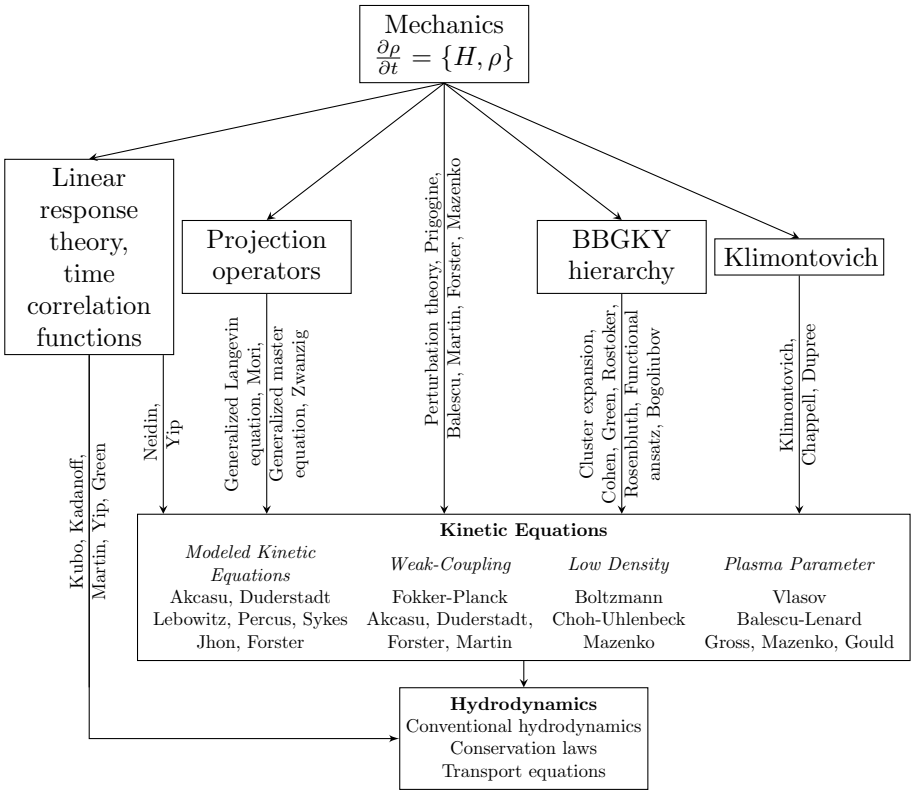


FIGURE 1.1. A map of nonequilibrium statistical mechanics. Taken from [**10**]

Traditionally, transport theorists have concerned themselves with transport processes such as neutron diffusion, radiative transfer, or rarefied gas dynamics -all of which are characterized by particle mean free paths that are many times greater

than the range over which a collision event takes place. That is, the particle transport process is dominated by particle streaming between collisions. The transport equation describing such processes in "dilute" systems is just the Boltzmann equation - either in its linear form for neutron or photon transport, or in its nonlinear form for the transport of molecules in a dilute gas.

However one can also consider transport processes in very dense systems in which the mean free path is comparable to the collision length. Such phenomena arise, for example, when we study the motions of molecules in a liquid or electrons in a plasma. Of course the physics of such transport processes is radically different, since now collision events dominate streaming behavior. Furthermore, the task of constructing a kinetic theory of such dense systems is far from trivial has yielded to the powerful tools of nonequilibrium statistical mechanics. Surprisingly enough, however, the mathematical features of the kinetic or transport equations that have been obtained for dense systems are remarkably similar to features more familiar from transport theories based on the Boltzmann or Boltzmann-like equations.

The chart in figure 1.1 indicates the place of transport in nonequilibrium statistical mechanics, starting with Hamiltonian mechanics.

The details of these theories are very well out of the scope of this thesis and this chart is provided only as a valuable "map" to locate different mathematical descriptions depending on the characteristics of the transport process being analyzed.

8. The one-speed approximation

Named in literature as "one-speed" approximation, it provides a model for the particle distribution that assumes all particles have the same speed. Assuming a non-relativistic speed, we can state $E = \frac{1}{2}mv^2$ and refer to this approximation as the "one-energy" or, as we will see later "one-group" approximation.

We begin by re-introducing the particle transport equation from the previous chapter, in the following time-independent form

$$\begin{aligned}\boldsymbol{\Omega}\cdot\nabla\varphi(\mathbf{x},E,\boldsymbol{\Omega})+\Sigma_t(E)\varphi(\mathbf{x},E,\boldsymbol{\Omega})\\ -\int_0^\infty\int_{4\pi}\Sigma_s(E'\to E,\boldsymbol{\Omega}'\to\boldsymbol{\Omega})\varphi(\mathbf{x},E',\boldsymbol{\Omega}')dE'+s(\mathbf{x},E,\boldsymbol{\Omega})=0.\end{aligned}$$

We change our notation to show only the dependencies in energy and integrate the equation over the energy variable to find

$$\begin{aligned}\boldsymbol{\Omega}\cdot\nabla\left(\int_0^\infty\varphi(E)dE\right)+\int_0^\infty\Sigma_t(E)\varphi(E)dE\\ -\int_0^\infty\int_{4\pi}\left(\int_0^\infty\Sigma_s(E'\to E)dE\right)\varphi(E')dE'+\int_0^\infty s(E)dE=0.\end{aligned}$$

$$\boldsymbol{\Omega}\cdot\nabla\left(\int_0^\infty\varphi(E)dE\right)+\frac{\int_0^\infty\Sigma_t(E)\varphi(E)dE}{\int_0^\infty\varphi(E)dE}\int_0^\infty\varphi(E)dE$$
$$-\int_0^\infty\int_{4\pi}\left(\frac{\int_0^\infty\Sigma_s(E'\to E)dE}{\int_0^\infty\varphi(E)dE}\int_0^\infty\varphi(E)dE\right)\varphi(E')dE'$$
$$+\int_0^\infty s(E)dE=0.$$

Following, we define the energy-averaged parameters

$$\overline{\Sigma}_t\coloneqq\frac{\int_0^\infty\Sigma_t(E)\varphi(\mathbf{x},E,\boldsymbol{\Omega})dE}{\int_0^\infty\varphi(\mathbf{x},E,\boldsymbol{\Omega})dE},$$
$$\overline{\Sigma}_s\coloneqq\frac{\int_0^\infty\Sigma_s(E'\to E,\boldsymbol{\Omega}'\to\boldsymbol{\Omega})dE}{\int_0^\infty\varphi(\mathbf{x},E,\boldsymbol{\Omega})dE},\text{ and}$$
$$\overline{s}\coloneqq\int_0^\infty s(\mathbf{x},E,\boldsymbol{\Omega})dE,$$

and the energy-integrated angular flux

$$\overline{\varphi}\coloneqq\int_0^\infty\varphi(\mathbf{x},E,\boldsymbol{\Omega})dE.$$

We can then write the energy-integrated transport equation as

$$\boldsymbol{\Omega}\cdot\nabla\overline{\varphi}(\mathbf{x},E,\boldsymbol{\Omega})+\overline{\Sigma}_t(\boldsymbol{\Omega})\overline{\varphi}(\mathbf{x},\boldsymbol{\Omega})-\int_{4\pi}\overline{\Sigma}_s(\boldsymbol{\Omega}'\to\boldsymbol{\Omega})\varphi(\mathbf{x},\boldsymbol{\Omega}')d\boldsymbol{\Omega}'+\overline{s}(\mathbf{x},\boldsymbol{\Omega})=0.$$

This equation, although correct from a mathematical standpoint, has only a formal significance since the averaged variables $\overline{\Sigma}_t$ and $\overline{\Sigma}_s$ depend on full knowledge of the solution $\varphi(\mathbf{x},E,\boldsymbol{\Omega})$.

In the field of nuclear reactor design, such equation can be used as a rough approximation since the energy distribution of neutrons is not far off from a Maxwell-Boltzmann distribution and that can be used as a guess flux. Other approximations can be done in other fields, but lacking a proper approximated shape of the spectrum, the equation must be solved in a nonlinear fashion. We will assume in this thesis that the averaged variables are available as input data.

The one-speed transport equation also arises in the study of radiative transfer. The general form of the steady-state radiative transfer equation for the radiant intensity $I_\nu(\mathbf{x},\boldsymbol{\Omega})$ in which scattering is included can be written as

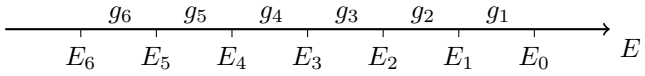
$$\boldsymbol{\Omega}\cdot\nabla I_\nu(\mathbf{x},\boldsymbol{\Omega})+\Sigma_{t\nu}I_\nu(\mathbf{x},\boldsymbol{\Omega})=\Sigma_{s\nu}\int f_\nu(\boldsymbol{\Omega}'\to\boldsymbol{\Omega})I_\nu(\mathbf{x},\boldsymbol{\Omega}')d\boldsymbol{\Omega}'+s_\nu(\mathbf{x},\boldsymbol{\Omega}),$$

where we have defined the photon interaction cross sections characterizing absorption and scattering $\Sigma_{a\nu}\coloneqq\rho\kappa_\nu$ and $\Sigma_{s\nu}\coloneqq\gamma_\nu\Sigma a\nu$. In certain applications (the so-called gray-atmosphere model) one can effectively ignore the frequency dependence of these cross sections. Then if we integrate over frequency, an equation identical to the one-speed transport model results:

$$\boldsymbol{\Omega}\cdot\nabla I+\Sigma_tI(\mathbf{x},\boldsymbol{\Omega})=\Sigma_s\int f(\boldsymbol{\Omega}'\to\boldsymbol{\Omega})I(\mathbf{x},\boldsymbol{\Omega}')d\boldsymbol{\Omega}'+s(\mathbf{x},\boldsymbol{\Omega}).$$

9. The multigroup approximation

In any accurate description of particle transport it is necessary to consider the particle energy dependence in more detail, and this usually requires a numerical approach. The most popular scheme [**11**, **12**] for treating the particle energy dependence involves a generalization of the one-speed or one-group model in which the energy range is broken in to a number of intervals or energy groups:



and then the transport equation is integrated over a given group, $E_g < E < E_{g-1}$. The "backward" group indexing is chosen to correspond to the fact that in most transport processes the particles lose energy in collisions, progressively scattering to lower energies. If we now define the angular flux characterizing a given group as

$$\varphi_g(\mathbf{x}, \mathbf{\Omega}) \coloneqq \int_{E_g}^{E_{g-1}} \varphi(\mathbf{x}, E, \mathbf{\Omega}) dE \coloneqq \int_g \varphi(E) dE,$$

and the corresponding energy-averaged cross sections or group constants characterizing that group as

$$\varphi_g \coloneqq \int_g \varphi(E) dE, \qquad s_g \coloneqq \int_g s(E) dE,$$

$$\Sigma_{t_g} \coloneqq \frac{\int_g \Sigma_t(E)\varphi(E) dE}{\int_g \varphi(E) dE}, \text{ and} \qquad \Sigma_{s_{g'g}} \coloneqq \frac{\int_g \int_{g'} \Sigma_s(E' \to E)\varphi(E) dE' dE}{\int_{g'} \varphi(E') dE'},$$

then it is apparent that we can integrate the transport equation over a given energy group g to find

$$\mathbf{\Omega} \cdot \nabla \varphi_g + \Sigma_{t_g} \varphi_g = \sum_{g'=1}^{G} \int \Sigma_{s_{g'g}}(\mathbf{\Omega}' \to \mathbf{\Omega}) \varphi_{g'} d\mathbf{\Omega}' + s_g(\mathbf{x}, \mathbf{\Omega}), \qquad g = 1, ..., G.$$

These equations are known as *multigroup transport equations* and such a multigroup approach is the most common method for obtaining accurate solutions to realistic (complex) transport problems in which energy-dependent effects must be included.

10. Multigroup local thermodynamic equilibrium

The multigroup discretization can be applied to the radiative transfer equation as well. We introduced a generalized version of the steady-state radiative transfer equation in local thermodynamic equilibrium with scattering in §6, the multigroup formulation of this equation reads:

$$\mathbf{\Omega} \cdot \nabla I_g + \rho(\kappa_{g,s} + \kappa_{g,a}) I_g - \sum_{g=1}^{G} \int_{4\pi} \rho \kappa_{g'g,s}(\mathbf{\Omega}' \to \mathbf{\Omega}) I_{g'}(\mathbf{\Omega}') d\mathbf{\Omega}' - \rho \kappa_{g,a} B_g = s_g$$

$$\sum_{g=1}^{G} \int_{4\pi} \rho \kappa_{g,a} I_g d\mathbf{\Omega} = \sum_{g=1}^{G} \int_{4\pi} \rho \kappa_{g,a} B_g d\mathbf{\Omega}$$

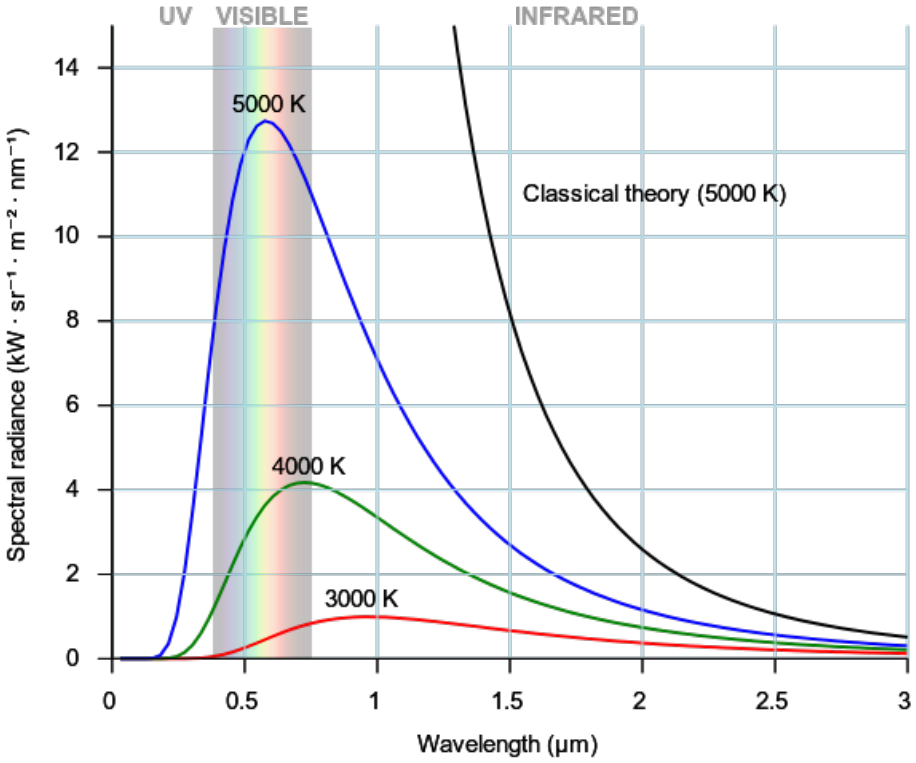


FIGURE 1.2. Planck's law for different temperatures

where the group variables are defined analogously to the previous section:

$$I_g := \int_g I_\nu d\nu, \qquad s_g := \int_g s(\nu) d\nu,$$

$$\kappa_{g,a} := \frac{\int_g \kappa_{\nu,a} I_\nu d\nu}{\int_g I_\nu d\nu}, \qquad \kappa_{g'g,s} := \frac{\int_g \int_{g'} \kappa_{\nu,s}(\nu' \to \nu) I_{\nu'} d\nu' d\nu}{\int_{g'} I_{\nu'} d\nu'},$$

$$\kappa_{g,s} := \frac{\int_g \int_0^\infty \kappa_{\nu,s}(\nu \to \nu') I_{\nu'} d\nu' d\nu}{\int_g I_\nu d\nu}, \text{ and} \qquad B_g := \int_g B_\nu d\nu.$$

11. Conclusion

We gave an introduction to the transport theory we will use throughout the thesis, including its place in the larger subject of kinetic theory and nonequilibrium statistical dynamics.

The notation of the equations changes depending on the area of study, particle energies are analogous to wave frequencies and *fluxes* are analogous to *intensities*. Our study of linear transport will mostly use the former, while the local thermodynamic equilibrium study will use the latter.

Time dependence will not be considered in the frame of this thesis, we will concentrate in the explicit direction and energy explicit solution of the equations, both linear and nonlinear.

We also described the multigroup approximation we will use throughout this thesis. The separation between energy bins has been used since the beginning of the XX century to model particle physics.

We will assume that the cross sections or *opacities* are available and are not a function of the solution, except of course for Planck's law that will define the non-linearities of local thermodynamic equilibrium we will address in the last chapter.

CHAPTER 2

The Diffusion Approximation

ABSTRACT. In this chapter we describe the diffusion approximation for transport processes in high scattering, isotropic regimes. The advection term of Boltzmann's linear transport equation is approximated by using Fick's laws of diffusion, obtaining an elliptic system with less degrees of freedom and easier to solve. We use a discontinuous Galerkin discretization of interior penalty type with multilevel preconditioners using cell-wise Schwarz smoothers and prove convergence estimates for a symmetric collision kernel, effectively generalizing the existing Schwarz theory by allowing for cell-wise subdomains on quadrilateral (hexahedral) meshes.

1. Monoenergetic transport model and diffusion limit

We considering radiation as a thin gas of small particles reacting with a background material but not with each other. Similarly to a Brownian motion process, a particle travels with constant momentum until one of the following events happens:

(1) after an average distance of $1/\widetilde{\sigma}_a$ (the mean free path of absorption) since the last event, the particle is absorbed by the background and removed from the system,
(2) after an average distance of $1/\widetilde{\sigma}_s$ (the mean free path of scattering) the particle is scattered by a background particle, changing its momentum and continuing its journey with a new momentum.

Additionally, particles may be emitted by the background with a certain probability. If the background is assumed stationary, absorption, emission and particle leakage from the system must be in balance.

The smallness assumption on the particles must be understood as the particle size being much smaller than either mean free path. For photons, this means that the wave length is negligible compared to these lengths. In addition, we make the modeling assumption that particles behave randomly. This implies small coherence length of optical radiation and in particular excludes lasers. Under this assumption, the motion of a sufficiently large sample of particles can be described by a linear form of the Boltzmann equation [**13**–**16**].

The basic quantity of radiation is a normalized density distribution function $\varphi(\mathbf{x}, \mathbf{\Omega})$ of particles in phase space. We focus on photons here, such that the momentum space of $\mathbb{R}^d$ reduces to the sphere S^{d-1} in dimensionless form. We obtain the equation

$$\mathbf{\Omega}\cdot\nabla\varphi(\mathbf{x}, \mathbf{\Omega}) + \widetilde{\sigma}_a(\mathbf{x})\varphi(\mathbf{x}, \mathbf{\Omega}) + \widetilde{\sigma}_s(\mathbf{x})\Sigma\varphi(\mathbf{x}, \mathbf{\Omega}) = \widetilde{q}(\mathbf{x}), \tag{1}$$

where

$$\Sigma\varphi(\mathbf{x},\mathbf{\Omega}) = \varphi(\mathbf{x},\mathbf{\Omega}) - \int_{S^{d-1}} H(\mathbf{\Omega}',\mathbf{\Omega})\varphi(\mathbf{x},\mathbf{\Omega}')\,\mathrm{d}\mathbf{\Omega}', \tag{2}$$

with a scattering kernel $H(\mathbf{\Omega}',\mathbf{\Omega})$ to be discussed below, see equations (10)–(12).

Originally, this equation is posed on the whole of $\mathbb{R}^d$. Artificial boundaries can be introduced to form a bounded domain of interest $\mathcal{D}$ for physical and numerical reasons, for instance if no radiation enters the domain of interest from the other side of such a boundary. Let $\mathbf{n}_{\mathcal{D}}$ be the outer unit normal vector to $\mathcal{D}$ at its boundary.

In the case of vacuum outside $\mathcal{D}$, the boundary condition is chosen as incident radiation $\varphi(\mathbf{x},\mathbf{\Omega}) = \varphi^{\mathrm{inc}}(\mathbf{x},\mathbf{\Omega})$ on the set

$$\Gamma_- = \big\{(\mathbf{x},\mathbf{\Omega}) \in \partial\mathcal{D}\times S^{d-1} \big| \mathbf{\Omega}\cdot\mathbf{n}_{\mathcal{D}} < 0\big\}. \tag{3}$$

Different boundary conditions are possible, see for instance [**17**] and below.

Let us define the derivative $\mathbf{\Omega}\cdot\nabla\varphi$ as the function with values $\mathbf{\Omega}\cdot\nabla\varphi(x,\mathbf{\Omega})$, that is, the derivative in space is taken with respect to the travel direction of the particle itself. Then, the natural solution space is W^0 defined by

$$W = \big\{\varphi \in L^2(\mathcal{D}\times S^{d-1}) \big| \mathbf{\Omega}\cdot\nabla\varphi \in L^2(\mathcal{D}\times S^{d-1})\big\} \quad \text{and} \tag{4}$$

$$W^0 = \big\{\varphi \in W \big|\ \varphi_{|\Gamma_-} = 0\big\}. \tag{5}$$

For details of this space, in particular the well-posedness of the boundary condition, see [**17**].

Two important derived quantities used below are the zeroth and first moments of φ, which are only functions of space and are defined as

$$\text{zeroth moment:} \qquad \overline{\varphi}(\mathbf{x}) = \int_{S^{d-1}} \varphi(\mathbf{x},\mathbf{\Omega})\,\mathrm{d}\mathbf{\Omega} \tag{6}$$

$$\text{first moment:} \qquad \mathbf{J}(\varphi)(\mathbf{x}) = \int_{S^{d-1}} \varphi(\mathbf{x},\mathbf{\Omega})\mathbf{\Omega}\,\mathrm{d}\mathbf{\Omega}. \tag{7}$$

Here, and throughout this article, we choose the measure on S^{d-1} such that

$$\int_{S^{d-1}} \mathrm{d}\mathbf{\Omega} = 1 \quad \text{and} \quad \forall \mathbf{x},\mathbf{y}\in\mathbb{R}^d : \int_{S^{d-1}} (\mathbf{x}\cdot\mathbf{\Omega})(\mathbf{y}\cdot\mathbf{\Omega})\,\mathrm{d}\mathbf{\Omega} = \frac{\mathbf{x}\cdot\mathbf{y}}{d}. \tag{8}$$

Note that this holds if we choose the usual measure divided by 4π. For the first moment, there holds the divergence identity

$$\int_{S^{d-1}} \mathbf{\Omega}\cdot\nabla\varphi\,\mathrm{d}\mathbf{\Omega} = \nabla\cdot\mathbf{J}(\varphi). \tag{9}$$

We make additional assumptions on the redistribution function $H(\mathbf{\Omega}',\mathbf{\Omega})$ in (2). First, we assume that H is smooth and bounded. Furthermore, we assume that scattering conserves particles and all absorption and emission processes are modeled by $\widetilde{\sigma_a}$ and $\widetilde{q}$ respectively. Therefore, we obtain

$$\int_{S^{d-1}}\int_{S^{d-1}} H(\mathbf{\Omega}',\mathbf{\Omega})\varphi(\mathbf{\Omega}')\,\mathrm{d}\mathbf{\Omega}'\,\mathrm{d}\mathbf{\Omega} = \int_{S^{d-1}} \varphi(\mathbf{\Omega})\,\mathrm{d}\mathbf{\Omega}, \tag{10}$$

for any sufficiently integrable function φ on the sphere. In particular, for constant functions

$$\iint_{S^{d-1}} H(\mathbf{\Omega}',\mathbf{\Omega})\,\mathrm{d}\mathbf{\Omega}'\,\mathrm{d}\mathbf{\Omega} = 1. \tag{11}$$

This implies that $\Sigma\varphi = 0$ if φ is independent of $\mathbf{\Omega}$. It also implies that there is a unique mean value free function $\breve{H}$ such that

$$
\begin{aligned}
H(\mathbf{\Omega}', \mathbf{\Omega}) &= 1 + \breve{H}(\mathbf{\Omega}', \mathbf{\Omega}), \\
\Sigma\varphi &= \varphi - \overline{\varphi} - \int_{S^{d-1}} \breve{H}(\mathbf{\Omega}', \mathbf{\Omega}) \varphi(\mathbf{\Omega}') \, \mathrm{d}\mathbf{\Omega}'.
\end{aligned} \tag{12}
$$

While the previous assumptions hold universally, we additionally assume that only functions independent of $\mathbf{\Omega}$ are in the kernel of Σ. This assumption is justified by many applications, e.g. thermal neutron transport in highly diffusive media.

There has been extensive research on the finite element approximation of solutions to equation (1) by several authors [**18**–**24**] under the assumption that $1/\widetilde{\sigma_s}$ is not small compared to the diameter of the domain of interest. In fact, their results are fairly complete and conclusive and that chapter is closed. Here, we point out in particular the analysis of the upwind discontinuous Galerkin (DG) method by Reed and Hill [**25**] in [**26**, **27**].

Nevertheless, in many important applications, very small mean free paths are of importance, and it has been pointed out in [**4**, **28**, **29**] that discretization with the upwind DG method suffers loss of accuracy in this case. In order to develop an efficient scheme for such a case, we first introduce the small parameter

$$
\varepsilon = \frac{\text{maximum mean free path of scattering}}{\text{diameter of domain of interest}} = \frac{1}{\operatorname{diam} \mathcal{D} \max_x \widetilde{\sigma_s}} \tag{13}
$$

and consider the limit for $\varepsilon \to 0$. Considering Brownian motion for $\widetilde{\sigma_s} \approx \varepsilon$, we realize that a particle stays within a given volume longer if the mean free path of scattering is reduced by a factor ε. Thus, in order to keep the absorption probability of a single particle equal, we scale σ_a and the source with ε. This leads to the scaled Boltzmann equation

$$
\mathbf{\Omega} \cdot \nabla \varphi(\mathbf{x}, \mathbf{\Omega}) + \varepsilon \sigma_a(\mathbf{x}) \varphi(\mathbf{x}, \mathbf{\Omega}) + \frac{\sigma_s(\mathbf{x})}{\varepsilon} \Sigma\varphi(\mathbf{x}, \mathbf{\Omega}) = \varepsilon q(\mathbf{x}). \tag{14}
$$

Its well-posed weak form is obtained by multiplying with a test function, integrating over the phase space $\mathcal{D} \times S^{d-1}$, integrating by parts and applying the boundary condition weakly (see [**17**]): find $\varphi \in W$ such that for all $w \in W$ there holds

$$
\begin{aligned}
&\left(-\varphi, \mathbf{\Omega} \cdot \nabla w\right)_{\mathcal{D} \times S^{d-1}} + \left\langle \mathbf{\Omega} \cdot \mathbf{n} \, \varphi, w \right\rangle_{\Gamma_+} \\
&\qquad + \varepsilon \left(\sigma_a \varphi, w\right)_{\mathcal{D} \times S^{d-1}} + \frac{1}{\varepsilon} \left(\sigma_s \Sigma\varphi, w\right)_{\mathcal{D} \times S^{d-1}} \\
&\qquad\qquad = \varepsilon \left(q, w\right)_{\mathcal{D} \times S^{d-1}} - \left\langle \mathbf{\Omega} \cdot \mathbf{n} \, \varphi^{\mathrm{inc}}, w \right\rangle_{\Gamma_-} .
\end{aligned} \tag{15}
$$

Here, we have adopted the inner product notation

$$
\left(f, g\right)_{\mathcal{D} \times S^{d-1}} = \int_{\mathcal{D}} \int_{S^{d-1}} f g \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{\Omega} \quad \text{and} \quad \left\langle f, g \right\rangle_{\Gamma} = \int_{\partial\mathcal{D}} \int_{S^{d-1}} f g \, \mathrm{d}s \, \mathrm{d}\mathbf{\Omega}.
$$

The outward radiation boundary is $\Gamma_+ = \mathcal{D} \times S^{d-1} \setminus \Gamma_-$. Into this equation, we insert the formal expansion

$$
\varphi = \varphi_0 + \varepsilon \varphi_1 + \varepsilon^2 \varphi_2 + \text{h.o.t.} \tag{16}
$$

Conditions for convergence of this sequence can be found in [**5**]. Here, we only point out the main steps of the formal asymptotic analysis from there and [**30**].

$$\begin{aligned} &\varepsilon^{-1}\Big(\qquad\qquad\qquad\quad \sigma_s\Sigma\varphi_0\Big) \\ &+\varepsilon^{0}\Big(\boldsymbol{\Omega}\cdot\nabla\varphi_0 \qquad\quad +\sigma_s\Sigma\varphi_1\Big) \\ &+\varepsilon^{1}\Big(\boldsymbol{\Omega}\cdot\nabla\varphi_1+\sigma_a\varphi_0+\sigma_s\Sigma\varphi_2\Big) = \varepsilon q + \text{h.o.t} \end{aligned} \tag{17}$$

Following [**5**], we first evaluate the term with power ε^{-1}, which in weak form reads

$$\big(\Sigma\varphi_0, w\big)_{\mathcal{D}\times S^{d-1}} = 0 \quad \forall w \in W. \tag{18}$$

Entering $w = \Sigma\varphi_0$, we obtain $\Sigma\varphi_0 = 0$ and thus φ_0 in the kernel of Σ. Thus, by the assumption below equation (11) φ_0 is isotropic, i.e. it is independent of $\boldsymbol{\Omega}$.

Let now $\mathbf{L}(x)$ be a vector valued function on $\mathcal{D}$. We test the weak formulation of the ε^0-term with $\boldsymbol{\Omega}\cdot\mathbf{L}$ and obtain by the integral relation (8) and the splitting (12)

$$\begin{aligned} 0 &= \big(\boldsymbol{\Omega}\cdot\nabla\varphi_0 + \sigma_s\Sigma\varphi_1, \boldsymbol{\Omega}\cdot\mathbf{L}\big)_{\mathcal{D}\times S^{d-1}} \\ &= \big(\tfrac{1}{d}\nabla\varphi_0, \mathbf{L}\big)_{\mathcal{D}} + \big(\sigma_s\mathbf{J}(\varphi_1), \mathbf{L}\big)_{\mathcal{D}} \\ &\quad - \sigma_s\int_{\mathcal{D}}\iint_{S^{d-1}} \breve{H}(\boldsymbol{\Omega}', \boldsymbol{\Omega})\varphi_1(\mathbf{x}, \boldsymbol{\Omega}')\boldsymbol{\Omega}\cdot\mathbf{L}\,\mathrm{d}\boldsymbol{\Omega}'\,\mathrm{d}\boldsymbol{\Omega}\,\mathrm{d}\mathbf{x}. \end{aligned} \tag{19}$$

Here, we will restrict the quantitative analysis to the case $\breve{H} \equiv 0$ and we immediately obtain the relation $\nabla\varphi_0 = d\sigma_0\mathbf{J}(\varphi_1)$. For the general case, we refer the reader to [**17**, Ch. XXI, §5, Lemma 1]; we point out though, that the second condition in (8) has to be augmented suitably when choosing an angular discretization.

Finally, we test the weak form of the ε^1-term with the isotropic test function w and obtain by the divergence identity (9), the implication of (10), and the fact that that $\Sigma\varphi_2$ is mean value free

$$\begin{aligned} q &= \big(\boldsymbol{\Omega}\cdot\nabla\varphi_1 + \sigma_a\varphi_0 + \sigma_s\Sigma\varphi_2, w\big)_{\mathcal{D}\times S^{d-1}} \\ &= \big(\nabla\cdot\mathbf{J}(\varphi), w\big)_{\mathcal{D}} + \big(\sigma_a\varphi_0, w\big)_{\mathcal{D}}. \end{aligned} \tag{20}$$

We now assume that $\varphi_0 \to \phi$ and $\mathbf{J}(\varphi) \to \mathbf{J}$. This is guaranteed for instance in the interior of a domain with smooth data where σ_s is uniformly positive. It does not necessarily hold at boundaries and material interfaces (see [**5**]) and we do not discuss this case here. Therefore, we refer to this step as "formal asymptotic analysis". We obtain the well-known diffusion limit (see e.g. [**14**, **17**])

$$\begin{aligned} d\sigma_s\mathbf{J} - \nabla\phi &= 0 \\ \nabla\cdot\mathbf{J} + \sigma_a\phi &= q, \end{aligned} \tag{21}$$

for isotropic scattering $\breve{H} \equiv 0$. For general H, we obtain a similar structure with a different factor in front of $\mathbf{J}$ in the first equation. Furthermore, eliminating the first moment we get

$$\nabla\cdot\left(\frac{1}{d\sigma_s}\nabla\phi\right) + \sigma_a\phi = q$$

defining $\delta = \frac{1}{d\sigma_s}$ we obtain the one-group diffusion equation

$$\nabla\cdot(\delta\nabla\phi) + \sigma_a\phi = q \tag{22}$$

2. Multigroup diffusion approximation

Retaking equation (22), we can apply a multigroup approach in the same fashion as we did in chapter 2, to obtain the multigroup diffusion equation

$$-\nabla\cdot(\delta_g\nabla u_g)+\frac{1}{\varepsilon}\sum_{g'=1}^{G}(\sigma_{gg'}u_g-\sigma_{g'g}u_{g'})=f_g \quad \text{in } \Omega \quad \text{with } g=1...G, \tag{23}$$

where g is the *group* index identifying each *substance* being modeled, δ_g is the diffusion coefficient for each group g, ε is a perturbation parameter defining the relative size of the reaction with respect to the diffusion term, Ω is a convex polyhedral domain in $\mathbb{R}^d$ with $d=2,3$ and f_g is the right hand side term which is given.

The equation is provided with the boundary conditions

$$u_g=0 \quad \text{on } \Gamma \quad \text{with } g=1...G.$$

where Γ is the boundary of Ω.

We assume $\delta_g,\sigma_{gg'}\in L^\infty(\Omega)$ and $\sigma_{gg'}\geq 0$, for all $g,g'=1...G$ and there exists $C>0$ such that $\delta_g\geq C$ almost everywhere in Ω, see [**31**, Chapter VII,§2.6].

Furthermore, to be able to use standard analytical tools, we assume that the scattering term is symmetric, in the sense that $\sigma_{gg'}=\sigma_{g'g}$.

We choose the Hilbert spaces

$$\mathcal{V}=(H_0^1(\Omega))^G \qquad\qquad \mathcal{H}=(L^2(\Omega))^G,$$

provided respectively with inner products

$$(u,v)_{\mathcal{V}}=\sum_{g=1}^{G}(\delta_g\nabla u_g,\nabla v_g)_{L^2(\Omega)} \qquad (u,v)_{\mathcal{H}}=\sum_{g=1}^{G}(u_g,v_g)_{L^2(\Omega)}. \tag{24}$$

and the norms

$$\|u\|_{\mathcal{V}}^2=(u,u)_{\mathcal{V}}, \qquad\qquad \|u\|_{\mathcal{H}}^2=(u,v)_{\mathcal{H}},$$

Introducing, on $\mathcal{V}$, the bilinear form:

$$\begin{aligned}\mathscr{A}(u,v):=&\sum_{g=1}^{G}\int_\Omega\delta_g\nabla u_g\cdot\nabla v_g dx\\ &+\frac{1}{\varepsilon}\left(\sum_{g=1}^{G}\int_\Omega\sigma_{gg'}u_{g'}\cdot v_g dx-\sum_{g'=1}^{G}\int_\Omega\sigma_{gg'}u_{g'}\cdot v_g dx\right)\\ &=(\boldsymbol{D}\nabla u,\nabla v)_{\mathcal{H}}+\frac{1}{\varepsilon}(\boldsymbol{\Sigma}u,v)_{\mathcal{H}}\end{aligned} \tag{25}$$

where

$$u=(u_1,...,u_G)^\intercal, \qquad\qquad v=(v_1,...,v_G)^\intercal,$$

$$\boldsymbol{D}=\operatorname{diag}((\delta_1,...,\delta_G)^\intercal), \qquad\qquad \boldsymbol{\Sigma}=\begin{pmatrix}\sum_{g\neq 1}\sigma_{1g} & \cdots & -\sigma_{G1}\\ \cdots & \cdots & \cdots\\ -\sigma_{1G} & \cdots & \sum_{g\neq G}\sigma_{Gg}\end{pmatrix}$$

According to our assumption above, the scattering operator $\boldsymbol{\Sigma}$ is a symmetric, weakly diagonally dominant M-matrix with zero column and row sum. By the Perron-Frobenius theorem, this implies $\boldsymbol{\Sigma}$ is singular with rank less than G and by the Geršgorin circle theorem, all eigenvalues are nonnegative. Physically, this

implies that the system can have no particle absorption. The presence of absorption would imply all eigenvalues are positive and the matrix is invertible.

Finally, problem (23) can be written in the form: Finding $u \in \mathcal{V}$ such that

$$\mathscr{A}(u,v) = (f,v)_{\mathcal{H}} \tag{26}$$

where $f \in \mathcal{H}(\Omega)$.

Under the assumptions on the parameters of equation (23), the bilinear form $\mathscr{A}(u,v)$ is continuous and $\mathcal{V}$-coercive relatively to $\mathcal{H}$ and from Lax-Milgram's theorem, the variational problem admits a unique solution in $\mathcal{V}$, i. e. there exist constants $\gamma_{\mathscr{A}}, C_{\mathscr{A}} > 0$ such that

$$\mathscr{A}(u,u) \geq \gamma_{\mathscr{A}} \|u\|^2_{\mathcal{H}} \qquad\qquad \mathscr{A}(u,v) \leq C_{\mathscr{A}} \|u\|_{\mathcal{H}} \|v\|_{\mathcal{H}}$$

We remark at this point, that even though $\gamma_{\mathscr{A}}$ is independent of ε, $C_{\mathscr{A}}$ is not. Hence, the system will need to be preconditioned in order to limit the growth of the condition number.

3. Discrete problem

We apply the IP discontinuous Galerkin formulation to the discretization of $\mathscr{A}(\cdot,\cdot)$ [**32**]. Let $\mathbb{T}$ be a mesh covering the domain Ω represented as a set of non-overlapping, convex polygons or polyhedra κ. Conformity of the faces of mesh cells is not required, but we assume local quasi-uniformity and shape regularity in the sense that the quotients of the diameter of a cell and each of its boundary faces f are uniformly bounded from above and below.

We denote $\left(H^m(\mathbb{T})\right)^G$ a G-vector valued *broken* Sobolev space consisting of functions in $\mathcal{H}$ whose restriction to each element $\kappa \in \mathbb{T}$ belongs to the Sobolev space $\left(H^m(\kappa)\right)^G$.

Let $V \subset \left(H^2(\mathbb{T})\right)^G$ be the space of discontinuous piecewise multivariate polynomials $\mathbb{Q}_p$ of mapped tensor product polynomials of degree up to p in each coordinate direction.

Let $\kappa_1, \kappa_2 \in \mathbb{T}$ be two mesh cells with a joint face f, and let u_1 and u_2 be the traces of functions u on f from κ_1 and κ_2 respectively. On face f, we define the averaging operators as

$$\{\!\{u\}\!\} = \frac{u_1 + u_2}{2}$$

The IP form for the Laplacian, as described in [**32**], reads

$$\begin{aligned} &\int_{\mathbb{T}} \nabla u \cdot \nabla v dx - \int_{\mathbb{F}^I \cup \mathbb{F}^B} 2\left(\{\!\{u\boldsymbol{n}\}\!\} \cdot \{\!\{\nabla v\}\!\} ds + \{\!\{\nabla u\}\!\} \cdot \{\!\{v\boldsymbol{n}\}\!\}\right) ds \\ &\quad + \int_{\mathbb{F}^I \cup \mathbb{F}^B} 4\delta_{\mathrm{IP}} \{\!\{u\boldsymbol{n}\}\!\} \cdot \{\!\{v\boldsymbol{n}\}\!\} ds \qquad\qquad \forall v \in V, \end{aligned} \tag{27}$$

where $u\boldsymbol{n} = (u_1\boldsymbol{n}, u_2\boldsymbol{n}, ..., u_G\boldsymbol{n})^\intercal$ and $\nabla u = (\nabla u_1, \nabla u_2, ..., \nabla u_G)^\intercal$. We have replaced the jump operator used in [**32**] for the equivalent expression: $2\{\!\{u\boldsymbol{n}\}\!\} = u_1\boldsymbol{n}_1 + u_2\boldsymbol{n}_2$. Coercivity and continuity have been proven in the aforementioned paper.

We define the bilinear form

(28)
$$a(u,v) = \int_{\mathbb{T}} \boldsymbol{D}\nabla u \cdot \nabla v dx - \int_{\mathbb{F}^I \cup \mathbb{F}^B} 2\left(\{\!\{u\boldsymbol{n}\}\!\} \cdot \{\!\{\boldsymbol{D}\nabla v\}\!\} ds + \{\!\{\boldsymbol{D}\nabla u\}\!\} \cdot \{\!\{v\boldsymbol{n}\}\!\}\right) ds$$
$$+ \int_{\mathbb{F}^I \cup \mathbb{F}^B} 4\delta_{\mathrm{IP}} \{\!\{\boldsymbol{D}(u\boldsymbol{n})\}\!\} \cdot \{\!\{v\boldsymbol{n}\}\!\} ds \qquad \forall v \in V,$$

Considering the assumptions made in the previous sections, $a(u,v)$ is coercive and continuous albeit with constants depending on $\boldsymbol{D}$.

Using the formulation for the Laplacian, our DG IP formulation for the singularly perturbed reaction diffusion problem reads: Find $u \in V$ such that

$$\mathcal{A}(u,v) = a(u,v) + \frac{1}{\varepsilon}\int_{\mathbb{T}} \boldsymbol{\Sigma} u \cdot v dx = \int_{\mathbb{T}} S \cdot v dx \quad \forall v \in V. \tag{29}$$

We observe that given the non-negativeness of $\boldsymbol{\Sigma}$, the coercivity constant for our problem coincides with the Laplacian case while the continuity constant is now dependent on ε^{-1}. Hence, in order to obtain a robust solver we will have to precondition the problem in order to eliminate the dependence of the spectral radius on ε^{-1}.

The linear system reads

$$\boldsymbol{A}\mathbf{u} = \mathbf{f},$$

where $\mathbf{u}$ and $\mathbf{f}$ are the coefficient vector of the representation of u and f respectively in terms of the chosen basis.

4. Solver

4.1. Krylov subspace methods. Given that some of the smoothers and preconditioners we use are not necessarily symmetric, we use a minimum residual method to solve our problem: GMRES. For any Krylov minimal residual method we have the following convergence estimate

LEMMA 4.1. Consider the linear system $Ax = b$ where $A \in \mathbb{R}^{n\times n}$ and x,$b \in \mathbb{R}^n$. Further suppose that all eigenvalues of A are positive. Then after k steps of any minimum residual method, the residual $r^{(k)} = b - Ax^{(k)}$

$$\left\|r^{(k)}\right\| \leq \left(1 - \frac{c_{\mathrm{MR}}}{C_{\mathrm{MR}}}\right)^{\frac{k}{2}} \left\|r^{(0)}\right\|$$

for some inner product $\|\cdot\|$ where c_{MR} and C_{MR} are lower bounds of the field of values of A and A^{-1} respectively. If A is positive real matrix, then $0 < c_{\mathrm{MR}} = \lambda_{\min}\left(\frac{A+A^{\intercal}}{2}\right)$ and $C_{\mathrm{MR}} \geq \|A\|^2 \geq 0$.

PROOF. See [**33**] Corollary 6.2 and [**34**]. □

From lemma 4.1 it can be seen that to bound c_{MR} and C_{MR} independently of ε, as described in §3, the matrix corresponding to the discrete problem alone would not be well conditioned. The following section describes our preconditioner choice.

4.2. Schwarz preconditioners. We choose Schwarz preconditioners for which there is a well-known framework and theory for symmetric positive definite problems (see [**35**], [**36**], [**37**], [**38**]). The following sections will provide the definitions needed to address convergence estimates in an abstract framework.

4.2.1. *Space decomposition.* Let

$$V_j \subset V, \qquad \text{for } j = 0, 1, 2, ..., J,$$

be a Hilbert space with norm $\|\cdot\|_{V_j}$. We note that V_0 is used to denote the so-called *coarse space* in the domain decomposition context. For $j = 0, 1, 2, ..., J$, let

$$\mathcal{R}_k^\intercal : V_j \to V$$

denote some *prolongation operator* for which the following relations hold:

$$\mathcal{R}_j^\intercal V_j \subset V \qquad V = \sum_{j=0}^{J} \mathcal{R}_j^\intercal V_j \qquad \text{for } j = 0, 1, 2, ..., J,$$

where $\mathcal{R}_j^\intercal V_j$ stands for the range of the linear operator $\mathcal{R}_j^\intercal$.

Associated with each local space V_j for $j = 0, 1, 2, ..., J$, we introduce local discrete bilinear forms $\mathcal{A}_j(\cdot, \cdot)$, defined on $V_j \times V_j$, as the restriction of global discrete bilinear form $\mathcal{A}(\cdot, \cdot)$ on $V_j \times V_j$ (i.e. *exact local solvers*).

Assumption 4.1. *Energy stability.* The spaces $\{V_j\}$ are said to provide a stable decomposition if there exists a constant C_V such that each $v \in V$ admits a decomposition

$$v = \sum_{j=0}^{J} \mathcal{R}_j^\intercal v_j,$$

with $v_j \in V_j$ such that

$$\sum_{j=0}^{J} \|v_j\|_{\mathcal{A}_j}^2 \leq C_V \|v\|_{\mathcal{A}}^2.$$

If $v \in \text{range}\,(\mathcal{I} - \mathcal{P}_0)$, $v \in V$ admits a stable decomposition without including the coarse space

$$\sum_{j=1}^{J} \|v_j\|_{\mathcal{A}_j}^2 \leq C_V \|v\|_{\mathcal{A}}^2.$$

Assumption 4.2. *Strengthened generalized Cauchy-Schwarz inequality.* There exist constants $\theta_j \in [0, 1]$ for $i, j = 0, 1, 2, ..., J$ such that

$$\mathcal{A}(\mathcal{R}_i^\intercal v_i, \mathcal{R}_j^\intercal v_j) \leq \theta_{ij} \mathcal{A}(\mathcal{R}_i^\intercal v_i, \mathcal{R}_i^\intercal v_i)^{\frac{1}{2}} \mathcal{A}(\mathcal{R}_j^\intercal v_j, \mathcal{R}_j^\intercal v_j)^{\frac{1}{2}}, \qquad \forall v_i \in V_i, v_j \in V_j.$$

Assumption 4.3. *Local compatibility.* A space V_j is said to be compatible with respect to $a_j(\cdot, \cdot)$ if the local problem of seeking $v_j \in V_j$ such that

$$\mathcal{A}_j(v_j, w_j) = f_j(w_j), \qquad \forall w_j \in V_j.$$

is uniquely solvable for any given bounded linear functional f_j on V_j. We use exact local solvers for the problems in the subdomains.

4.2.2. *Local projections.* We now introduce a set of *projection-like* operators $\widetilde{\mathcal{P}}_j : V \to V_j$ for $j = 0, 1, 2, ..., J$. These projection-like operators will serve as the building blocks for the constructions of Schwarz methods. For any fixed $v, w \in V$, define $\widetilde{\mathcal{P}}_j v \in V_j$ by

$$\mathcal{A}_j(\widetilde{\mathcal{P}}_j v, w_j) := a(v, \mathcal{R}_j^\intercal w_j), \qquad \forall w_j \in V_j.$$

We note that the well posedness of the global problem ensures $\widetilde{\mathcal{P}}_j$ is well defined for $j = 0, 1, 2, ..., J$.

Since V_j may not be a subspace of V, $\widetilde{\mathcal{P}}_j v$ may not belong to V. To pull them back to the global discrete space V, we appeal to the *prolongation operator* $\mathcal{R}_j^\intercal$ for help. Define the composite operator

$$\mathcal{P}_j := \mathcal{R}_j^\intercal \circ \widetilde{\mathcal{P}}_j, \qquad \text{for } j = 0, 1, 2, ..., J.$$

Trivially, we have $\mathcal{P}_j : V \to V$ for $j = 0, 1, 2, ..., J$.

4.2.3. *Additive version.* We define the additive Schwarz preconditioned system as

$$\mathcal{P}_{\text{ad}} := \mathcal{P}_0 + \mathcal{P}_1 + \mathcal{P}_2 + ... + \mathcal{P}_J$$

We use the operator notation for the bilinear forms $\mathcal{A}$ and $\mathcal{A}_j$ to obtain the following expression for the local projections

$$\mathcal{A}_j \widetilde{\mathcal{P}}_j v := \mathcal{R}_j^\intercal \mathcal{A} v \qquad \forall v \in V$$

Thus,

$$\widetilde{\mathcal{P}}_j = \mathcal{A}_j^{-1} \mathcal{R}_j \mathcal{A}, \qquad \mathcal{P}_j = \mathcal{R}_j^\intercal \mathcal{A}_j^{-1} \mathcal{R}_j \mathcal{A},$$

Finally, our additive Schwarz preconditioned system reads

$$\mathcal{P}_{\text{ad}} = \mathcal{R}_0^\intercal \mathcal{A}_0^{-1} \mathcal{R}_0 \mathcal{A} + \sum_{j=1}^{J} \mathcal{R}_j^\intercal \mathcal{A}_j^{-1} \mathcal{R}_j \mathcal{A}$$

4.2.4. *Multiplicative version.*

Definition 4.1. We introduce the following *error propagation* operators

$$\mathcal{E}_{\text{mu}} = (\mathcal{I} - \mathcal{P}_N) \circ (\mathcal{I} - \mathcal{P}_{N-1}) \circ ... \circ (\mathcal{I} - \mathcal{P}_0)$$

where $\mathcal{I}$ denotes the identity operator on V. Using $\mathcal{E}_{\text{mu}}$ we define the multiplicative Schwarz preconditioner

$$\mathcal{P}_{\text{mu}} = \mathcal{I} - \mathcal{E}_{\text{mu}}$$

where $\mathcal{I}$ denotes the identity operator on V.

4.2.5. *Hybrid version.*

Definition 4.2. We introduce the projection

$$\mathcal{T}_i = (\mathcal{I} - \mathcal{P}_0)\mathcal{P}_i(\mathcal{I} - \mathcal{P}_0), \qquad i = 1, ..., N$$

where $\mathcal{I}$ denotes the identity operator on V.

Define

$$\mathcal{P}_{\text{hy}} = \mathcal{P}_0 + \sum_{i=1}^{N} \mathcal{T}_i$$

as the hybrid Schwarz preconditioned system.

4.3. Abstract convergence theory. In this section we establish condition number estimates for the operators $\mathcal{P}_{\rm ad}$ and $\mathcal{P}_{\rm mu}$ in order to use lemma 4.1. We will use the previous definitions to estimate $c_{\rm MR}$ and $C_{\rm MR}$ as

$$c_{\rm MR} = \inf_{v\neq 0} \frac{\mathcal{A}(\mathcal{T}v, v)}{\|v\|^2_{\mathcal{A}}} \qquad\qquad C_{\rm MR} = \left(\sup_{v\neq 0} \frac{\|\mathcal{T}v\|_{\mathcal{A}}}{\|v\|_{\mathcal{A}}} \right)^2$$

for $\mathcal{T} = \mathcal{P}_{\rm ad}$, $\mathcal{T} = \mathcal{P}_{\rm mu}$ and $\mathcal{T} = \mathcal{P}_{\rm hy}$.

We use the general abstract convergence theory of Schwarz methods given in [**35**, §2]. We shall do so by verifying that a set of four *assumptions* are satisfied and by estimating the constants involved in terms of the parameters of our method.

4.3.1. *Application to the discrete problem.* In this section we will verify the assumptions provided in definitions 4.1, 4.2 and 4.3 to use the abstract Schwarz convergence theory developed above on the discrete problem in eqn. (29).

We use nonoverlapping, cell based subdomains in order to define the decomposition

$$\sum_{j=0}^{J} \mathcal{R}_j^{\intercal} V_j.$$

We will use the following lemmas from [**38**]

LEMMA 4.2. For any $v \in V_D = \prod_{K\in\mathbb{T}_H} \mathcal{V}(K)$, there holds the trace inequality

$$\|v\|^2_{\partial D} \leq c \left[\frac{1}{H} \|v\|^2_{\mathcal{H}(D)} + H \|v\|^2_{\mathcal{V}(D)} \right].$$

PROOF. See [**38**, Lemma 3.1]. □

LEMMA 4.3. Suppose D is a convex domain. For any $v \in V_D$, let $\overline{u} = \frac{1}{\text{meas}(D)} \int_D v dx$ be the average value of v over D. Then we can write a Poincaré inequality as follows

$$\|v - \overline{v}\|_{\mathcal{H}(D)} \leq c \ \text{diam}(D) \|u\|_{\mathcal{V}(D)} \text{ on } D.$$

In particular, if $D \in \mathbb{T}_H$

$$\|v - \overline{v}\|_{\mathcal{H}(D)} \leq cH \|u\|_{\mathcal{V}(D)} \text{ on } D.$$

PROOF. See [**38**, Lemma 3.2] □

LEMMA 4.4. Let $v, w \in V$, let $v_j, w_j \in V_j$, $j = 1, ..., J$, be given (uniquely) by $v = \sum_{j=1}^J v_j$, $w = \sum_{j=1}^J w_j$. Then the following identity holds:

$$a(v, w) = \sum_{j=1}^{J} a_j(v_j, w_j) + I(v, w),$$

where $I(\cdot, \cdot) : V \times V \to \mathbb{R}$ comprises all terms located outside the block diagonal of the bilinear from $a(v, w)$, connecting different subdomains.

PROOF. See [**38**, Lemma 4.2]. □

Furthermore, we obtain the following interface estimate for cell-wise subdomains

LEMMA 4.5. There exists a constant c such that

$$|I(v,v)| \leq c \left[\frac{1}{h^2} \sum_{K \in \mathbb{T}} \|v\|^2_{\mathcal{H}(K)} + a(v,v) \right].$$

PROOF. We prove a particular case of [**38**, Lemma 4.3] for cell-wise subdomains.

Taking the following estimate from [**38**, Eq. (4.20)], for the case of subdomains comprising only one cell

$$|I(v,v)| \leq c \left(a(v,v) + \frac{1}{h} \sum_{\partial K \in (\mathbb{F}^I \cup \mathbb{F}^B)} \|v\|^2_{\partial K} \right),$$

where $\|\cdot\|_{\partial K}$ is the L^2-inner product on the faces of cell K of the fine mesh.

We use the following trace inequality: $\|v\|^2_{\partial K} = c \left[\frac{1}{h} \|v\|^2_K + h \|\nabla v\|^2_K \right]$ from [**38**, Eq. (3.9)]

$$|I(v,v)| \leq c \left(a(v,v) + \frac{1}{h} \sum_{K \in \mathbb{T}} \left[\frac{1}{h} \|v\|^2_{\mathcal{H}(K)} + h \|\nabla v\|^2_{\mathcal{H}(K)} \right] \right).$$

The result follows from observing that $\sum_{K \in \mathbb{T}} \|\nabla v\|^2_{\mathcal{H}(K)} \leq c\, a(v,v)$. □

Finally, we achieve the following stable decomposition

LEMMA 4.6. Every $v \in V$ admits a decomposition of the form $v = \sum_{j=0}^{J} \mathcal{R}_j^{\intercal} V_j$, $v_j \in V_j$, $j = 0, ..., J$ which satisfies the bound

$$\sum_{i=0}^{J} a_j (v_j, v_j) \leq C_{V,\Delta} a(v,v)$$

with $C_{V,\Delta} = c \frac{H^2}{h^2}$, where h and H denote the granularity used in the fine and coarse meshes respectively.

PROOF. Let $u_0 \in V_0$ be the piecewise constant average of v on the coarse mesh $\mathbb{T}_H$, let $w = v - v_0$, we decompose w in nonoverlapping cell-wise subdomains as follows

$$w = \sum_{j=1}^{J} v_j,$$

where $v_1, ..., v_J$ are uniquely determined. From (4.4) we have

$$\begin{aligned} a(w,w) &= \sum_{j=1}^{J} a_j(v_j, v_j) + I(w,w) \\ a(v - v_0, v - v_0) &= \sum_{j=1}^{J} a_j(v_j, v_j) + I(v - v_0, v - v_0), \end{aligned}$$

reordering and estimating the interface term by its absolute value we obtain

$$\sum_{j=1}^{J} a_j(v_j, v_j) \leq a(v - v_0, v - v_0) + |I(v - v_0, v - v_0)|,$$

using lemma 4.5 we have

$$\begin{aligned}\sum_{j=1}^{J} a_j(v_j, v_j) \leq & c\left(a(v-v_0, v-v_0) + \frac{1}{h^2}\sum_{K\in\mathbb{T}} \|v-v_0\|^2_{\mathcal{H}(K)}\right)\\ = & c\left(a(v-v_0, v-v_0) + \frac{1}{h^2}\sum_{D\in\mathbb{T}_H} \|v-v_0\|^2_{\mathcal{H}(D)}\right)\\ \leq & c\left(a(v-v_0, v-v_0) + \frac{H^2}{h^2}\|v\|^2_{\mathcal{V}}\right),\end{aligned}$$

where we used lemma 4.3 on the last term. Using the triangle and Young inequalities we obtain

$$\sum_{j=1}^{J} a_j(v_j, v_j) \leq c\left(\frac{H^2}{h^2}a(v,v) + a(v_0, v_0)\right).$$

It remains to bound $a(v_0, v_0)$. Since v_0 is piecewise constant on $\mathbb{T}_H$, and hence also on $\mathbb{T}_h$,

$$a(v_0, v_0) = \delta_{\text{IP}} \sum_{\partial K\in\mathbb{F}^I} \frac{1}{h}\|v_0^+ - v_0^-\|^2_{\partial K} + \delta_{\text{IP}} \sum_{\partial K\in\mathbb{F}^B} \frac{1}{h}\|v_0^+\|^2_{\partial K}.$$

Now adding and subtracting v gives

$$\begin{aligned}a(v_0, v_0) \leq c\delta_{\text{IP}} \Bigg(&\sum_{\partial K\in\mathbb{F}^I} \frac{1}{h}\|(v-v_0)^+ - (v-v_0)^-\|^2_{\partial K}\\ &+ \sum_{\partial K\in\mathbb{F}^B} \frac{1}{h}\|(v-v_0)^+\|^2_{\partial K}\\ &+ \sum_{\partial K\in\mathbb{F}^I} \frac{1}{h}\|v^+ - v^-\|^2_{\partial K} + \sum_{\partial K\in\mathbb{F}^B} \frac{1}{h}\|v^+\|^2_{\partial K}\Bigg).\end{aligned}$$

The last two terms are obviously bounded by $a(v,v)$. Also, since u_0 is piecewise constant on each $D \in \mathbb{T}_H$, $\|(v-v_0)^+ - (v-v_0)^-\|_{\partial K} = \|v^+ - v^-\|_{\partial K}$ whenever ∂K is in the interior of some $D \in \mathbb{T}_H$. Thus,

$$\begin{aligned}&\sum_{\partial K\in\mathbb{F}^I} \frac{1}{h}\|(v-v_0)^+ - (v-v_0)^-\|^2_{\partial K} + \sum_{\partial K\in\mathbb{F}^B} \frac{1}{h}\|(v-v_0)^+\|^2_{\partial K}\\ &\qquad = \sum_{D\in\mathbb{T}_H} \Bigg(\sum_{\partial K\subset D} \|v^+ - v^-\|_{\partial K}\\ &\quad + \sum_{\partial K\in\partial D} \frac{1}{h}\|(v-v_0)^+ - (v-v_0)^-\|^2_{\partial K} + \sum_{\partial K\subset\partial D\in\mathbb{F}^B} \frac{1}{h}\|(v-v_0)^+\|^2_{\partial K}\Bigg)\\ &\qquad \leq ca(v,v) + c\sum_{D\in\mathbb{T}_H} \frac{1}{h}\|v-v_0\|^2_{\partial D}.\end{aligned}$$

Now using the trace inequality in lemma 4.2, we obtain

$$\sum_{D\in\mathbb{T}_H} \frac{1}{h}\|v-v_0\|^2_{\partial D} \leq c\sum_{D\in\mathbb{T}_H} \frac{1}{h}\left[\frac{1}{H}\|v-v_0\|^2_{\mathcal{H}(D)} + H\|v-v_0\|^2_{\mathcal{V}(D)}\right].$$

Also note that $\|v - v_0\|^2_{\mathcal{V}(D)} = \|v\|^2_{\mathcal{V}(D)}$. Hence, applying the approximation result from lemma 4.3 to $\|v - v_0\|_{\mathcal{H}(D)}$ we obtain

$$a(v_0, v_0) \leq c\frac{H}{h}a(v, v),$$

hence, the result falls from the fact that $\frac{H}{h} \leq \frac{H^2}{h^2}$. □

LEMMA 4.7. *Energy stability.* The spaces V_j provide a stable decomposition of V, with respect to the bilinear form $\mathcal{A}(\cdot, \cdot)$, in the sense of definition 4.1.

PROOF. Let $C_{V,\Delta}$ be the stable decomposition constant for the Laplacian, as deduced in lemma 4.6, we then have

$$\sum_{i=0}^{J} a_j\left(v_j, v_j\right) \leq C_{V,\Delta} a\left(v, v\right)$$

$$\sum_{i=0}^{J} a_j\left(v_j, v_j\right) + \frac{1}{\varepsilon}\left(\mathbf{\Sigma} v, v\right)_{\mathcal{H}} \leq C_{V,\Delta} a\left(v, v\right) + \frac{1}{\varepsilon}\left(\mathbf{\Sigma} v, v\right)_{\mathcal{H}}$$

$$\sum_{i=0}^{J}\left\{a_j\left(v_j, v_j\right) + \frac{1}{\varepsilon}\left(\mathbf{\Sigma} v_j, v_i\right)_{\mathcal{H}}\right\} \leq \max\left\{C_{V,\Delta}, 1\right\}\left\{a\left(v, v\right) + \frac{1}{\varepsilon}\left(\mathbf{\Sigma} v, v\right)_{\mathcal{H}}\right\}$$

$$\sum_{i=0}^{J} \mathcal{A}_j\left(v_j, v_j\right) \leq C_{V,\Delta} \mathcal{A}\left(v, v\right).$$

It follows that the V_j decomposition for our reaction-diffusion problem is *energy stable* with $C_V = C_{V,\Delta} = \mathcal{O}\left(\frac{H^2}{h^2}\right)$ where H and h are the largest and smallest cell diameters respectively. □

PROPOSITION 4.1. *Strengthened generalized Cauchy-Schwarz inequality.* There exists a strengthened Cauchy-Schwarz inequality in the sense of definition 4.2.

PROOF. (See [**38**, §4.2]). Verifying this inequality consists of obtaining a bound for the spectral radius $\rho(\Theta)$ of the $J \times J$ matrix $\Theta = [\theta_{ij}]_{j=0}^{J}$.

That such values exist is a consequence of the Cauchy-Schwarz inequality. The important thing, however, is to obtain a small bound on ρ. To do so, we observe that $a(v_i, v_j) = 0$ if the supports of v_i and v_j do not share a face f_{ij}. For the remaining cases, we take $\theta_{ij} = 1$. It follows at once from Gershgorin's circle theorem that

$$\rho(\Theta) \leq \max_m \operatorname{card}\left\{k | f_{mk} \neq 0 \text{ almost everywhere}\right\} + 1 \qquad f_{mk} \in \mathbb{F}^I \cup \mathbb{F}^B,$$

i.e., $\rho(\Theta)$ is bounded by 1 plus the maximum number of adjacent subdomains a given subdomain can have. In practice this number 4 in 2D and 6 in 3D. Even for "unusual" subdomain partitions, this number is not expected to be large. □

4.3.2. *Convergence estimates.* To apply the general abstract convergence theory of Schwarz methods given in [**35**, §2.3], we need verify a set of assumptions with the lemmas and propositions stated before.

THEOREM 4.1. Let the assumptions 4.1, 4.2 and 4.3 hold, then the following bounds hold for the additive Schwarz preconditioned system

$$c_{\text{ad}} \geq \frac{1}{C_V}, \qquad C_{\text{ad}} \leq \rho(\Theta) + 1.$$

Where $c_{\mathrm{ad}}, C_{\mathrm{ad}}$ are the constants to be used in place of $c_{\mathrm{MR}}, C_{\mathrm{MR}}$ in lemma 4.1.

PROOF. We verify the assumptions required in [**35**, §2.3]. Energy stability was proved in lemma 4.7, a strengthened generalized Cauchy-Schwarz inequality was shown in proposition 4.1 and the use of local solvers delivers local compatibility (assumption 4.3).

A detailed proof is provided in appendix A for the convenience of the reader. □

THEOREM 4.2. Let the assumptions 4.1, 4.2 and 4.3 hold, then the following bounds hold for the hybrid Schwarz preconditioned system

$$c_{\mathrm{hy}} \geq \min\left\{1, \frac{1}{C_V}\right\}, \qquad\qquad C_{\mathrm{hy}} \leq \max\left\{1, \rho(\Theta)\right\}.$$

Where $c_{\mathrm{hy}}, C_{\mathrm{hy}}$ are the constants to be used in place of $c_{\mathrm{MR}}, C_{\mathrm{MR}}$ in lemma 4.1.

PROOF. The assumptions proven in theorem 4.1 deliver this result as well. □

The multiplicative operator is not symmetric and we will consider a simple Richardson iteration applied to the corresponding preconditioned system and provide an upper bound for the norm of the error propagation operator.

THEOREM 4.3. Let the assumptions in definitions 4.1, 4.2 and 4.3 hold, then the following bounds hold for the multiplicative Schwarz preconditioned system

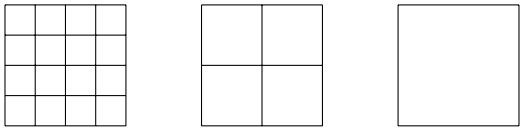
$$\|\mathcal{E}_{\mathrm{mu}}\| \leq 1 - \frac{1}{(2\rho^2(\Theta)+1)\, C_V} \leq 1.$$

PROOF. The proven assumptions enumerated in theorem 4.1 deliver this result as well. □

4.4. Multigrid V-cycle preconditioner. The preconditioners developed in the previous section are easily implemented as smoothers for multigrid preconditioners. In this section we provide convergence estimates for the multigrid V-cycle.

4.4.1. *Functional setting.* Let $\{\mathbb{T}\}_{\ell=0,\ldots,L}$ be a hierarchy of meshes of quadrilateral and hexahedral cells in two and three dimensions, respectively. In view of multilevel methods, the index ℓ refers to the mesh level defined as follows: let a coarse mesh $\mathbb{T}_0$ be given. The mesh hierarchy is defined recursively, such that the cells of $\mathbb{T}_{\ell+1}$ are obtained by splitting each cell of $\mathbb{T}_\ell$ into 2^d children by connecting edge and face midpoints (refinement). These meshes are nested in the sense that every cell of $\mathbb{T}_\ell$ is equal to the union of its four (respectively eight) children. We define the mesh size h_ℓ as the maximum of the diameters of the cells of $\mathbb{T}_\ell$. Due to the refinement process, we have $h_\ell \approx \frac{h_{\ell-1}}{2}$.

Due to the nestedness of mesh cells, the finite element spaces associated with these meshes are nested as well:



$$\mathbb{T}_0 \sqsubset \mathbb{T}_1 \sqsubset \ \ldots \sqsubset \mathbb{T}_\ell \sqsubset \ \ldots \sqsubset \mathbb{T}_L$$

$$V_0 \subset V_1 \subset \ldots \subset V_L.$$

The nestedness of the spaces implies that there is a sequence of projections and interpolation operators of the form

$$\mathcal{Q}_{\ell-1} : V_\ell \to V_{\ell-1}$$
$$\mathcal{Q}_{\ell-1}^\intercal : V_{\ell-1} \to V_\ell.$$

$\mathcal{Q}_{\ell-1}$ is the H-projection from $V_\ell \to V_{\ell-1}$ with

$$(\mathcal{Q}_{\ell-1} v_\ell, w_{\ell-1})_H = \left(v_\ell, \mathcal{Q}_{\ell-1}^\intercal w_{\ell-1}\right)_H \qquad \forall v_{\ell-1} \in V_{\ell-1}, w_{\ell-1} \in V_{\ell-1}$$

Let $\mathcal{A}_{\ell-1}(\cdot,\cdot)$ be the bilinear form defined in §3 on the mesh $\mathbb{T}_{\ell-1}$. The bilinear forms corresponding to coarser meshes are obtained from equation 29 by choosing the δ_{IP} according to the mesh granularity.

$$\begin{aligned}\mathcal{A}_{\ell-1}(\cdot,\cdot) : V_{\ell-1} \times V_{\ell-1} &\longrightarrow \mathbb{R} \\ (v_{\ell-1}, w_{\ell-1}) &\longrightarrow a_\ell\left(\mathcal{Q}_{\ell-1}^\intercal u_{\ell-1}, \mathcal{Q}_{\ell-1}^\intercal v_{\ell-1}\right)\end{aligned}$$

we define the operator $\mathcal{A}_{\ell-1} : V_{\ell-1} \longrightarrow V_{\ell-1}$ such that

$$\mathcal{A}_{\ell-1}(u_{\ell-1}, v_{\ell-1}) = (\mathcal{A}_{\ell-1} u_{\ell-1}, v_{\ell-1})_H$$

and we verify

$$\mathcal{A}_{\ell-1} = \mathcal{Q}_{\ell-1} \mathcal{A}_\ell \mathcal{Q}_{\ell-1}^\intercal.$$

The $A_{\ell-1}$-orthogonal projection $\mathcal{P}_{\ell-1} : V_\ell \longrightarrow V_{\ell-1}$ is defined by

$$\mathcal{A}_{\ell-1}(\mathcal{P}_{\ell-1} u_\ell, v_{\ell-1}) = \mathcal{A}_\ell(u_\ell, \mathcal{Q}_{\ell-1}^\intercal v_{\ell-1}), \qquad \forall u_\ell \in V_\ell, v_{\ell-1} \in V_{\ell-1},$$

which implies $\mathcal{A}_{\ell-1}\mathcal{P}_{\ell-1} = \mathcal{Q}_{\ell-1}\mathcal{A}_\ell$.

4.4.2. *Multigrid algorithm.* We define the multigrid operator $\mathcal{M}_L$ by induction.

Let $\mathcal{B}_\ell$ be a smoother defined as the preconditioner operator on the preconditioned systems presented in section 4.2 without including the coarse space, i.e.

$$\mathcal{B}_{\ell,\text{ad}} = \sum_{i=1}^{N_\ell} \mathcal{P}_{\ell,i} \mathcal{A}_\ell^{-1} = \sum_{i=1}^{N_\ell} \mathcal{R}_{\ell,i}^\intercal \mathcal{A}_{\ell,i}^{-1} \mathcal{R}_{\ell,i} \quad \text{and} \quad \mathcal{B}_{\ell,\text{mu}} = \left(\mathcal{I} - \prod_{i=N_\ell}^{1} \mathcal{P}_{\ell,i}\right) \mathcal{A}_\ell^{-1}.$$

Taking $\mathcal{M}_0 = \mathcal{A}_0^{-1}$, for $1 \le \ell \le L$ we define $\mathcal{M}_\ell g$ in terms of $\mathcal{M}_{\ell-1}$, for $g \in V_\ell$

(1) Set $x_0 = 0$, $q_0 = 0$.
(2) Define x_1 by

$$x_1 = x_0 + \mathcal{B}_\ell\left(g - \mathcal{A}_\ell x_0\right)$$

(3) Define $y_1 = x_1 + q_1$, where q_1 is defined by

$$q_1 = q_0 + \mathcal{Q}_{\ell-1}^\intercal \mathcal{M}_{\ell-1}\left(\mathcal{Q}_{\ell-1}\left(g - \mathcal{A}_\ell x\right) - \mathcal{A}_{\ell-1}\mathcal{Q}_{\ell-1} q_0\right)$$

(4) Define y_2 by

$$y_2 = y_1 + \mathcal{B}_\ell\left(g - \mathcal{A}_\ell x_1\right)$$

(5) Set $\mathcal{M}_\ell g = y_2$

4.4.3. *Convergence estimate.* The theory from [**39**] and [**40**] can be thoroughly applied to our additive Schwarz smoother case, for which we need to prove the following smoothing property lemma.

LEMMA 4.8. Given the non-overlapping subspace decomposition of V_ℓ in §4.2 and the interior penalty bilinear form $a_\ell(u,v)$ in equation (27). Assume δ_{IP} is chosen sufficiently large, the following estimate holds on each level ℓ. Then, there is a constant C_0 which is independent of the multigrid level such that for any $u \in V_\ell$ holds

$$\sum_{i=1}^{N_\ell} a_\ell\left(u_\ell, v_\ell\right) \leq c_S a\left(u,v\right) \qquad \Leftrightarrow \qquad \sum_{i=1}^{N_\ell} \mathcal{A}_\ell\left(u_\ell, v_\ell\right) \leq c_S \mathcal{A}\left(u,v\right).$$

PROOF. For the Laplacian and a fixed L, the penalty constant δ_{IP} is $\frac{\delta_0}{h_L}$ which is greater than $\frac{\delta_0}{h_\ell}$. For the latter, this is a standard result: the proof and details on the choice of δ_0 can be found in [**38**, §4].

For the reaction-diffusion system the result falls from the same proof as lemma 4.7. □

THEOREM 4.4. Let $\mathcal{B}_\ell$ be a $\mathcal{A}_\ell$-self-adjoint smoother satisfying $\forall v_\ell \in V_\ell$

$$c_S \mathcal{A}_\ell\left(v_\ell, v_\ell\right) \leq \mathcal{A}_\ell\left(v_\ell, \mathcal{B}_\ell \mathcal{A}_\ell v_\ell\right),$$

where $0 < c_S < 1$, let $\mathcal{M}_\ell$ be the operators defined in §4.4.2, then there exist $0 \leq c_{\text{MG}} \leq 1$ and $0 \leq C_{\text{MG}} \leq 1$ such that

$$(1 - C_{\text{MG}})\mathcal{A}_\ell\left(v_\ell, v_\ell\right) \leq \mathcal{A}_\ell\left(\left(\mathcal{I} - \mathcal{M}_\ell \mathcal{A}_\ell\right) v_\ell, v_\ell\right) \leq (1 - c_{\text{MG}})\mathcal{A}_\ell\left(v_\ell, v_\ell\right), \qquad \forall v_\ell \in V_\ell.$$

From lemma 4.1, if $c_{\text{MG}} < C_{\text{MG}}$, the constants for the convergence estimate of the multigrid preconditioned system are

$$c_{\text{MG}} \geq \frac{2c_{\text{S}}}{1 + 2c_{\text{S}}}, \qquad\qquad C_{MG} \leq 1.$$

PROOF. The assumption is proven in lemma 4.8, and we defer the proof to appendix 2 where we use the theory developed in [**40**] and [**39**]. □

5. Numerical experiments

5.1. Poisson's equation. As a baseline for further experiments we show the results for Poisson's equation using different preconditioners, for $\delta_0 = 2$ and $h_\ell / h_{\ell-1} = 1/2$. We use a uniform mesh over the domain $[0,1] \times [0,1]$ and discretize the Laplacian using an interior penalty DG discretization, as shown in equation (27).

Table 2.1 shows the results for different mesh sizes, represented by the amount of levels in the multigrid case. For the unpreconditioned and two-level Schwarz methods, the finest mesh is the same as the finest mesh from the multigrid cases.

Results show that all preconditioner methods achieve a flat iteration count after a certain refining, with multiplicative methods performing significantly better, at a higher computational cost due to their sequential nature.

levels	U	2AS	2HS	2MS	MGAS	MGMS
2	3	3	3	4	3	4
3	10	10	6	6	6	6
4	22	18	9	7	10	7
5	43	24	11	7	12	8
6	85	26	11	7	13	8
7	> 100	25	11	7	14	8
8	> 100	25	11	7	14	8

TABLE 2.1. CG (respectively GMRES for 2MS and MGMS) Iterations using a unit source to reduce the residual by 10^{-8} for $\Sigma = 0$. Where U is unpreconditioned; 2AS, 2HS, 2MS are two level additive, hybrid and multiplicative Schwarz respectively; MGAS, MGMS are multigrid with additive and multiplicative Schwarz smoothers respectively.

5.2. 2 groups. In the case of a two groups problem, because of the conservation condition of zero column sum for Σ, the diagonal elements are equal. Should we calculate using a unit source for all groups, both the problem and the residual would be completely symmetric and the convergence equivalent to the Poisson equation case.

To illustrate, in this case the scattering matrix is

$$\Sigma = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix},$$

which will be pre-multiplied by $1/\varepsilon$, leading to a reaction term:

$$\frac{1}{\varepsilon}\Sigma v = \frac{1}{\varepsilon}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} v.$$

	MGAS					MGMS	2AS	2HS	2MS
ε / levels	1.0	10^{-1}	10^{-2}	10^{-3}	10^{-4}	w.c.	w.c.	w.c.	w.c.
2	5	5	4	4	4	4	6	5	4
3	8	8	6	6	6	6	14	8	6
4	10	10	10	10	10	7	22	10	7
5	12	12	12	12	12	8	25	11	7
6	13	13	13	13	13	8	25	11	7
7	14	14	14	14	14	8	25	11	7
8	14	14	14	14	14	8	25	11	7
9	14	14	14	14	14	8	25	11	7

TABLE 2.2. CG Iterations using a source $(\varepsilon, 0)$ or $(0, \varepsilon)$ to reduce the residual by 10^{-8}, where "w.c." is the maximum amount of iterations for different ε.

We show results in table 2.2. We observe that the iteration count flattens as well, with a very slightly higher count in coarser mesh cases, but asymptotically converging to the same iteration count as in the Poisson's equation case.

5.3. Multigroup. We devise a reaction matrix with a *contrast* between coefficients in different groups that is inversely proportional to different powers of ε as follows:

$$\Sigma = \begin{pmatrix} \alpha_1 & -1 & -\varepsilon^{-1} & -\varepsilon^{-2} & -\varepsilon^{-3} & ... \\ -1 & \alpha_2 & -1 & -1 & -1 & ... \\ -\varepsilon^{-1} & -1 & \alpha_3 & -\varepsilon^{-1} & -\varepsilon^{-2} & ... \\ -\varepsilon^{-2} & -1 & -\varepsilon^{-1} & \alpha_4 & -\varepsilon^{-1} & ... \\ -\varepsilon^{-3} & -1 & -\varepsilon^{-2} & -\varepsilon^{-1} & \alpha_5 & ... \\ ... & ... & ... & ... & ... & ... \end{pmatrix},$$

where $\alpha_g = -\sum_{g' \neq g} \Sigma_{g,g'} = 1 + \varepsilon + \varepsilon^2 + \varepsilon^3 + ...$. We remark that the elements in the diagonal are such that the matrix has zero column sum. We will consider the top left block of this matrix as the scattering matrix in the following tests.

In table 2.3 we show results for 5 groups. In this case the columns are shown only up to $\varepsilon = 0.01$ to avoid a floating point under/overflow.

	MGAS			MGMS	2AS	2HS	2MS
levels \ ε	1.0	0.1	0.01	w.c.	w.c.	w.c.	w.c.
2	5	5	4	4	9	5	4
3	8	7	6	6	15	8	6
4	10	10	10	7	22	10	7
5	12	12	12	8	25	11	7
6	13	13	13	8	26	11	7
7	14	14	14	8	25	11	7
8	14	14	14	8	25	11	7
9	14	14	14	8	25	11	7

TABLE 2.3. CG Iterations to reduce the residual by 10^{-8} for a 5 groups calculation, where "w.c." is the maximum amount of iterations for different ε. Sources tested: $(\varepsilon, 0, \varepsilon, 0, \varepsilon), (0, \varepsilon, 0, \varepsilon, 0), (0, \varepsilon, \varepsilon, \varepsilon, 0)$ or $(\varepsilon, 0, 0, 0, \varepsilon)$

Results show that increasing the amount of groups do not modify the asymptotic iteration count, effectively showing the preconditioner is insensitive to the size of the mesh and the relative size of the scattering.

5.4. Space dependent scattering. We modify the matrix used in the previous section using the following function:

$$f_i(x,y) = \begin{cases} (x,y) \in \mathbf{\Omega}_i & \sin^2(2\pi x)\sin^2(2\pi y) \\ (x,y) \notin \mathbf{\Omega}_i & 0 \end{cases},$$

where $\mathbf{\Omega}_i$, with $i = 0, 1, 2, 3$ are the four quadrants of the square domain. We choose this shape in order to test variable coefficients, that vary proportionally to $\frac{1}{\varepsilon}$ and

also with different shape for each group.

$$\Sigma = \begin{pmatrix} \alpha_1 & -f_0 & -\varepsilon^{-1}f_1 & -\varepsilon^{-2}f_2 & -\varepsilon^{-3}f_3 & \dots \\ -f_0 & \alpha_2 & -f_0 & -f_0 & -f_0 & \dots \\ -\varepsilon^{-1}f_1 & -f_0 & \alpha_3 & -\varepsilon^{-1}f_1 & -\varepsilon^{-2}f_2 & \dots \\ -\varepsilon^{-2}f_2 & -f_0 & -\varepsilon^{-1}f_1 & \alpha_4 & -\varepsilon^{-1}f_1 & \dots \\ -\varepsilon^{-3}f_3 & -f_0 & -\varepsilon^{-2}f_2 & -\varepsilon^{-1}f_1 & \alpha_5 & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \end{pmatrix}$$

	MGAS			MGMS	2AS	2HS	2MS
levels \ ε	1.0	0.1	0.01	w.c.	w.c.	w.c.	w.c.
2	6	7	6	4	19	7	4
3	9	10	9	6	22	10	6
4	11	12	12	7	25	11	7
5	13	13	13	8	27	12	8
6	13	14	14	8	28	12	8
7	14	14	15	8	28	13	8
8	14	15	15	9	27	12	8
9	14	15	15	9	27	12	8
10	15	15	15	9	27	12	8
11	15	15	15	9	27	12	8
12	15	15	15	9	27	12	8

TABLE 2.4. CG Iterations to reduce the residual by 10^{-8} for a 5 groups calculation, where "w.c." is the maximum amount of iterations for different ε. Sources tested: $(\varepsilon, 0, \varepsilon, 0, \varepsilon), (0, \varepsilon, 0, \varepsilon, 0), (0, \varepsilon, \varepsilon, \varepsilon, 0)$ or $(\varepsilon, 0, 0, 0, \varepsilon)$

Results show a slight change in the iteration count, but still keeping a flat distribution asymptotically as the mesh is refined, effectively showing that the preconditioner can deal with varying coefficients without major difficulties.

6. Conclusion

We have shown the efficiency of Schwarz smoothers for multilevel preconditioners for a discontinuous Galerkin discretization of the multigroup diffusion (reaction-diffusion) problem. The behavior of the solver is very similar to the case of Poisson's equation, implying that the standard theory of Schwarz methods is applicable.

In order to increase parallelization, we use minimal subdomains composed of only one cell, effectively eliminating the standard assumption that subdomains have to be equal or larger than the coarse mesh, at the cost of changing the condition number from $\mathcal{O}\left(\frac{H}{h}\right)$ to $\mathcal{O}\left(\frac{H^2}{h^2}\right)$. This result is consistent with the one shown in [**41**] and [**42**] for simplices, making use of the interpolation operator from the DG space onto the space of linear non-conforming elements.

In the following chapter we will address a nonsymmetric scattering matrix, for which we have similar numerical results.

CHAPTER 3

Fourier analysis of multigroup diffusion

Abstract. In this chapter, we perform a Fourier analysis of the additive Schwarz smoothers we used in the previous chapter. The advantage of this analysis lies on the precision at which the effect of the smoother on the error operator is described, provided that the expressions for the error operator eigenvalues and numerical range can be obtained, the estimators obtained can be very sharp. We perform a field of values analysis as well, for the case of the nonsymmetric scattering matrix, obtaining some interesting results on the behavior of Schwarz smoothers for the nonsymmetric problem.

1. Two domains problem

We begin the study with an illustrating example of the application of Fourier analysis to the diffusion problem, the purpose of this section is to provide a meaningful introduction to some of the tools we will use in later sections when addressing the nonsymmetric case.

This section closely follows the theory of optimized Schwarz methods in [**43**].

The reaction diffusion equation, considering an unitary diffusion reads

$$-\Delta u + \frac{1}{\varepsilon}\Sigma u = f, \text{ on } \Omega,$$

where Σ is in this case a positive scalar.

We decompose the 2D domain Ω into two subdomains separated by 0

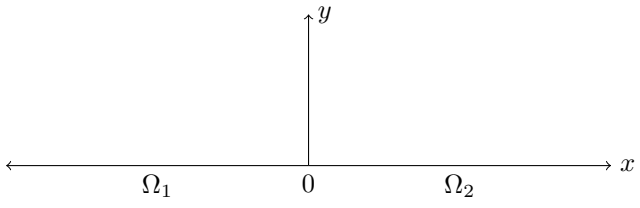
We study a Jacobi-Schwarz method for the two subdomains of the problem using a Robin-Robin coupling as follows

$$\left(\frac{1}{\varepsilon}\Sigma + \Delta\right) u_1^n = f \text{ in } \Omega_1, \tag{30}$$

$$\left.\frac{\partial u_1^n}{\partial x}\right|_{(0,y)} + \delta_1 u_1^n(0,y) = \left.\frac{\partial u_2^{n-1}}{\partial x}\right|_{(0,y)} + \delta_2 u_2^{n-1}(0,y), \tag{31}$$

$$\left(\frac{1}{\varepsilon}\Sigma + \Delta\right) u_2^n = f \text{ in } \Omega_2, \tag{32}$$

$$\left.\frac{\partial u_2^n}{\partial x}\right|_{(0,y)} - \delta_2 u_2^n(0,y) = \left.\frac{\partial \delta_1 u_1^{n-1}}{\partial x}\right|_{(0,y)} - \delta_1 u_1^{n-1}(0,y), \tag{33}$$



and we require the iterates to decay at infinity. By linearity it suffices to consider only the case $f = 0$ and analyze convergence to the zero solution. Our analysis is based on the Fourier transform

$$\begin{aligned}\widehat{f}(k) =& \mathcal{F}(f) \coloneqq \int_{-\infty}^{\infty} e^{-ikx} f(x) dx,\\ f(x) =& \mathcal{F}^{-1}(\widehat{f}) \coloneqq \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ikx} \widehat{f}(k) dk, \qquad\qquad k \in \mathbb{R}.\end{aligned}$$

Taking a Fourier transform of the Schwarz algorithm in the y direction, and using the property of the Fourier transform that derivatives in y become multiplications by ik, we obtain

$$\begin{aligned}\left(\frac{\Sigma}{\varepsilon} + k^2 - \frac{\partial^2}{\partial x^2}\right) u_1^n &= 0 \text{ in } \Omega_1,\\ \left.\frac{\partial u_1^n}{\partial x}\right|_{(0,k)} + \delta_1(k) u_1^n(0,k) &= \left.\frac{\partial u_2^{n-1}}{\partial x}\right|_{(0,k)} + \delta_2(k) u_2^{n-1}(0,k),\\ \left(\frac{\Sigma}{\varepsilon} + k^2 - \frac{\partial^2}{\partial x^2}\right) u_2^n &= 0 \text{ in } \Omega_2,\\ \left.\frac{\partial u_2^n}{\partial x}\right|_{(0,k)} - \delta_2(k) u_2^n(0,k) &= \left.\frac{\partial u_1^{n-1}}{\partial x}\right|_{(0,k)} - \delta_1(k) u_1^{n-1}(0,k),\end{aligned}$$

hence subdomain solutions in the Fourier transformed domain are of the form

$$\widehat{u}_j^n(x,k) = A_j^n(k) e^{\lambda_1(k)x} + B_j^n(k) e^{\lambda_2(k)x},$$

taking into account the condition on $u = 0$ at infinity we get

$$\widehat{u}_1^n(x,k) = A_1^n(k) e^{\lambda(k)x}, \qquad\qquad \widehat{u}_2^n(x,k) = A_2^n(k) e^{-\lambda(k)x},$$

where $\lambda = \sqrt{k^2 + \frac{\Sigma}{\varepsilon}}$.

Replacing in equation 31 we obtain

$$\begin{aligned}\left.\frac{\partial u_1^n}{\partial x}\right|_{(0,y)} + \delta_1(k) u_1^n(0,y) &= \left.\frac{\partial u_2^{n-1}}{\partial x}\right|_{(0,y)} + \delta_2(k) u_2^{n-1}(0,y),\\ A_1^n(k)\left(\lambda(k) + \delta_1(k)\right) &= A_2^{n-1}(k)\left(-\lambda(k) + \delta_2(k)\right),\end{aligned}$$

likewise

$$\begin{aligned}\left.\frac{\partial u_2^n}{\partial x}\right|_{(0,y)} - \delta_2(k) u_2^n(0,y) &= \left.\frac{\partial u_1^{n-1}}{\partial x}\right|_{(0,y)} - \delta_1(k) u_1^{n-1}(0,y),\\ A_2^n(k)\left(-\lambda(k) - \delta_2(k)\right) &= A_1^{n-1}(k)\left(\lambda(k) - \delta_1(k)\right),\end{aligned}$$

we deduce

$$\begin{aligned}A_1^n(k)\left(\lambda(k) + \delta_1(k)\right) &= A_2^{n-1}(k)\left(\delta_2(k) - \lambda(k)\right)\\ A_1^n(k)\left(\lambda(k) + \delta_1(k)\right) &= A_1^{n-2}(k)\frac{\delta_1(k) - \lambda(k)}{\delta_2(k) + \lambda(k)}\left(\delta_2(k) - \lambda(k)\right)\\ \rho(k) = \frac{A_1^n(k)}{A_1^{n-2}(k)} &= \frac{\delta_1(k) - \lambda(k)}{\delta_2(k) + \lambda(k)}\frac{\delta_2(k) - \lambda(k)}{\delta_1(k) + \lambda(k)}, \qquad (34)\end{aligned}$$

where $\rho(k)$ indicates convergence factor, the relative size between two iterations for the eigenvector corresponding to the eigenvalue k.

REMARK 3.1. Since we set the solution at infinity to be zero and we have set the right hand side to be zero as well, the solution of the equation is identically zero. This analysis is useful since it gives a measure on how fast the solution converges from the initial guess to the zero solution, hence providing convergence properties as a function of the frequency and, most importantly, as a function of the parameter δ of the method.

Expression (34) shows that by choosing the constants of the method, we can choose the frequency with the most residual reduction, therefore allowing us to optimize the algorithm to perform best as a *smoother*, reducing the high frequencies faster than the lower ones that will be addressed by using a coarse space solver. The constants cannot be chosen completely arbitrarily however, since it has to conserve the coercivity and continuity of the bilinear form.

We observe that we can choose positive values for δ_1 and δ_2, we simplify further by making them equal and note their value as δ, then

$$\rho(k) = \left(\frac{\delta(k) - \sqrt{k^2 + \frac{\Sigma}{\varepsilon}}}{\delta(k) + \sqrt{k^2 + \frac{\Sigma}{\varepsilon}}} \right)^2 .$$

Should we choose to solve exactly for the frequency k then we must choose $\delta(k) = \sqrt{k^2 + \frac{\Sigma}{\varepsilon}}$.

The analysis we have performed shows that parameters can be chosen in order to optimize the Schwarz methods convergence, depending on the constants involved in the problem. We address the multigroup diffusion problem in the following sections.

2. Preconditioned 1D, one group problem on a mesh

We use the definition of the bilinear form for the multigroup diffusion system in equation (29) for one group, a unitary diffusion coefficient and scaled reaction term as follows

$$\begin{aligned} \mathcal{A}(u,v) := & \int_{\mathbb{T}} \nabla u \nabla v dx + \frac{1}{\varepsilon} \int_{\mathbb{T}} uv dx - \int_{\mathbb{F}^I \cup \mathbb{F}^B} \left([\![u]\!] \{\!\{\nabla v\}\!\} ds + \{\!\{\nabla u\}\!\} [\![v]\!] \right) ds \\ & + \int_{\mathbb{F}^I \cup \mathbb{F}^B} \delta [\![u]\!] [\![v]\!] ds = \int_{\mathbb{T}} f v dx. \end{aligned}$$

We analyze the 1D problem using a discretization with nodal linear functions with support only on one cell as shown in figure 3.1.

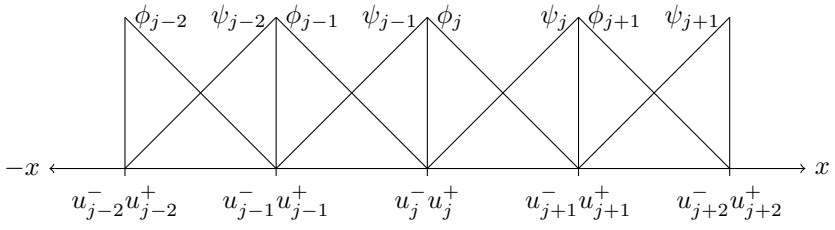


FIGURE 3.1. Discretization for a 1D reaction-diffusion problem.

We begin stating the DG bilinear form in one spatial dimension

$$\mathcal{A}(u,v) := \int_{\mathbb{T}} \frac{\partial u}{\partial x}\frac{\partial v}{\partial x}dx + \frac{1}{\varepsilon}\int_{\mathbb{T}} uvdx - \int_{\mathbb{F}^I\cup\mathbb{F}^B}\left(\llbracket u\rrbracket \left\{\!\!\left\{\frac{\partial v}{\partial x}\right\}\!\!\right\} ds + \left\{\!\!\left\{\frac{\partial u}{\partial x}\right\}\!\!\right\}\llbracket v\rrbracket\right) ds + \int_{\mathbb{F}^I\cup\mathbb{F}^B}\delta\llbracket u\rrbracket\llbracket v\rrbracket ds.$$

We set the origin at node j and identify the rest of the nodes as multiples of h, the bilinear form expression for cell j is

$$\begin{aligned}
\mathcal{A}(\phi_j,\phi_j) &= \int_0^h \frac{\partial\phi_j}{\partial x}\frac{\partial\phi_j}{\partial x}dx + \frac{1}{\varepsilon}\int_0^h \phi_j\phi_j dx - (0-\phi_j(0))\frac{0+\phi_j'(0)}{2}\\
&\quad -\frac{0+\phi_j'(0)}{2}(0-\phi_j(0)) + \delta(0-\phi_j(0))(0-\phi_j(0)),\\
\mathcal{A}(\phi_j,\psi_j) &= \int_0^h \frac{\partial\phi_j}{\partial x}\frac{\partial\psi_j}{\partial x}dx + \frac{1}{\varepsilon}\int_0^h \phi_j\psi_j dx - (0-\phi_j(0))\frac{0+\psi_j'(0)}{2}\\
&\quad -\frac{\phi_j'(h)+0}{2}(\psi_j(h)-0),\\
\mathcal{A}(\psi_j,\phi_j) &= \int_0^h \frac{\partial\psi_j}{\partial x}\frac{\partial\phi_j}{\partial x}dx + \frac{1}{\varepsilon}\int_0^h \psi_j\phi_j dx - (\psi_j(h)-0)\frac{\phi_j'(h)-0}{2}\\
&\quad -\frac{\psi_j'(0)+0}{2}(0-\phi_j(0)),\\
\mathcal{A}(\psi_j,\psi_j) &= \int_0^h \frac{\partial\psi_j}{\partial x}\frac{\partial\psi_j}{\partial x}dx + \frac{1}{\varepsilon}\int_0^h \psi_j\psi_j dx - (0-\psi_j(0))\frac{0+\psi_j'(0)}{2}\\
&\quad -\frac{0+\psi_j'(0)}{2}(0-\psi_j(0)) + \delta(0-\psi_j(0))(0-\psi_j(0)).
\end{aligned}$$

This leads to cell and face matrices connecting neighboring cells as follows

$$C = \begin{pmatrix} \delta+\frac{h}{3\varepsilon} & \frac{h}{6\varepsilon}\\ \frac{h}{6\varepsilon} & \delta+\frac{h}{3\varepsilon}\end{pmatrix}, \qquad F_u = \begin{pmatrix} -\frac{1}{2h} & 0\\ -\delta+\frac{1}{h} & -\frac{1}{2h}\end{pmatrix}, \qquad F_l = \begin{pmatrix} -\frac{1}{2h} & -\delta+\frac{1}{h}\\ 0 & -\frac{1}{2h}\end{pmatrix},$$

which, after assembly, compose the following Toeplitz linear system matrix for the discretized 1D problem

$$Au = \begin{pmatrix}
\delta+\frac{h}{3\varepsilon} & \frac{h}{6\varepsilon} & -\frac{1}{2h} & 0 & 0 & 0 & 0 & 0\\
\frac{h}{6\varepsilon} & \delta+\frac{h}{3\varepsilon} & -\delta+\frac{1}{h} & -\frac{1}{2h} & 0 & 0 & 0 & 0\\
-\frac{1}{2h} & -\delta+\frac{1}{h} & \delta+\frac{h}{3\varepsilon} & \frac{h}{6\varepsilon} & -\frac{1}{2h} & 0 & 0 & 0\\
0 & -\frac{1}{2h} & \frac{h}{6\varepsilon} & \delta+\frac{h}{3\varepsilon} & -\delta+\frac{1}{h} & -\frac{1}{2h} & 0 & 0\\
0 & 0 & -\frac{1}{2h} & -\delta+\frac{1}{h} & \delta+\frac{h}{3\varepsilon} & \frac{h}{6\varepsilon} & -\frac{1}{2h} & 0\\
0 & 0 & 0 & -\frac{1}{2h} & \frac{h}{6\varepsilon} & \delta+\frac{h}{3\varepsilon} & -\delta+\frac{1}{h} & -\frac{1}{2h}\\
0 & 0 & 0 & 0 & -\frac{1}{2h} & -\delta+\frac{1}{h} & \delta+\frac{h}{3\varepsilon} & \frac{h}{6\varepsilon}\\
0 & 0 & 0 & 0 & 0 & -\frac{1}{2h} & \frac{h}{6\varepsilon} & \delta+\frac{h}{3\varepsilon}
\end{pmatrix}
\begin{pmatrix} u_{j-2}^+\\ u_{j-1}^-\\ u_{j-1}^+\\ u_j^-\\ u_j^+\\ u_{j+1}^-\\ u_{j+1}^+\\ u_{j+2}^-\end{pmatrix} =
\begin{pmatrix} f_{j-2}^+\\ f_{j-1}^-\\ f_{j-1}^+\\ f_j^-\\ f_j^+\\ f_{j+1}^-\\ f_{j+1}^+\\ f_{j+2}^-\end{pmatrix}.$$

We perform a simple test to the matrix, we calculate the result of $u = A^{-1}f$, setting $h = 0.25$, $\varepsilon = \infty$, $\delta = 10/h$, $f = 1$ in a unitary domain composed by four cells, the result can be seen in figure 3.2. It can be observed, that the result is the inverse of a Laplacian, with a maximum of 1/8 at 1/2. We do not observe the discontinuity of the solution space, since the choice of $\delta = 10/h$ penalizes the jumps strongly, we illustrate with such an example to test the correctness of the approximation with respect to the expected shape of the exact solution.

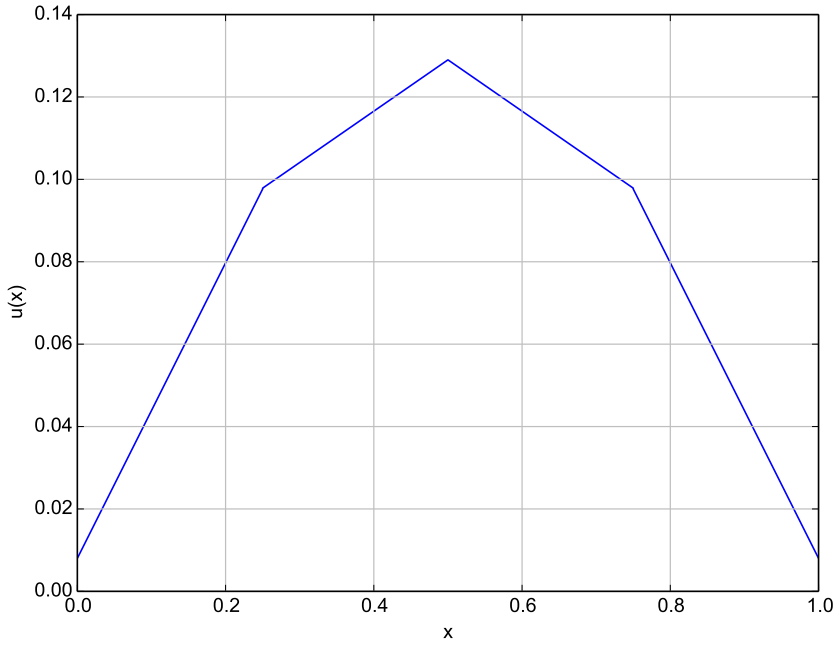


FIGURE 3.2. Solution to a Poisson problem with a 1D DG discretization

In order to study the behavior of the bilinear form in a similar fashion to what we did in the previous section, we will propose an oscillatory behavior for the solutions at the faces, u^- and u^+ separately, that allows us to represent all possible DG functions (see figure 3.3).

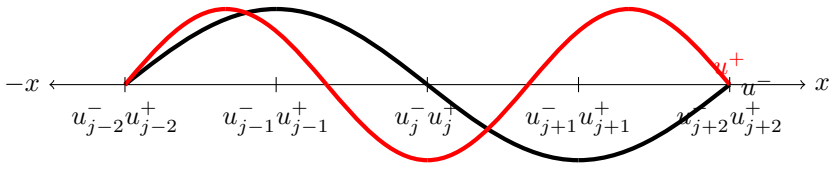


FIGURE 3.3. Oscillatory Fourier basis to represent functions in the DG space.

We proceed to reorder the variables separating u^- and u^+ and we get the following block-Toeplitz matrix

$$A_r u = \begin{pmatrix} \delta+\frac{h}{3\varepsilon} & -\frac{1}{2h} & 0 & 0 & \frac{h}{6\varepsilon} & -\delta+\frac{1}{h} & 0 & 0 \\ -\frac{1}{2h} & \delta+\frac{h}{3\varepsilon} & -\frac{1}{2h} & 0 & 0 & \frac{h}{6\varepsilon} & -\delta+\frac{1}{h} & 0 \\ 0 & -\frac{1}{2h} & \delta+\frac{h}{3\varepsilon} & -\frac{1}{2h} & 0 & 0 & \frac{h}{6\varepsilon} & -\delta+\frac{1}{h} \\ 0 & 0 & -\frac{1}{2h} & \delta+\frac{h}{3\varepsilon} & 0 & 0 & 0 & \frac{h}{6\varepsilon} \\ \frac{h}{6\varepsilon} & 0 & 0 & 0 & \delta+\frac{h}{3\varepsilon} & -\frac{1}{2h} & 0 & 0 \\ -\delta+\frac{1}{h} & \frac{h}{6\varepsilon} & 0 & 0 & -\frac{1}{2h} & \delta+\frac{h}{3\varepsilon} & -\frac{1}{2h} & 0 \\ 0 & -\delta+\frac{1}{h} & \frac{h}{6\varepsilon} & 0 & 0 & -\frac{1}{2h} & \delta+\frac{h}{3\varepsilon} & -\frac{1}{2h} \\ 0 & 0 & -\delta+\frac{1}{h} & \frac{h}{6\varepsilon} & 0 & 0 & -\frac{1}{2h} & \delta+\frac{h}{3\varepsilon} \end{pmatrix} \begin{pmatrix} u^-_{j-1} \\ u^-_j \\ u^-_{j+1} \\ u^-_{j+2} \\ u^+_{j-2} \\ u^+_{j-1} \\ u^+_j \\ u^+_{j+1} \end{pmatrix} = \begin{pmatrix} f^-_{j-1} \\ f^-_j \\ f^-_{j+1} \\ f^-_{j+2} \\ f^+_{j-2} \\ f^+_{j-1} \\ f^+_j \\ f^+_{j+1} \end{pmatrix}.$$

As stated before, we propose the following behavior for u^-_j and u^+_j

$$u^-_j = C^-_\omega e^{i\omega x_j}, \qquad\qquad u^+_j = C^+_\omega e^{i\omega x_j},$$

using the assumption that the space grid is homogeneous we can express the coordinate of any face with a multiple of h as follows

$$A_r u = A_r \begin{pmatrix} u^-_{j-1} \\ u^-_j \\ u^-_{j+1} \\ u^-_{j+2} \\ u^+_{j-2} \\ u^+_{j-1} \\ u^+_j \\ u^+_{j+1} \end{pmatrix} = A_r \begin{pmatrix} C^-_\omega e^{-i\omega h} \\ C^-_\omega \\ C^-_\omega e^{i\omega h} \\ C^-_\omega e^{2i\omega h} \\ C^+_\omega e^{-2i\omega h} \\ C^+_\omega e^{-i\omega h} \\ C^+_\omega \\ C^+_\omega e^{i\omega h} \end{pmatrix} e^{i\omega x_j}$$

$$= A_r \begin{pmatrix} C^-_\omega e^{-i\omega h} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & C^-_\omega & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & C^-_\omega e^{i\omega h} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & C^-_\omega e^{2i\omega h} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & C^+_\omega e^{-2i\omega h} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & C^+_\omega e^{-i\omega h} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & C^+_\omega & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & C^+_\omega e^{i\omega h} \end{pmatrix} e^{i\omega x_j}.$$

We then proceed to extract the system of equations at node j

$$\widehat{A}_j = \begin{pmatrix} \delta+\frac{h}{3\varepsilon}-\frac{\cos(h\omega)}{h} & \left(-\delta+\frac{1}{h}\right)e^{\imath h\omega}+\frac{h}{6\varepsilon} \\ \left(-\delta+\frac{1}{h}\right)e^{-\imath h\omega}+\frac{h}{6\varepsilon} & \delta+\frac{h}{3\varepsilon}-\frac{\cos(h\omega)}{h} \end{pmatrix} \begin{pmatrix} C^-_\omega \\ C^+_\omega \end{pmatrix},$$

in the same fashion we obtain the block diagonal of A, noted D as

$$D = \begin{pmatrix} \delta+\frac{h}{3\varepsilon} & \frac{h}{6\varepsilon} & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{h}{6\varepsilon} & \delta+\frac{h}{3\varepsilon} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \delta+\frac{h}{3\varepsilon} & \frac{h}{6\varepsilon} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{h}{6\varepsilon} & \delta+\frac{h}{3\varepsilon} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \delta+\frac{h}{3\varepsilon} & \frac{h}{6\varepsilon} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{h}{6\varepsilon} & \delta+\frac{h}{3\varepsilon} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \delta+\frac{h}{3\varepsilon} & \frac{h}{6\varepsilon} \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{h}{6\varepsilon} & \delta+\frac{h}{3\varepsilon} \end{pmatrix},$$

after reordering this gives

$$D_r = \begin{pmatrix} \delta+\frac{h}{3\varepsilon} & 0 & 0 & 0 & \frac{h}{6\varepsilon} & 0 & 0 & 0 \\ 0 & \delta+\frac{h}{3\varepsilon} & 0 & 0 & 0 & \frac{h}{6\varepsilon} & 0 & 0 \\ 0 & 0 & \delta+\frac{h}{3\varepsilon} & 0 & 0 & 0 & \frac{h}{6\varepsilon} & 0 \\ 0 & 0 & 0 & \delta+\frac{h}{3\varepsilon} & 0 & 0 & 0 & \frac{h}{6\varepsilon} \\ \frac{h}{6\varepsilon} & 0 & 0 & 0 & \delta+\frac{h}{3\varepsilon} & 0 & 0 & 0 \\ 0 & \frac{h}{6\varepsilon} & 0 & 0 & 0 & \delta+\frac{h}{3\varepsilon} & 0 & 0 \\ 0 & 0 & \frac{h}{6\varepsilon} & 0 & 0 & 0 & \delta+\frac{h}{3\varepsilon} & 0 \\ 0 & 0 & 0 & \frac{h}{6\varepsilon} & 0 & 0 & 0 & \delta+\frac{h}{3\varepsilon} \end{pmatrix},$$

and we extract the equation on node j, which naturally is a block of matrix D

$$\widehat{D}_j = \begin{pmatrix} \delta+\frac{h}{3\varepsilon} & \frac{h}{6\varepsilon} \\ \frac{h}{6\varepsilon} & \delta+\frac{h}{3\varepsilon} \end{pmatrix},$$

following we obtain the explicit inverse

$$\widehat{D}_j^{-1} = \frac{1}{\left(\delta+\frac{h}{3\varepsilon}\right)^2-\left(\frac{h}{6\varepsilon}\right)^2} \begin{pmatrix} \delta+\frac{h}{3\varepsilon} & -\frac{h}{6\varepsilon} \\ -\frac{h}{6\varepsilon} & \delta+\frac{h}{3\varepsilon} \end{pmatrix}.$$

In order to express the error operator of the block-Jacobi preconditioned system in our smoother, we write $I - \alpha\widehat{D}_j^{-1}\widehat{A}_j$, where α is a relaxation parameter.

We can calculate the eigenvalues of $I-\alpha\widehat{D}_j^{-1}\widehat{A}_j$ explicitly as well, the expression reads

(35)

$$\lambda_{\pm}\left(I-\alpha\widehat{D}_j^{-1}\widehat{A}_j\right) = 1-\alpha\left(1-\frac{\frac{2}{h}\left(\delta+\frac{h}{3\varepsilon}\right)+\frac{h}{3\varepsilon}\left(\frac{1}{h}-\delta\right)}{\left(\delta+\frac{h}{2\varepsilon}\right)\left(2\delta+\frac{h}{3\varepsilon}\right)}\cos(h\omega) \pm \frac{1}{\left(\delta+\frac{h}{2\varepsilon}\right)\left(2\delta+\frac{h}{3\varepsilon}\right)} \sqrt{4\left(\frac{1}{h}-\delta\right)^2\left(\delta+\frac{h}{3\varepsilon}\right)^2+\left(\frac{4}{3\varepsilon}\left(\frac{1}{h}-\delta\right)\left(\delta+\frac{h}{3\varepsilon}\right)+\frac{1}{9\varepsilon^2}\right)\cos^2(h\omega)-\frac{h^2}{9\varepsilon^2}\left(\frac{1}{h}-\delta\right)^2\sin^2(h\omega)}\right).$$

In the following sections it will be our objective to provide an explicit expression for the parameter values that make the block-Jacobi algorithm a good *smoother*, as opposed to a good *solver*.

A better smoother would provide a higher residual reduction in high frequencies, and therefore leaving the task of eliminate the error for lower frequencies to the coarse solver.

2.1. Test on Poisson's equation. We can now test the expression for the eigenvalue obtained in equation (35) for our DG method on a 1D Poisson's problem, by evaluating the expression at $\varepsilon \to \infty$ and $\delta = \delta_0/h$ with $\delta_0 > 1$ to get

$$\lambda\left(I-\alpha\widehat{D}_j^{-1}\widehat{A}_j\right) = 1-\frac{\alpha}{\delta_0}(1-\cos(h\omega))$$

The eigenvalue becomes unique since the two elements in the diagonal of $I-\alpha\widehat{D}_j^{-1}\widehat{A}_j$ are equal. We show the behavior on figure 3.4, the optimal value for $\frac{\alpha}{\delta_0} = 2/3$ is achieved by minimizing the maximum value of $|\lambda(\omega)|$ for $\omega \in \left[\frac{\pi}{2}, \pi\right]$, using the equioscillation theorem, the minimum is attained when $\lambda\left(\omega = \frac{\pi}{2}\right) = \lambda(\omega = \pi)$.

2.2. Test on scalar reaction diffusion. We analyze now the case depending on ε while setting $\delta = 1/h$, evaluating equation (35) we obtain

$$\lambda\left(I-\alpha\widehat{D}_j^{-1}\widehat{A}_j\right) = 1-\alpha\left(1-\frac{\frac{2}{h}\left(\frac{1}{h}+\frac{h}{3\varepsilon}\right)\pm\frac{1}{3\varepsilon}}{\left(\frac{1}{h}+\frac{h}{2\varepsilon}\right)\left(\frac{2}{h}+\frac{h}{3\varepsilon}\right)}\cos(h\omega)\right),$$

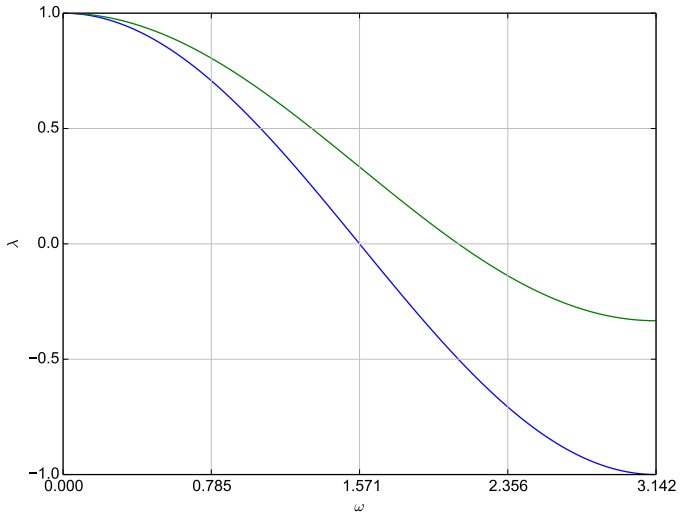


FIGURE 3.4. Eigenvalues of the error operator for Poisson's equation for $\alpha/\delta_0 = 1$ (blue) and $\alpha/\delta_0 = 2/3$ (green).

we apply the same technique as in the previous section, namely the equioscillation theorem, to obtain an optimal α and look for the value such that $\lambda\left(\omega = \frac{\pi}{2}\right) = \lambda\left(\omega = \pi\right)$ to get

$$\alpha = \frac{2\varepsilon^2 + \frac{4\varepsilon h^2}{3} + \frac{h^4}{6}}{3\varepsilon^2 + \frac{11\varepsilon h^2}{6} + \frac{h^4}{6}},$$

where we see that as $h \to 0$ and/or $\varepsilon \to \infty$ we come back to the optimal value of Poisson's equation, $\alpha = 2/3$ (see figure 3.5).

3. General expression of the optimal relaxation parameter

Analyzing the general expression for the eigenvalues depending on ε and taking $\delta = \delta_0/h$, with $\delta_0 > 1$ we follow the same procedure as in the case for $\delta = 1/h$ to obtain

$$\alpha\left(\delta_0, \varepsilon, h\right) = \left\{-2\left(\frac{\delta_0}{h} + \frac{h}{3\varepsilon}\right)^2 + \frac{h^2}{18\varepsilon^2}\right\} \Big/ \left\{\frac{h^2}{18\varepsilon^2} - \frac{1-\delta_0}{6\varepsilon} - 2\left(\frac{\delta_0}{h} + \frac{h}{3\varepsilon}\right)^2 - \frac{1}{h}\left(\frac{\delta_0}{h} + \frac{h}{3\varepsilon}\right) + \frac{1-\delta_0}{h}\sqrt{\left(\frac{\delta_0}{h} + \frac{h}{3\varepsilon}\right)^2 - \frac{h^2}{36\varepsilon^2}} + \sqrt{\frac{1}{h^2}\left(1-\delta_0\right)^2\left(\frac{\delta_0}{h} + \frac{h}{3\varepsilon}\right)^2 + \frac{1}{3\varepsilon h}\left(1-\delta_0\right)\left(\frac{\delta_0}{h} + \frac{h}{3\varepsilon}\right) + \frac{1}{36\varepsilon^2}}\right\}.$$

Inspecting the last expression we observe:

$$\lim_{\varepsilon\to\infty} \alpha\left(\delta_0, \varepsilon, h\right) = \lim_{h\to 0} \alpha\left(\delta_0, \varepsilon, h\right) = \frac{2\delta_0}{2\delta_0 + 1},$$

which is consistent with Poisson's case, as when $\delta_0 = 1$, $\alpha = 2/3$.

This expression still leaves us with a free parameter $\delta_0 > 1$, varying this coefficient modifies the behaviour of the eigenvalues as shown in figure 3.6.

The figure shows that in order to minimize the eigenvalues in $[\pi/2, \pi]$, and hence improve the smoothing properties, we must look for a difference between

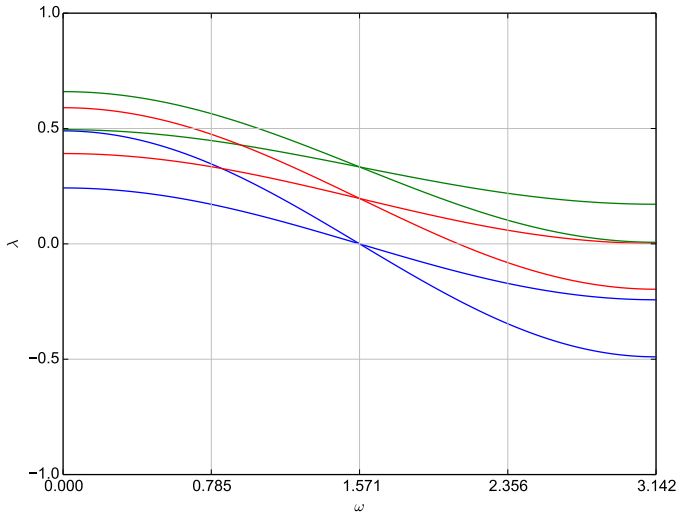


FIGURE 3.5. Eigenvalues of the error operator for the scalar reaction diffusion equation with $\alpha = 1$ (blue), $\alpha = 2/3$ (green) and $\alpha = \frac{2\varepsilon^2+\frac{4\varepsilon h^2}{3}+\frac{h^4}{6}}{3\varepsilon^2+\frac{11\varepsilon h^2}{6}+\frac{h^4}{6}}$ (red). Here plotted for $\delta = 1/h$, $\varepsilon = 0.01$ and $h = 0.25$.

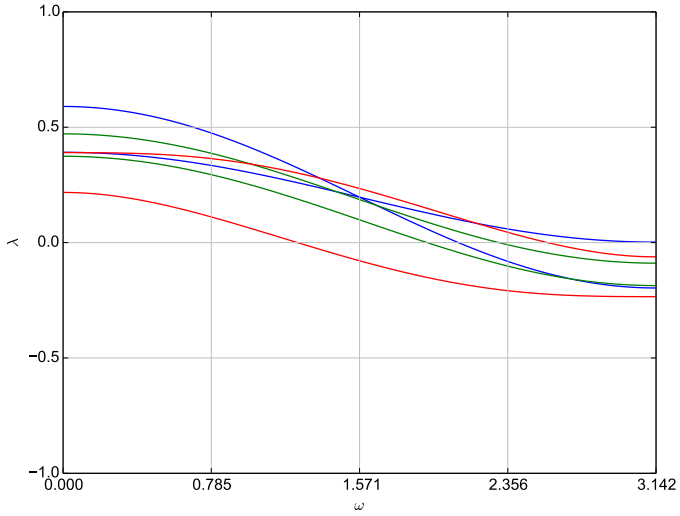


FIGURE 3.6. Eigenvalues of the error operator for the scalar reaction diffusion equation with $\delta_0 = 1$ (blue), $\delta_0 = 1.25$ (green) and $\delta_0 = 1.6$ (red). Here plotted for $\varepsilon = 0.01$ and $h = 0.25$.

both eigenvalues that does not depend on ω. Inspecting the general expression for the eigenvalues, a discriminant independent of ω occurs when

$$\frac{4}{3h\varepsilon}(1-\delta_0)\left(\delta+\frac{h}{3\varepsilon}\right)+\frac{1}{9\varepsilon^2}=-\frac{h}{9\varepsilon^2}(1-\delta_0)^2$$
$$\delta_0=\frac{6\varepsilon-h^2+\sqrt{(6\varepsilon+h^2)(6\varepsilon+5h^2)}}{12\varepsilon+h^2},$$

where we have chosen the root such that $\delta_0>1$ as required for coercivity of the original problem. We observe that both for $\varepsilon\to\infty$ and $h\to 0$, we obtain $\delta_0\to 1$, however, when $\varepsilon\to 0$, $\delta_0\to\sqrt{5}-1\approx 1.24>1$, $\forall h$.

Finally, the optimal values of λ to maximize its smoothing capacities are

$$\lambda_{\text{opt}}=1-\alpha_{\text{opt}}\left(1-\frac{\frac{2}{h}\left(\frac{\delta_{\text{opt}}}{h}+\frac{h}{3\varepsilon}\right)+\frac{1}{3\varepsilon}(1-\delta_{\text{opt}})}{\left(\frac{\delta_{\text{opt}}}{h}+\frac{h}{2\varepsilon}\right)\left(\frac{2\delta_{\text{opt}}}{h}+\frac{h}{3\varepsilon}\right)}\cos(h\omega)\pm\frac{(1-\delta_{\text{opt}})\sqrt{\frac{4}{h}\left(\frac{\delta_{\text{opt}}}{h}+\frac{h}{3\varepsilon}\right)^2-\frac{h}{9\varepsilon^2}}}{\left(\frac{\delta_{\text{opt}}}{h}+\frac{h}{2\varepsilon}\right)\left(\frac{2\delta_{\text{opt}}}{h}+\frac{h}{3\varepsilon}\right)}\right),$$

where

$$\alpha_{\text{opt}}=\left\{-2\left(\frac{\delta_{\text{opt}}}{h}+\frac{h}{3\varepsilon}\right)^2+\frac{h^2}{18\varepsilon^2}\right\}\Big/\left\{\frac{h^2}{18\varepsilon^2}-\frac{1-\delta_{\text{opt}}}{6\varepsilon}-2\left(\frac{\delta_{\text{opt}}}{h}+\frac{h}{3\varepsilon}\right)^2-\frac{1}{h}\left(\frac{\delta_{\text{opt}}}{h}+\frac{h}{3\varepsilon}\right)\right.$$
$$+\frac{1-\delta_{\text{opt}}}{h}\sqrt{\left(\frac{\delta_{\text{opt}}}{h}+\frac{h}{3\varepsilon}\right)^2-\frac{h^2}{36\varepsilon^2}}$$
$$\left.+\sqrt{\frac{1}{h^2}(1-\delta_{\text{opt}})^2\left(\frac{\delta_{\text{opt}}}{h}+\frac{h}{3\varepsilon}\right)^2+\frac{1}{3\varepsilon h}(1-\delta_{\text{opt}})\left(\frac{\delta_{\text{opt}}}{h}+\frac{h}{3\varepsilon}\right)+\frac{1}{36\varepsilon^2}}\right\},$$

and

$$\delta_{\text{opt}}=\frac{6\varepsilon-h^2+\sqrt{(6\varepsilon+h^2)(6\varepsilon+5h^2)}}{12\varepsilon+h^2}.$$

The behaviour of these expressions is shown in figure 3.7.

We obtain the exact same behaviour increasing ε than decreasing h suggests both variables are inversely proportional.

4. Verification of the general expression

Taking the original definitions at the beginning of this section of the operators A and D we can calculate the eigenvalues of $I-\alpha D^{-1}A$ explicitly.

We provide verification of the values of λ_{opt}, ordered from largest to smallest in figure 3.8 and a plot of the solution in figure 3.9. Figure 3.9 shows the discontinuities close to the boundary, albeit in a very small size.

To verify that the frequencies for which the eigenvalues have been plotted are correct we provide a glimpse of the eigenvectors corresponding to the most positive eigenvalue (figure 3.10) and the most negative eigenvalue (figure 3.11), we point out that the behaviour of the eigenvalues has to be interpreted at the faces separately for u^+ and u^-.

We observe that in the eigenvector with the largest eigenvalue, u^+ and u^- oscilate half a period in the whole domain while in the eigenvector with the smallest eigenvalue they oscilate as fast as it is possible to represent in the mesh.

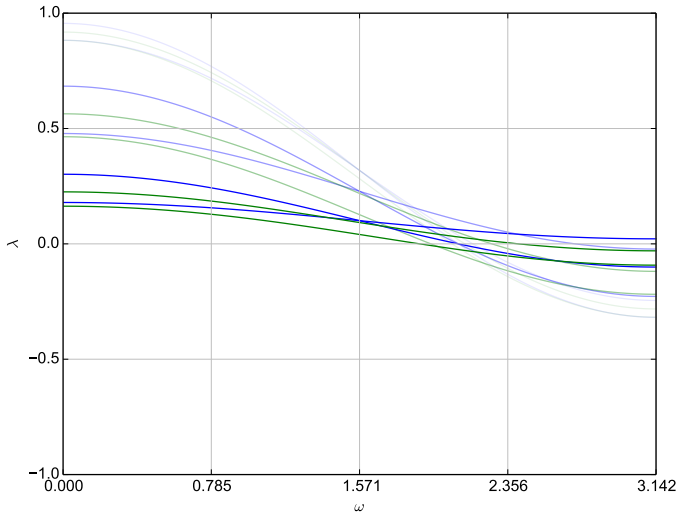


FIGURE 3.7. Scalar reaction diffusion equation optimal eigenvalues (green), compared with setting $\delta_0 = 1$ (blue). For $\varepsilon = 0.1$, $\varepsilon = 0.015$ and $\varepsilon = 0.003$ (from clear to bold line) . Here plotted for $h = 0.25$.

5. Preconditioned 1D, 2 groups problem on a mesh

In this section, we concentrate on a 2-group multigroup diffusion problem, in an attempt to expand what we have learned in previous sections with a more simplified 1-group model. The 2-group interior penalty DG diffusion equation reads

$$\begin{aligned}
\mathcal{A}_1(u_1, v_1) := &\int_{\mathbb{T}} \nabla u_1 \nabla v_1 dx + \frac{1}{\varepsilon}\int_{\mathbb{T}} \sigma_1 u_1 v_1 dx - \frac{1}{\varepsilon}\int_{\mathbb{T}} \sigma_2 u_2 v_2 dx \\
&- \int_{\mathbb{F}^I \cup \mathbb{F}^B} \left([\![u_1]\!] \{\!\{\nabla v_1\}\!\} ds + \{\!\{\nabla u_1\}\!\} [\![v_1]\!] \right) ds \\
&+ \int_{\mathbb{F}^I \cup \mathbb{F}^B} \delta [\![u_1]\!] [\![v_1]\!] ds = \int_{\mathbb{T}} f v dx, \\
\mathcal{A}_2(u_2, v_2) := &\int_{\mathbb{T}} \nabla u_2 \nabla v_2 dx + \frac{1}{\varepsilon}\int_{\mathbb{T}} \sigma_2 u_2 v_2 dx - \frac{1}{\varepsilon}\int_{\mathbb{T}} \sigma_1 u_1 v_1 dx \\
&- \int_{\mathbb{F}^I \cup \mathbb{F}^B} \left([\![u_2]\!] \{\!\{\nabla v_2\}\!\} ds + \{\!\{\nabla u_2\}\!\} [\![v_2]\!] \right) ds \\
&+ \int_{\mathbb{F}^I \cup \mathbb{F}^B} \delta [\![u_2]\!] [\![v_2]\!] ds = \int_{\mathbb{T}} f v dx.
\end{aligned}$$

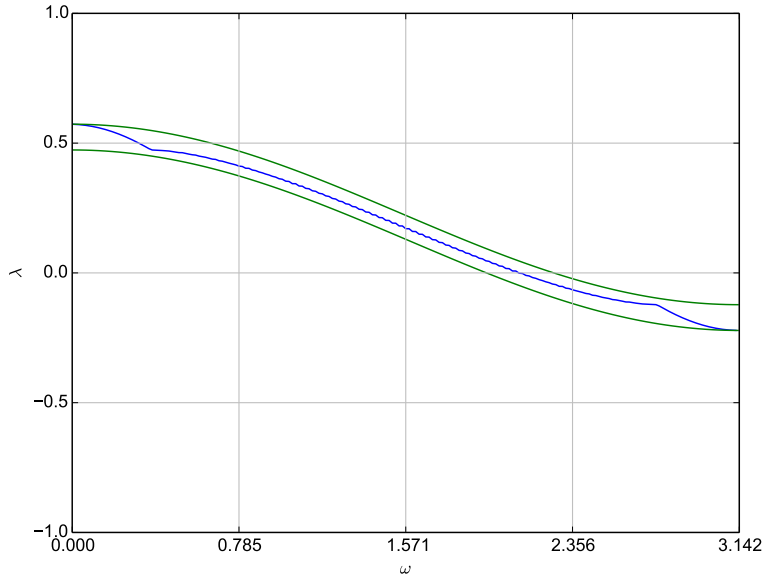


FIGURE 3.8. Verification of optimal eigenvalues from the Fourier analysis (green), with an explicit calculation of $\lambda_{\text{opt}}\left(I - \alpha_{\text{opt}}\left(\delta_{\text{opt}}\right) D\left(\delta_{\text{opt}}\right)^{-1} A\left(\delta_{\text{opt}}\right)\right)$ (blue). Here shown for $h = 1/100$, $\varepsilon = 0.000025$ in $[0,1]$, i.e. 100 cells.

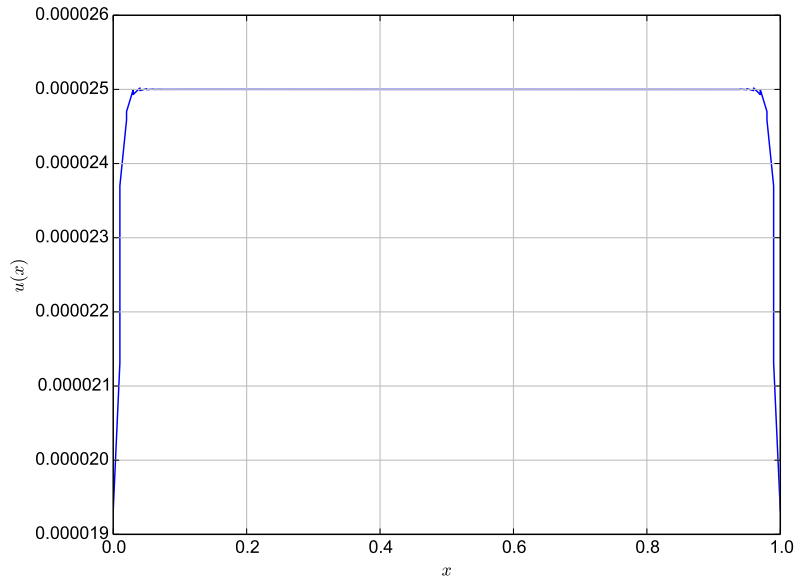


FIGURE 3.9. Solution for $h = 1/100$, $\varepsilon = 0.000025$ in $[0,1]$, i.e. 100 cells.

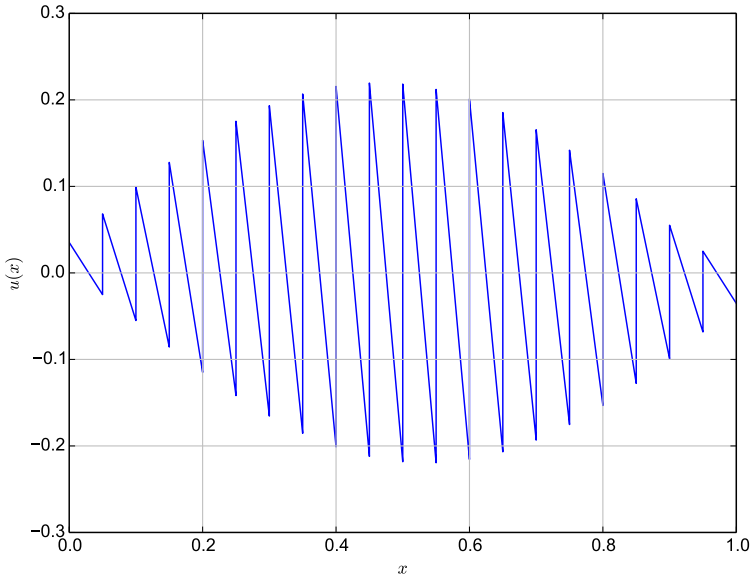


FIGURE 3.10. Eigenvector with the most positive eigenvalue. Here shown for $h = 1/20$, $\varepsilon = 0.000025$ in $[0, 1]$, i.e. 20 cells.

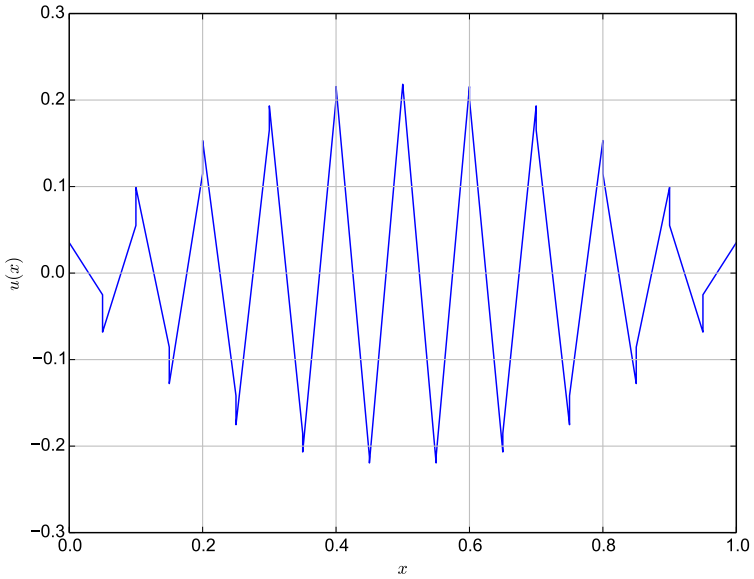


FIGURE 3.11. Eigenvector with the most negative eigenvalue. Here shown for $h = 1/20$, $\varepsilon = 0.000025$ in $[0, 1]$, i.e. 20 cells.

This leads to the following cell and face matrices

$$C = \begin{pmatrix} \delta_1 + \frac{h\sigma_1}{3\varepsilon} & \frac{h\sigma_1}{6\varepsilon} & -\frac{h\sigma_2}{3\varepsilon} & -\frac{h\sigma_2}{6\varepsilon} \\ \frac{h\sigma_1}{6\varepsilon} & \delta_1 + \frac{h\sigma_1}{3\varepsilon} & -\frac{h\sigma_2}{6\varepsilon} & -\frac{h\sigma_2}{3\varepsilon} \\ -\frac{h\sigma_1}{3\varepsilon} & -\frac{h\sigma_1}{6\varepsilon} & \delta_2 + \frac{h\sigma_2}{3\varepsilon} & \frac{h\sigma_2}{6\varepsilon} \\ -\frac{h\sigma_1}{6\varepsilon} & -\frac{h\sigma_1}{3\varepsilon} & \frac{h\sigma_2}{6\varepsilon} & \delta_2 + \frac{h\sigma_2}{3\varepsilon} \end{pmatrix},$$

$$F_u = \begin{pmatrix} -\frac{1}{2h} & 0 & 0 & 0 \\ -\delta_1 + \frac{1}{h} & -\frac{1}{2h} & 0 & 0 \\ 0 & 0 & -\frac{1}{2h} & 0 \\ 0 & 0 & -\delta_2 + \frac{1}{h} & -\frac{1}{2h} \end{pmatrix},$$

$$F_l = \begin{pmatrix} -\frac{1}{2h} & -\delta_1 + \frac{1}{h} & 0 & 0 \\ 0 & -\frac{1}{2h} & 0 & 0 \\ 0 & 0 & -\frac{1}{2h} & -\delta_2 + \frac{1}{h} \\ 0 & 0 & 0 & -\frac{1}{2h} \end{pmatrix}.$$

We assemble cell and face matrix intro a system matrix, again a Toeplitz type, shown below for two cells.

$$\begin{pmatrix} \delta_1 + \frac{h\sigma_1}{3\varepsilon} & \frac{h\sigma_1}{6\varepsilon} & -\frac{h\sigma_2}{3\varepsilon} & -\frac{h\sigma_2}{6\varepsilon} & -\frac{1}{2h} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{h\sigma_1}{6\varepsilon} & \delta_1 + \frac{h\sigma_1}{3\varepsilon} & -\frac{h\sigma_2}{6\varepsilon} & -\frac{h\sigma_2}{3\varepsilon} & -\delta_1 + \frac{1}{h} & -\frac{1}{2h} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -\frac{h\sigma_1}{3\varepsilon} & -\frac{h\sigma_1}{6\varepsilon} & \delta_2 + \frac{h\sigma_2}{3\varepsilon} & \frac{h\sigma_2}{6\varepsilon} & 0 & 0 & -\frac{1}{2h} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -\frac{h\sigma_1}{6\varepsilon} & -\frac{h\sigma_1}{3\varepsilon} & \frac{h\sigma_2}{6\varepsilon} & \delta_2 + \frac{h\sigma_2}{3\varepsilon} & 0 & 0 & -\delta_2 + \frac{1}{h} & -\frac{1}{2h} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -\frac{1}{2h} & -\delta_1 + \frac{1}{h} & 0 & 0 & \delta_1 + \frac{h\sigma_1}{3\varepsilon} & \frac{h\sigma_1}{6\varepsilon} & -\frac{h\sigma_2}{3\varepsilon} & -\frac{h\sigma_2}{6\varepsilon} & -\frac{1}{2h} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -\frac{1}{2h} & 0 & 0 & \frac{h\sigma_1}{6\varepsilon} & \delta_1 + \frac{h\sigma_1}{3\varepsilon} & -\frac{h\sigma_2}{6\varepsilon} & -\frac{h\sigma_2}{3\varepsilon} & -\delta_1 + \frac{1}{h} & -\frac{1}{2h} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -\frac{1}{2h} & -\delta_2 + \frac{1}{h} & -\frac{h\sigma_1}{3\varepsilon} & -\frac{h\sigma_1}{6\varepsilon} & \delta_2 + \frac{h\sigma_2}{3\varepsilon} & \frac{h\sigma_2}{6\varepsilon} & 0 & 0 & -\frac{1}{2h} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -\frac{1}{2h} & -\frac{h\sigma_1}{6\varepsilon} & -\frac{h\sigma_1}{3\varepsilon} & \frac{h\sigma_2}{6\varepsilon} & \delta_2 + \frac{h\sigma_2}{3\varepsilon} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -\frac{1}{2h} & 0 & 0 & 0 & \delta_1 + \frac{h\sigma_1}{3\varepsilon} & \frac{h\sigma_1}{6\varepsilon} & -\frac{h\sigma_2}{3\varepsilon} & -\frac{h\sigma_2}{6\varepsilon} & -\frac{1}{2h} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -\frac{1}{2h} & -\delta_1 + \frac{1}{h} & 0 & 0 & \frac{h\sigma_1}{6\varepsilon} & \delta_1 + \frac{h\sigma_1}{3\varepsilon} & -\frac{h\sigma_2}{6\varepsilon} & -\frac{h\sigma_2}{3\varepsilon} & -\delta_1 + \frac{1}{h} & -\frac{1}{2h} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{2h} & -\delta_2 + \frac{1}{h} & -\frac{h\sigma_1}{3\varepsilon} & -\frac{h\sigma_1}{6\varepsilon} & \delta_2 + \frac{h\sigma_2}{3\varepsilon} & \frac{h\sigma_2}{6\varepsilon} & 0 & 0 & -\frac{1}{2h} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{2h} & -\delta_2 + \frac{1}{h} & -\frac{h\sigma_1}{6\varepsilon} & -\frac{h\sigma_1}{3\varepsilon} & \frac{h\sigma_2}{6\varepsilon} & \delta_2 + \frac{h\sigma_2}{3\varepsilon} & 0 & -\frac{1}{2h} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{2h} & -\frac{h\sigma_1}{3\varepsilon} & -\frac{h\sigma_1}{6\varepsilon} & -\frac{h\sigma_2}{3\varepsilon} & -\frac{h\sigma_2}{6\varepsilon} & -\frac{1}{2h} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{2h} & -\frac{h\sigma_1}{6\varepsilon} & -\frac{h\sigma_2}{6\varepsilon} & -\frac{h\sigma_2}{3\varepsilon} & 0 & -\frac{1}{2h} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{2h} & -\frac{h\sigma_1}{3\varepsilon} & -\frac{h\sigma_1}{6\varepsilon} & -\frac{h\sigma_2}{6\varepsilon} & -\frac{h\sigma_2}{3\varepsilon} \end{pmatrix} \begin{pmatrix} u^+_{1,j-2} \\ u_{1,j-1} \\ u^+_{2,j-2} \\ u_{2,j-1} \\ u^+_{1,j-1} \\ u_{1,j} \\ u^+_{2,j-1} \\ u_{2,j} \\ u^+_{1,j} \\ u_{1,j+1} \\ u^+_{2,j} \\ u_{2,j+1} \\ u^+_{1,j+1} \\ u_{1,j+2} \\ u^+_{2,j+1} \\ u_{2,j+2} \end{pmatrix}.$$

We reorder the matrix as follows

$$\begin{pmatrix} \delta_1 + \frac{h\sigma_1}{3\varepsilon} & -\frac{1}{2h} & 0 & 0 & -\frac{h\sigma_2}{3\varepsilon} & 0 & 0 & 0 & \frac{h\sigma_1}{6\varepsilon} & -\delta_1 + \frac{1}{h} & 0 & 0 & -\frac{h\sigma_2}{6\varepsilon} & 0 & 0 & 0 \\ -\frac{1}{2h} & \delta_1 + \frac{h\sigma_1}{3\varepsilon} & -\frac{1}{2h} & 0 & 0 & -\frac{h\sigma_2}{3\varepsilon} & 0 & 0 & \frac{h\sigma_1}{6\varepsilon} & -\delta_1 + \frac{1}{h} & 0 & 0 & -\frac{h\sigma_2}{6\varepsilon} & 0 & 0 & 0 \\ 0 & -\frac{1}{2h} & \delta_1 + \frac{h\sigma_1}{3\varepsilon} & -\frac{1}{2h} & 0 & 0 & -\frac{h\sigma_2}{3\varepsilon} & 0 & 0 & \frac{h\sigma_1}{6\varepsilon} & -\delta_1 + \frac{1}{h} & 0 & 0 & -\frac{h\sigma_2}{6\varepsilon} & 0 & 0 \\ 0 & 0 & -\frac{1}{2h} & \delta_1 + \frac{h\sigma_1}{3\varepsilon} & 0 & 0 & 0 & -\frac{h\sigma_2}{3\varepsilon} & 0 & 0 & \frac{h\sigma_1}{6\varepsilon} & -\delta_1 + \frac{1}{h} & 0 & 0 & -\frac{h\sigma_2}{6\varepsilon} & 0 \\ -\frac{h\sigma_2}{3\varepsilon} & 0 & 0 & 0 & \delta_2 + \frac{h\sigma_2}{3\varepsilon} & -\frac{1}{2h} & 0 & 0 & -\frac{h\sigma_1}{6\varepsilon} & 0 & 0 & 0 & \frac{h\sigma_2}{6\varepsilon} & -\delta_2 + \frac{1}{h} & 0 & 0 \\ 0 & -\frac{h\sigma_2}{3\varepsilon} & 0 & 0 & -\frac{1}{2h} & \delta_2 + \frac{h\sigma_2}{3\varepsilon} & -\frac{1}{2h} & 0 & 0 & -\frac{h\sigma_1}{6\varepsilon} & 0 & 0 & \frac{h\sigma_2}{6\varepsilon} & -\delta_2 + \frac{1}{h} & 0 & 0 \\ 0 & 0 & -\frac{h\sigma_2}{3\varepsilon} & 0 & 0 & -\frac{1}{2h} & \delta_2 + \frac{h\sigma_2}{3\varepsilon} & -\frac{1}{2h} & 0 & 0 & -\frac{h\sigma_1}{6\varepsilon} & 0 & 0 & \frac{h\sigma_2}{6\varepsilon} & -\delta_2 + \frac{1}{h} & 0 \\ 0 & 0 & 0 & -\frac{h\sigma_2}{3\varepsilon} & 0 & 0 & -\frac{1}{2h} & \delta_2 + \frac{h\sigma_2}{3\varepsilon} & 0 & 0 & 0 & -\frac{h\sigma_1}{6\varepsilon} & 0 & 0 & \frac{h\sigma_2}{6\varepsilon} & 0 \\ \frac{h\sigma_1}{6\varepsilon} & 0 & 0 & 0 & -\frac{h\sigma_2}{6\varepsilon} & 0 & 0 & 0 & \delta_1 + \frac{h\sigma_1}{3\varepsilon} & -\frac{1}{2h} & 0 & 0 & -\frac{h\sigma_2}{3\varepsilon} & 0 & 0 & 0 \\ -\delta_1 + \frac{1}{h} & \frac{h\sigma_1}{6\varepsilon} & 0 & 0 & 0 & -\frac{h\sigma_2}{6\varepsilon} & 0 & 0 & -\frac{1}{2h} & \delta_1 + \frac{h\sigma_1}{3\varepsilon} & 0 & 0 & 0 & -\frac{h\sigma_2}{3\varepsilon} & 0 & 0 \\ 0 & -\delta_1 + \frac{1}{h} & \frac{h\sigma_1}{6\varepsilon} & 0 & 0 & 0 & -\frac{h\sigma_2}{6\varepsilon} & 0 & 0 & -\frac{1}{2h} & \delta_1 + \frac{h\sigma_1}{3\varepsilon} & 0 & 0 & 0 & -\frac{h\sigma_2}{3\varepsilon} & 0 \\ 0 & 0 & -\delta_1 + \frac{1}{h} & \frac{h\sigma_1}{6\varepsilon} & 0 & 0 & 0 & -\frac{h\sigma_2}{6\varepsilon} & 0 & 0 & -\frac{1}{2h} & \delta_1 + \frac{h\sigma_1}{3\varepsilon} & 0 & 0 & 0 & -\frac{h\sigma_2}{3\varepsilon} \\ -\frac{h\sigma_1}{6\varepsilon} & 0 & 0 & 0 & \frac{h\sigma_2}{6\varepsilon} & 0 & 0 & 0 & -\frac{h\sigma_1}{3\varepsilon} & 0 & 0 & 0 & \delta_2 + \frac{h\sigma_2}{3\varepsilon} & -\frac{1}{2h} & 0 & 0 \\ 0 & -\frac{h\sigma_1}{6\varepsilon} & 0 & 0 & 0 & -\frac{h\sigma_2}{6\varepsilon} & 0 & 0 & 0 & -\frac{h\sigma_1}{3\varepsilon} & 0 & 0 & -\frac{1}{2h} & \delta_2 + \frac{h\sigma_2}{3\varepsilon} & -\frac{1}{2h} & 0 \\ 0 & 0 & -\frac{h\sigma_1}{6\varepsilon} & 0 & 0 & 0 & 0 & -\frac{h\sigma_2}{6\varepsilon} & 0 & 0 & -\frac{h\sigma_1}{3\varepsilon} & 0 & 0 & -\frac{1}{2h} & \delta_2 + \frac{h\sigma_2}{3\varepsilon} & -\frac{1}{2h} \\ 0 & 0 & 0 & -\frac{h\sigma_1}{6\varepsilon} & 0 & 0 & -\delta_2 + \frac{1}{h} & \frac{h\sigma_2}{6\varepsilon} & 0 & 0 & 0 & -\frac{h\sigma_1}{3\varepsilon} & 0 & 0 & -\frac{1}{2h} & \delta_2 + \frac{h\sigma_2}{3\varepsilon} \end{pmatrix} \begin{pmatrix} u^-_{1,j-1} \\ u^-_{1,j} \\ u^-_{1,j+1} \\ u^-_{1,j+2} \\ u^-_{2,j-1} \\ u^-_{2,j} \\ u^-_{2,j+1} \\ u^-_{2,j+2} \\ u^+_{1,j-1} \\ u^+_{1,j} \\ u^+_{1,j+1} \\ u^+_{1,j+2} \\ u^+_{2,j-1} \\ u^+_{2,j} \\ u^+_{2,j+1} \\ u^+_{2,j+2} \end{pmatrix}.$$

and extract the system of equations at node j, shown below

$$\widehat{A}_j = \begin{pmatrix} \delta_1 - \frac{\cos(h\omega)}{h} + \frac{h\sigma_1}{3\varepsilon} & -\frac{h\sigma_2}{3\varepsilon} & \left(-\delta_1 + \frac{1}{h}\right) e^{ih\omega} + \frac{h\sigma_1}{6\varepsilon} & -\frac{h\sigma_2}{6\varepsilon} \\ -\frac{h\sigma_1}{3\varepsilon} & \delta_2 - \frac{\cos(h\omega)}{h} + \frac{h\sigma_2}{3\varepsilon} & -\frac{h\sigma_1}{6\varepsilon} & \left(-\delta_2 + \frac{1}{h}\right) e^{ih\omega} + \frac{h\sigma_2}{6\varepsilon} \\ \left(-\delta_1 + \frac{1}{h}\right) e^{-ih\omega} + \frac{h\sigma_1}{6\varepsilon} & -\frac{h\sigma_2}{6\varepsilon} & \delta_1 - \frac{\cos(h\omega)}{h} + \frac{h\sigma_1}{3\varepsilon} & -\frac{h\sigma_2}{3\varepsilon} \\ -\frac{h\sigma_1}{6\varepsilon} & \left(-\delta_2 + \frac{1}{h}\right) e^{-ih\omega} + \frac{h\sigma_2}{6\varepsilon} & -\frac{h\sigma_1}{3\varepsilon} & \delta_2 - \frac{\cos(h\omega)}{h} + \frac{h\sigma_2}{3\varepsilon} \end{pmatrix} \begin{pmatrix} C^-_{\omega,1} \\ C^-_{\omega,2} \\ C^+_{\omega,1} \\ C^+_{\omega,2} \end{pmatrix}.$$

We proceed now to define the diagonal, lower and upper triangular block matrices as follows

$$\widehat{D}_j = \begin{pmatrix} \delta_1 + \frac{h\sigma_1}{3\varepsilon} & -\frac{h\sigma_2}{3\varepsilon} & \frac{h\sigma_1}{6\varepsilon} & -\frac{h\sigma_2}{6\varepsilon} \\ -\frac{h\sigma_1}{3\varepsilon} & \delta_2 + \frac{h\sigma_2}{3\varepsilon} & -\frac{h\sigma_1}{6\varepsilon} & \frac{h\sigma_2}{6\varepsilon} \\ \frac{h\sigma_1}{6\varepsilon} & -\frac{h\sigma_2}{6\varepsilon} & \delta_1 + \frac{h\sigma_1}{3\varepsilon} & -\frac{h\sigma_2}{3\varepsilon} \\ -\frac{h\sigma_1}{6\varepsilon} & \frac{h\sigma_2}{6\varepsilon} & -\frac{h\sigma_1}{3\varepsilon} & \delta_2 + \frac{h\sigma_2}{3\varepsilon} \end{pmatrix},$$

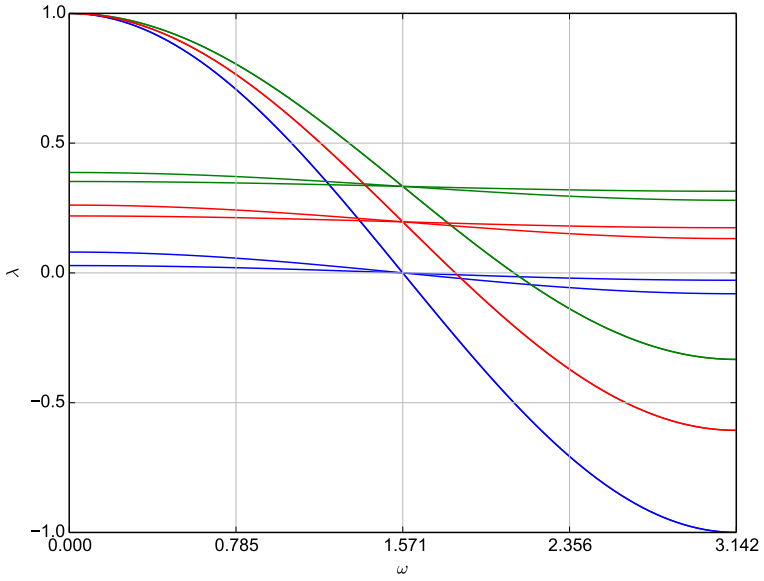


FIGURE 3.12. Eigenvalues with the maximum absolute value for $\alpha = 1$ (blue), $\alpha = 2/3$ (green) and $\alpha = \frac{2\varepsilon^2 + \frac{4\varepsilon h^2}{3} + \frac{h^4}{6}}{3\varepsilon^2 + \frac{11\varepsilon h^2}{6} + \frac{h^4}{6}}$ (red). Here plotted for $\sigma_1 = 1$, $\sigma_2 = 10$, $\varepsilon = 0.01$ and $h = 0.25$.

$$\widehat{L}_j = \begin{pmatrix} -\frac{e^{-ih\omega}}{2h} & 0 & 0 & 0 \\ 0 & -\frac{e^{-ih\omega}}{2h} & 0 & 0 \\ \left(-\delta_1 + \frac{1}{h}\right) e^{-ih\omega} & 0 & -\frac{e^{-ih\omega}}{2h} & 0 \\ 0 & \left(-\delta_2 + \frac{1}{h}\right) e^{-ih\omega} & 0 & -\frac{e^{-ih\omega}}{2h} \end{pmatrix},$$

$$\widehat{U}_j = \begin{pmatrix} -\frac{e^{ih\omega}}{2h} & 0 & \left(-\delta_1 + \frac{1}{h}\right) e^{ih\omega} & 0 \\ 0 & -\frac{e^{ih\omega}}{2h} & 0 & \left(-\delta_2 + \frac{1}{h}\right) e^{ih\omega} \\ 0 & 0 & -\frac{e^{ih\omega}}{2h} & 0 \\ 0 & 0 & 0 & -\frac{e^{ih\omega}}{2h} \end{pmatrix}.$$

With this information and in order to obtain an explicit expression of the eigenvalues, we must obtain the inverse of $\widehat{D}_j$ and multiply by $\widehat{L}_j + \widehat{U}_j$, the expressions are fairly complicated we were not able to find a closed form of the eigenvalues for the general case. We could however obtain some results by making proper assumptions, we show those results later on.

We provide the shape of the relevant matrices in appendix B.

5.1. Numerical evaluation of eigenvalues. Setting $\delta_1 = \delta_2 = 1/h$ we show the eigenvalues $\lambda(\omega)$ for the relaxation parameters analyzed in the previous section.

Two of the eigenvalues correspond to the pure laplacian mode, so we see three different lines: pure laplacian mode, and one scalar reaction diffusion curve.

We see that the error reduction is dominated by the behaviour of the elements in the kernel of the scattering operator, which are solutions to Poisson's equation.

In this case, a value of α adapted to the scalar reaction diffusion equation worsens the result for the pure laplacian mode.

The result indicates that the performance of Richardson's iteration is dominated by the Laplacian mode and the choice of $\delta = 1/h$ and $\alpha = 2/3$ is optimal.

Even though we cannot do better for the laplacian mode, we can do better for each of the other modes, by addressing them independently as we do in the following section.

6. Optimal stabilisation

Observing the results obtained in the previous section, we analyzed the usage of a stabilization matrix instead of a stabilization constant in order to address the stabilization of different modes independently.

We modify the previously introduced 2-group interior penalty DG reaction diffusion equation by considering a stabilisation operator as follows

$$\begin{aligned}
\mathcal{A}_1(u_1, v_1) := &\int_{\mathbb{T}} \nabla u_1 \nabla v_1 dx + \frac{1}{\varepsilon}\int_{\mathbb{T}} \sigma_1 u_1 v_1 dx - \frac{1}{\varepsilon}\int_{\mathbb{T}} \sigma_2 u_2 v_2 dx \\
&- \int_{\mathbb{F}^I \cup \mathbb{F}^B} \left([\![u_1]\!] \{\!\{\nabla v_1\}\!\} ds + \{\!\{\nabla u_1\}\!\} [\![v_1]\!] \right) ds \\
&+ \int_{\mathbb{F}^I \cup \mathbb{F}^B} \delta_{11} [\![u_1]\!][\![v_1]\!] ds + \int_{\mathbb{F}^I \cup \mathbb{F}^B} \delta_{21} [\![u_1]\!][\![v_2]\!] ds = \int_{\mathbb{T}} f v dx, \\
\mathcal{A}_2(u_2, v_2) := &\int_{\mathbb{T}} \nabla u_2 \nabla v_2 dx + \frac{1}{\varepsilon}\int_{\mathbb{T}} \sigma_2 u_2 v_2 dx - \frac{1}{\varepsilon}\int_{\mathbb{T}} \sigma_1 u_1 v_1 dx \\
&- \int_{\mathbb{F}^I \cup \mathbb{F}^B} \left([\![u_2]\!] \{\!\{\nabla v_2\}\!\} ds + \{\!\{\nabla u_2\}\!\} [\![v_2]\!] \right) ds \\
&+ \int_{\mathbb{F}^I \cup \mathbb{F}^B} \delta_{12} [\![u_2]\!][\![v_1]\!] ds + \int_{\mathbb{F}^I \cup \mathbb{F}^B} \delta_{22} [\![u_2]\!][\![v_2]\!] ds = \int_{\mathbb{T}} f v dx,
\end{aligned}$$

in vector form this expression is

$$\begin{aligned}
\mathcal{A}(u, v) := &\int_{\mathbb{T}} \nabla u \nabla v dx + \frac{1}{\varepsilon}\int_{\mathbb{T}} \boldsymbol{\sigma} u v dx - \int_{\mathbb{F}^I \cup \mathbb{F}^B} \left([\![u]\!] \{\!\{\nabla v\}\!\} ds + \{\!\{\nabla u\}\!\} [\![v]\!] \right) ds \\
&+ \int_{\mathbb{F}^I \cup \mathbb{F}^B} \boldsymbol{\delta} [\![u]\!][\![v]\!] ds = \int_{\mathbb{T}} f v dx,
\end{aligned}$$

where both $\boldsymbol{\sigma}$ and $\boldsymbol{\delta}$ are matrices.

We proceed in the same way we did previously, leading to the following cell and face matrices

$$C = \begin{pmatrix} \delta_{11} + \frac{h\sigma_1}{3\varepsilon} & \frac{h\sigma_1}{6\varepsilon} & \delta_{21} - \frac{h\sigma_2}{3\varepsilon} & -\frac{h\sigma_2}{6\varepsilon} \\ \frac{h\sigma_1}{6\varepsilon} & \delta_{11} + \frac{h\sigma_1}{3\varepsilon} & -\frac{h\sigma_2}{6\varepsilon} & \delta_{21} - \frac{h\sigma_2}{3\varepsilon} \\ \delta_{12} - \frac{h\sigma_1}{3\varepsilon} & -\frac{h\sigma_1}{6\varepsilon} & \delta_{22} + \frac{h\sigma_2}{3\varepsilon} & \frac{h\sigma_2}{6\varepsilon} \\ -\frac{h\sigma_1}{6\varepsilon} & \delta_{12} - \frac{h\sigma_1}{3\varepsilon} & \frac{h\sigma_2}{6\varepsilon} & \delta_{22} + \frac{h\sigma_2}{3\varepsilon} \end{pmatrix},$$

$$F_u = \begin{pmatrix} -\frac{1}{2h} & -\delta_{11} + \frac{1}{h} & 0 & -\delta_{21} \\ 0 & -\frac{1}{2h} & 0 & 0 \\ 0 & -\delta_{12} & -\frac{1}{2h} & -\delta_{22} + \frac{1}{h} \\ 0 & 0 & 0 & -\frac{1}{2h} \end{pmatrix},$$

$$F_l = \begin{pmatrix} -\frac{1}{2h} & 0 & 0 & 0 \\ -\delta_{11} + \frac{1}{h} & -\frac{1}{2h} & -\delta_{21} & 0 \\ 0 & 0 & -\frac{1}{2h} & 0 \\ -\delta_{12} & 0 & -\delta_{22} + \frac{1}{h} & -\frac{1}{2h} \end{pmatrix},$$

after assembly we get the following Toeplitz matrix and linear system

$$\begin{pmatrix} \delta_{11} + \frac{h\sigma_1}{3\varepsilon} & \frac{h\sigma_1}{6\varepsilon} & \delta_{21} - \frac{h\sigma_2}{3\varepsilon} & -\frac{h\sigma_2}{6\varepsilon} & -\frac{1}{2h} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{h\sigma_1}{6\varepsilon} & \delta_{11} + \frac{h\sigma_1}{3\varepsilon} & -\frac{h\sigma_2}{6\varepsilon} & \delta_{21} - \frac{h\sigma_2}{3\varepsilon} & -\delta_{11} + \frac{1}{h} & -\frac{1}{2h} & -\delta_{21} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \delta_{12} - \frac{h\sigma_1}{3\varepsilon} & -\frac{h\sigma_1}{6\varepsilon} & \delta_{22} + \frac{h\sigma_2}{3\varepsilon} & \frac{h\sigma_2}{6\varepsilon} & 0 & 0 & -\frac{1}{2h} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -\frac{h\sigma_1}{6\varepsilon} & \delta_{12} - \frac{h\sigma_1}{3\varepsilon} & \frac{h\sigma_2}{6\varepsilon} & \delta_{22} + \frac{h\sigma_2}{3\varepsilon} & -\delta_{12} & 0 & -\delta_{22} + \frac{1}{h} & -\frac{1}{2h} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -\frac{1}{2h} & -\delta_{11} + \frac{1}{h} & 0 & -\delta_{21} & \delta_{11} + \frac{h\sigma_1}{3\varepsilon} & \frac{h\sigma_1}{6\varepsilon} & \delta_{21} - \frac{h\sigma_2}{3\varepsilon} & -\frac{h\sigma_2}{6\varepsilon} & -\frac{1}{2h} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -\frac{1}{2h} & 0 & 0 & \frac{h\sigma_1}{6\varepsilon} & \delta_{11} + \frac{h\sigma_1}{3\varepsilon} & -\frac{h\sigma_2}{6\varepsilon} & \delta_{21} - \frac{h\sigma_2}{3\varepsilon} & -\delta_{11} + \frac{1}{h} & -\frac{1}{2h} & -\delta_{21} & 0 & 0 & 0 & 0 & 0 \\ 0 & -\delta_{12} & -\frac{1}{2h} & -\delta_{22} + \frac{1}{h} & \delta_{12} - \frac{h\sigma_1}{3\varepsilon} & -\frac{h\sigma_1}{6\varepsilon} & \delta_{22} + \frac{h\sigma_2}{3\varepsilon} & \frac{h\sigma_2}{6\varepsilon} & 0 & 0 & -\frac{1}{2h} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -\frac{1}{2h} & -\frac{h\sigma_1}{6\varepsilon} & \delta_{12} - \frac{h\sigma_1}{3\varepsilon} & \frac{h\sigma_2}{6\varepsilon} & \delta_{22} + \frac{h\sigma_2}{3\varepsilon} & -\delta_{12} & 0 & -\delta_{22} + \frac{1}{h} & -\frac{1}{2h} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -\frac{1}{2h} & -\frac{h\sigma_1}{6\varepsilon} & 0 & 0 & -\delta_{21} & \delta_{11} + \frac{h\sigma_1}{3\varepsilon} & \frac{h\sigma_1}{6\varepsilon} & -\frac{h\sigma_1}{6\varepsilon} & -\frac{1}{2h} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -\frac{1}{2h} & 0 & 0 & 0 & \frac{h\sigma_1}{6\varepsilon} & \delta_{11} + \frac{h\sigma_1}{3\varepsilon} & -\frac{h\sigma_1}{6\varepsilon} & \delta_{21} - \frac{h\sigma_2}{3\varepsilon} & -\delta_{11} + \frac{1}{h} & -\frac{1}{2h} & -\delta_{21} \\ 0 & 0 & 0 & 0 & 0 & -\delta_{12} & -\frac{1}{2h} & -\delta_{22} + \frac{1}{h} & 0 & -\frac{h\sigma_1}{6\varepsilon} & \frac{h\sigma_1}{6\varepsilon} & \delta_{22} + \frac{h\sigma_2}{3\varepsilon} & 0 & 0 & -\frac{1}{2h} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{2h} & 0 & -\frac{h\sigma_1}{6\varepsilon} & \delta_{12} - \frac{h\sigma_1}{3\varepsilon} & \frac{h\sigma_2}{6\varepsilon} & -\delta_{12} & 0 & -\delta_{22} + \frac{1}{h} & -\frac{1}{2h} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{2h} & 0 & 0 & -\frac{h\sigma_1}{6\varepsilon} & 0 & -\delta_{11} + \frac{1}{h} & -\delta_{21} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{2h} & 0 & 0 & \frac{h\sigma_1}{6\varepsilon} & \delta_{11} + \frac{h\sigma_1}{3\varepsilon} & -\frac{h\sigma_2}{6\varepsilon} & \delta_{21} - \frac{h\sigma_2}{3\varepsilon} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{2h} & -\frac{h\sigma_1}{6\varepsilon} & -\delta_{12} & -\frac{1}{2h} & -\delta_{22} + \frac{1}{h} & \delta_{12} - \frac{h\sigma_1}{3\varepsilon} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{2h} & 0 & 0 & -\frac{h\sigma_1}{6\varepsilon} & \delta_{12} - \frac{h\sigma_1}{3\varepsilon} \end{pmatrix} \begin{pmatrix} u^+_{i,j-2} \\ u^-_{i,j-1} \\ u^-_{i,j-2} \\ u^+_{2,j-1} \\ u^+_{i,j-1} \\ u^-_{i,j} \\ u^+_{2,j-1} \\ u^-_{2,j} \\ u^+_{i,j} \\ u^-_{i,j+1} \\ u^-_{2,j} \\ u^+_{2,j+1} \\ u^+_{i,j+1} \\ u^-_{i,j+2} \\ u^+_{2,j+1} \\ u^+_{2,j+2} \end{pmatrix},$$

reordering we get

$$\begin{pmatrix} \delta_{11} + \frac{h\sigma_1}{3\varepsilon} & -\frac{1}{2h} & 0 & 0 & \delta_{21} - \frac{h\sigma_2}{3\varepsilon} & 0 & 0 & 0 & \frac{h\sigma_1}{6\varepsilon} & -\delta_{11} + \frac{1}{h} & 0 & 0 & -\frac{h\sigma_2}{6\varepsilon} & -\delta_{21} & 0 & 0 \\ -\frac{1}{2h} & \delta_{11} + \frac{h\sigma_1}{3\varepsilon} & -\frac{1}{2h} & 0 & 0 & \delta_{21} - \frac{h\sigma_2}{3\varepsilon} & 0 & 0 & 0 & \frac{h\sigma_1}{6\varepsilon} & -\delta_{11} + \frac{1}{h} & 0 & 0 & -\frac{h\sigma_2}{6\varepsilon} & -\delta_{21} & 0 \\ 0 & -\frac{1}{2h} & \delta_{11} + \frac{h\sigma_1}{3\varepsilon} & -\frac{1}{2h} & 0 & 0 & \delta_{21} - \frac{h\sigma_2}{3\varepsilon} & 0 & 0 & 0 & \frac{h\sigma_1}{6\varepsilon} & -\delta_{11} + \frac{1}{h} & 0 & 0 & -\frac{h\sigma_2}{6\varepsilon} & -\delta_{21} \\ 0 & 0 & -\frac{1}{2h} & \delta_{11} + \frac{h\sigma_1}{3\varepsilon} & 0 & 0 & 0 & \delta_{21} - \frac{h\sigma_2}{3\varepsilon} & 0 & 0 & 0 & \frac{h\sigma_1}{6\varepsilon} & 0 & 0 & 0 & -\frac{h\sigma_2}{6\varepsilon} \\ \delta_{12} - \frac{h\sigma_1}{3\varepsilon} & 0 & 0 & 0 & \delta_{22} + \frac{h\sigma_2}{3\varepsilon} & -\frac{1}{2h} & 0 & 0 & -\frac{h\sigma_1}{6\varepsilon} & -\delta_{12} & 0 & 0 & \frac{h\sigma_2}{6\varepsilon} & -\delta_{22} + \frac{1}{h} & 0 & 0 \\ 0 & \delta_{12} - \frac{h\sigma_1}{3\varepsilon} & 0 & 0 & -\frac{1}{2h} & \delta_{22} + \frac{h\sigma_2}{3\varepsilon} & -\frac{1}{2h} & 0 & 0 & -\frac{h\sigma_1}{6\varepsilon} & -\delta_{12} & 0 & 0 & \frac{h\sigma_2}{6\varepsilon} & -\delta_{22} + \frac{1}{h} & 0 \\ 0 & 0 & \delta_{12} - \frac{h\sigma_1}{3\varepsilon} & 0 & 0 & -\frac{1}{2h} & \delta_{22} + \frac{h\sigma_2}{3\varepsilon} & -\frac{1}{2h} & 0 & 0 & -\frac{h\sigma_1}{6\varepsilon} & -\delta_{12} & 0 & 0 & \frac{h\sigma_2}{6\varepsilon} & -\delta_{22} + \frac{1}{h} \\ 0 & 0 & 0 & \delta_{12} - \frac{h\sigma_1}{3\varepsilon} & 0 & 0 & -\frac{1}{2h} & \delta_{22} + \frac{h\sigma_2}{3\varepsilon} & 0 & 0 & 0 & -\frac{h\sigma_1}{6\varepsilon} & 0 & 0 & 0 & \frac{h\sigma_2}{6\varepsilon} \\ \frac{h\sigma_1}{6\varepsilon} & 0 & 0 & 0 & -\frac{h\sigma_2}{6\varepsilon} & 0 & 0 & 0 & \delta_{11} + \frac{h\sigma_1}{3\varepsilon} & -\frac{1}{2h} & 0 & 0 & \delta_{21} - \frac{h\sigma_2}{3\varepsilon} & 0 & 0 & 0 \\ -\delta_{11} + \frac{1}{h} & \frac{h\sigma_1}{6\varepsilon} & 0 & 0 & -\delta_{21} & -\frac{h\sigma_2}{6\varepsilon} & 0 & 0 & -\frac{1}{2h} & \delta_{11} + \frac{h\sigma_1}{3\varepsilon} & -\frac{1}{2h} & 0 & \delta_{21} - \frac{h\sigma_2}{3\varepsilon} & 0 & 0 & 0 \\ 0 & -\delta_{11} + \frac{1}{h} & \frac{h\sigma_1}{6\varepsilon} & 0 & 0 & -\delta_{21} & -\frac{h\sigma_2}{6\varepsilon} & 0 & 0 & -\frac{1}{2h} & \delta_{11} + \frac{h\sigma_1}{3\varepsilon} & -\frac{1}{2h} & 0 & \delta_{21} - \frac{h\sigma_2}{3\varepsilon} & 0 & 0 \\ 0 & 0 & -\delta_{11} + \frac{1}{h} & \frac{h\sigma_1}{6\varepsilon} & 0 & -\delta_{21} & -\frac{h\sigma_2}{6\varepsilon} & 0 & 0 & 0 & -\frac{1}{2h} & \delta_{11} + \frac{h\sigma_1}{3\varepsilon} & 0 & 0 & 0 & \delta_{21} - \frac{h\sigma_2}{3\varepsilon} \\ -\frac{h\sigma_1}{6\varepsilon} & 0 & 0 & 0 & \frac{h\sigma_1}{6\varepsilon} & 0 & 0 & 0 & \delta_{12} - \frac{h\sigma_1}{3\varepsilon} & 0 & 0 & 0 & \delta_{22} + \frac{h\sigma_2}{3\varepsilon} & -\frac{1}{2h} & 0 & 0 \\ -\delta_{12} & -\frac{h\sigma_1}{6\varepsilon} & 0 & 0 & -\delta_{22} + \frac{1}{h} & \frac{h\sigma_1}{6\varepsilon} & 0 & 0 & 0 & \delta_{12} - \frac{h\sigma_1}{3\varepsilon} & 0 & 0 & -\frac{1}{2h} & \delta_{22} + \frac{h\sigma_2}{3\varepsilon} & -\frac{1}{2h} & 0 \\ 0 & -\delta_{12} & -\frac{h\sigma_1}{6\varepsilon} & 0 & 0 & -\delta_{22} + \frac{1}{h} & \frac{h\sigma_1}{6\varepsilon} & 0 & 0 & 0 & \delta_{12} - \frac{h\sigma_1}{3\varepsilon} & 0 & 0 & -\frac{1}{2h} & \delta_{22} + \frac{h\sigma_2}{3\varepsilon} & -\frac{1}{2h} \\ 0 & 0 & -\delta_{12} & -\frac{h\sigma_1}{6\varepsilon} & 0 & 0 & -\delta_{22} + \frac{1}{h} & \frac{h\sigma_1}{6\varepsilon} & 0 & 0 & 0 & \delta_{12} - \frac{h\sigma_1}{3\varepsilon} & 0 & 0 & -\frac{1}{2h} & \delta_{22} + \frac{h\sigma_2}{3\varepsilon} \end{pmatrix} \begin{pmatrix} u^-_{i,j-1} \\ u^+_{i,j} \\ u^-_{i,j+1} \\ u^+_{i,j+2} \\ u^-_{2,j-1} \\ u^+_{2,j} \\ u^-_{2,j+1} \\ u^+_{i,j+1} \\ u^-_{i,j+2} \\ u^+_{2,j+2} \\ u^-_{i,j-2} \\ u^+_{i,j-1} \\ u^-_{i,j} \\ u^+_{i,j+1} \\ u^-_{2,j-2} \\ u^+_{2,j-1} \\ u^-_{2,j} \\ u^+_{2,j+1} \end{pmatrix},$$

We then proceed to extract the system of equations at node j

$$\widehat{A}_j = \begin{pmatrix} \delta_{11} - \frac{\cos(h\omega)}{h} + \frac{hs_1}{3\varepsilon} & \delta_{21} - \frac{hs_2}{3\varepsilon} & \left(-\delta_{11} + \frac{1}{h}\right) e^{\imath h\omega} + \frac{hs_1}{6\varepsilon} & -\delta_{21} e^{\imath h\omega} - \frac{hs_2}{6\varepsilon} \\ \delta_{12} - \frac{hs_1}{3\varepsilon} & \delta_{22} - \frac{\cos(hw)}{h} + \frac{hs_2}{3\varepsilon} & -\delta_{12} e^{\imath h\omega} - \frac{hs_1}{6\varepsilon} & \left(-\delta_{22} + \frac{1}{h}\right) e^{\imath h\omega} + \frac{hs_2}{6\varepsilon} \\ \frac{-\delta_{11} + \frac{1}{h}}{e^{\imath h\omega}} + \frac{hs_1}{6\varepsilon} & -\frac{\delta_{21}}{e^{\imath h\omega}} - \frac{hs_2}{6\varepsilon} & \delta_{11} - \frac{\cos(hw)}{h} + \frac{hs_1}{3\varepsilon} & \delta_{21} - \frac{hs_2}{3\varepsilon} \\ -\frac{\delta_{12}}{e^{\imath h\omega}} - \frac{hs_1}{6\varepsilon} & \frac{-\delta_{22} + \frac{1}{h}}{e^{\imath h\omega}} + \frac{hs_2}{6\varepsilon} & \delta_{12} - \frac{hs_1}{3\varepsilon} & \delta_{22} - \frac{\cos(hw)}{h} + \frac{hs_2}{3\varepsilon} \end{pmatrix} \begin{pmatrix} C^-_{\omega,1} \\ C^-_{\omega,2} \\ C^+_{\omega,1} \\ C^+_{\omega,2} \end{pmatrix},$$

we proceed in the same fashion to obtain

$$\widehat{D}_j = \begin{pmatrix} \delta_{11} + \frac{h\sigma_1}{3\varepsilon} & \delta_{21} - \frac{h\sigma_2}{3\varepsilon} & \frac{h\sigma_1}{6\varepsilon} & -\frac{h\sigma_2}{6\varepsilon} \\ \delta_{12} - \frac{h\sigma_1}{3\varepsilon} & \delta_{22} + \frac{h\sigma_2}{3\varepsilon} & -\frac{h\sigma_1}{6\varepsilon} & \frac{h\sigma_2}{6\varepsilon} \\ \frac{h\sigma_1}{6\varepsilon} & -\frac{h\sigma_2}{6\varepsilon} & \delta_{11} + \frac{h\sigma_1}{3\varepsilon} & \delta_{21} - \frac{h\sigma_2}{3\varepsilon} \\ -\frac{h\sigma_1}{6\varepsilon} & \frac{h\sigma_2}{6\varepsilon} & \delta_{12} - \frac{h\sigma_1}{3\varepsilon} & \delta_{22} + \frac{h\sigma_2}{3\varepsilon}, \end{pmatrix}$$

$$\widehat{L}_j = \begin{pmatrix} -\frac{1}{2he^{\imath h\omega}} & 0 & 0 & 0 \\ 0 & -\frac{1}{2he^{\imath h\omega}} & 0 & 0 \\ \frac{-\delta_{11}+\frac{1}{h}}{e^{\imath h\omega}} & -\frac{\delta_{21}}{e^{\imath h\omega}} & -\frac{1}{2he^{\imath h\omega}} & 0 \\ -\frac{\delta_{12}}{e^{\imath h\omega}} & \frac{-\delta_{22}+\frac{1}{h}}{e^{\imath h\omega}} & 0 & -\frac{1}{2he^{\imath h\omega}} \end{pmatrix},$$

$$\widehat{U}_j = \begin{pmatrix} -\frac{e^{\imath h\omega}}{2h} & 0 & \left(-\delta_{11}+\frac{1}{h}\right)e^{\imath h\omega} & -\delta_{21}e^{\imath h\omega} \\ 0 & -\frac{e^{\imath h\omega}}{2h} & -\delta_{12}e^{\imath h\omega} & \left(-\delta_{22}+\frac{1}{h}\right)e^{\imath h\omega} \\ 0 & 0 & -\frac{e^{\imath h\omega}}{2h} & 0 \\ 0 & 0 & 0 & -\frac{e^{\imath h\omega}}{2h} \end{pmatrix}.$$

We would like to use our definitions of δ_{opt} and α_{opt} to obtain λ_{opt} in each group.

We will note the scalar reaction-diffusion optimized relaxation and stabilization with non-bold α_{opt} and δ_{opt} as opposed to the vector reaction-diffusion optimized relaxation and stabilization *matrices* $\boldsymbol{\alpha}_{\text{opt}}$ and $\boldsymbol{\delta}_{\text{opt}}$

We remark, as it was explained before, that the mode having the least reaction will be the bottleneck of Richardson's iteration and we know that we cannot do better than reducing the high frequency error to 1/3 in each iteration when the pure laplacian mode is present.

The idea of this section is to optimize the error reduction in all other modes as well as the one with the least reaction.

In the scalar case we have $\delta = \delta(\varepsilon, h)$ so there is a dependence between the stabilisation parameter and the magnitude of the reaction term.

We deduced an optimal value δ_{opt} for the scalar case and we would like to use those values in the vector case by proceeding as follows:

(1) Obtain the singular value decomposition (SVD) of $\boldsymbol{\sigma}$ into $\boldsymbol{\sigma} = USV^{\dagger}$.
(2) Deduce the operator $\boldsymbol{\delta}_{\text{opt}}$ by changing basis to the column basis of $\boldsymbol{\sigma}$ using the left matrix of the SVD, U as follows

$$\boldsymbol{\delta}_{\text{opt}} = U \begin{pmatrix} \delta_{\text{opt}}\left(\varepsilon/S_{11}, h\right) & 0 \\ 0 & \delta_{\text{opt}}\left(\varepsilon/S_{22}, h\right) \end{pmatrix} U^{\dagger}.$$

(3) Deduce the operator $\boldsymbol{\alpha}_{\text{opt}}$ also by changing to the column basis, addressing each group with its optimal relaxation

$$\boldsymbol{\alpha}_{\text{opt}} = \begin{pmatrix} U \begin{pmatrix} \alpha_{\text{opt}}\left(\varepsilon/S_{11}, h\right) & 0 \\ 0 & \alpha_{\text{opt}}\left(\varepsilon/S_{22}, h\right) \end{pmatrix} U^{\dagger} & \begin{matrix} 0 & 0 \\ 0 & 0 \end{matrix} \\ \begin{matrix} 0 & 0 \\ 0 & 0 \end{matrix} & U \begin{pmatrix} \alpha_{\text{opt}}\left(\varepsilon/S_{11}, h\right) & 0 \\ 0 & \alpha_{\text{opt}}\left(\varepsilon/S_{22}, h\right) \end{pmatrix} U^{\dagger} \end{pmatrix}.$$

We continue with a symmetric and a nonsymmetric example of the above methodology.

6.1. Symmetric $\boldsymbol{\sigma}$. The eigenvalues of $I - \boldsymbol{\alpha}_{\text{opt}} \widehat{D}\left(\boldsymbol{\delta}_{\text{opt}}\right)_j^{-1} \widehat{A}\left(\boldsymbol{\delta}_{\text{opt}}\right)_j$ are shown in figure 3.13 for $\sigma_1 = \sigma_2 = 1$, $\varepsilon = 0.04$ and $h = 0.25$.

We deduce $S_{11} = 2$ and $S_{22} = 0$ which means the S_{22} mode is the pure laplacian with $\alpha_{\text{opt}}\left(\varepsilon/S_{22}, h\right) = 2/3$.

The solution obtained by inverting $A(\boldsymbol{\delta}_{\text{opt}})$ with a unitary right hand side is shown in figure 3.14. We observe the solution corresponding to a finite volume method and it is equal for both groups because of the definition of $\boldsymbol{\sigma}$, its symmetry

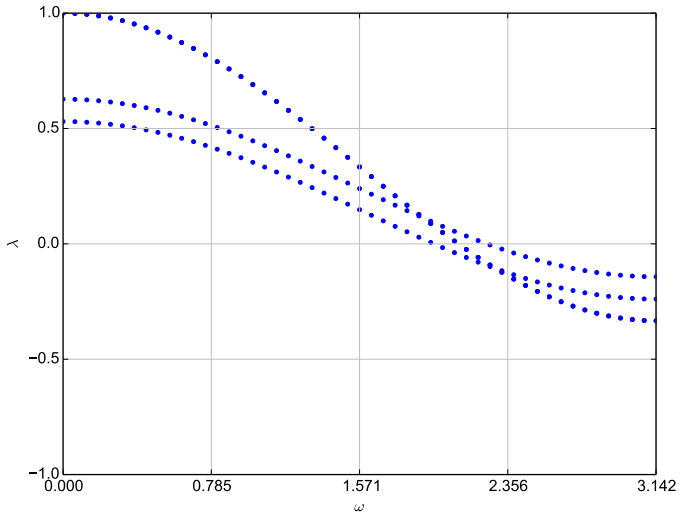


FIGURE 3.13. Eigenvalues of $I - \boldsymbol{\alpha}_{\text{opt}} \widehat{D} \left(\boldsymbol{\delta}_{\text{opt}}\right)_j^{-1} \widehat{A} \left(\boldsymbol{\delta}_{\text{opt}}\right)_j$.

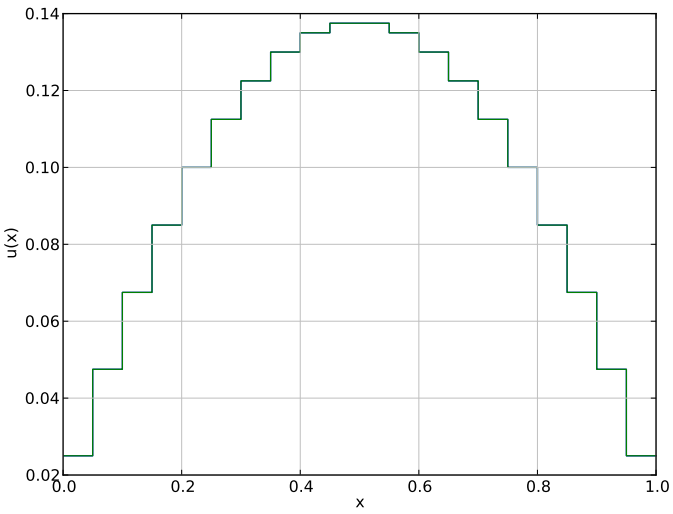


FIGURE 3.14. Solution of the symmetric 2 group problem using $\boldsymbol{\delta}_{\text{opt}}$.

implies that every rate of reaction from u_1 to u_2 is compensated with an equivalent rate from u_2 to u_1.

6.2. Non-symmetric $\boldsymbol{\sigma}$. In the non-symmetric case the the error matrix can have complex eigenvalues, in the 2-group case however, given the zero-column sum

structure we have one eigenvalue equal to zero, which forces the other to be real given that eigenvalues come in complex conjugate pairs.

We show the eigenvalues in 3D plots in order to show their imaginary part and its field of values, though the spectrum has no imaginary part, for different choices of $\boldsymbol{\sigma}$ and $\boldsymbol{\delta}$ in 2 groups.

We show the behaviour of the eigenvalues and field of values of

$$I - \widehat{D}\left(\delta = 1/h\right)_j^{-1} \widehat{A}\left(\delta = 1/h\right)_j,$$

for the unoptimized case in figure 3.15, for the optimization $\boldsymbol{\delta}_{\text{opt}}$ which aligns the eigenvalues of the non-zero reaction mode and for the optimization of both $\boldsymbol{\delta}_{\text{opt}}$ and $\boldsymbol{\alpha}_{\text{opt}}$ which reduces the field of values and the spectrum in the region $\omega \in [\pi/2, \pi]$.

In figure 3.18 we show the shape of the solution for 2 groups where, again, corresponds to a finite volume method.

We observe that we cannot make any statement just based on the spectrum in the non-symmetric case. Even when the spectrum is restricted to values between -1 and 1 the field of values is larger and the operator can indeed enlarge the error.

We perform a proper field of values analysis in the next section.

7. Field of values analysis

In this section we will use some terminology associated to the field of values that we describe hereafter.

Definition 7.1. The *numerical range* or *field of values* of a complex $n \times n$ matrix A is the set

$$W(A) = \left\{ \frac{v^* A v}{v^* v} : v \in \mathbb{C}^n, v \neq 0 \right\},$$

where v^* denotes the conjugate transpose of the vector v.

Additional definitions characterize $W(A)$:

(1) the *numerical radius*

$$\mu(A) = \sup_{z \in W(A)} |z| \leq \|A\| \leq 2\mu(A),$$

where $\|\cdot\|$ is the operator norm,

(2) the real part of $W(A)$

$$W^{\mathbb{R}}(A) = \Re\left(W(A)\right) = W(A^H) \subseteq W(A),$$

(3) and the least and largest eigenvalues of $W^{\mathbb{R}}(A)$

$$\lambda_{\min}\left(A^H\right) \leq W^{\mathbb{R}}(A) \leq \lambda_{\max}\left(A^H\right),$$

where if $A \in \mathbb{R}^{n \times n}$, $\nu\left(A\right) = \lambda_{\min}\left(A^H\right)$ and $\mu\left(A\right) = \lambda_{\max}\left(A^H\right)$.

We begin by studying the effect of increasing the non-symmetry of Σ. We use the theory and scripts from [**44**].

The symmetric case was shown in figure 3.13 where we remark that for symmetric matrices the field of values is equal to the convex hull of the spectrum.

A first non-symmetry example was shown in figure 3.17 where we can see the field of values growing larger than the convex hull of the spectrum.

We show the effect of increasing the non-symmetry to a larger extent in figure 3.19 and 3.20.

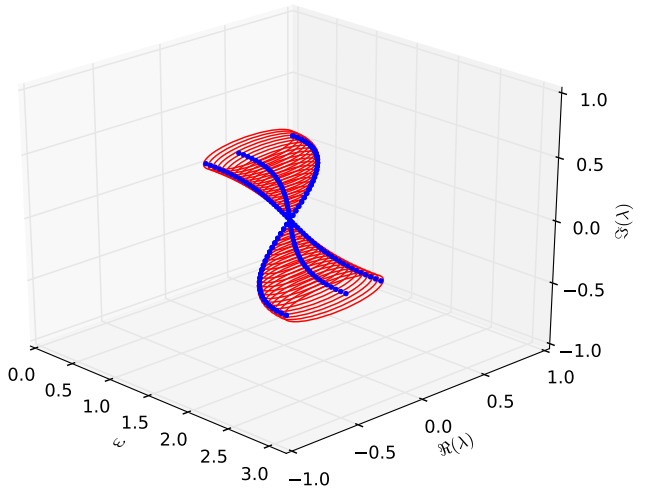


FIGURE 3.15. Eigenvalues and field of values of $I - \widehat{D}\,(\delta = 1/h)_j^{-1}\,\widehat{A}\,(\delta = 1/h)_j$ for $\sigma_1 = 1$, $\sigma_2 = 2$, $\varepsilon = 0.04$ and $h = 0.25$.

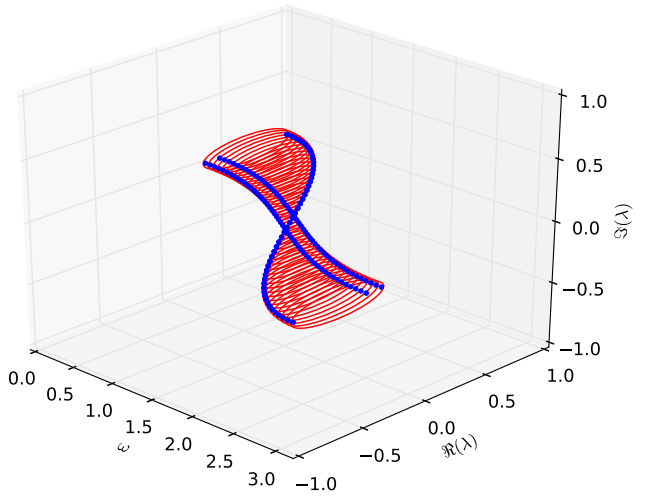


FIGURE 3.16. Eigenvalues and field of values of $I - \widehat{D}\,(\boldsymbol{\delta}_{\text{opt}})_j^{-1}\,\widehat{A}\,(\boldsymbol{\delta}_{\text{opt}})_j$ for $\sigma_1 = 1$, $\sigma_2 = 2$, $\varepsilon = 0.04$ and $h = 0.25$.

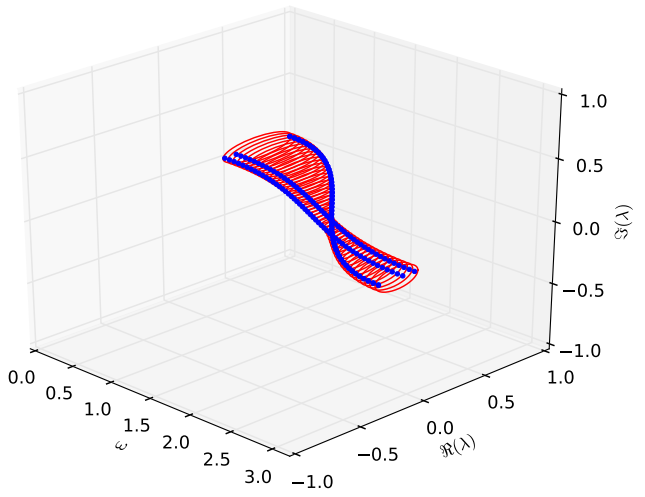


FIGURE 3.17. Eigenvalues and field of values of $I - \boldsymbol{\alpha}_{\text{opt}} \widehat{D}\left(\boldsymbol{\delta}_{\text{opt}}\right)_j^{-1} \widehat{A}\left(\boldsymbol{\delta}_{\text{opt}}\right)_j$ for $\sigma_1 = 1$, $\sigma_2 = 2$, $\varepsilon = 0.04$ and $h = 0.25$.

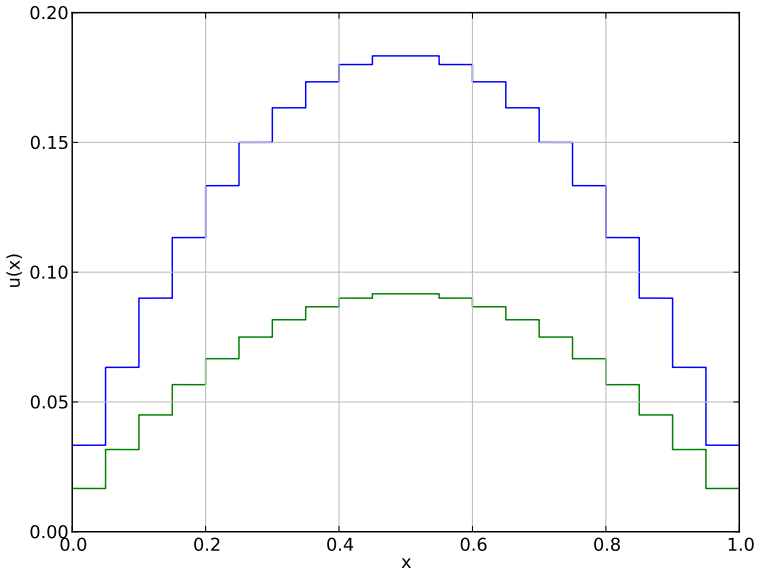


FIGURE 3.18. Solution of the non-symmetric 2 group problem using $\boldsymbol{\delta}_{\text{opt}}$.

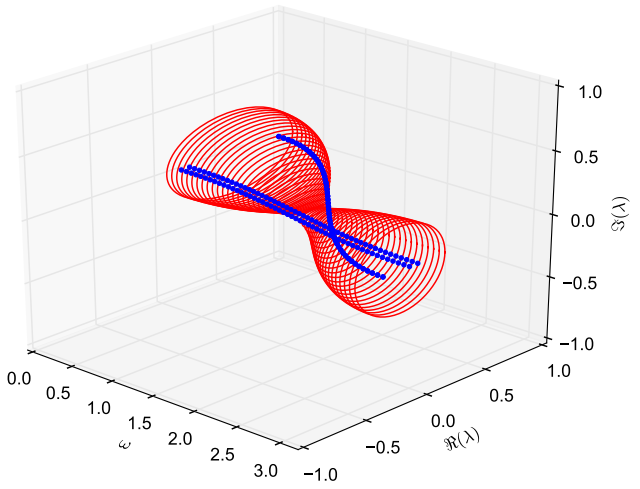


FIGURE 3.19. Eigenvalues and field of values of $I - \boldsymbol{\alpha}_{\text{opt}} \widehat{D}\,(\boldsymbol{\delta}_{\text{opt}})_j^{-1}\, \widehat{A}\,(\boldsymbol{\delta}_{\text{opt}})_j$ for $\sigma_1 = 1$, $\sigma_2 = 10$, $\varepsilon = 0.04$ and $h = 0.25$.

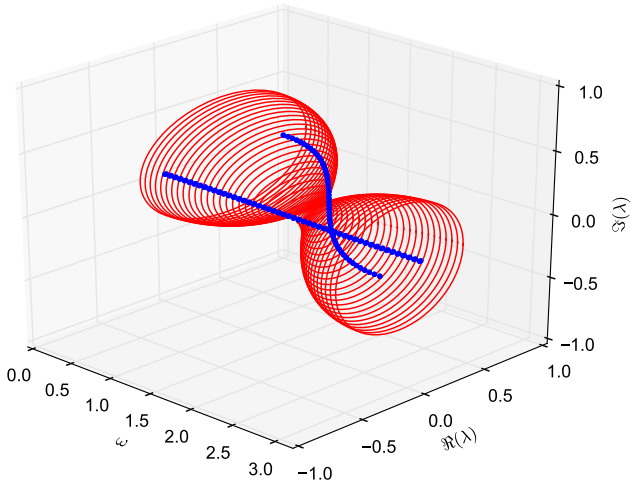


FIGURE 3.20. Eigenvalues and field of values of $I - \boldsymbol{\alpha}_{\text{opt}} \widehat{D}\,(\boldsymbol{\delta}_{\text{opt}})_j^{-1}\, \widehat{A}\,(\boldsymbol{\delta}_{\text{opt}})_j$ for $\sigma_1 = 1$, $\sigma_2 = 100$, $\varepsilon = 0.04$ and $h = 0.25$.

It can be seen that increasing the non-symmetry increases the size of the field of values while keeping the spectrum bounded between the eigenvalue of the laplacian and zero.

We interest ourselves in the effect of decreasing ε, knowing already that increasing it will cluster the eigenvalues around two points: the laplacian eigenvalue and zero.

A first non-symmetry example will be again figure 3.17 where we used $\varepsilon = 0.04$, we show the effect of decreasing ε in figures 3.21 and 3.22. We observe that the field of values grows following the growth of the spectrum.

We remark the three characteristics of the behaviour of

$$I - \boldsymbol{\alpha}_{\text{opt}} \widehat{D} \left(\boldsymbol{\delta}_{\text{opt}}\right)_j^{-1} \widehat{A} \left(\boldsymbol{\delta}_{\text{opt}}\right)_j ,$$

(1) The field of values W includes the spectrum Λ when Σ is non-symmetric (and W is *equal* to the convex hull of Λ when Σ is symmetric). This is a proven fact for all matrices.
(2) $W \setminus \Lambda$ is larger, the larger $|\sigma_2 - \sigma_1|$ is.
(3) For a fixed $|\sigma_2 - \sigma_1| \neq 0$, $W \setminus \Lambda$ grows with $1/\varepsilon$

This suggests that the size of $I - \boldsymbol{\alpha}_{\text{opt}} \widehat{D} \left(\boldsymbol{\delta}_{\text{opt}}\right)_j^{-1} \widehat{A} \left(\boldsymbol{\delta}_{\text{opt}}\right)_j$ will be critical when $|\sigma_2 - \sigma_1| \to \infty$ and $\varepsilon \to 0$.

We turn our view to the analysis of the matrices involved to explain the behaviour observed in the next sections.

7.1. Analysis of the critical case $\varepsilon \to 0$ and $|\sigma_2 - \sigma_1| \to \infty$. For this case, we observe that $I - \boldsymbol{\alpha}_{\text{opt}} \widehat{D} \left(\boldsymbol{\delta}_{\text{opt}}\right)_j^{-1} \widehat{A} \left(\boldsymbol{\delta}_{\text{opt}}\right)_j$ tends to be block diagonal, with the following block repeating twice:

$$B_{\varepsilon \to 0} = \begin{bmatrix} \frac{-\sigma_1 \cos(h\omega) + \sigma_1 + 5\sigma_2 \cos(h\omega) + \sigma_2}{6\sigma_1 + 6\sigma_2} & \frac{-\sigma_1 \cos(h\omega) + \sigma_1 + 5\sigma_2 \cos(h\omega) + \sigma_2}{6\sigma_1 + 6\sigma_2} \\ \frac{5\sigma_1 \cos(h\omega) + \sigma_1 - \sigma_2 \cos(h\omega) + \sigma_2}{6\sigma_1 + 6\sigma_2} & \frac{5\sigma_1 \cos(h\omega) + \sigma_1 - \sigma_2 \cos(h\omega) + \sigma_2}{6\sigma_1 + 6\sigma_2} \end{bmatrix} .$$

Then, in the limit of $\varepsilon \to 0$, $I - \boldsymbol{\alpha}_{\text{opt}} \widehat{D} \left(\boldsymbol{\delta}_{\text{opt}}\right)_j^{-1} \widehat{A} \left(\boldsymbol{\delta}_{\text{opt}}\right)_j$ has two sets of equal eigenvalues which are stable if σ_1 and σ_2 are switched. This is expected since in the reaction-diffusion equation it would result on the switch of the solutions u_1 and u_2.

It is known that for all 2-by-2 matrices the field of values is a (possibly degenerated) ellipse, its shape was determined in one of the (many) works on field of values by C.R. Johnson [**45**]. I our case for real matrices the result can be obtained just by taking the symmetric and skew-symmetric parts of $B_{\varepsilon \to 0}$, and computing their largest and smallest eigenvalues, which will define the smallest rectangular box including the field of values.

We will be interested in the largest axis of the field of values ellipse for each ω, as it will define an upper bound for the numerical radius of $B_{\varepsilon \to 0}$, which is equal to the numerical radius of $\lim_{\varepsilon \to 0} I - \boldsymbol{\alpha}_{\text{opt}} \widehat{D} \left(\boldsymbol{\delta}_{\text{opt}}\right)_j^{-1} \widehat{A} \left(\boldsymbol{\delta}_{\text{opt}}\right)_j$.

We use the formulas in [**45**] to obtain the upper limit of the field of values μ_+ and the lower limit μ_-

$$\mu_\pm = \frac{1}{6} \left(1 + 2\cos(h\omega) \pm \sqrt{9 \left(\frac{\sigma_2 - \sigma_1}{\sigma_2 + \sigma_1} \right)^2 \cos^2(h\omega) + (2\cos(h\omega) + 1)^2} \right) .$$

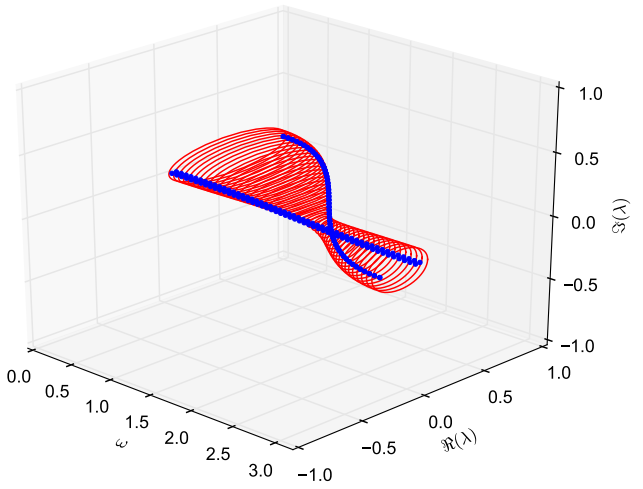


FIGURE 3.21. Eigenvalues and field of values of $I - \boldsymbol{\alpha}_{\text{opt}} \widehat{D}\,(\boldsymbol{\delta}_{\text{opt}})_j^{-1}\, \widehat{A}\,(\boldsymbol{\delta}_{\text{opt}})_j$ for $\sigma_1 = 1$, $\sigma_2 = 2$, $\varepsilon = 0.004$ and $h = 0.25$.

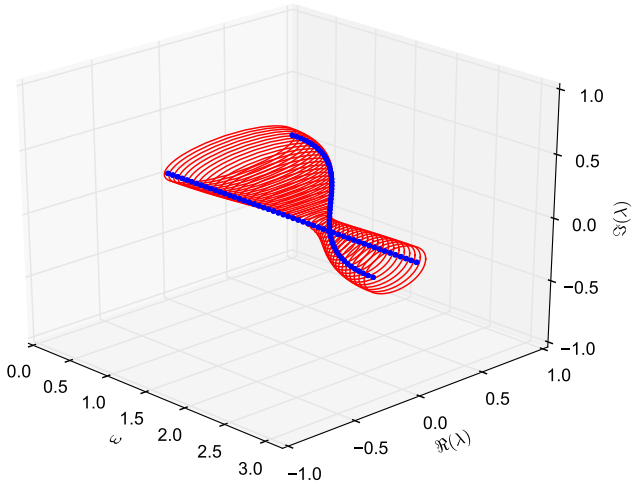


FIGURE 3.22. Eigenvalues and field of values of $I - \boldsymbol{\alpha}_{\text{opt}} \widehat{D}\,(\boldsymbol{\delta}_{\text{opt}})_j^{-1}\, \widehat{A}\,(\boldsymbol{\delta}_{\text{opt}})_j$ for $\sigma_1 = 1$, $\sigma_2 = 2$, $\varepsilon = 0.0004$ and $h = 0.25$.

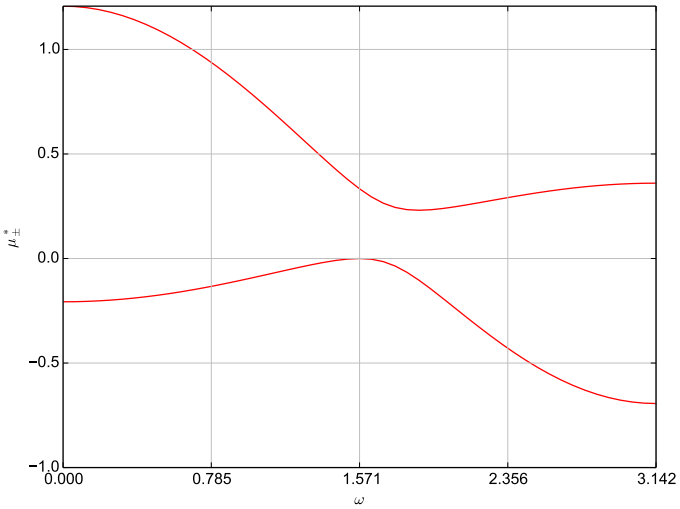


FIGURE 3.23. $\mu^*_\pm$ vs. ω

This formula shows that the leading quantity here is $\left(\frac{\sigma_2-\sigma_1}{\sigma_2+\sigma_1}\right)^2$, it defines a measure of non-symmetry that characterises the field of values W in the limit $\varepsilon \to 0$.

Figure 3.23 shows the values of $\mu^*_\pm$.

As ε gets smaller the spectrum Λ grows inside the field of values when a non-symmetry is present (in the symmetric case W is equal to the convex hull of Λ) and the field of values W grows with it, so we interest ourselves in the limit case of $|\sigma_2 - \sigma_1| \to \infty$ to obtain:

$$\mu^*_\pm = \frac{1}{6}\left(1 + 2\cos(h\omega) \pm \sqrt{9\cos^2(h\omega) + (2\cos(h\omega)+1)^2}\right).$$

This suggests that in the limit $\varepsilon \to 0$ and $|\sigma_2 - \sigma_1| \to \infty$, a V-cycle multigrid algorithm could have a critical number of levels after which the method will fail to converge.

We show some numerical results on the next section which show, however, that the method achieves indeed a flat iteration count.

8. Numerical results

8.1. Multigroup. We use a reaction matrix with a *contrast* between coefficients in different groups that is proportional to different powers of ε, as we did in the previous chapter for the symmetric case.

In this case however, we use a nonsymmetric version as follows

$$\Sigma = \begin{pmatrix} \alpha_1 & -1 & -\varepsilon & -\varepsilon^2 & -\varepsilon^3 & ... \\ -1 & \alpha_2 & -1 & -1 & -1 & ... \\ -\varepsilon^{-1} & -1 & \alpha_3 & -\varepsilon & -\varepsilon^2 & ... \\ -\varepsilon^{-2} & -1 & -\varepsilon^{-1} & \alpha_4 & -\varepsilon & ... \\ -\varepsilon^{-3} & -1 & -\varepsilon^{-2} & -\varepsilon^{-1} & \alpha_5 & ... \\ ... & ... & ... & ... & ... & ... \end{pmatrix},$$

where $\alpha_g = -\sum_{g' \neq g} \Sigma_{g,g'}$. We remark that the elements in the diagonal are such that the matrix has zero column sum. We will consider the top left block of this matrix as the scattering matrix in the following tests.

REMARK 3.2. The choice of the scattering matrix is very similar to the one in the previous chapter, however the nonsymmetry grows with $\frac{1}{\varepsilon}$, instead of only the *contrast*.

Results for a 5 groups calculation are shown in table 3.1. In this case the columns are shown only up to $\varepsilon = 0.01$ to avoid a floating point under/overflow.

	MGAS			MGMS		
levels \ ε	1.0	0.1	0.01	1.0	0.1	0.01
2	5	5	4	4	4	4
3	8	7	7	6	6	6
4	10	10	10	7	7	7
5	12	12	12	8	8	8
6	13	14	13	8	8	8
7	14	14	14	8	9	9
8	14	14	14	9	9	9
9	14	15	15	9	9	9
10	14	15	15	9	9	9
11	15	15	15	9	9	9

TABLE 3.1. CG Iterations to reduce the residual by 10^{-8} for a 5 groups calculation, where "w.c." is the maximum amount of iterations for different ε. Sources tested: $(\varepsilon, 0, \varepsilon, 0, \varepsilon), (0, \varepsilon, 0, \varepsilon, 0), (0, \varepsilon, \varepsilon, \varepsilon, 0)$ or $(\varepsilon, 0, 0, 0, \varepsilon)$

9. Conclusions

The analysis of the interaction between the reaction and the diffusion term of our equations show a very interesting complexity, in particular for the non-symmetric reaction case. The field of values analysis shows that the block-Jacobi type smoothers is indeed useful, as experiments show, because the higher frequencies are addressed with high error reductions in each Richardson step; on the other hand, lower frequency error can be indeed incremented.

Numerical results show that the behavior is satisfactory for quite a significant amount of levels, suggesting that there may be a bound to the error growth for lower frequencies, that coarse solvers can take care of.

CHAPTER 4

Linear Transport Discretization

Abstract. In this chapter, we address the discretization of the linear transport operator, this work was published in [**46**] and it includes a review of previous work in [**7**] and [**5**]. In addition to the existing work, we performed a detailed study on the behavior of the method with respect to the boundary and interior stabilization parameters, effectively minimizing the iteration count in a relatively large range of calculations and when using non-isotropic scattering. We also tested the effect of combining different smoother types in different levels.

1. Introduction

We review recent advances in discontinuous Galerkin (DG) methods for radiation transport in dense, scattering media and shows extensions to improve the boundary approximation and behavioral studies for nonisotropic scattering. Its focus is as in [**5**, **30**], on the reliable discretization and efficient multigrid solvers in regimes where the mesh size of the discretization is considerably greater than the mean free path. It has already been observed in [**4**, **47**] that the upwind DG method [**25**–**27**] suffers from loss of approximation in such regimes. This is exemplified in the top row of Figure 4.1, where a simple model problem was solved with this method and the solutions converge to zero in an nonphysical way. The mechanism behind this loss was analyzed in [**4**, **5**] and a new, robust DG scheme (referred to below as RGK) was proposed by Ragusa, Guermond, and Kanschat in [**30**]. As it is seen in the second row of Figure 4.1, it has the correct convergence behavior. This scheme was improved and an efficient multigrid method was proposed in [**7**]. In both publications [**7**, **30**] the DG method was not modified at the boundary. Analysis and numerical tests were confined to isotropic scattering.

We proceed as follows: in chapter 2 we reviewed the linear Boltzmann equation of monochromatic radiation transport and the asymptotic analysis leading to its diffusion limit. In Section 2, we derive the asymptotic expansion of the DG discretization and discuss an improved boundary model in 2.4. In Section 3, we review the derivation of an efficient multilevel Schwarz method and demonstrate its performance with the new boundary fluxes and for nonisotropic scattering.

2. Discretization and diffusion limit

The domain of computation $\mathcal{D} \times S^{d-1}$ is the tensor product of two spaces with operators of very different properties. Therefore, we discretize separately, such that the resulting discretization space is the tensor product of a spatial space V_ℓ and an angular space S_ℓ to be defined in the following subsections. Alternatively, see [**48**–**50**] for a sparse phase space discretization which is not a plain tensor

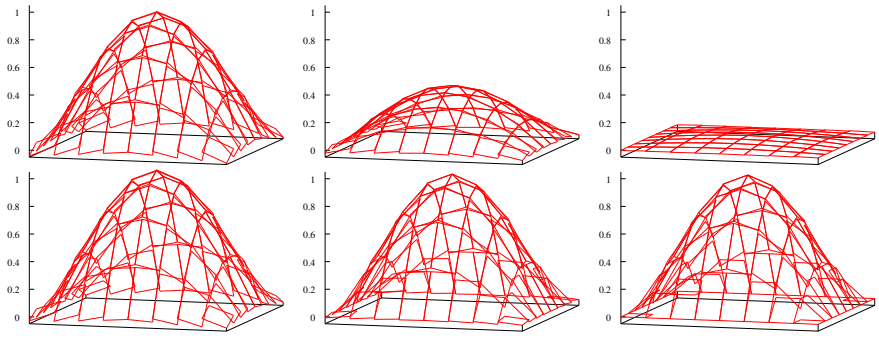


FIGURE 4.1. Solutions of a radiation transport problem with upwind DG method (top) and RGK method [**30**] (bottom); scattering cross sections 2^6, 2^{10}, and 2^{14} (left to right)

product of spatial and angular spaces. Let us assume abstractly that there is a set of basis functions $\{v_i(\mathbf{x})\}$ for V_ℓ and a second set of basis functions $\{\vartheta_j(\mathbf{\Omega})\}$ of S_ℓ. Then, we define the discrete solution space and its basis as

$$W_\ell = V_\ell \otimes S_\ell = \operatorname{span}\{\varphi_{ij}\}, \quad \varphi_{ij}(\mathbf{x},\mathbf{\Omega}) = v_i(\mathbf{x})\vartheta_j(\mathbf{\Omega}). \tag{36}$$

2.1. Discretization of the scattering operator. We begin discretizing the scattering term approximating the integral operator by a quadrature formula. In order to maintain energy conservation and the correct diffusion limit, both integral identities of equation (8) must hold for the quadrature formula. Additionally, in order to keep the odd/even arguments in the asymptotic analysis simple, we require that a quadrature formula with each angle $\mathbf{\Omega}$ shall contain the angle $-\mathbf{\Omega}$ with the same weight. High accuracy formulas with these properties are discussed in [**51**].

After choosing a set of quadrature points $\{\mathbf{\Omega}_j\}$, $j = 1, \ldots, m$, we discretize the space S by collocation. To this end, we define basis functions ϑ_k for S_ℓ by the standard interpolation conditions $\vartheta_k(\mathbf{\Omega}_j) = \delta_{jk}$. Since only these values are used, we do not investigate any further into the actual definition of ϑ_k as a function.

We will avoid the notational overhead of replacing integrals by quadrature sums and just assume a discrete measure on S^{d-1} from now on, that is

$$\int_{S^{d-1}} \vartheta(\mathbf{\Omega}) \, \mathrm{d}\mathbf{\Omega} \equiv \sum_{j=1}^{m} \omega_j \vartheta(\mathbf{\Omega}_j) \tag{37}$$

with quadrature points $\mathbf{\Omega}_j$ and quadrature weights ω_j. We show an example of their distribution in figure 4.2.

2.2. Discretization of the transport term. We intend to use discontinuous Galerkin finite element methods for the discretization of (14), in spite of the notion developed in [**4**,**28**,**29**] that they are not suited. Indeed, the analysis in [**5**] suggested a solution along the lines of [**30**], which was further investigated in [**7**]. But, let us first follow [**5**] and investigate the standard upwind method [**25**, **26**].

To this end, we cover the domain $\mathcal{D}$ by a mesh $\mathbb{T}_\ell$ which may consist of arbitrary, nonoverlapping, convex polygons or polyhedra T. Conformity of the faces of mesh cells is not required, but we assume shape regularity in the sense that the quotients

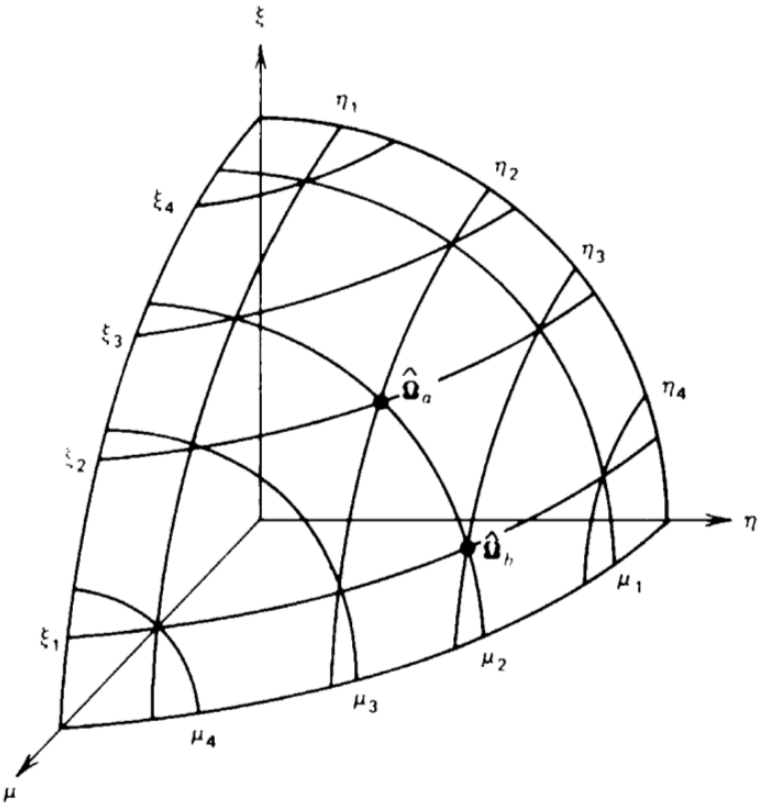


FIGURE 4.2. Symmetric quadrature point arrangement on one octant of the unit sphere [**10**].

of the diameter of a cell and each of its boundary faces are uniformly bounded from above and below independent of the mesh parameter ℓ.

All intersections F of two cells with non-vanishing surface measure form the set $\mathbb{F}_\ell^i$ of interior faces. The intersection F of a cell with the boundary $\partial\mathcal{D}$ is a boundary face and $\mathbb{F}_\ell^\partial$ is the set of all boundary faces. Furthermore, $\mathbb{F}_\ell = \mathbb{F}_\ell^i \cup \mathbb{F}_\ell^\partial$. For simplicity, we introduce the inner product notation

$$\begin{aligned} \big(f,g\big)_T &\equiv \int_T f(\mathbf{x}) \cdot g(\mathbf{x})\,\mathrm{d}\mathbf{x}, & \big(f,g\big)_{\mathbb{T}_\ell} &\equiv \sum_{T\in\mathbb{T}_\ell} \big(f,g\big)_T \\ \big(f,g\big)_F &\equiv \int_F f(\mathbf{x}) \cdot g(\mathbf{x})\,\mathrm{d}\mathbf{s}, & \big(f,g\big)_{\mathbb{F}_\ell} &\equiv \sum_{F\in\mathbb{F}_\ell} \big(f,g\big)_F\,, \end{aligned}$$

where $f \cdot g$ denotes the product of two scalars or the inner product of two vectors as applicable. Extensions to product domains like $\mathbb{T}_\ell \times S^{d-1}$ are obvious and make use of (37).

We obtain the spatial discretization space V_ℓ in the usual fashion of discontinuous Galerkin methods, namely we let

$$V_\ell = \big\{ v \in L^2(\mathcal{D}) \big| \forall T \in \mathbb{T}_\ell : v_{|T} \in V_T \big\}, \tag{38}$$

where V_T is a polynomial space on the cell T, typically the space $\mathbb{P}_k$ of multivariate polynomials of degree up to k or the space $\mathbb{Q}_k$ of mapped tensor product polynomials of degree up to k in each coordinate direction. The actual spaces will be specified with the numerical results.

We ignore the angular dependence for a moment and focus on the transport term $\mathbf{\Omega}\cdot\nabla\varphi$. Integration by parts on each cell T yields

$$\left(\mathbf{\Omega}\cdot\nabla\varphi, w\right)_T = \left(-\varphi, \nabla\cdot(\mathbf{\Omega}w)\right)_T + \left\langle\varphi, (\mathbf{\Omega}\cdot n)w\right\rangle_{\partial T}. \tag{39}$$

The value of a piecewise polynomial function φ on an interface is not well defined, since the values of φ may be different from both cells adjacent to the interface. Therefore, φ is replaced by a so called numerical flux $\widehat{\varphi}$. The flux originally suggested in [**25**, **26**] is the upwind flux

$$\widehat{\varphi}(\mathbf{x}) = \varphi^{\uparrow}(\mathbf{x}) = \lim_{\varepsilon\searrow 0}\varphi(\mathbf{x}-\varepsilon\mathbf{\Omega}), \tag{40}$$

namely the value of φ on the cell in upwind direction $-\mathbf{\Omega}$ from the interface at the point $\mathbf{x}$, or the given boundary value at the part of the boundary, where $\mathbf{\Omega}$ points inwards to $\mathcal{D}$. For instance, in [**52**] it was suggested to split the upwind flux into its consistency part and its stabilization part

$$\varphi^{\uparrow} = \{\!\{\varphi\}\!\} + \{\!\{\operatorname{sign}(\mathbf{\Omega}\cdot\mathbf{n})\varphi\}\!\}, \tag{41}$$

where for any discontinuous function f at the interface

$$\{\!\{f\}\!\}(\mathbf{x}) = \frac{f^{\uparrow}+f^{\downarrow}}{2}, \tag{42}$$

where $f^{\downarrow}(\mathbf{x}) = \lim_{\varepsilon\searrow 0}\varphi(\mathbf{x}+\varepsilon\mathbf{\Omega})$ is the downwind value. Note that this implies in particular

$$\{\!\{\operatorname{sign}(\mathbf{\Omega}\cdot\mathbf{n})\varphi\}\!\} = \frac{\operatorname{sign}(\mathbf{\Omega}\cdot\mathbf{n}^{\uparrow})}{2}\left(\varphi^{\uparrow}-\varphi^{\downarrow}\right) = \frac{\operatorname{sign}(\mathbf{\Omega}\cdot\mathbf{n}^{\downarrow})}{2}\left(\varphi^{\downarrow}-\varphi^{\uparrow}\right).$$

On the boundary, we define $\{\!\{\varphi\}\!\}(\mathbf{x})$ as the limit value from inside $\mathcal{D}$.

Adding up cell contributions in equation (39), we obtain the bilinear form of the standard upwind DG method

$$c_{\ell,\mathbf{\Omega}}(\varphi, w) = \left(-\varphi, \nabla\cdot(\mathbf{\Omega}w)\right)_{\mathbb{T}_\ell} + \left\langle\{\!\{\varphi\}\!\} + \{\!\{\operatorname{sign}(\mathbf{\Omega}\cdot\mathbf{n})\varphi\}\!\}, \{\!\{\mathbf{\Omega}\cdot\mathbf{n}w\}\!\}\right\rangle_{\mathbb{F}_\ell}. \tag{43}$$

Reintroducing $\mathbf{\Omega}$ as an independent variable, we define the upwind DG transport operator

$$c_\ell(\varphi, w) = \left(-\varphi, \nabla\cdot(\mathbf{\Omega}w)\right)_{\mathbb{T}_\ell\times S^{d-1}} + \left\langle\{\!\{\varphi\}\!\} + \{\!\{\operatorname{sign}(\mathbf{\Omega}\cdot\mathbf{n})\varphi\}\!\}, \{\!\{\mathbf{\Omega}\cdot\mathbf{n}w\}\!\}\right\rangle_{\mathbb{F}_\ell\times S^{d-1}}. \tag{44}$$

In order to understand the failure of the upwind DG method in the diffusion limit, we replace the transport operator in the zero order term of equation (17) by its discretization $c_\ell(.,.)$ and test with an isotropic test function (cf. [**5**]):

$$\begin{aligned}
0 &= c_\ell(\varphi_0, w) + \left(\Sigma\varphi_1, w\right)_{\mathbb{T}_\ell\times S^{d-1}} \\
&= \left(-\varphi_0, \nabla\cdot(\mathbf{\Omega}w)\right)_{\mathbb{T}_\ell\times S^{d-1}} + \left\langle\{\!\{\varphi_0\}\!\} + \{\!\{\operatorname{sign}(\mathbf{\Omega}\cdot\mathbf{n})\varphi_0\}\!\}, \{\!\{\mathbf{\Omega}\cdot\mathbf{n}w\}\!\}\right\rangle_{\mathbb{F}_\ell\times S^{d-1}} \\
&= \left\langle\{\!\{\varphi_0\}\!\}, \{\!\{\mathbf{\Omega}\cdot\mathbf{n}w\}\!\}\right\rangle_{\mathbb{F}_\ell\times S^{d-1}} + \left\langle\{\!\{\operatorname{sign}(\mathbf{\Omega}\cdot\mathbf{n})\varphi_0\}\!\}, \{\!\{\mathbf{\Omega}\cdot\mathbf{n}w\}\!\}\right\rangle_{\mathbb{F}_\ell\times S^{d-1}} \\
&= \left\langle\{\!\{\operatorname{sign}(\mathbf{\Omega}\cdot\mathbf{n})\varphi_0\}\!\}, \{\!\{\mathbf{\Omega}\cdot\mathbf{n}w\}\!\}\right\rangle_{\mathbb{F}_\ell\times S^{d-1}}
\end{aligned} \tag{45}$$

where each term drops out because it is the integral over S^{d-1} of the product of a mean value free function with a constant. Since φ_0 is isotropic as well, we let $w = \varphi_0$ to get the equation

$$\left(\int_{S^{d-1}} |\mathbf{\Omega}\cdot\mathbf{n}|\,\mathrm{d}\mathbf{\Omega}\right)\left(\frac{1}{4}\left\langle\varphi_0^{\uparrow}-\varphi_0^{\downarrow},\varphi_0^{\uparrow}-\varphi_0^{\downarrow}\right\rangle_{\mathbb{F}_\ell^i}+\left\langle\varphi_0,\varphi_0\right\rangle_{\mathbb{F}_\ell^\partial}\right)=0. \tag{46}$$

Since the integral on the left is clearly nonzero and positive, φ_0 is continuous and (assuming zero inward radiation condition) zero at the boundary. Given that the diffusion limit ϕ is the limit of φ_0 as $\varepsilon\to 0$ yields ϕ continuous as well. Thus, the choice of the shape of mesh cells and polynomials on these cells is limited to the choices for H^1-conforming finite elements and locking occurs in every other case. Furthermore, it was pointed out in [**5**] that the limit equation converges to an ill-posed boundary value problem as $h\to 0$, if the incoming radiation is not isotropic.

In order to facilitate the discussion, we abstractly write the weak DG version of equation (14) as

$$a_\ell(\varphi,w)=\tau_\ell(\varphi,w)+b_{\ell,i}(\varphi,w)+b_{\ell,b}(\varphi,w)=\varepsilon\left(q,w\right)_{\mathcal{D}\times S^{d-1}}, \tag{47}$$

where

$$\tau_\ell(\varphi,w)=\left(-\varphi,\mathbf{\Omega}\cdot\nabla w\right)_{\mathcal{D}\times S^{d-1}}+\varepsilon\left(\sigma_a\varphi,w\right)_{\mathcal{D}\times S^{d-1}}+\tfrac{1}{\varepsilon}\left(\sigma_s\Sigma\varphi,w\right)_{\mathcal{D}\times S^{d-1}}. \tag{48}$$

The interface forms $b_{\ell,i}(\varphi,w)$ and $b_{\ell,b}(\varphi,w)$ on interior and boundary faces are of the general form

$$b_{\ell,i}(\varphi,w)=\left\langle\{\!\{\varphi\}\!\}+\varepsilon\gamma_{h,i}\,\{\!\{\mathrm{sign}(\mathbf{\Omega}\cdot\mathbf{n})\varphi\}\!\},\{\!\{\mathbf{\Omega}\cdot\mathbf{n}w\}\!\}\right\rangle_{\mathbb{F}_\ell^i\times S^{d-1}}, \tag{49}$$

$$b_{\ell,b}(\varphi,w)=\left\langle\varphi+\varepsilon\gamma_{h,b}\,\mathrm{sign}(\mathbf{\Omega}\cdot\mathbf{n})\varphi,\mathbf{\Omega}\cdot\mathbf{n}w\right\rangle_{\mathbb{F}_\ell^b\times S^{d-1}}, \tag{50}$$

where now $\gamma_{h,i}$ and $\gamma_{h,b}$ are parameters of the method, possibly depending on the (local) mesh size h and the coefficients of the equation. The upwind method is obtained for $\gamma_{h,i}=\gamma_{h,b}=1/\varepsilon$.

2.3. Numerical fluxes at interior interfaces. As discussed in [**7**, **30**], the key to a solution of the problem of locking is the understanding that upwinding, while physically motivated, is not a necessity, and that equation (41) suggests that, while the consistency part is fixed, the stabilization part of the upwind flux can be modified such that it does not appear in (45). To this end, we replace the numerical flux in equation (40) by

$$\widehat{\varphi}(\mathbf{x})=\{\!\{\varphi\}\!\}(\mathbf{x})+\varepsilon\gamma_{h,i}(\mathbf{x})\,\{\!\{\mathrm{sign}(\mathbf{\Omega}\cdot\mathbf{n})\varphi\}\!\}(\mathbf{x}), \tag{51}$$

with

$$\gamma_{h,i}=\frac{1}{\varepsilon}\min\left\{1,\frac{\varepsilon\gamma_i}{h\sigma_s(\mathbf{x})}\right\}=\min\left\{\frac{1}{\varepsilon},\frac{\gamma_i}{h\sigma_s(\mathbf{x})}\right\}, \tag{52}$$

where h is the diameter of the face on which this term is evaluated (we assumed shape regularity). While these fluxes in principle coincide with those in [**7**], we have chosen a presentation which more clearly exhibits the role of ε and the free parameter γ_i We set $\gamma_{h,i}=1/\varepsilon$ whenever the denominator is zero. This choice of γ_h has an immediate physical interpretation, since it compares the scattering mean free path ε/σ_s to the diameter h of the mesh cell. If the latter is greater by at least a factor of γ_i, we use the modified scheme and else the standard upwind scheme. This choice thus corresponds to the old observation that discretization of

radiation transport becomes difficult if the mean free path is much smaller than the cell diameter.

In particular, we have $\gamma_{h,i}$ independent of ε, as long as $\varepsilon \ll h$, the case of the diffusion limit. As an immediate result, the last remaining term of equation (45) will only appear in the equation for ε^1, and thus equation (45) reduces to the tautology $0 = 0$.

By entering the new DG formulation into equations (19) and (20) for ε^0 and ε^1, it was shown in [**7**] that the discrete diffusion limit with the modified upwind flux is the LDG discretization of Poisson's equation [**53**]:

$$\begin{array}{rcll} \big(\nabla\phi, \mathbf{L}\big) + \big\langle \{\!\{\mathbf{L}\}\!\}, [\![\phi\mathbf{n}]\!] \big\rangle + & \big(d\sigma_s \mathbf{J}, \mathbf{L}\big) & = & 0 \quad \forall \mathbf{L} \in (V_h)^d \\ \big(\nabla\cdot\mathbf{J}, w\big) + \big\langle \{\!\{\mathbf{J}\}\!\}, [\![w\mathbf{n}]\!] \big\rangle + c\big\langle \gamma[\![\phi]\!], [\![w]\!] \big\rangle & = \big(q, w\big) & & \quad \forall w \in V_h \end{array} \tag{53}$$

The diffusion limit is thus a well-posed and convergent method for diffusion problems independent of the choice of the parameter γ_0. Nevertheless, it has been pointed out already in [**54**, **55**] that the choice of such a parameter might very well have considerable impact on the discretization accuracy and solver performance. These questions will be investigated in the following sections.

2.4. Modeling the boundary condition. In this section, we discuss the modeling and discretization of boundary conditions, which has been neglected in the discussion of weakly penalized DG methods so far. Indeed, in [**7**, **30**] the discretization at the boundary remained unchanged from the original upwind method, which corresponds to the assumption of a void exterior domain (item two in the list below). Here, we change this assumption to a boundary which cuts through a scattering domain. To this end, we have to discuss the various possible boundary conditions for Boltzmann equations like (1). These are either reflecting or artificial. Reflecting boundary conditions originate from the fact that the boundary acts as a physical mirror or diffusive surface by sending a part or all the incident radiation back into the domain, absorbing the remainder. It has the following form: for $(\mathbf{x}, \mathbf{\Omega}) \in \Gamma_-$:

$$\varphi(\mathbf{x}, \mathbf{\Omega}) = \int_{\mathbf{\Omega}'\cdot\mathbf{n}(\mathbf{x})>0} R(\mathbf{\Omega}', \mathbf{\Omega}) \varphi(\mathbf{x}, \mathbf{\Omega}') \, \mathrm{d}\mathbf{\Omega}'. \tag{54}$$

Here, R is the reflection function with integral not greater than one. No smoothness is required on R and it can degenerate to a reflection of a Dirac functional on $\mathbf{\Omega}'$ in the case of a perfect mirror. In another limit, $R \equiv 0$ amounts to a perfectly absorbing boundary condition where particles leaving the domain are lost while none enters and if the domain is convex also models the artificial vacuum boundary condition. A boundary condition of reflective type also results from the fact that the domain was cut off at a symmetry plane to save computational cost.

All other boundary conditions are artificial, meaning that the convex domain $\mathcal{D}$ was cut out of $\mathbb{R}^d$ arbitrarily, splitting the latter into $\mathcal{D}$ and the exterior domain $\mathbb{R}^d \setminus \mathcal{D}$. In this case, there may exist inward radiation from an external source. A popular modeling assumption states that all radiation leaving the domain is lost forever. Nevertheless, if the exterior domain is diffusive this assumption is not true, since diffusive areas reflect radiation. Thus, for a diffusive domain $\mathcal{D}$, where $\widetilde{\sigma_s}$ is uniformly large, the artificial boundary condition may be a combination of three limit cases:

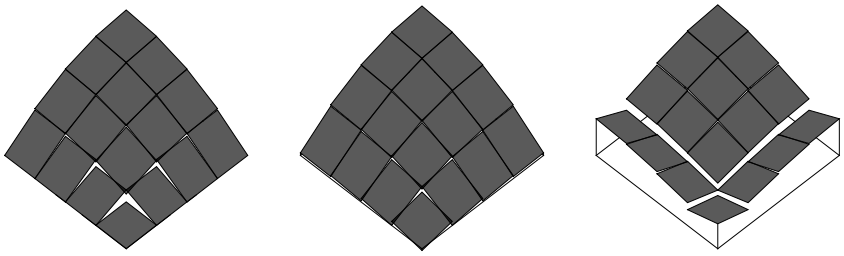


FIGURE 4.3. Solutions with upwind (left) and diffusive boundary flux $\gamma_b = 16$ (center) and $\gamma_b = 1$ (right), respectively. Zoom into the boundary of a 16-by-16 mesh, linear elements.

(1) The exterior domain is diffusive with $\widetilde{\sigma_s}$ continuous at $\partial\mathcal{D}$. In this case, reflection happens. The diffusion limit (21) is a good approximation beyond $\partial\mathcal{D}$ and thus appropriate numerical fluxes for the diffusive case should be applied.
(2) The exterior domain is void with $\widetilde{\sigma_s} = 0$. In this case, the boundary is also a material interface between diffusive and non-diffusive radiation. To our knowledge, no satisfactory numerical fluxes have been investigated for this case. The upwind flux used in [**30**] is a possibility, but not justified by any analysis. To the contrary, if we follow the reasoning in [**56**], the flux in the diffusive region —and thus the solution there— should rather ignore the advective part.
(3) If the scattering cross section $\widetilde{\sigma_s}$ is small inside $\mathcal{D}$ close to its boundary, then $\widetilde{\sigma_s} = 0$ in the exterior domain is a natural assumption and upwind fluxes are a natural choice, if we assume continuity of coefficients across the boundary.

The last case is of minor mathematical interest, and we will compare the first two in this section. First, we consider the discretization aspect of this choice. The effect on Schwarz preconditioners is discussed in the following section. To this end, we choose the following setup for numerical experiments: let $\mathcal{D} = [-1, 1]^2$. All our experiments are currently two-dimensional, albeit the theoretical discussion in this chapter is not. In fact, we choose quadrature formulas which are the projections of three-dimensional formulas into two dimensions. As right hand side in (14), we choose the constant function $q = 1$. Furthermore, $\sigma_a = 0$ and $\sigma_s = 1$ constantly and all inflow boundary values are zero. Thus, for $\varepsilon \to 0$, the solutions to the continuous radiation transport problem converge to the solution of $-\Delta u = 3$ with homogeneous Dirichlet boundary condition. In Figure 4.3, we show a zoom into a corner of a 16-by-16 mesh and compare different boundary stabilizations for linear elements.

Clearly, the limit boundary condition of the upwind method on the left forces the solution to be almost zero everywhere in the corner cell, while it approximates the diffusive boundary solution in the corner in the center image, where we chose $\gamma_b = 16$. The graph on the right shows the result for $\gamma_b = 1$, which is clearly not sufficient to obtain a qualitatively correct result.

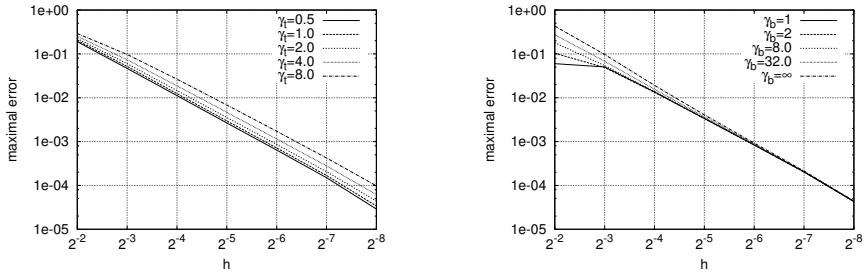


FIGURE 4.4. Maximal difference to the solution of the diffusion limit for different γ_i (left) and γ_b (right).

In Figure 4.4, we explore the influence of the stabilization parameters on the error measured at the center of the domain. We compare to the solution $\phi_0(0) = 0.884056$ of the limit diffusion problem, which was computed by over-refinement down to 13 levels. The scattering parameter of the transport problem is chosen as $\varepsilon = 10^{-5}$. The values displayed are $\max_{\mathbf{x}} \left| \sum_{j=1}^{m} \omega_j \varphi(\mathbf{x}, \mathbf{\Omega}_j) - \phi_0(\mathbf{x}) \right|$ and the figure clearly shows second order convergence over a range of stabilization parameters. We conclude that while the parameters may not be chosen too small or too large, they do not strongly influence solution accuracy over a wide range. This is definitely true for γ_b, which hardly influences accuracy at all, while a factor of 16 in γ_i may cause a difference in error by a factor of 5. Our transport solver using the values which are found optimal for the solver below converges to the value 0.88406, which is consistent with our expectations that the error should be bounded close to ε.

3. Multilevel Schwarz methods

The choice of a suitable solver was discussed in [**7**], such that we can rely on the conclusion there and describe the multilevel Schwarz method for equation (47). The method consists of a standard multigrid V-cycle, combined with a smoother based on domain decomposition and local solvers. Since the multigrid V-cycle algorithm has been described over and over again [**35**, **39**, **40**], we focus here on the subspace decomposition used for the smoother.

We choose a nonoverlapping decomposition of the domain $\mathcal{D}$ into subdomains corresponding to the cells T of the mesh $\mathbb{T}_\ell$, and thus a decomposition

$$W_\ell = \bigoplus_{T \in \mathbb{T}_h} W_T = \bigoplus_{T \in \mathbb{T}_h} V_T \otimes S_\ell. \tag{55}$$

We introduce the projection operator $P_T : W_\ell \to W_T$ defined by (see for instance [**35**, Chapter 2])

$$a_\ell(P_T \varphi, w) = a_\ell(\varphi, w), \quad \forall w \in W_T. \tag{56}$$

Identifying with $a_\ell(.,.)$ an operator $A_\ell : W_\ell \to W_\ell$ through

$$a_\ell(\varphi, w) = \left(A_\ell \varphi, w\right)_{\mathcal{D} \times S^{d-1}}, \quad \forall w \in W_\ell, \tag{57}$$

we define the additive Schwarz smoother $R_{J,\ell} : W_\ell \to W_\ell$ as

$$R_\ell^{\mathrm{a}} = \varrho \sum_{T\in\mathbb{T}_h} P_T A_\ell^{-1}. \tag{58}$$

Here, $\varrho \in (0,1)$ is a relaxation parameter. This is indeed the block Jacobi method where each inverted diagonal block corresponds to a cell matrix.

Note that since $a_\ell(.,.)$ is not a symmetric bilinear form, the projection P_T is not orthogonal and thus the standard analysis of Schwarz methods does not apply. Nevertheless, the fact that $a_\ell(.,.)$ is well approximated by a diffusion operator if $\varepsilon \ll 1$ suggests that R_ℓ^{a} might be an effective smoother in a V-cycle algorithm in that case. This is verified later in Table 4.2, where we use the V-cycle with this smoother as a preconditioner in a GMRES iteration.

If the scattering cross section is small, it is well known that an additive Schwarz method is not sufficient and we have to employ a multiplicative Schwarz method with downwind ordering, see for instance [**57**]. To this end, we number the mesh cells as $T_1, \dots T_n$, where n is the number of cells in $\mathbb{T}_h$. If now o is a permutation of the set $\{1,\dots,n\}$, the multiplicative Schwarz smoother with respect to the ordering o is defined by

$$R_\ell^o = (I - E_\ell^o)A_\ell^{-1} \text{ with } E_\ell^o = (I - P_{T_{o(n)}})(I - P_{T_{o(n-1)}})\cdots(I - P_{T_{o(1)}}). \tag{59}$$

For a given vector $\boldsymbol{\Theta} \in \mathbb{R}^d$, generate an ordering $o_{\boldsymbol{\Theta}}$ by the condition that for the cell centers C_T there holds

$$o(i) \le o(j) \iff \left(C_{T_{o(j)}} - C_{T_{o(i)}}\right)\cdot\boldsymbol{\Theta} \le 0. \tag{60}$$

Clearly, $o_{\boldsymbol{\Theta}}$ is a downwind ordering only for vectors in a cone around $\boldsymbol{\Theta}$ and not for instance for $-\boldsymbol{\Theta}$. Experience shows that at least on regular meshes choosing 2^d vectors as the diagonals of all orthants of $\mathbb{R}^d$ is sufficient for covering all angles $\boldsymbol{\Omega}$ of an angular quadrature (37) (see figure 4.5). Thus, with downwind orderings $o_1, o_2, \dots, o_{2^d}$ we define the full sweep smoother

$$\begin{aligned} R_\ell^s =& (I - E_\ell^s)A_\ell^{-1} \\ E_\ell^s =& (I - P_{T_{o_1(n)}})(I - P_{T_{o_1(n-1)}})\cdots(I - P_{T_{o_1(1)}}) \\ & (I - P_{T_{o_2(n)}})(I - P_{T_{o_2(n-1)}})\cdots(I - P_{T_{o_2(1)}}) \\ & \cdots \\ & (I - P_{T_{o_{2^d}(n)}})(I - P_{T_{o_{2^d}(n-1)}})\cdots(I - P_{T_{o_{2^d}(1)}}). \end{aligned} \tag{61}$$

It was demonstrated in [**7**] that R_ℓ^s is a direct solver in the case $\widetilde{\sigma_s} = 0$.

Even though in the case of the full sweep, each sweep is applied sequentially, we tested a parallel version in which all the sweeps are performed at the same time (see figure 4.6).

Hereafter, we focus on studying its performance for non-isotropic scattering, its dependence on the stabilization parameters $\gamma_{h,i}$ and $\gamma_{h,b}$ and combining different smoothers between levels.

3.1. Influence of stabilization parameters. While in principle the stability of the method is guaranteed for any stabilization parameter, we study here the influence of these parameters on the performance of the multilevel solver. To this end, we test the full sweep and block-Jacobi algorithms as preconditioners to the GMRES iteration to a problem with constant coefficients $\sigma_s = 1$ and $\sigma_a = 0$ in a

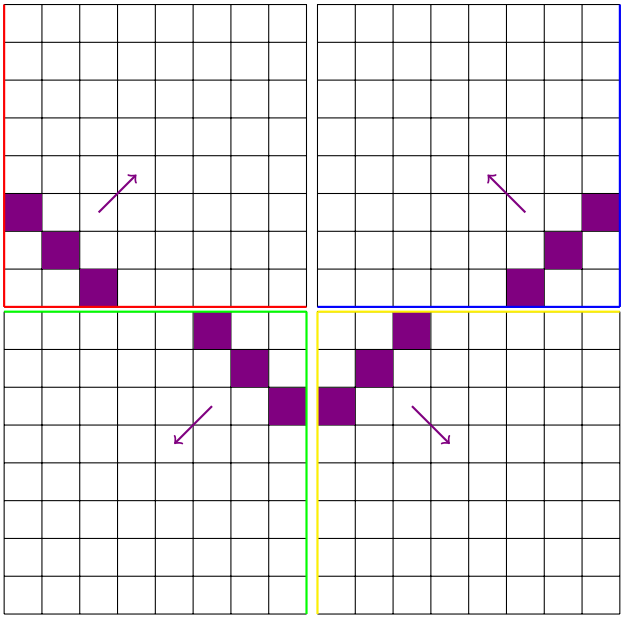


FIGURE 4.5. Full sweep in 2 spatial dimensions.

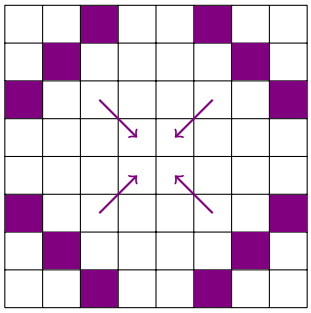


FIGURE 4.6. Parellel full sweep in 2 spatial dimensions.

Cartesian 2D domain $(x, y) \in [-1, 1]^2$. It is known that the additive block-Jacobi smoother may require a relaxation parameter $\varrho < 1$ to converge [**58**]: we use a relaxation of $\varrho = 0.7$.

We measure performance by the number of iteration counts to reduce the residual by 10^{-8} for different combinations of the stabilization parameters γ_i and γ_b, and we show that it becomes constant as we tend to a diffusive regime or the mesh is refined and for different angle quadratures with both isotropic and anisotropic scattering redistribution functions.

We start by looking for an optimal combination of stabilization parameters that minimize the number of iterations. We select a test case with isotropic scattering, $\varepsilon = 10^{-5}$ on a 256-by-256 mesh and Gauss-Legendre-Chebyshev (TGLC1) angular quadrature from [**51**] with four angles for both the full sweep and block-Jacobi smoothers.

Results in Tables 4.1 and 4.2 show that the choice of γ_i and γ_b strongly affects the convergence rate (Parameter values ∞ correspond to the upwind method). In fact, the iteration does not converge within 100 steps if the upwind method is used in the interior or if γ_i is too small.

	γ_b						
γ_i	∞	32	16	8	4	2	1
∞	∞	∞	∞	∞	∞	∞	∞
8	12	8	7	8	8	11	18
4	8	6	6	6	6	8	11
2	6	5	5	5	5	6	8
1	5	5	5	5	6	6	7
1/2	17	17	17	17	18	19	22
1/4	∞	∞	∞	∞	∞	∞	∞

TABLE 4.1. Number of GMRES steps needed to reduce the initial residual by 10^{-8}. 4 angular quadrature points, isotropic, $h = 2^{-7}$ and $\varepsilon = 10^{-5}$ using a full sweep smoother. The box indicates parameters with best average residual reduction. ∞ indicates no convergence in 100 steps.

	γ_b						
γ_i	∞	32	16	8	4	2	1
∞	∞	∞	∞	∞	∞	∞	∞
8	28	19	21	24	30	35	35
4	23	16	16	17	22	28	31
2	20	15	15	15	17	23	30
1	22	19	18	18	18	21	29
1/2	∞	∞	∞	∞	∞	∞	∞

TABLE 4.2. Number of GMRES steps needed to reduce the initial residual by 10^{-8}. 4 angular quadrature points, isotropic, $h = 2^{-7}$ and $\varepsilon = 10^{-5}$ using a block-Jacobi smoother. The box indicates parameters with best average residual reduction. ∞ indicates no convergence in 100 steps.

Finally, we see a common minimum for both smoothers at $\gamma_i = 2$ and $\gamma_b = 16$, we select it between the lowest iteration counts as the calculation that achieved the highest residual reduction and we use it for the rest of the calculations to show that iteration counts are independent of the mesh size in different settings.

3.2. Isotropic scattering. Tables 4.3, 4.4 and 4.5 show the iteration counts for isotropic scattering, using a TGLC1 quadrature for the optimal values of γ_i and γ_b found above, we observe that the amount of iterations becomes independent of the mesh size for both smoothers, having a higher iteration count for the block-Jacobi smoother, which is expected.

levels \ ε	1	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
3	3	4	4	4	4	4
4	3	4	5	5	5	5
5	3	4	5	5	5	5
6	3	4	4	5	5	5
7	3	4	4	5	5	5
8	2	4	4	4	5	5
9	2	4	4	4	4	4

TABLE 4.3. Number of GMRES steps needed to reduce the initial residual by 10^{-8}. 4 angular quadrature points, isotropic, $\gamma_i = 2$ and $\gamma_b = 16$ using a full sweep smoother.

levels \ ε	1	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
3	4	5	5	5	5	5
4	5	5	6	6	6	6
5	5	5	6	6	6	6
6	5	5	5	6	6	6
7	5	5	5	5	5	5
8	5	5	5	5	5	5
9	5	5	5	5	5	5

TABLE 4.4. Number of GMRES steps needed to reduce the initial residual by 10^{-8}. 4 angular quadrature points, isotropic, $\gamma_i = 2$ and $\gamma_b = 16$ using a parallel full sweep smoother.

levels \ ε	1	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
3	9	10	10	9	8	7
4	15	12	14	14	14	14
5	21	12	15	15	15	15
6	30	14	15	15	15	15
7	45	18	13	15	15	15
8	65	24	12	15	15	15
9	96	34	12	14	14	14

TABLE 4.5. Number of GMRES steps needed to reduce the initial residual by 10^{-8}. 4 angular quadrature points, isotropic, $\gamma_i = 2$ and $\gamma_b = 16$ using a block-Jacobi smoother

Tables 4.6 and 4.7 show results for a 12 angle quadrature (TGLC2) which are very similar to the results shown previously for the TGLC1 quadrature and with some iteration counts being slightly lower.

Results for a 24 angle quadrature (TGLC3) show no impact in the iteration counts seen for TGLC2 from which we infer that the iteration counts become independent of the angle quadrature size for isotropic scattering.

levels \ ε	1	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
3	3	3	4	4	4	4
4	3	4	4	4	4	4
5	2	4	4	5	5	5
6	2	4	4	5	5	5
7	2	4	4	4	4	4
8	2	4	4	4	4	4
9	2	4	4	4	4	4

TABLE 4.6. Number of GMRES steps needed to reduce the initial residual by 10^{-8}. 12 quadrature angles, isotropic, $\gamma_i = 2$ and $\gamma_b = 16$ using a full sweep smoother.

levels \ ε	1	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
3	10	10	10	9	8	8
4	13	11	14	14	14	14
5	19	11	15	15	15	15
6	26	12	14	15	15	15
7	36	14	12	15	15	15
8	57	20	11	15	15	15
9	76	27	11	14	14	14

TABLE 4.7. Number of GMRES steps needed to reduce the initial residual by 10^{-8}
TGLC2, isotropic, $\gamma_i = 2$ and $\gamma_b = 16$ using a block-Jacobi smoother.

3.3. Nonisotropic scattering. Tables 4.8 and 4.9 show results for a test case using a TGLC1 quadrature and redistribution functions in (2) of the form

$$H(\mathbf{\Omega}', \mathbf{\Omega}) = 1 + \alpha \mathbf{\Omega}' \cdot \mathbf{\Omega}. \tag{62}$$

The shape of this functions is shown in Figure 4.7. We see no differences with moderate values of α so we compare with extreme values $\alpha = 0.9$ and $\alpha = 1$, the latter excluding back-scattering. Tables 4.3 and 4.5 show that adding anisotropy does not significantly change the behavior of the solver, whether we apply the Jacobi or the full sweep smoother.

This holds for both quadratures TGLC1 and TGLC2. We note that it even works for the limit case $\alpha = 1$, although the kernel of the scattering operator Σ in this case contains not only isotropic functions, which is an underlying assumption of the asymptotic analysis.

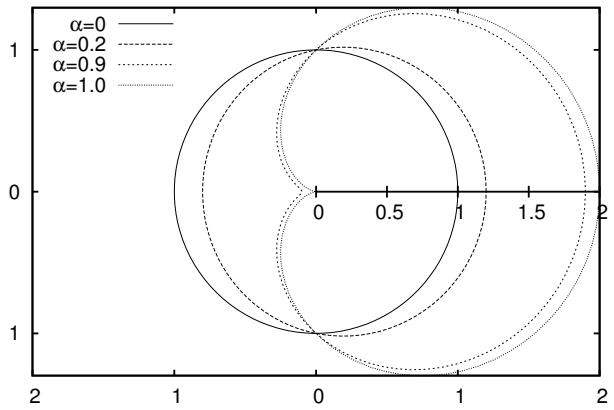


FIGURE 4.7. Shape of the redistribution functions $H(\mathbf{\Omega}', \mathbf{\Omega}) = 1 + \alpha \mathbf{\Omega}' \cdot \mathbf{\Omega}$ for different values of α. Shown is H as a function of the angle $\theta = \arccos(\mathbf{\Omega}' \cdot \mathbf{\Omega})$ in polar coordinates.

		ε					
$n_{\mathbf{\Omega}}$	α	1	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
4	0.0	2	4	4	4	4	4
	0.9	2	3	5	4	5	5
	1.0	2	3	5	5	5	5
12	0.0	2	4	4	4	4	4
	0.9	2	3	5	4	5	5
	1.0	2	3	5	5	5	5

TABLE 4.8. Number of GMRES steps needed to reduce the initial residual by 10^{-8}. TGLC1 ($n_{\mathbf{\Omega}} = 4$) and TGLC2 ($n_{\mathbf{\Omega}} = 12$) angular quadrature, nonisotropic scattering according to equation (62), 9 levels, $\gamma_i = 2$ and $\gamma_b = 16$ using a full sweep smoother.

		ε					
$n_{\mathbf{\Omega}}$	α	1	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
4	0.0	96	34	12	14	14	14
	0.9	89	37	12	15	16	17
	1.0	89	38	12	16	17	17
12	0.0	76	27	11	14	14	14
	0.9	75	27	10	15	17	17
	1.0	75	27	10	16	18	18

TABLE 4.9. Number of GMRES steps needed to reduce the initial residual by 10^{-8}. TGLC1 ($n_{\mathbf{\Omega}} = 4$) and TGLC2 ($n_{\mathbf{\Omega}} = 12$) angular quadrature, nonisotropic scattering, 9 levels, $\gamma_i = 2$ and $\gamma_b = 16$ using a block-Jacobi smoother.

3.4. Combination of smoothers. In figure 4.10 we present some results regarding using a parallel full sweep smoother only for the finest mesh and a block-Jacobi for the rest of the multigrid levels.

levels \ ε	1	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
3	4	5	5	5	5	5
4	5	5	6	6	6	6
5	5	6	7	7	7	7
6	6	7	8	8	8	8
7	6	7	9	9	9	9
8	6	7	9	10	10	10
9	6	7	9	10	10	10
10	6	7	9	10	10	10

TABLE 4.10. Number of GMRES steps needed to reduce the initial residual by 10^{-8}. 4 angular quadrature points, isotropic, $\gamma_i = 2$ and $\gamma_b = 16$ using a parallel full sweep smoother for the finest mesh and block-Jacobi smoother for the coarser meshes.

We observe that we retain a flattening iteration count, however at a higher overall count. These results show the interest in the parallel full sweep since it is relatively cheaper than the classical full sweep and reduces the calculation time by a factor of almost 4 in 2D and almost 8 in 3D.

3.5. Other dependencies of the stabilization parameters. Additionally, we present examples for the dependence of the optimal stabilization on other parameters. First, we consider geometry. To this end, we run simulations on the disk domain shown in figure 4.8. The mesh size is refined 5 times from the one shown. Here, the cells are not quadratic anymore, they even do not have an affine transformation to the reference square. Nevertheless, the results in Table 4.11 show, that the optimum is attained at the same number of iteration steps as in Table 4.1. Furthermore, convergence is also robust for a wide range of stabilization parameters.

So far, we have only computed for $\sigma_a = 0$, since it is the most challenging case. In Table 4.12, we report results for $\sigma_s = 10^5$ and $\sigma_a = 1$. As expected, adding absorption increases the solver efficiency considerably to the point, where the other parameters almost do not matter anymore. Nevertheless, the upwind DG method with $\gamma_i = \infty$ still requires a lot more steps.

Finally, we increase the polynomial degree to two. According to Table 4.13, the optimal values for γ_i and γ_b double from those for degree one, a result consistent with observations for DG methods for elliptic problems, see for instance [**59**].

4. Implementation remarks

Through the deal.II library, we use distributed meshes with the p4est library [**60**] and parallelize vectors using MPI [**61**]. The operator was written in a matrix-free fashion, thereby reducing its memory footprint. Smoother's local matrices are assembled and inverted explicitly using a LAPACK [**62**] singular value decomposition. The coarse matrix is assembled and solved using a GMRES solver with an

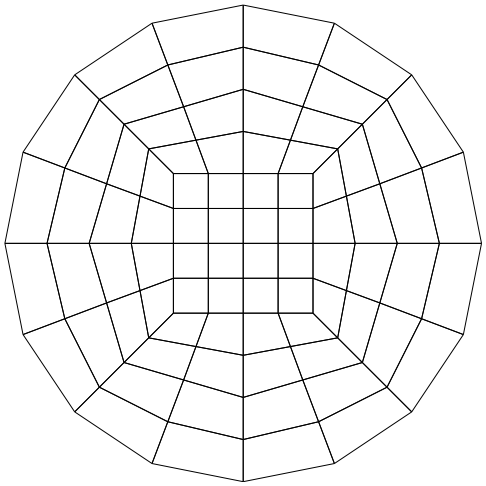


FIGURE 4.8. Disk domain triangulation (shown here for level 2)

				γ_b			
γ_i	∞	32	16	8	4	2	1
∞	∞	∞	∞	∞	∞	∞	∞
8	9	8	9	9	11	16	26
4	6	6	6	7	8	10	15
2	6	5	$\boxed{5}$	6	6	7	9
1	∞	∞	∞	∞	∞	∞	∞

TABLE 4.11. Number of GMRES steps needed to reduce the initial residual by 10^{-8} on a disk domain with radius 1. TGLC1 angular quadrature, isotropic, $h \approx 2^{-7}$ and $\varepsilon = 10^{-5}$ using a full sweep smoother. The box indicates parameters with best average residual reduction. ∞ indicates no convergence in 100 steps.

				γ_b			
γ_i	∞	32	16	8	4	2	1
∞	24	24	23	22	21	21	20
8	2	2	2	2	2	2	2
4	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2
1	2	2	2	2	2	2	2
1/2	2	2	2	2	2	2	2
1/4	2	2	2	2	2	2	2

TABLE 4.12. Number of GMRES steps needed to reduce the initial residual by 10^{-8} with $\sigma_a/\sigma_s = 10^{-5}$. TGLC1 angular quadrature, isotropic, $h = 2^{-7}$ and $\varepsilon = 10^{-5}$ using a full sweep smoother.

AMG preconditioner using block Gauss-Seidel smoothers, provided by the Trilinos library [**63**].

				γ_b			
γ_i	∞	32	16	8	4	2	1
∞	∞	∞	∞	∞	∞	∞	∞
8	6	5	5	5	6	8	13
4	4	4	**4**	5	5	7	9
2	6	7	7	7	8	8	16
1	∞	∞	∞	∞	∞	∞	∞

TABLE 4.13. Number of GMRES steps needed to reduce the initial residual by 10^{-8} with polynomial degree 2. TGLC1 angular quadrature, isotropic, $h = 2^{-7}$ and $\varepsilon = 10^{-5}$ using a full sweep smoother. The box indicates parameters with best average residual reduction. ∞ indicates no convergence in 100 steps.

We show in figures 4.9 and 4.10 scaling results for the additive Schwarz and parallel full sweep only on the finest mesh (coarser meshes using additive Schwarz) smoothers.

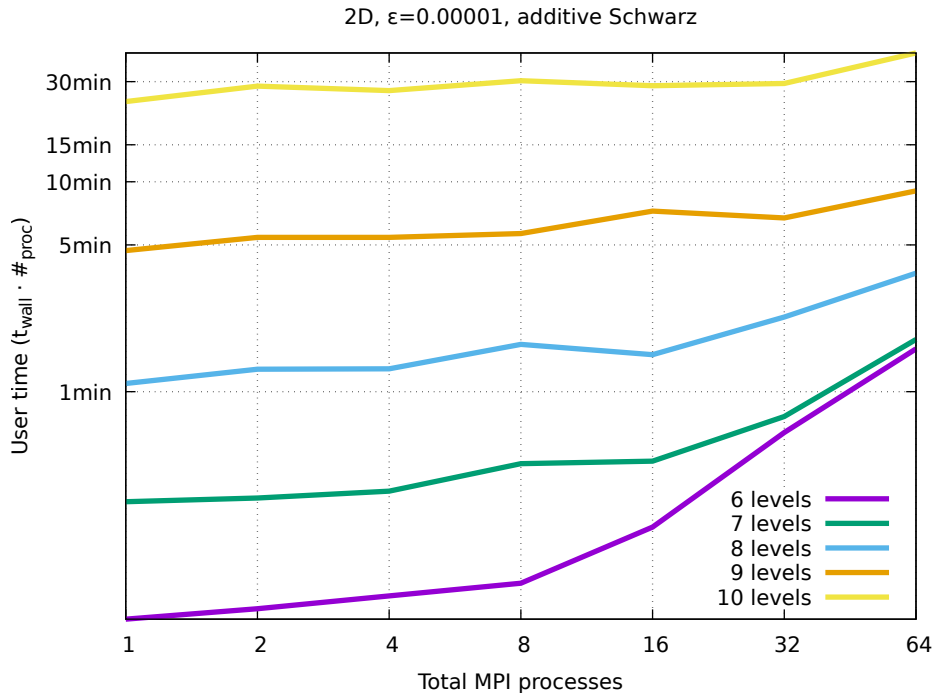


FIGURE 4.9. Scaling using additive Schwarz smoothers.

Results using additive Schwarz smoothers have a relatively good behavior due to the independence of the local problems in each smoothing step, therefore improving parallelization. We observe a flattening of the user time from 10 levels on.

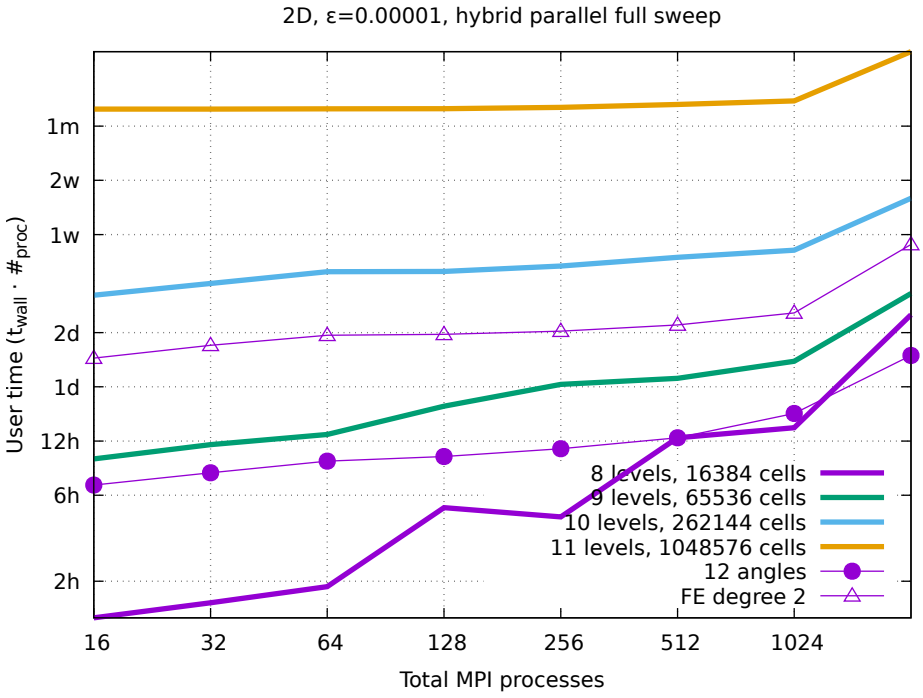


FIGURE 4.10. Scaling using parallel a full sweep smoother in the finest mesh and additive Schwarz smoothers in the coarser meshes.

Parallel full sweep smoothers only on the fine mesh show good results starting from 11 levels, the increase in the last step is due to an increase in the number of processes per node.

Even though the parallel full sweep has a relatively flat user time for fine enough meshes, we observe on the user time axis that the absolute time being used is significantly larger, this indicates that our implementation of the full sweep will be convenient in very hyperbolic cases where the iteration counts of the additive Schwarz method is very large.

4.1. Parallel implementation of the full sweep. The implementation of the full sweep when using distributed meshes is done following a procedure of determining local and global *generations* of cells that can be calculated in parallel.

We show the procedure in figure 4.11. We identify the first generation for a sweep (in this case a sweep in the top right direction) for each submesh. As it can be seen in the figure, more than one first generation cell can exist in the same submesh.

Following, we identify the next generation of cells by executing a recursive procedure of identifying a cell that depends on the result of the previous generation by its shared face, setting the next generation number in it, and re-execute the procedure on the new cell, until no shared faces exist (i.e. faces are shared with upwind cells or boundaries of the submesh).

Finally, we share the last and first generation of submeshes that are connected by faces in the downwind direction (last image in figure 4.11 to the right). We have then obtained a global ordering of cells and the generations that can be calculated in parallel.

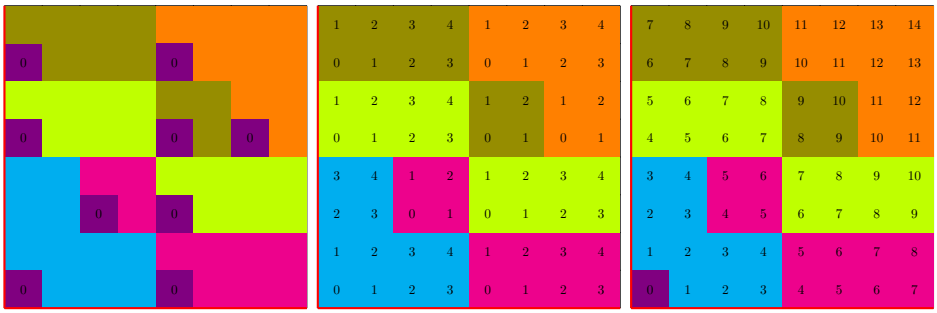


FIGURE 4.11. Parallelization of the full sweep in distributed meshes

5. Conclusions

We discussed various boundary models and resulting discretization options for radiation transport. In our results, the choice of the actual model has only a minor influence on discretization accuracy.

We showed in examples that the choice of stabilization parameters can have a considerable effect on the performance of non-overlapping Schwarz smoothers in a multigrid preconditioner. While the solver behaves robustly for a considerably wide range of parameters, a bad choice can cause deterioration of iteration counts to the point where the method becomes infeasible.

In a final discussion, we exhibited that our scheme performs as well in the case of nonisotropic scattering.

CHAPTER 5

Local Thermodynamic Equilibrium Transport

Abstract. With the knowledge developed in previous chapters concerning the efficiency of different algorithms to solve particle transport problems and their approximations, we apply the algorithms to the problem of Local Thermodynamic Equilibrium (LTE) particle transport including scattering. We implement a Newton solver, preconditioned with an additive Schwarz nonlinear preconditioner, to solve the multigroup transport equation with the LTE constraint.

1. Discretization

We retake the LTE transport equations developed in chapter 1 in steady state

$$\begin{aligned}&\boldsymbol{\Omega}\cdot\nabla I_\nu(\mathbf{x},\nu,\boldsymbol{\Omega})+\rho(\kappa_{\nu,s}+\kappa_{\nu,a})I_\nu(\mathbf{x},\nu,\boldsymbol{\Omega})\\&-\int_0^\infty\int_{4\pi}\rho\kappa_{\nu,s}(\nu'\to\nu,\boldsymbol{\Omega}'\to\boldsymbol{\Omega})I_\nu(\mathbf{x},\boldsymbol{\Omega}',\nu')d\boldsymbol{\Omega}'d\nu'-\rho\kappa_{\nu,a}B_\nu=S_\nu(\mathbf{x},\nu,\boldsymbol{\Omega}),\end{aligned}$$

where all the absorbed photons are re-emitted with a Planck's spectrum as follows

$$\int_0^\infty\int_{4\pi}\rho\kappa_{\nu,a}I_\nu d\boldsymbol{\Omega}d\nu-\int_0^\infty\int_{4\pi}\rho\kappa_{\nu,a}B_\nu d\boldsymbol{\Omega}d\nu=0,$$

$\kappa_{\nu,a}$ is the absorption opacity and $\kappa_{\nu,s}$ is the scattering opacity, with

$$\kappa_{\nu,s}(\mathbf{x},\nu,\boldsymbol{\Omega})=\int_0^\infty\int_{4\pi}\kappa_{\nu,s}(\nu\to\nu',\boldsymbol{\Omega}\to\boldsymbol{\Omega}')d\boldsymbol{\Omega}'d\nu'.$$

The equation including the LTE constraint provides closure to the system by having an equation for $T(\mathbf{x})$, we remark that the photons being absorbed are recovered in the first equation with a blackbody spectrum. In this manner, the temperature equation acts as a *loop*, similarly to the scattering phenomena, relocating photons from one frequency to the rest of possible frequencies, albeit with a predefined spectrum depending on temperature.

We apply a multigroup discretization to the frequency variable, using the definitions in chapter 1, §10 to obtain

$$\begin{aligned}\boldsymbol{\Omega}\cdot\nabla I_g+\rho(\kappa_{g,s}+\kappa_{g,a})I_g-\sum_{g=1}^G\int_{4\pi}\rho\kappa_{g'g,s}(\boldsymbol{\Omega}'\to\boldsymbol{\Omega})I_{g'}(\boldsymbol{\Omega}')d\boldsymbol{\Omega}'-\rho\kappa_{g,a}B_g=s_g\\\sum_{g=1}^G\int_{4\pi}\rho\kappa_{g,a}I_g d\boldsymbol{\Omega}-\sum_{g=1}^G\int_{4\pi}\rho\kappa_{g,a}B_g d\boldsymbol{\Omega}=0.\end{aligned}$$

Following the discrete ordinates collocation method we used to discretize $\mathbf{\Omega}$ we applied in chapter 4, we replace integrals with quadrature sums as follows

(63)
$$\mathbf{\Omega}_j \cdot \nabla I_{g,j} + \rho(\kappa_{g,j,s} + \kappa_{g,j,a}) I_{g,j} - \sum_{g=1}^{G}\sum_{j=1}^{m} \omega_j \rho \kappa_{g'g,j'j,s} I_{g',j'} - \rho \kappa_{g,j,a} B_g = s_{g,j}$$
$$\sum_{g=1}^{G}\sum_{j=1}^{m} \rho \kappa_{g,j,a} \omega_j I_{g,j} - \sum_{g=1}^{G}\sum_{j=1}^{m} \omega_j \rho \kappa_{g,j,a} B_g = 0,$$

where the subscript j represents the photon direction being considered.

The system is nonlinear in the temperature variable since Planck's law is

$$B(T(\mathbf{x}), \nu) = \frac{2h\nu^3}{c^2\left(e^{\frac{h\nu}{k_B T(\mathbf{x})}} - 1\right)},$$

where in the multigroup discretization we need the integral of Planck's law in a group:

$$B_g(T(\mathbf{x})) = \int_{\nu_{g+1}}^{\nu_g} \frac{2h\nu^3}{c^2\left(e^{\frac{h\nu}{k_B T(\mathbf{x})}} - 1\right)} d\nu.$$

To perform the integrals needed of Planck's law we use the following series [**64**]:

$$\int_x^\infty B_{(T,\nu)} d\nu = 2\frac{k^4 T^4}{h^3 c^2} \sum_{n=1}^{\infty} \left(\frac{x^3}{n} + \frac{3x^2}{n^2} + \frac{6x}{n^3} + \frac{6}{n^4}\right) e^{-nx}, \; x = \frac{h\nu}{kT}. \tag{64}$$

We consider the frequency and angle components of $I_{g,j}$ plus the temperature T stacked as follows:

$$v = (I_{1,1}(\mathbf{x}), I_{1,2}(\mathbf{x}) ... I_{2,1}(\mathbf{x}), I_{2,2}(\mathbf{x}) ... I_{g,j} ... I_{G,m}(\mathbf{x}), T(\mathbf{x}))^\intercal,$$
$$s = (s_{1,1}(\mathbf{x}), s_{1,2}(\mathbf{x}) ... s_{2,1}(\mathbf{x}), s_{2,2}(\mathbf{x}) ... s_{g,j} ... s_{G,m}(\mathbf{x}), 0)^\intercal.$$

The system (63) can then be represented as

$$A(v)v = s, \tag{65}$$

where we observe that the dependence on v of A is only given by the temperature block of v, which acts as another type of *redistribution* (aside from the scattering frequency and direction redistribution) that appears in every Newton step.

2. Nonlinear solver

In order to undertake the solution of the nonlinear system (65), we use a classical Newton method. The homogeneous nonlinear system reads

$$F(v) = A(v)v - s = 0,$$

where we identify the solution to the system with the intensity $v = I_\nu(\mathbf{x}, \nu, \mathbf{\Omega})$.

The Newton iteration can therefore be written as

$$\begin{aligned} v_{n+1} =& v_n - (F'(v_n))^{-1} F(v_n) \\ =& v_n + (A'(v_n))^{-1} (s - A(v_n)v_n), \end{aligned}$$

where $A'(v_n)$ is calculated using the derivative of formula (64) for Planck's law.

The evaluation of $(A'(v_n))^{-1}$ is of course a linear problem of the form

$$A'(v_n)\,(v_{n+1} - v_n) = s - A(v_n)v_n,$$

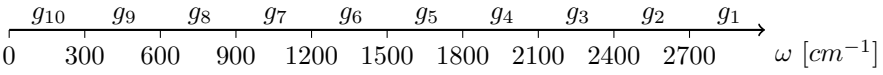
that we address by using the techniques developed in previous chapters. We apply a geometric multigrid preconditioned GMRES solver, using cell-wise Schwarz smoothers of additive or multiplicative (full sweep) type.

This approach allows us to solve a complete linear multigroup transport problem in each cell to calculate the inverse of $A'(v_n)$, therefore obtaining the performance we showed in previous chapter for these preconditioner types.

3. Numerical Experiments with Newton's method

We calculated a scaled LTE problem, similar to the experiments for the linear solvers in chapter 4 using $\rho = 1$, a source equal to 1000ε, an absorption opacity equal to ε and a scattering opacity equal to $\frac{1}{\varepsilon}$. The choice of an increased source is motivated by the necessity of keeping the temperature above a certain threshold in order for the series in equation (64) to provide meaningful results.

The 2D domain is $[0,1] \times [0,1]$, with 4 velocity directions and 10 frequency groups with the following grid:



where $\omega = \nu/c$ is the wavenumber and c is the speed of light.

We assume a full frequency redistribution in the sense that $\kappa_{gg',jj',s} = \kappa_{g'g,j'j,s}$, $\forall g, g', j, j'$.

The results in the tables below are the iteration counts to reduce the Newton's method residual by a factor 10^{-8} for different distributions of scattering and LTE absorption. The tolerance of the linear solver is a residual reduction of 10^{-10} for the coarse, local and global linear problems.

The temperature guess is a flat distribution at 500 K and the corresponding flux is also flat, at 100 $W/m^2/sr/cm^{-1}$ for each direction and frequency group.

We show the iteration count in the subsections below, linear iterations are shown on the left of the slash while nonlinear iterations are shown on the right. The experiments were performed using both an additive Schwarz and a multiplicative Schwarz (full sweep).

3.1. Additive Schwarz. The results for the use of an additive smoother for the linear problem can be seen in tables 5.1, 5.2 and 5.3.

Table 5.1 shows the case when $\kappa_a = 0$, where we retrieve the results of the linear problems we solved in chapter 4, i.e. linear transport problems. The behavior is satisfactory as long as the scattering is large enough and the problem has a more elliptic behavior. On the other hand, as we can see in the first column, when the problem has a transport component, comparable with scattering, and therefore a larger hyperbolic component, the additive Schwarz preconditioner requires a high amount of linear iterations for convergence.

Table 5.2 shows the result of including a relatively small amount of absorption into the system. We observe that the linear preconditioner still has a limitation when addressing problems with a large transport component, as it can be seen in the first column.

levels \ ε	1	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
2	6/1	8/1	9/1	7/1	8/1	9/1
3	13/1	13/1	16/1	15/1	15/1	15/1
4	21/1	16/1	20/1	20/1	20/1	20/1
5	31/1	22/1	21/1	22/1	22/1	22/1
6	50/1	30/1	19/1	23/1	23/1	23/1
7	79/1	44/1	17/1	23/1	23/1	23/1
8	$> 100/1$	66/1	19/1	22/1	23/1	23/1

TABLE 5.1. $\frac{\kappa_a}{\kappa_s+\kappa_a} = 0$

We note that we need more Newton iterations in order to reduce the residual, but less linear iterations per nonlinear solve. The inclusion of absorption into the system enlarges the diagonal of the redistribution matrix and shifts the field of values of the operator to the right of the complex plane, effectively making the linear preconditioner more effective.

levels \ ε	1	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
2	17/2	24/4	25/5	22/5	36/6	48/6
3	13/1	24/3	33/5	37/6	47/6	77/7
4	20/1	29/3	37/5	40/6	55/7	81/7
5	31/1	36/3	40/5	42/6	51/7	88/8
6	47/1	47/3	37/4	47/6	53/7	81/8
7	76/1	65/3	35/4	49/6	57/7	76/8
8	$> 100/1$	83/2	38/4	48/6	59/7	79/8

TABLE 5.2. $\frac{\kappa_a}{\kappa_s+\kappa_a} = 10^{-7}$

Table 5.3 shows result for a full LTE problem, we see a general increase in the nonlinear iterations and a slight reduction in the linear iterations per nonlinear solve. We observe that there is an increase in the nonlinear iterations with increasing absorption, which is expected from a larger nonlinear interaction, and a slight increase with refining as well.

levels \ ε	1	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
2	29/4	49/6	44/7	38/7	37/7	46/7
3	48/4	67/7	59/8	50/7	47/7	68/7
4	70/4	82/7	71/8	57/8	67/8	87/8
5	103/4	113/7	92/9	71/9	78/9	107/9
6	155/4	159/7	101/9	81/10	83/10	116/10
7	237/4	234/7	103/9	96/11	83/10	106/10
8	$> 100/1$	348/7	128/9	103/11	89/11	114/11

TABLE 5.3. $\frac{\kappa_a}{\kappa_s+\kappa_a} = 1$

The behavior of the solver is satisfactory in that the increase in nonlinear iterations with increasing absorption is relatively small. The total amount of linear iterations is, however, rather large, specially for small redistribution, with a hyperbolic behavior of the linear system.

To improve the behavior for the hyperbolic case we will apply the full sweep as a smoother, effectively reduce the amount of linear iterations needed per nonlinear solve. We shall do so keeping for the same cases in order to compare in section below.

3.2. Multiplicative Schwarz. The results for the use of the full sweep for the linear problem can be seen in tables 5.4, 5.5 and 5.6.

Table 5.4 shows the linear case for $\kappa_a = 0$, the iteration numbers are fully compatible with the ones obtained in chapter 4. We observe a very good behavior in terms of number of iterations for a hyperbolic, transport-dominant regime, but also the linear iterations are kept low for the elliptic, scattering dominated case.

levels \ ε	1	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
2	2/1	3/1	4/1	4/1	4/1	4/1
3	3/1	4/1	5/1	5/1	5/1	5/1
4	3/1	5/1	6/1	6/1	6/1	6/1
5	3/1	4/1	6/1	6/1	6/1	6/1
6	3/1	4/1	5/1	6/1	6/1	6/1
7	3/1	4/1	6/1	6/1	6/1	6/1
8	3/1	4/1	5/1	6/1	6/1	6/1

TABLE 5.4. $\frac{\kappa_a}{\kappa_s+\kappa_a} = 0$

Table 5.5 shows the results when including a relatively small amount of absorption into the system. The number of nonlinear iterations is kept with respect to the additive Schwarz smoother case, however the linear iterations per nonlinear solve are greatly reduced.

levels \ ε	1	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
2	3/2	6/4	8/5	8/5	14/6	16/6
3	3/1	6/3	9/5	11/6	15/6	24/7
4	3/1	7/3	10/5	11/6	17/7	24/7
5	3/1	6/3	10/5	11/6	15/7	21/8
6	3/1	6/3	9/4	13/6	15/7	21/8
7	3/1	6/3	9/4	13/6	16/7	22/8
8	3/1	5/2	8/4	12/6	16/7	21/8

TABLE 5.5. $\frac{\kappa_a}{\kappa_s+\kappa_a} = 10^{-7}$

Finally, table 5.6 shows the full LTE problem. We see a very strong reduction of the number of linear iterations with respect to the additive Schwarz smoother case. The number of nonlinear iterations is kept equal, as we have only improved the behavior of the linear solver.

levels \ ε	1	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
2	5/4	13/6	9/7	9/7	10/7	10/7
3	6/4	15/7	12/8	10/7	10/7	10/7
4	6/4	10/7	17/8	12/8	11/8	11/8
5	6/4	10/7	20/9	13/9	12/9	14/9
6	6/4	10/7	20/9	14/10	13/10	15/10
7	6/4	10/7	13/9	21/11	13/10	14/10
8	6/4	10/7	13/9	23/11	14/11	15/11

TABLE 5.6. $\frac{\kappa_a}{\kappa_s+\kappa_a} = 1$

The results with a Newton solver are quite satisfactory, especially for the full sweep smoother. We should note, however, that this is a relatively expensive preconditioner that is highly sequential, as the error reduction in one side of the domain is key for the error reduction elsewhere.

A part of the sweeps can indeed be parallelized, leaving the issue of load balancing for distributed meshes since the sweeps must pass by the whole domain. Some good results have been reported by Ghaddar and Ragusa [**65**] on the load balancing issue, however, the additive Schwarz smoother is still significantly cheaper for diffusive (high scattering) regimes.

In the next section we show a glance of the frequency spectrum obtained, in particular for the most interesting case of full LTE.

3.3. Energy spectrum. We show an example of the calculated frequency spectrum obtained in figure 5.1 for the case with no scattering, where we can compare directly with Planck's law.

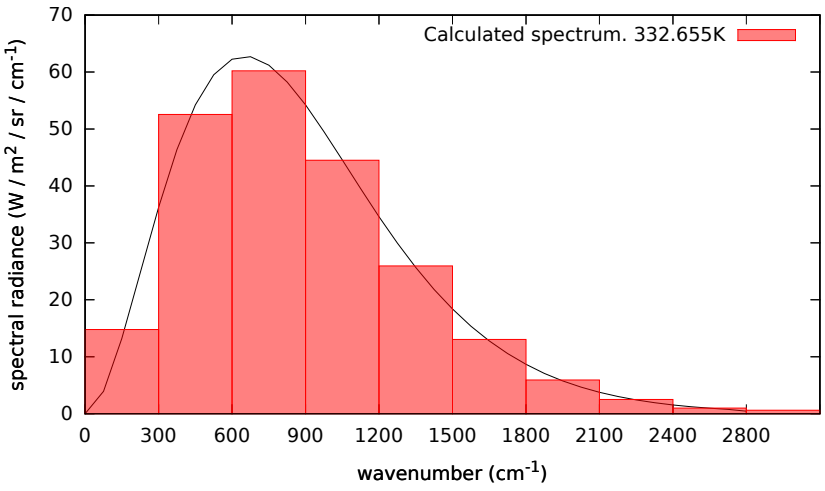


FIGURE 5.1. Calculated 10 group spectrum (red) and black body spectrum (black) for 332.655K and emissivity=1

4. Nonlinear preconditioner

Following the ideas in [**66**], we apply a nonlinear preconditioner to our algorithm similar to the additive Schwarz smoother we used for solving the linear system.

Using the same notation to denote the restriction and prolongation operators that we have been using throughout the text, defined in chapter 4,§4.2, we define the cell-wise local nonlinear systems as

$$\begin{aligned} v_{1,m+1} &= v_{1,m} + \left(A_1'(v_{1,m})\right)^{-1}\left(s - A_1(v_{1,m})v_{1,m}\right) \\ &\dots \\ v_{j,m+1} &= v_{j,m} + \left(A_j'(v_{j,m})\right)^{-1}\left(s - A_j(v_{j,m})v_{j,m}\right) \\ &\dots \\ v_{J,m+1} &= v_{J,m} + \left(A_J'(v_{J,m})\right)^{-1}\left(s - A_J(v_{J,m})v_{J,m}\right), \end{aligned}$$

where $v_{j,0} = \mathcal{R}_j v_n$, $A_j(v_{j,0}) = \mathcal{R}_j A(v_n)\mathcal{R}_j^\intercal$ and $A_j'(v_{j,0}) = \mathcal{R}_j A'(v_n)\mathcal{R}_j^\intercal$ and the subdomains are composed of one cell. After the first local Newton iteration, given that the local block of the system $A(v_n)$ only depends on $v_{j,0}$ we can calculate $A_j(v_{j,m})$ and $A_j'(v_{j,m})$ only by using $v_{j,m}$.

We remark that these local nonlinear problems are defined in an analogous way to the local linear smoothers used for the geometric multigrid method used before.

After the nonlinear preconditioning we apply a standard Newton iteration on the global space

$$v_{n+1} = v_n + \left(A'(v_n)\right)^{-1}\left(s - A(v_n)v_n\right).$$

The preconditioner, used in this manner, acts as a provider of an *educated* guess before the application of a global Newton step. We next describe the algorithm in more detail.

Algorithm 1 Nonlinear preconditioned Newton method

$n \leftarrow 0$
$r_n \leftarrow s - A(v_n)v_n$
while $\|r_n\| >$ tolerance **do**
  for $j = 1 \dots J$ **do** ▷ Start of the nonlinear preconditioning step
    $s_j \leftarrow \mathcal{R}_j s$
    $m \leftarrow 0$
    $r_{j,m} \leftarrow \mathcal{R}_j r_n$
    $v_{j,m} \leftarrow \mathcal{R}_j v_n + \left(A_j'(v_{j,m})\right)^{-1} r_{j,m}$
    while $\|r_{j,m}\| >$ tolerance **do** ▷ Local parallel Newton solvers
      $v_{j,m+1} = v_{j,m} + \left(A_j'(v_{j,m})\right)^{-1} r_{j,m}$
      $r_{j,m+1} \leftarrow s - A(v_{j,m+1})v_{j,m+1}$
      $m \leftarrow m + 1$
    $v_{j,\text{conv}} \leftarrow v_{j,m}$
  $v_n \leftarrow \sum_{j=1}^{J} \mathcal{R}_j^\intercal v_{j,\text{conv}}$ ▷ Cell-wise additive, nonoverlapping
  $v_{n+1} = v_n + \left(A'(v_n)\right)^{-1}\left(s - A(v_n)v_n\right)$ ▷ Global Newton step
  $r_{n+1} \leftarrow s - A(v_{n+1})v_{n+1}$
  $n \leftarrow n + 1$

In the following section we show some results with the preconditioned Newton iteration for the solution of the nonlinear problem.

5. Numerical Experiments with a preconditioned Newton's method

We calculated the same LTE problem as previously, with the unpreconditioned Newton method. We now use a non-linear preconditioner as explained in the previous section.

As it was used before, the tables show the number of linear iterations and nonlinear iterations as a function of ε and the number of levels (mesh refining) used for the linear geometric multigrid preconditioner.

We did not recalculate the case of $\frac{\kappa_a}{\kappa_s+\kappa_a} = 0$ since the problem is linear and the results are equal.

5.1. Additive Schwarz. The results for the use of an additive Schwarz method for the linear preconditioner are shown in tables 5.7 and 5.8.

Table 5.7 shows the case of $\frac{\kappa_a}{\kappa_s+\kappa_a} = 10^{-7}$, a problem having a relatively small nonlinearity. We observe a significant reduction of the nonlinear iterations with respect to the unpreconditioned case, accompanied by an important reduction on the linear iterations as well.

The linear preconditioner has a similar problem with cases where transport is dominant, as it can be seen in the last row.

levels \ ε	1	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
2	5/1	11/2	11/2	11/2	13/2	17/2
3	12/1	17/2	20/2	20/2	22/2	25/2
4	20/1	24/2	23/2	25/2	27/2	31/2
5	30/1	28/2	26/2	28/2	30/2	34/2
6	48/1	38/2	25/2	29/2	31/2	34/2
7	76/1	55/2	24/2	30/2	32/2	35/2
8	$> 100/1$	$> 100/1$	37/2	29/2	32/2	35/2

TABLE 5.7. $\frac{\kappa_a}{\kappa_s+\kappa_a} = 10^{-7}$

Table 5.8 shows the case of $\frac{\kappa_a}{\kappa_s+\kappa_a} = 1$, i.e. full LTE. We observe a slight increase on the linear iterations with respect to the $\frac{\kappa_a}{\kappa_s+\kappa_a} = 10^{-7}$ case. The number of iterations remain controlled, however, stabilizing with respect to the mesh refinement and covering a relatively large range of ε, while keeping a very low amount of nonlinear iterations.

The transport dominant case is still a problem for the linear solver, which justifies our analysis of the full sweep in the following section.

5.2. Multiplicative Schwarz. The results for the use of an full sweep multiplicative Schwarz method for the linear preconditioner are shown in tables 5.9 and 5.10.

Table 5.9 shows results for $\frac{\kappa_a}{\kappa_s+\kappa_a} = 10^{-7}$ for using the multiplicative full sweep smoother for our geometric multigrid linear preconditioner. We see a strong decrease in the linear iterations with respect to the additive Schwarz smoother in the

levels \ ε	1	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
2	14/2	17/2	16/2	14/2	16/2	19/2
3	25/2	24/2	24/2	22/2	26/2	34/2
4	38/2	28/2	28/2	27/2	32/2	40/2
5	57/2	38/2	32/2	32/2	37/2	46/2
6	86/2	53/2	32/2	33/2	39/2	48/2
7	134/2	80/2	30/2	35/2	39/2	46/2
8	$> 100/1$	$> 100/1$	35/2	34/2	37/2	46/2

TABLE 5.8. $\frac{\kappa_a}{\kappa_s+\kappa_a} = 1$

previous section, including the case for strong transport where it is particularly effective.

levels \ ε	1	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
2	2/1	5/2	5/2	5/2	6/2	6/2
3	3/1	5/2	6/2	6/2	7/2	8/2
4	3/1	6/2	7/2	7/2	8/2	9/2
5	3/1	5/2	7/2	7/2	8/2	9/2
6	3/1	5/2	7/2	7/2	8/2	9/2
7	3/1	5/2	7/2	7/2	8/2	9/2
8	3/1	5/2	6/2	7/2	8/2	8/2

TABLE 5.9. $\frac{\kappa_a}{\kappa_s+\kappa_a} 10^{-7}$

Finally, table 5.10 shows results for a full LTE case, again where it can be observed that the linear iterations are very well controlled by the full sweep smoother.

levels \ ε	1	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
2	3/2	6/2	5/2	5/2	5/2	7/2
3	4/2	6/2	6/2	6/2	6/2	7/2
4	4/2	6/2	7/2	7/2	7/2	8/2
5	4/2	5/2	7/2	7/2	7/2	8/2
6	4/2	5/2	7/2	7/2	7/2	8/2
7	4/2	5/2	6/2	8/2	7/2	8/2
8	4/2	5/2	6/2	8/2	7/2	7/2

TABLE 5.10. $\frac{\kappa_a}{\kappa_s+\kappa_a} = 1$

Results show that the nonlinear preconditioner is very effective for all the regimes tested, keeping nonlinear iterations very low. The linear iterations can be controlled depending on the relative size of the transport and scattering terms, with additive or multiplicative Schwarz methods.

All in all, the method shows very satisfactory results being able to deal with problems with a wide range of parameters. In the next section we test variable coefficients and rectangular domains, both in 2D and 3D.

6. Numerical Experiments with a density distribution

6.1. 2D case. In order to test the capacities of the LTE code, we set up a two-dimensional full LTE problem on a rectangle $[-500, 500] \times [-50, 50]$, representing a radial cut of a dust cloud, typically found in astrophysics applications (see [**67**]).

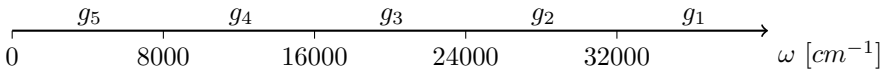
The density is represented by a Gaussian vertical profile and a power-law distribution for the radial profile as follows:

$$\rho(r,z) = \begin{cases} \frac{1}{\varepsilon} \frac{e^{-y^2}}{(0.15 \cdot 50)^2} & \text{for } x \leq 50 \\ \frac{1}{\varepsilon} \left(\frac{x}{50}\right)^{-2.5} \frac{e^{-y^2}}{(0.15x)^2} & \text{for } x > 50 \end{cases}.$$

We set up a spherically symmetric source, also with a Gaussian profile

$$s(r,z) = \varepsilon e^{-\left(\frac{\sqrt{x^2+y^2}}{50}\right)^2}.$$

We perform a calculation with a 12-angle Gauss-Legendre-Chebyshev (TGLC2) and 5 frequency groups, where the frequency mesh is



that we chose by performing a 1 group, 4 angles calculation to estimate a temperature guess. We use a flat flux and temperature guess of $4 \cdot 10^6 W/m^2/sr/cm^{-1}$ and $5500K$ respectively.

The results are plotted for selected groups and angles and including the temperature in figure 5.2

We show the iterations needed to converge in table 5.11 where we compare different smoother types. It can be observed that the full sweep reduces significantly the amount of linear iterations needed, while the nonlinear iterations remain constant.

smoother \ ε	1	10^{-1}	10^{-2}
additive	86/2	59/2	58/2
full sweep	7/2	8/2	8/2

TABLE 5.11. Linear/Nonlinear iterations for a 2D full LTE calculation with a density distribution, with 12 angles and 5 groups.

6.2. 3D case. We setup a 3D case of a calculation with a density distribution with the following shape:

$$\rho(r,z) = \begin{cases} \frac{1}{\varepsilon} \frac{e^{-x^2-y^2}}{(0.15 \cdot 50)^2} & \text{for } \sqrt{x^2+y^2} \leq 50 \\ \frac{1}{\varepsilon} \left(\frac{\sqrt{x^2+y^2}}{50}\right)^{-2.5} \frac{e^{-z^2}}{(0.15\sqrt{x^2+y^2})^2} & \text{for } \sqrt{x^2+y^2} > 50 \end{cases}.$$

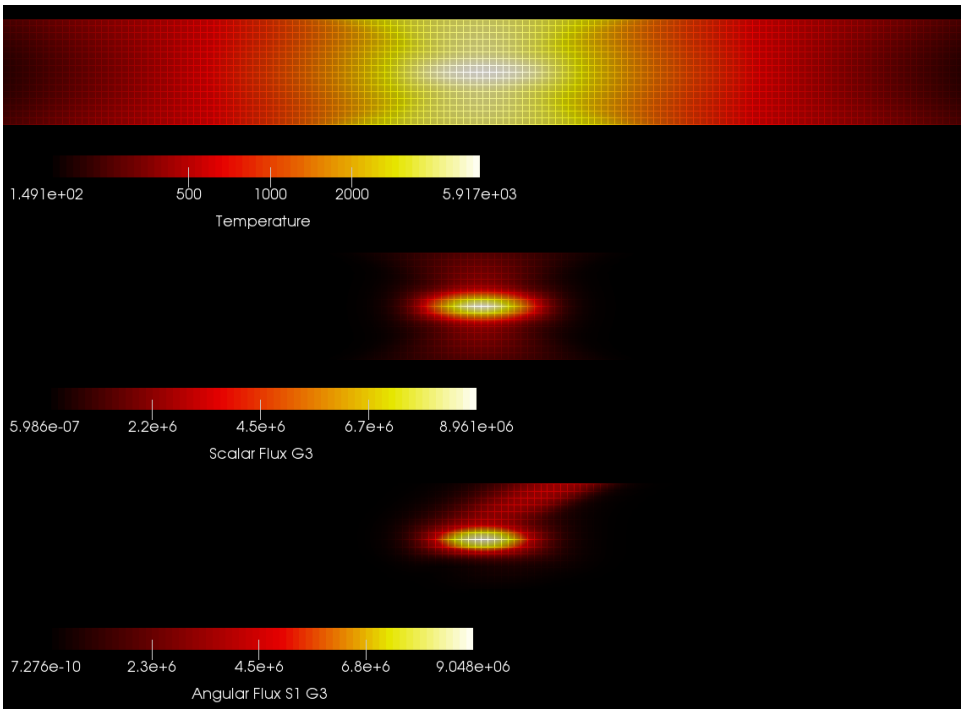


FIGURE 5.2. Temperature and fluxes for group g3 and angle s1 (0.87,0.36) for 2D case

We set up the same spherically symmetric source as in the 2D case

$$s(r,z) = \varepsilon e^{-\left(\frac{\sqrt{x^2+y^2+z^2}}{50}\right)^2}.$$

We use a flat flux and temperature guess of $1.2 \cdot 10^6 W/m^2/sr/cm^{-1}$ and $4500K$ respectively.

A glance to the solution obtained can be seen in figure 5.3.

We show the iterations needed to converge in table 5.12 for an additive Schwarz smoother. The difference seen for the case of $\varepsilon = 10^{-2}$ is produced by the inadequacy of the initial guess for the temperature, we kept the initial guess constant in order to compare between different regimes.

smoother \ ε	1	10^{-1}	10^{-2}
additive	64/2	58/2	95/2

TABLE 5.12. Linear/Nonlinear iterations for a 3D full LTE calculation with a density distribution, with 8 angles and 5 groups.

7. Remarks on implementation

The implementation of the linear solver is equivalent to the one described in chapter 4, with the exception that for these calculations we used iterative solvers

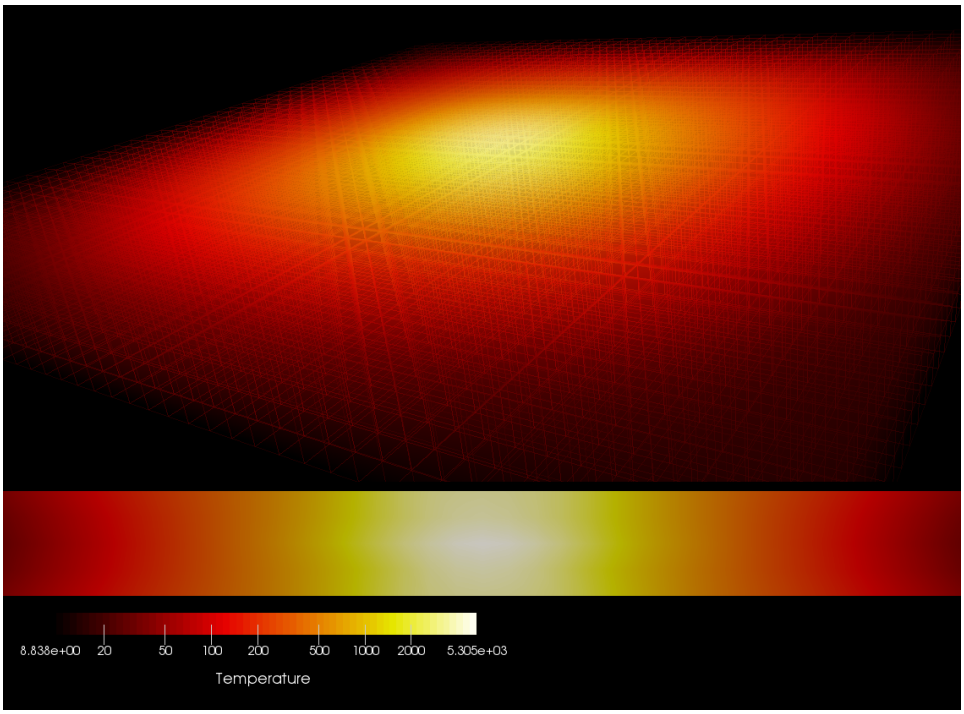


FIGURE 5.3. Temperature for a 3D case (above). Vertical cut through the center (below).

for the smoothers local problems as well, with an algebraic multigrid solver as the one used for the coarse space, given that the size of the local matrix made the LAPACK explicit solver too slow.

We performed tests using iterative solvers, using matrix-free local and coarse operators, therefore drastically reducing the memory footprint, but the problems become stiffer with mesh refining, which shows the need of a preconditioner for the local problems or approximate tensor-product preconditioners as the ones proposed in [**68**].

8. Conclusion

We have shown a discontinuous Galerkin, multigroup, discrete ordinates discretization of the radiative transfer equation in local thermodynamic equilibrium including scattering. We used the techniques developed in previous chapters to solver the linear transport equations with a discontinuous Galerkin discretization, preconditioned with a multigrid V-cycle using Schwarz smoothers of additive type for high scattering and of multiplicative type for high mean free path. We used a nonlinear preconditioner and achieved a drastic reduction of the nonlinear iterations needed for a Newton method to converge, while keeping linear iterations controlled.

We performed experiments with different density distributions, including cases with solutions having very high gradients and elongated domains, using typical density distribution shapes of dust around protoplanetary disks available from literature.

Results show that our combination of solvers and preconditioners is able control both the linear and nonlinear iteration counts for a large variety of cases, while allowing the parallelization of the code on distributed meshes and reducing the memory footprint by using matrix-free formulations of our operators.

APPENDIX A

Abstract convergence theory

1. Two level additive Schwarz

1.1. Properties of the projections.

1.1.1. *Operator norm estimates.*

LEMMA 1.1. (See [**35**], Lemma 11.6). Using previous definitions, for $v \in V$

$$\frac{1}{C_V}\|v\|_{\mathcal{A}}^2 \leq \sum_{i=0}^{N} \|\mathcal{P}_i v\|_{\mathcal{A}}^2 \leq (\rho(\Theta)+1)\,\|v\|_V^2$$

PROOF. An upper bound for $\|\mathcal{P}_0 v\|_V^2$ is obtained using coercivity and continuity of $\mathcal{A}(\cdot,\cdot)$ as follows

$$\|\mathcal{P}_0 v\|_{\mathcal{A}}^2 \leq \mathcal{A}(\mathcal{P}_0 v, \mathcal{P}_0 v) = \mathcal{A}_0(\widetilde{\mathcal{P}}_0 v, \widetilde{\mathcal{P}}_0 v) = \mathcal{A}(v, \mathcal{P}_0 v) \leq \|v\|_{\mathcal{A}} \|\mathcal{P}_0 v\|_{\mathcal{A}}$$

therefore $\|\mathcal{P}_0 v\|_{\mathcal{A}}^2 \leq \|v\|_{\mathcal{A}}^2$.

Then we estimate the rest of the sum by using coercivity and continuity of $a(\cdot,\cdot)$ as well

$$\begin{aligned}
\sum_{i=1}^{N} \|\mathcal{P}_i v\|_{\mathcal{A}}^2 \leq & \sum_{i=1}^{N} \mathcal{A}(\mathcal{P}_i v, \mathcal{P}_i v) = \sum_{i=1}^{N} \mathcal{A}_i(\widetilde{\mathcal{P}}_i v, \widetilde{\mathcal{P}}_i v) = \mathcal{A}\left(v, \sum_{i=1}^{N} \mathcal{P}_i v\right) \\
\leq & \|v\|_{\mathcal{A}} \left\| \sum_{i=1}^{N} \mathcal{P}_i v \right\|_{\mathcal{A}} \leq \|v\|_{\mathcal{A}} \mathcal{A}\left(\sum_{i=1}^{N} \mathcal{P}_i v, \sum_{j=1}^{N} \mathcal{P}_j v\right)^{\frac{1}{2}} \\
\leq & \|v\|_{\mathcal{A}} \left(\sum_{i,j=1}^{N} \mathcal{A}(\mathcal{P}_i v, \mathcal{P}_j v)\right)^{\frac{1}{2}} \leq \|v\|_{\mathcal{A}} \left(\sum_{i,j=1}^{N} \theta_{ij} \|P_i v\|_{\mathcal{A}} \|P_j v\|_{\mathcal{A}}\right)^{\frac{1}{2}} \\
\leq & \rho(\Theta)^{\frac{1}{2}} \|v\|_{\mathcal{A}} \left(\sum_{i=1}^{N} \|\mathcal{P}_i v\|_{\mathcal{A}}^2\right)^{\frac{1}{2}}
\end{aligned}$$

therefore $\sum_{i=1}^{N} \|\mathcal{P}_i v\|_{\mathcal{A}}^2 \leq \rho(\Theta)\|v\|_{\mathcal{A}}^2$.

Following, we obtain a lower bound

$$\begin{aligned}\|v\|_{\mathcal{A}}^2 \leq & \mathcal{A}(v,v) = \mathcal{A}(v, \sum_{i=0}^{N} \mathcal{R}_i^\intercal v_i) = \sum_{i=0}^{N} \mathcal{A}_i(\widetilde{\mathcal{P}}_i v, v_i) = \sum_{i=0}^{N} \mathcal{A}(\mathcal{P}_i v, \mathcal{R}_i^\intercal v_i) \\ & \leq \sum_{i=0}^{N} \|\mathcal{P}_i v\|_{\mathcal{A}} \|\mathcal{R}_i^\intercal v_i\|_{\mathcal{A}} \leq \left(\sum_{i=0}^{N} \|\mathcal{P}_i v\|_{\mathcal{A}}^2 \right)^{\frac{1}{2}} \left(\sum_{i=0}^{N} \|\mathcal{R}_i^\intercal v_i\|_{\mathcal{A}}^2 \right)^{\frac{1}{2}} \\ & \leq C_V^{\frac{1}{2}} \left(\sum_{i=0}^{N} \|\mathcal{P}_i v\|_{\mathcal{A}}^2 \right)^{\frac{1}{2}} \|v\|_{\mathcal{A}}\end{aligned}$$

therefore $\frac{1}{C_V}\|v\|_{\mathcal{A}}^2 \leq \sum_{i=0}^{N} \|\mathcal{P}_i v\|_{\mathcal{A}}^2$. □

1.1.2. *Eigenvalue estimates.*

Lemma 1.2. Using previous definitions, the following estimate holds for all $v \in V$

$$\mathcal{A}(\mathcal{P}_i v, v) \geq \|\mathcal{P}_i v\|_{\mathcal{A}}^2$$

Proof. We estimate the projections locally as follows

$$\begin{aligned}\mathcal{A}(\mathcal{P}_i v, \mathcal{P}_i v) = & \mathcal{A}_i(\widetilde{\mathcal{P}}_i v_i, \widetilde{\mathcal{P}}_i v) = \mathcal{A}(v, \mathcal{P}_i v) \\ \mathcal{A}(v, \mathcal{P}_i v) = & \mathcal{A}(\mathcal{P}_i v, \mathcal{P}_i v) \geq \|\mathcal{P}_i v\|_{\mathcal{A}}^2\end{aligned}$$

hence the result. □

1.2. Proof of theorem 4.1.

Proof. We use the estimates developed in appendix 1.1. We begin with C_{ad}

$$\|\mathcal{P}_{\text{ad}} v\|_a^2 = \left\| \sum_{i=0}^{N} \mathcal{P}_i v \right\|_{\mathcal{A}}^2 \leq \sum_{i=0}^{N} \|\mathcal{P}_i v\|_{\mathcal{A}}^2 \leq (\rho(\Theta) + 1) \|v\|_{\mathcal{A}}^2$$

We have therefore $C_{\text{ad}} \leq \rho(\Theta) + 1$.

For the second estimate we estimate the projections locally as follows

$$\mathcal{A}(v, \sum_{i=0}^{N} \mathcal{P}_i v) = \sum_{i=0}^{N} \mathcal{A}(v, \mathcal{P}_i v) \geq \sum_{i=0}^{N} \|\mathcal{P}_i v\|_{\mathcal{A}}^2 \geq \frac{1}{C_V} \|v\|_{\mathcal{A}}^2$$

Therefore $c_{\text{ad}} \geq \frac{1}{C_V}$. □

2. Multigrid

2.1. Proof of theorem 4.4.

Proof. The error after one V-cycle is

$$\begin{aligned}v^{k+1} - v_L &= v^k - v_L + \mathcal{M}_L(\mathcal{A}_L v_L - \mathcal{A}_L v^k) \\ e^{k+1} &= e^k - \mathcal{M}_L \mathcal{A}_L e^k\end{aligned}$$

$$e^{k+1} = (\mathcal{I} - \mathcal{M}_L \mathcal{A}_L) e^k$$

Furthermore, we can express the error in each step of the multigrid iteration

(1) Pre-smoothing: begin with $x_0 = 0$ and let

$$e_1 = (\mathcal{I} - \mathcal{B}_\ell \mathcal{A}_\ell) e_0,$$

(2) Coarse grid correction:

$$e_2 = (\mathcal{I} - \mathcal{Q}_{\ell-1}^\intercal \mathcal{M}_{\ell-1} \mathcal{Q}_{\ell-1} \mathcal{A}_\ell) e_1$$

(3) Post-smoothing:

$$e_3 = (\mathcal{I} - \mathcal{B}_\ell \mathcal{A}_\ell) e_2$$

In each level of the V-cycle, a compact expression of the action of the multigrid iteration on each level can be written as:

$$e_\ell^{k+1} = (\mathcal{I} - \mathcal{M}_\ell \mathcal{A}_\ell) e_\ell^k = \mathcal{K}_\ell (\mathcal{I} - \mathcal{Q}_{\ell-1}^\intercal \mathcal{M}_{\ell-1} \mathcal{A}_{\ell-1} \mathcal{P}_{\ell-1}) \mathcal{K}_\ell e_\ell^k. \tag{66}$$

We define $\mathcal{K}_\ell = \mathcal{I} - \mathcal{B}_\ell \mathcal{A}_\ell$ as the pre and post-smoothing operators respectively. We want to show by induction on ℓ that

$$(1 - C_{\mathrm{MG}})\, \mathcal{A}_\ell (v_\ell, v_\ell) \leq \mathcal{A}_\ell (v_\ell, (\mathcal{I} - \mathcal{M}_\ell \mathcal{A}_\ell)\, v_\ell) \leq (1 - c_{\mathrm{MG}})\, \mathcal{A}_\ell (v_\ell, v_\ell)$$

For $\ell = 0$ it is trivial since $\mathcal{M}_0 = \mathcal{A}_0^{-1} \implies \mathcal{A}_0 (v_\ell, (\mathcal{I} - \mathcal{M}_0 A_0)\, v_\ell) = 0$. We suppose that the result holds for $\ell - 1$ and using equation (66):

$$\begin{aligned}
(\mathcal{I} - \mathcal{M}_\ell \mathcal{A}_\ell) &= \mathcal{K}_\ell \left[\mathcal{I} - \mathcal{Q}_{\ell-1}^\intercal \mathcal{M}_{\ell-1} \mathcal{A}_{\ell-1} \mathcal{P}_{\ell-1}\right] \mathcal{K}_\ell^1 \\
&= \mathcal{K}_\ell \left[\left(\mathcal{I} - \mathcal{Q}_{\ell-1}^\intercal \mathcal{P}_{\ell-1}\right) + \mathcal{Q}_{\ell-1}^\intercal \left(\mathcal{I} - \mathcal{M}_{\ell-1} \mathcal{A}_{\ell-1}\right) \mathcal{P}_{\ell-1}\right] \mathcal{K}_\ell
\end{aligned}$$

Using this expression we have

$$\begin{aligned}
\mathcal{A}_\ell (v_\ell, (\mathcal{I} - \mathcal{M}_\ell \mathcal{A}_\ell)\, v_\ell) =& \mathcal{A}_\ell \left(v_\ell, \mathcal{K}_\ell \left[\mathcal{I} - \mathcal{Q}_{\ell-1}^\intercal \mathcal{P}_{\ell-1}\right] \mathcal{K}_\ell v_\ell\right) \\
&+ \mathcal{A}_\ell \left(v_\ell, \mathcal{K}_\ell \mathcal{Q}_{\ell-1}^\intercal \left[\mathcal{I} - \mathcal{M}_{\ell-1} \mathcal{A}_{\ell-1}\right] \mathcal{P}_{\ell-1} \mathcal{K}_\ell v_\ell\right) \\
=& \mathcal{A}_\ell \left(\mathcal{K}_\ell v_\ell, \left[\mathcal{I} - \mathcal{Q}_{\ell-1}^\intercal \mathcal{P}_{\ell-1}\right] \mathcal{K}_\ell v_\ell\right) \\
&+ \mathcal{A}_{\ell-1} \left(\mathcal{P}_{\ell-1} \mathcal{K}_\ell v_\ell, \left[\mathcal{I} - \mathcal{M}_{\ell-1} \mathcal{A}_{\ell-1}\right] \mathcal{P}_{\ell-1} \mathcal{K}_\ell v_\ell\right)
\end{aligned}$$

Observing that the operators $\mathcal{Q}_{\ell-1}\mathcal{P}_{\ell-1}$ and $\mathcal{M}_{\ell-1}\mathcal{P}_{\ell-1}$ are $\mathcal{A}_\ell$-self adjoint and idempotent, then $0 < \mathcal{A}_\ell (v_\ell, (\mathcal{I} - \mathcal{M}_\ell \mathcal{A}_\ell)\, v_\ell)$ by induction and therefore $C_{\mathrm{MG}} \leq 1$.

In regards to the upper estimate we have

$$\begin{aligned}
\mathcal{A}_\ell (v_\ell, (\mathcal{I} - \mathcal{M}_\ell \mathcal{A}_\ell)\, v_\ell) \leq& \mathcal{A}_\ell \left(\mathcal{K}_\ell v_\ell, \left[\mathcal{I} - \mathcal{Q}_{\ell-1}^\intercal \mathcal{P}_{\ell-1}\right] \mathcal{K}_\ell v_\ell\right) \\
&+ (1 - c_{\mathrm{MG}})\, \mathcal{A}_{\ell-1} \left(\mathcal{P}_{\ell-1} \mathcal{K}_\ell v_\ell, \mathcal{P}_{\ell-1} \mathcal{K}_\ell v_\ell\right) \\
=& \mathcal{A}_\ell \left(\mathcal{K}_\ell v_\ell, \left[\mathcal{I} - \mathcal{Q}_{\ell-1}^\intercal \mathcal{P}_{\ell-1}\right] \mathcal{K}_\ell v_\ell\right) \\
&+ (1 - c_{\mathrm{MG}})\, \mathcal{A}_\ell \left(\mathcal{K}_\ell v_\ell, \mathcal{Q}_{\ell-1}^\intercal \mathcal{P}_{\ell-1} \mathcal{K}_\ell v_\ell\right) \\
=& \mathcal{A}_\ell \left(\mathcal{K}_\ell v_\ell, \left[\mathcal{I} - \mathcal{Q}_{\ell-1}^\intercal \mathcal{P}_{\ell-1}\right] \mathcal{K}_\ell v_\ell\right) \\
&- (1 - c_{\mathrm{MG}})\, \mathcal{A}_\ell \left(\mathcal{K}_\ell v_\ell, \mathcal{K}_\ell v_\ell\right) + (1 - c_{\mathrm{MG}})\, \mathcal{A}_\ell \left(\mathcal{K}_\ell v_\ell, \mathcal{K}_\ell v_\ell\right) \\
&+ (1 - c_{\mathrm{MG}})\, \mathcal{A}_\ell \left(\mathcal{K}_\ell v_\ell, \mathcal{Q}_{\ell-1}^\intercal \mathcal{P}_{\ell-1} \mathcal{K}_\ell v_\ell\right) \\
=& \mathcal{A}_\ell \left(\mathcal{K}_\ell v_\ell, \left[\mathcal{I} - \mathcal{Q}_{\ell-1}^\intercal \mathcal{P}_{\ell-1}\right] \mathcal{K}_\ell v_\ell\right) \\
&- (1 - c_{\mathrm{MG}})\, \mathcal{A}_\ell \left(\mathcal{K}_\ell v_\ell, \left[\mathcal{I} - \mathcal{Q}_{\ell-1}^\intercal \mathcal{P}_{\ell-1}\right] \mathcal{K}_\ell v_\ell\right) \\
&+ (1 - c_{\mathrm{MG}})\, \mathcal{A}_\ell \left(\mathcal{K}_\ell v_\ell, \mathcal{K}_\ell v_\ell\right) \\
\mathcal{A}_\ell (v_\ell, (\mathcal{I} - \mathcal{M}_\ell \mathcal{A}_\ell)\, v_\ell) \leq& c_{\mathrm{MG}} \mathcal{A}_\ell \left(\mathcal{K}_\ell v_\ell, \left[\mathcal{I} - \mathcal{Q}_{\ell-1}^\intercal \mathcal{P}_{\ell-1}\right] \mathcal{K}_\ell v_\ell\right) \\
&+ (1 - c_{\mathrm{MG}})\, \mathcal{A}_\ell \left(\mathcal{K}_\ell v_\ell, \mathcal{K}_\ell v_\ell\right)
\end{aligned} \tag{67}$$

then

$$\mathcal{A}_\ell\left(\mathcal{K}_\ell v_\ell, \left[\mathcal{I}-\mathcal{Q}_{\ell-1}^{\intercal}\mathcal{P}_{\ell-1}\right]\mathcal{K}_\ell v_\ell\right) = \\ \mathcal{A}_\ell\left(\left[\mathcal{I}-\mathcal{Q}_{\ell-1}^{\intercal}\mathcal{P}_{\ell-1}\right]\mathcal{K}_\ell v_\ell, \left[\mathcal{I}-\mathcal{Q}_{\ell-1}^{\intercal}\mathcal{P}_{\ell-1}\right]\mathcal{K}_\ell v_\ell\right),$$

using the Cauchy-Schwarz inequality we obtain

$$\mathcal{A}_\ell\left(\mathcal{K}_\ell v_\ell, \left[\mathcal{I}-\mathcal{Q}_{\ell-1}^{\intercal}\mathcal{P}_{\ell-1}\right]\mathcal{K}_\ell v_\ell\right) \leq \\ \mathcal{A}_\ell\left(\mathcal{K}_\ell v_\ell, \mathcal{K}_\ell v_\ell\right)^{\frac{1}{2}} \mathcal{A}_\ell\left(\left[\mathcal{I}-\mathcal{Q}_{\ell-1}^{\intercal}\mathcal{P}_{\ell-1}\right]\mathcal{K}_\ell v_\ell, \left[\mathcal{I}-\mathcal{Q}_{\ell-1}^{\intercal}\mathcal{P}_{\ell-1}\right]\mathcal{K}_\ell v_\ell\right)^{\frac{1}{2}},$$

therefore

$$\begin{aligned}\mathcal{A}_\ell\left(\mathcal{K}_\ell v_\ell, \left[\mathcal{I}-\mathcal{Q}_{\ell-1}^{\intercal}\mathcal{P}_{\ell-1}\right]\mathcal{K}_\ell v_\ell\right) \leq & \mathcal{A}_\ell\left(\mathcal{K}_\ell v_\ell, \mathcal{K}_\ell v_\ell\right) \\ \leq & \frac{1}{c_S}\mathcal{A}_\ell\left(\mathcal{K}_\ell v_\ell, \mathcal{B}_\ell A_\ell \mathcal{K}_\ell v_\ell\right) \\ = & \frac{1}{c_S}\mathcal{A}_\ell\left(\mathcal{K}_\ell v_\ell, \left(\mathcal{I}-\mathcal{K}_\ell\right)\mathcal{K}_\ell v_\ell\right).\end{aligned}$$

The hypothesis on the smoother implies that the spectrum of $\mathcal{K}_\ell$ is contained in $[0,1]$, and therefore using the inequality of arithmetic and geometric means we obtain $\mathcal{K}_\ell^2 \leq \sqrt{\mathcal{K}_\ell} \leq \frac{\mathcal{I}+\mathcal{K}_\ell}{2} = \frac{\mathcal{I}-\mathcal{K}_\ell^2}{\mathcal{I}-\mathcal{K}_\ell}$, hence we deduce

$$\begin{aligned}\mathcal{A}_\ell\left(v_\ell, \left(\mathcal{I}-\mathcal{M}_\ell\mathcal{A}_\ell\right)v_\ell\right) \leq & \frac{c_{\mathrm{MG}}}{c_S}\mathcal{A}_\ell\left(v_\ell, \left(\mathcal{I}-\mathcal{K}_\ell\right)\mathcal{K}_\ell^2 v_\ell\right) \\ & + \left(1-c_{\mathrm{MG}}\right)\mathcal{A}_\ell\left(\mathcal{K}_\ell v_\ell, \mathcal{K}_\ell v_\ell\right) \\ \leq & \frac{c_{\mathrm{MG}}}{2c_S}\mathcal{A}_\ell\left(v_\ell, \left[\mathcal{I}-\mathcal{K}_\ell^2\right]v_\ell\right) \\ & + \left(1-c_{\mathrm{MG}}\right)\mathcal{A}_\ell\left(\mathcal{K}_\ell v_\ell, \mathcal{K}_\ell v_\ell\right) \\ = & \frac{c_{\mathrm{MG}}}{2c_S}\mathcal{A}_\ell\left(v_\ell, v_\ell\right) \\ & + \left[\left(1-c_{\mathrm{MG}}\right) - \frac{c_{\mathrm{MG}}}{2c_S}\right]\mathcal{A}_\ell\left(\mathcal{K}_\ell v_\ell, \mathcal{K}_\ell v_\ell\right)\end{aligned}$$

The result follows by choosing $c_{\mathrm{MG}} = \dfrac{2c_S}{1+2c_S} \leq 1$. □

APPENDIX B

Fourier analysis matrices

Inverse of the diagonal

$$\widehat{D}_j^{-1} = \frac{1}{12\delta_1^2\delta_2^2\varepsilon^2 + 8\delta_1^2\delta_2\varepsilon h\sigma_2 + \delta_1^2h^2\sigma_2^2 + 8\delta_1\delta_2^2\varepsilon h\sigma_1 + 2\delta_1\delta_2h^2\sigma_1\sigma_2 + \delta_2^2h^2\sigma_1^2}$$

$$\begin{pmatrix} 12\delta_1\delta_2^2\varepsilon^2 + 8\delta_1\delta_2\varepsilon h\sigma_2 + \delta_1h^2\sigma_2^2 + 4\delta_2^2\varepsilon h\sigma_1 + \delta_2h^2\sigma_1\sigma_2 & h\sigma_2\left(4\delta_1\delta_2\varepsilon + \delta_1h\sigma_2 + \delta_2h\sigma_1\right) & -2\delta_2^2\varepsilon h\sigma_1 & 2\delta_1\delta_2\varepsilon h\sigma_2 \\ h\sigma_1\left(4\delta_1\delta_2\varepsilon + \delta_1h\sigma_2 + \delta_2h\sigma_1\right) & 12\delta_1^2\delta_2\varepsilon^2 + 4\delta_1^2\varepsilon h\sigma_2 + 8\delta_1\delta_2\varepsilon h\sigma_1 + \delta_1h^2\sigma_1\sigma_2 + \delta_2h^2\sigma_1^2 & 2\delta_1\delta_2\varepsilon h\sigma_1 & -2\delta_1^2\varepsilon h\sigma_2 \\ -2\delta_2^2\varepsilon h\sigma_1 & 2\delta_1\delta_2\varepsilon h\sigma_2 & 12\delta_1\delta_2^2\varepsilon^2 + 8\delta_1\delta_2\varepsilon h\sigma_2 + \delta_1h^2\sigma_2^2 + 4\delta_2^2\varepsilon h\sigma_1 + \delta_2h^2\sigma_1\sigma_2 & h\sigma_2\left(4\delta_1\delta_2\varepsilon + \delta_1h\sigma_2 + \delta_2h\sigma_1\right) \\ 2\delta_1\delta_2\varepsilon h\sigma_1 & -2\delta_1^2\varepsilon h\sigma_2 & h\sigma_1\left(4\delta_1\delta_2\varepsilon + \delta_1h\sigma_2 + \delta_2h\sigma_1\right) & 12\delta_1^2\delta_2\varepsilon^2 + 4\delta_1^2\varepsilon h\sigma_2 + 8\delta_1\delta_2\varepsilon h\sigma_1 + \delta_1h^2\sigma_1\sigma_2 + \delta_2h^2\sigma_1^2 \end{pmatrix}$$

Iteration matrix

$$\widehat{D}_j^{-1}\left(-\widehat{L}_j - \widehat{U}_j\right) = \frac{1}{24\delta_1^2\delta_2^2\varepsilon^2h + 16\delta_1^2\delta_2\varepsilon h^2\sigma_2 + 2\delta_1^2h^3\sigma_2^2 + 16\delta_1\delta_2^2\varepsilon h^2\sigma_1 + 4\delta_1\delta_2h^3\sigma_1\sigma_2 + 2\delta_2^2h^3\sigma_1^2}$$

$$\begin{pmatrix} -4\delta_2^2\varepsilon h\sigma_1\left(\delta_1h-1\right)e^{-ih\omega} + \left(24\delta_1\delta_2^2\varepsilon^2 + 16\delta_1\delta_2\varepsilon h\sigma_2 + 2\delta_1h^2\sigma_2^2 + 8\delta_2^2\varepsilon h\sigma_1 + 2\delta_2h^2\sigma_1\sigma_2\right)\cos\left(h\omega\right) & 4\delta_1\delta_2^2\varepsilon h^2\sigma_2e^{-ih\omega} + 4\delta_1\delta_2\varepsilon h\sigma_2e^{ih\omega} + \left(2\delta_1h^2\sigma_2^2 + 2\delta_2h^2\sigma_1\sigma_2\right)\cos\left(h\omega\right) & -4\delta_2^2\varepsilon h\sigma_1\cos\left(h\omega\right) + \left(\delta_1h-1\right)\left(24\delta_1\delta_2^2\varepsilon^2 + 16\delta_1\delta_2\varepsilon h\sigma_2 + 2\delta_1h^2\sigma_2^2 + 8\delta_2^2\varepsilon h\sigma_1 + 2\delta_2h^2\sigma_1\sigma_2\right)e^{ih\omega} & 4\delta_1\delta_2\varepsilon h\sigma_2\cos\left(h\omega\right) + \left(\delta_1h-1\right)\left(8\delta_1\delta_2\varepsilon h\sigma_2 + 2\delta_1h^2\sigma_2^2 + 2\delta_2h^2\sigma_1\sigma_2\right)e^{ih\omega} \\ 4\delta_1^2\delta_2\varepsilon h^2\sigma_1e^{-ih\omega} + 4\delta_1\delta_2\varepsilon h\sigma_1e^{ih\omega} + \left(2\delta_1h^2\sigma_1\sigma_2 + 2\delta_2h^2\sigma_1^2\right)\cos\left(h\omega\right) & -4\delta_1^2\varepsilon h\sigma_2\left(\delta_2h-1\right)e^{-ih\omega} + \left(24\delta_1^2\delta_2\varepsilon^2 + 8\delta_1^2\varepsilon h\sigma_2 + 16\delta_1\delta_2\varepsilon h\sigma_1 + 2\delta_1h^2\sigma_1\sigma_2 + 2\delta_2h^2\sigma_1^2\right)\cos\left(h\omega\right) & 4\delta_1\delta_2\varepsilon h^2\sigma_1\cos\left(h\omega\right) + \left(\delta_1h-1\right)\left(8\delta_1\delta_2\varepsilon h\sigma_1 + 2\delta_1h^2\sigma_1\sigma_2 + 2\delta_2h^2\sigma_1^2\right)e^{ih\omega} & -4\delta_1^2\varepsilon h\sigma_2\cos\left(h\omega\right) + \left(\delta_2h-1\right)\left(24\delta_1^2\delta_2\varepsilon^2 + 8\delta_1^2\varepsilon h\sigma_2 + 16\delta_1\delta_2\varepsilon h\sigma_1 + 2\delta_1h^2\sigma_1\sigma_2 + 2\delta_2h^2\sigma_1^2\right)e^{ih\omega} \\ -4\delta_2^2\varepsilon h\sigma_1\cos\left(h\omega\right) + \left(\delta_1h-1\right)\left(24\delta_1\delta_2^2\varepsilon^2 + 16\delta_1\delta_2\varepsilon h\sigma_2 + 2\delta_1h^2\sigma_2^2 + 8\delta_2^2\varepsilon h\sigma_1 + 2\delta_2h^2\sigma_1\sigma_2\right)e^{-ih\omega} & 4\delta_1\delta_2\varepsilon h\sigma_2\cos\left(h\omega\right) + \left(\delta_2h-1\right)\left(8\delta_1\delta_2\varepsilon h\sigma_2 + 2\delta_1h^2\sigma_2^2 + 2\delta_2h^2\sigma_1\sigma_2\right)e^{-ih\omega} & -4\delta_2^2\varepsilon h\sigma_1\left(\delta_1h-1\right)e^{ih\omega} + \left(24\delta_1\delta_2^2\varepsilon^2 + 16\delta_1\delta_2\varepsilon h\sigma_2 + 2\delta_1h^2\sigma_2^2 + 8\delta_2^2\varepsilon h\sigma_1 + 2\delta_2h^2\sigma_1\sigma_2\right)\cos\left(h\omega\right) & 4\delta_1\delta_2^2\varepsilon h^2\sigma_2e^{ih\omega} + 4\delta_1\delta_2\varepsilon h\sigma_2e^{-ih\omega} + \left(2\delta_1h^2\sigma_2^2 + 2\delta_2h^2\sigma_1\sigma_2\right)\cos\left(h\omega\right) \\ 4\delta_1\delta_2\varepsilon h\sigma_1\cos\left(h\omega\right) + \left(\delta_1h-1\right)\left(8\delta_1\delta_2\varepsilon h\sigma_1 + 2\delta_1h^2\sigma_1\sigma_2 + 2\delta_2h^2\sigma_1^2\right)e^{-ih\omega} & -4\delta_1^2\varepsilon h\sigma_2\cos\left(h\omega\right) + \left(\delta_2h-1\right)\left(24\delta_1^2\delta_2\varepsilon^2 + 8\delta_1^2\varepsilon h\sigma_2 + 16\delta_1\delta_2\varepsilon h\sigma_1 + 2\delta_1h^2\sigma_1\sigma_2 + 2\delta_2h^2\sigma_1^2\right)e^{-ih\omega} & 4\delta_1^2\delta_2\varepsilon h^2\sigma_1e^{ih\omega} + 4\delta_1\delta_2\varepsilon h\sigma_1e^{-ih\omega} + \left(2\delta_1h^2\sigma_1\sigma_2 + 2\delta_2h^2\sigma_1^2\right)\cos\left(h\omega\right) & -4\delta_1^2\varepsilon h\sigma_2\left(\delta_2h-1\right)e^{ih\omega} + \left(24\delta_1^2\delta_2\varepsilon^2 + 8\delta_1^2\varepsilon h\sigma_2 + 16\delta_1\delta_2\varepsilon h\sigma_1 + 2\delta_1h^2\sigma_1\sigma_2 + 2\delta_2h^2\sigma_1^2\right)\cos\left(h\omega\right) \end{pmatrix}$$

If we use $\delta_1 = \delta_2 = 1/h$ the matrix has real coefficients as follows:

$$\widehat{D}_j^{-1}\left(-\widehat{L}_j - \widehat{U}_j\right) = \frac{\cos\left(h\omega\right)}{\frac{24\varepsilon^2}{h^2} + 16\varepsilon\sigma_1 + 16\varepsilon\sigma_2 + 2h^2\sigma_1^2 + 4h^2\sigma_1\sigma_2 + 2h^2\sigma_2^2}$$

$$\begin{pmatrix} \frac{24\varepsilon^2}{h^2} + 8\varepsilon\sigma_1 + 16\varepsilon\sigma_2 + 2h^2\sigma_1\sigma_2 + 2h^2\sigma_2^2 & 8\varepsilon\sigma_2 + 2h^2\sigma_1\sigma_2 + 2h^2\sigma_2^2 & -4\varepsilon\sigma_1 & 4\varepsilon\sigma_2 \\ 8\varepsilon\sigma_1 + 2h^2\sigma_1^2 + 2h^2\sigma_1\sigma_2 & \frac{24\varepsilon^2}{h^2} + 16\varepsilon\sigma_1 + 8\varepsilon\sigma_2 + 2h^2\sigma_1^2 + 2h^2\sigma_1\sigma_2 & 4\varepsilon\sigma_1 & -4\varepsilon\sigma_2 \\ -4\varepsilon\sigma_1 & 4\varepsilon\sigma_2 & \frac{24\varepsilon^2}{h^2} + 8\varepsilon\sigma_1 + 16\varepsilon\sigma_2 + 2h^2\sigma_1\sigma_2 + 2h^2\sigma_2^2 & 8\varepsilon\sigma_2 + 2h^2\sigma_1\sigma_2 + 2h^2\sigma_2^2 \\ 4\varepsilon\sigma_1 & -4\varepsilon\sigma_2 & 8\varepsilon\sigma_1 + 2h^2\sigma_1^2 + 2h^2\sigma_1\sigma_2 & \frac{24\varepsilon^2}{h^2} + 16\varepsilon\sigma_1 + 8\varepsilon\sigma_2 + 2h^2\sigma_1^2 + 2h^2\sigma_1\sigma_2 \end{pmatrix}$$

Bibliography

[1] D. Arndt, W. Bangerth, D. Davydov, T. Heister, L. Heltai, M. Kronbichler, M. Maier, J.-P. Pelteret, B. Turcksin, and D. Wells. The `deal.II` library, version 8.5. *Journal of Numerical Mathematics*, 25(3):137–146, 2017.

[2] W. Bangerth, R. Hartmann, and G. Kanschat. deal.II – a general purpose object oriented finite element library. *ACM Trans. Math. Softw.*, 33(4):24/1–24/27, 2007.

[3] E. Meinkohn, G. Kanschat, R. Rannacher, and R. Wehrse. *Numerical methods for multidimensional radiative transfer*. Springer-Verlag, 2009.

[4] M. L. Adams. Discontinuous finite element transport solutions in thick diffusive problems. *Nuclear Sci. Engng.*, 3(137):298–333, 2001.

[5] J.-L. Guermond and G. Kanschat. Asymptotic analysis of upwind DG approximation of the radiative transport equation in the diffusive limit. *SIAM J. Numer. Anal.*, 1(48):53–78, 2010.

[6] J.C. Ragusa, J.-L. Guermond, and G. Kanschat. A robust sn-dg-approximation for radiation transport in optically thick and diffusive regimes. *Journal of Computational Physics*, 231(4):1947–1962, 2012.

[7] G. Kanschat and J.-C. Ragusa. A robust multigrid preconditioner for $s_n$dg approximation of monochromatic, isotropic radiation transport problems. *SIAM Journal on Scientific Computing*, 36(5):A2326–A2345, 2014.

[8] Xiao chuan Cai, David E. Keyes, and David P. Young. A nonlinear additive schwarz preconditioned inexact Newton method for shocked duct flow. In *Proceedings of the 13th International Conference on Domain Decomposition Methods*, pages 1463–1470, 2001.

[9] Xiao-Chuan Cai and David E. Keyes. Nonlinearly preconditioned inexact Newton algorithms. *SIAM Journal on Scientific Computing*, 24(1):183–200, 2002.

[10] J.J. Duderstadt and W.R. Martin. *Transport Theory*. Wiley-Interscience Publications. John Wiley & Sons, Inc., 1979.

[11] J. J. Duderstadt and L. J. Hamilton. *Nuclear Reactor Analysis*. Wiley, New York, 1976.

[12] G. Bell and S. Glasstone. *Nuclear Reactor Theory*. Van Nostrand, Princeton, N. J., 1970.

[13] K. M. Case and P. F. Zweifel. *Linear Transport Theory*. Addison-Wesley, Reading, Mass., 1967.

[14] S. Chandrasekhar. *Radiative Transfer*. Oxford University Press, 1950.

[15] D. Mihalas and B. Weibel-Mihalas. *Foundations of Radiation Hydrodynamics*. 1984, Dover (reprint).

[16] J. Oxenius. *Kinetic Theory of Particles and Photons*. Springer, 1986.

[17] R. Dautray and J.-L. Lions. *Mathematical Analysis and Numerical Methods for Science and Technology*, volume 6. Springer, 2000.

[18] M. Asadzadeh. Analysis of a fully discrete scheme for neutron transport in two-dimensional geometry. *SIAM J. Numer. Anal.*, (23):543–561, 1986.

[19] S. Larsson M. Asadzadeh, P. Kumlin. The discrete ordinates method for the neutron transport equation in an infinite cylindrical domain. *Math. Models Methods Appl. Sci.*, 3(2):317–338, 1992.

[20] C. Johnson and J. Pitkäranta. Convergence of a fully discrete scheme for two-dimensional neutron transport. *SIAM J. Numer. Anal.*, (20):951–966, 1983.

[21] J. Pitkäranta. Approximate solution of the transport equation by methods of Galerkin type. *J. Math. Anal. Appl.*, 1(60):186–210, 1977.

[22] J. Pitkäranta. On the variational approximation of the transport operator. *J. Math. Anal. Appl.*, 2(54):419–440, 1976.

[23] J. Pitkäranta. A non-self-adjoint variational procedure for the finite-element approximation of the transport equation. *Transport Theory Statist. Phys.*, 1(4):1–24, 1975.

[24] J. Pitkäranta and R. L. Scott. Error estimates for the combined spatial and angular approximations of the transport equation for slab geometry. *SIAM J. Numer. Anal.*, 5(20):922–950, 1983.

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

[26] P. LeSaint and P.-A. Raviart. On a finite element method for solving the neutron transport equation. *Mathematical aspects of finite elements in partial differential equations*, pages 89–123, 1974.

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

[28] E. W. Larsen. The asymptotic diffusion limit of discretized transport problems. *Nucl. Sci. Eng.*, (112):336–346, 1992.

[29] E. W. Larsen and J. E. Morel. Asymptotic solutions of numerical transport problems in optically thick, diffusive regimes. II. *J. Comput. Phys.*, 1(83):212–236, 1989.

[30] G. Kanschat J. Ragusa, J.-L. Guermond. A robust S_n-DG-approximation for radiation transport in optically thick and diffusive regimes. *J. Comput. Phys.*, 4(231):1947–1962, 2012.

[31] Robert Dautray and Jacques-Louis Lions. *Mathematical analysis and numerical methods for science and technology. Volume 2. , Functional and Variational Methods*. Springer-Verlag, Berlin Heidelberg New York London Paris Tokyo, 1985.

[32] Douglas N. Arnold. An interior penalty finite element method with discontinuous elements. *SIAM J. Numer. Anal.*, 19(4):742–760, 1982.

[33] M. Eiermann and O. G. Ernst. Geometric aspects of the theory of krylov subspace methods. *Acta Numerica*, pages 251–312, 2001.

[34] H. C. Elman. Iterative methods for large sparse nonsymmetric systems of linear equations. *Ph.D. thesis, Computer Science Dept., Yale Univ., New Haven, CT*, 1982.

[35] Andrea Toselli and Olof B. Widlund. *Domain decomposition methods : algorithms and theory*. Springer series in computational mathematics. Springer, Berlin, 2005.

[36] S. C. Brenner and L. R. Scott. *The Mathematical Theory of Finite Element Methods*. Springer, New York, 2002.

[37] Barry Francis Smith, Petter E. Bjørstad, and William D. Gropp. *Domain decomposition : parallel multilevel methods for elliptic partial differential equations*. Cambridge University Press, Cambridge, 1996.

[38] X. Feng and O. Karakashian. Two-level non-overlapping schwarz methods for a discontinuous galerkin method. *SIAM J. Numer. Anal.*, 39(4):1343–1365, 2001.

[39] J. H. Bramble. *Multigrid Methods*. 294. Pitman Research Notes in Mathematics, Longman Scientific, 1993.

[40] D. Braess and W. Hackbusch. A new convergence proof for the multigrid method including the v-cycle. *SIAM J. Sci. Comput.*, 5(20):967–975, 1983.

[41] Maksymilian Dryja and Piotr Krzyżanowski. A massively parallel nonoverlapping additive schwarz method for discontinuous galerkin discretization of elliptic problems. *Numerische Mathematik*, 132(2):347–367, February 2016.

[42] Krzyżanowski Piotr. On a nonoverlapping additive schwarz method for h-p discontinuous galerkin discretization of elliptic problems. *Numerical Methods for Partial Differential Equations*, 32(6):1572–1590.

[43] Martin J. Gander. Optimized schwarz methods. *SIAM Journal on Numerical Analysis*, 44(2):699–731, 2006.

[44] Carl Cowen and Elad Harel. An effective algorithm for computing the numerical range. September 1995.

[45] C.R. Johnson. Computation of the field of values of a 2 x 2 matrix. *JOURNAL OF RESEARCH of the National Bureau of Standards - B. Matthematical Sciences*, 78B(3):105–107, 1974.

[46] Guido Kanschat and José Pablo Lucero Lorca. A weakly penalized discontinuous galerkin method for radiation in dense, scattering media. 16, October 2016.

[47] E. W. Larsen, J. E. Morel, and Jr. W. F. Miller. Asymptotic solutions of numerical transport problems in optically thick, diffusive regimes. *J. Comput. Phys.*, 2(69):283–324, 1987.

[48] K. Grella. Sparse tensor phase space Galerkin approximation for radiative transport. *SpringerPlus*, 1(3):3–230, 2014.

[49] K. Grella and C. Schwab. Sparse tensor spherical harmonics approximation in radiative transfer. *J. Comput. Phys.*, 23(230):8452–8473, 2011.

[50] C. Schwab G. Widmer, R. Hiptmair. Sparse adaptive finite elements for radiative transfer. *J. Comput. Phys.*, 12(227):6071–6105, 2008.

[51] R. Sanchez and J. Ragusa. On the construction of Galerkin angular quadratures. *Nucl. Sci. Eng.*, (169):133–154, 2011.

[52] B. Ayuso and L. D. Marini. Discontinuous Galerkin methods for advection-diffusion-reaction problems. *SIAM J. Numer. Anal.*, 2(47):1391–1420, 2009.

[53] P. Castillo, B. Cockburn, I. Perugia, and D. Schötzau. An a priori error estimate of the local discontinuous Galerkin method for elliptic problems. *SIAM J. Numer. Anal.*, 5(38):1676–1706, 2000.

[54] G. Kanschat. Preconditioning methods for local discontinuous Galerkin discretizations. *SIAM J. Sci. Comput.*, 3(25):815–831, 2003.

[55] G. Kanschat. Block preconditioners for LDG discretizations of linear incompressible flow problems. *J. Sci. Comput.*, 1(22):381–394, 2005.

[56] D. A. D. Pietro, A. Ern, and J.-L. Guermond. Discontinuous Galerkin methods for anisotropic semidefinite diffusion with advection. *SIAM J. Numer. Anal.*, 2(46):805–831, 2008.

[57] W. Hackbusch and T. Probst. Downwind Gauss-Seidel smoothing for convection dominated problems. *Numer. Linear Algebra Appl.*, 2(4):85–102, 1997.

[58] J. Xu. Iterative methods by space decomposition and subspace correction. *SIAM Review*, 4(34):581–613, 1992.

[59] G. Kanschat. *Discontinuous Galerkin Methods for Viscous Flow*. Deutscher Universitätsverlag, Wiesbaden, 2007.

[60] Carsten Burstedde, Lucas C. Wilcox, and Omar Ghattas. `p4est`: Scalable algorithms for parallel adaptive mesh refinement on forests of octrees. *SIAM Journal on Scientific Computing*, 33(3):1103–1133, 2011.

[61] Edgar Gabriel, Graham E. Fagg, George Bosilca, Thara Angskun, Jack J. Dongarra, Jeffrey M. Squyres, Vishal Sahay, Prabhanjan Kambadur, Brian Barrett, Andrew Lumsdaine, Ralph H. Castain, David J. Daniel, Richard L. Graham, and Timothy S. Woodall. Open MPI: Goals, concept, and design of a next generation MPI implementation. In *Proceedings, 11th European PVM/MPI Users' Group Meeting*, pages 97–104, Budapest, Hungary, September 2004.

[62] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen. *LAPACK Users' Guide*. Society for Industrial and Applied Mathematics, Philadelphia, PA, third edition, 1999.

[63] Michael A. Heroux, Roscoe A. Bartlett, Vicki E. Howle, Robert J. Hoekstra, Jonathan J. Hu, Tamara G. Kolda, Richard B. Lehoucq, Kevin R. Long, Roger P. Pawlowski, Eric T. Phipps, Andrew G. Salinger, Heidi K. Thornquist, Ray S. Tuminaro, James M. Willenbring, Alan Williams, and Kendall S. Stanley. An overview of the trilinos project. *ACM Trans. Math. Softw.*, 31(3):397–423, 2005.

[64] W. K. Widger and M. P. Woodall. Integration of the Planck blackbody radiation function. *Bulletin of the Am. Meteorological Society*, 10(57):1217–1219, 1976.

[65] J. C. Ragusa T. H. Ghaddar. An approach for load balancing massively parallel transport sweeps on unstructured grids. *International Conference on Mathematics & Computational Methods Applied to Nuclear Science & Engineering*, April 2017.

[66] V. Dolean, M. J. Gander, W. Kheriji, F. Kwok, and R. Masson. Nonlinear preconditioning: How to use a nonlinear schwarz method to precondition Newton's method. *SIAM Journal on Scientific Computing*, 38(6):A3357–A3380, 2016.

[67] Pinte, C., Harries, T. J., Min, M., Watson, A. M., Dullemond, C. P., Woitke, P., Ménard, F., and Durán-Rojas, M. C. Benchmark problems for continuum radiative transfer - high optical depths, anisotropic scattering, and polarisation. *A&A*, 498(3):967–980, 2009.

[68] Will Pazner and Per-Olof Persson. Approximate tensor-product preconditioners for very high order discontinuous galerkin methods. *Journal of Computational Physics*, 354:344–369, 2018.